

Hierarchical QSAR technology based on the Simplex representation of molecular structure

V. E. Kuz'min · A. G. Artemenko · E. N. Muratov

Received: 16 August 2007 / Accepted: 10 January 2008 / Published online: 6 February 2008
© Springer Science+Business Media B.V. 2008

Abstract This article is about the hierarchical quantitative structure–activity relationship technology (HiT QSAR) based on the Simplex representation of molecular structure (SiRMS) and its application for different QSAR/QSP(property)R tasks. The essence of this technology is a sequential solution (with the use of the information obtained on the previous steps) to the QSAR problem by the series of enhanced models of molecular structure description [from one dimensional (1D) to four dimensional (4D)]. It is a system of permanently improved solutions. In the SiRMS approach, every molecule is represented as a system of different simplexes (tetraatomic fragments with fixed composition, structure, chirality and symmetry). The level of simplex descriptors detailing increases consecutively from the 1D to 4D representation of the molecular structure. The advantages of the approach reported here are the absence of “molecular alignment” problems, consideration of different physical–chemical properties of atoms (e.g. charge, lipophilicity, etc.), the high adequacy and good interpretability of obtained models and clear ways for molecular design. The efficiency of the HiT QSAR approach is demonstrated by comparing it 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 predictive virtual screening tools and their ability to serve as

the base of directed drug design was validated by subsequent synthetic and biological experiments, among others. The HiT QSAR is realized as a complex of computer programs known as HiT QSAR software that also includes a powerful statistical block and a number of useful utilities.

Keywords Anticancer activity · Antiviral activity · Applicability domain · Drug design · Mechanistic interpretation of QSAR · Molecular descriptors · 1D–4D QSAR/QSPR · Selectivity · Toxicity · Virtual screening

Abbreviations

| | |
|----------|---|
| AI/EVS | Automatic/interactive/evolutionary variables selection |
| ACE | Angiotensin converting enzyme |
| 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 |

V. E. Kuz'min · A. G. Artemenko · E. N. Muratov (✉)
A.V. Bogatsky Physico-Chemical Institute, National Academy
of Sciences of Ukraine, Lustdorfskaya doroga 86, Odessa 65080,
Ukraine
e-mail: murik@ccmsi.us; 00dqsar@mail.ru

| | |
|---------------------|---|
| QSAR/ | Quantitative structure–activity/property |
| QSPR | 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 the molecular structure QSAR approach |
| TV | Trend-vector statistical method |

Introduction

The development of a new medicine costs more than one billion dollars, and the price of this process is increasing steadily [1]. Different theoretical approaches are currently used to facilitate and accelerate the process involved in creating new drugs; these are not only very expensive, but are also multi-stepped and long-termed [2]. The choice of approaches depends on the presence or absence of information on a biological target and those substances that interact with it. The most common situation that arises is one in which there are a set of biologically active compounds (ligands) and no information on the biological target (receptor). Different QSAR (quantitative structure–activity relationship) approaches are used in this case. For many years, QSAR has been successfully used to analyze a large variety of parameters, including antiviral and anticancer activity, toxicity, among others [3–14]. Its staying power may be attributed to the strength of its initial postulate, which is 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].

A large number of different QSAR methods [16–20] have been developed since the second half of the last century, and new and new techniques and improvements are still being created nowadays [21]. These approaches differ mainly by the principles and levels of representation and the description of molecular structure. The degree of adequacy in terms of molecular structure models varies from the one-dimensional (1D) to four-dimensional (4D) level.

One-dimensional models consider only the gross-formula of a molecule (for example, alanine: $\text{C}_3\text{H}_7\text{NO}_2$). However, such models reflect only the composition of a molecule, and it is obviously impossible to solve adequately the “structure–activity” tasks using such approaches. Consequently, these models usually have an auxiliary role only, although they can occasionally be used as independent virtual screening tools.

Two-dimensional models contain information on the structure of a compound and are based on its structural formula [20]; as such, these models reflect only the topology of the molecule. These models are very popular [3, 22]. The capacity of such approaches is that the topology model of the molecular structure in an implicit form contains information on the possible conformations of the compound. Our operational experience shows that the 2D level of representation of the molecular structure is adequate to solve more than 90% of existing QSAR/QSPR tasks.

Three-dimensional QSAR models [16, 17, 19, 20] provide a full structural information, taking into account the composition, topology and spatial shape of molecule for one conformer only. These models are widespread. However, the choice of the analyzed conformer is mostly accidental.

The description of the molecular structure is realized by 4-D QSAR models [10, 23] more adequately. These models are similar to 3D models, but in comparison to the latter, the structural information in 4D models is considered for a set of conformers (conditionally, 4D) instead of for one fixed conformation.

Hierarchical QSAR technology

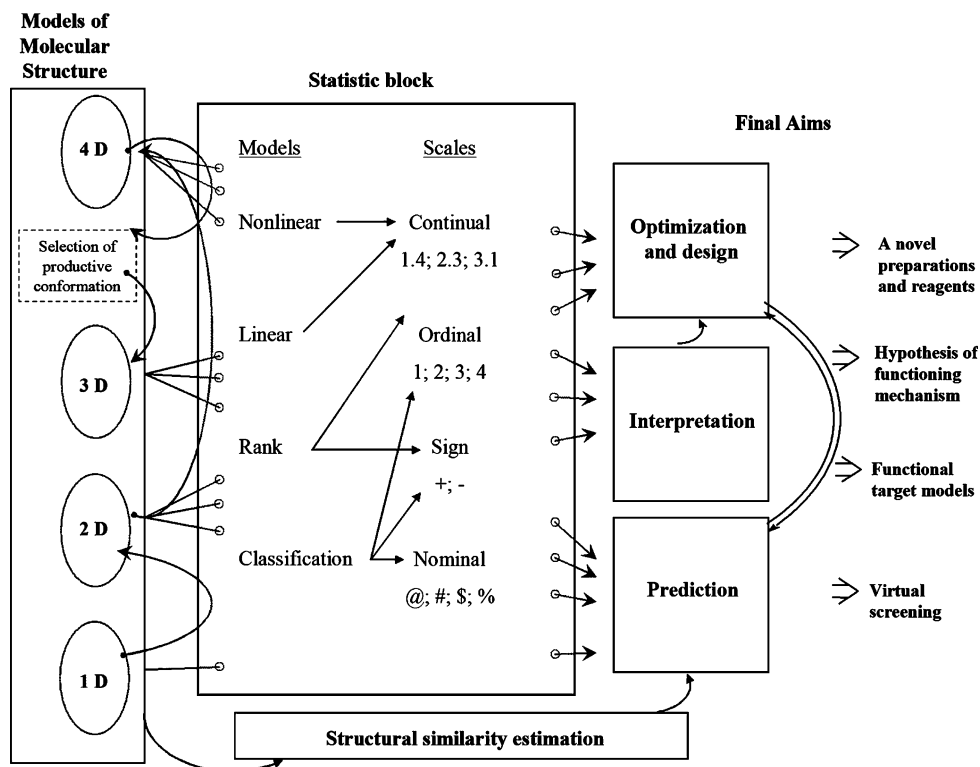
Multi-hierarchical strategy

We consider here the hierarchic QSAR technology (HiT QSAR) [24, 25] based on the Simplex representation of the molecular structure (SiRMS). This method has proved to be efficient in numerous studies for solving different “structure–activity/property” problems [3, 10–12, 26–33]. The essence of the strategy is based on the solution of QSAR problems via the sequence of permanently improved molecular structure models (from 1D to 4D) (Fig. 1). Thus, during 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, we propose that this approach deals with a system of permanently improved solutions and leads to a more effective interpretation of the obtained QSAR models because it enables the molecular fragments/models to be revealed, and this knowledge is important for the detailed development of structure.

The main features of our strategy consists of the multiple-aspect hierarchy, as related to (Fig. 1):

- models of the molecular structure description (1D → 2D → 3D → 4D);
- scales of activity estimation (binomial → nominal → ordinal → continual);

Fig. 1 Scheme of hierarchical quantitative structure–activity relationship (QSAR) technology



- mathematical methods used for the establishing of structure–activity relationship [pattern recognition → rank correlation → multivariate regression → partial least squares (PLS)];
- final aims of the solution the QSAR task (prediction → interpretation → structure optimization → molecular design).

The set of different QSAR models that are supplementing each other is a result of HiT QSAR application. Taken together, these models, in complex, solve problems of virtual screening and evaluate the influence of structural factors on activity the modification of known molecular structures and design of new high-performing potential antiviral agents or other compounds with desired properties.

The scheme of the HiT QSAR is shown on Fig. 1. In this figure, the information on QSAR models of the lowest level has been transferred (curved arrows) to the models of the highest level after the corresponding statistical processing (“Statistic block” on Fig. 1), during which the most significant structural parameters have been chosen. It is necessary to note that after the 2D step, the QSAR task is solved on the 4D level because there is no a priori available information about a “productive” conformation (the conformer that interacts with a biological target most effectively) for the 3D QSAR models. This information appears only after the construction of 4D QSAR models and the calculations of activity for all of the conformers considered. The information about the “productive” (the

most active) conformation is transferred to the 3D QSAR level. It distinguishes HiT QSAR from ordinary 3D QSAR approaches where the investigated conformers have mostly been chosen randomly. Given the investigated activity mostly determined by the interaction of the exact “productive” conformation (not by the set of conformers) with the 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 for obtaining of the QSAR relations (“Statistic block” on Fig. 1).

The principle feature of the HiT QSAR is its multi-hierarchy—i.e. not only the hierarchy of different models but also the hierarchy of the purposes are taken into account (Fig. 1, unit: “Final Aims”).

It is clearly very difficult to obtain a model that can solve all of the problems related to the influence of the structure of the molecules being studied on the property being examined. As such, it is necessary to develop a set of different QSAR models for solving each concrete task, with some of these being more suitable for the prognosis of the studied property, others for interpretation of the obtained relations and still other ones for a molecular design. Taken together, these models, in complex, solve the problem of the creation of the new prospective compounds and substances with the desired set of properties. The important feature of such an approach is that the general results obtained by few different independent models are

always more relevant. It is necessary to note that these resulting QSAR models have been chosen in accordance with all of QSAR Organization for Economic Co-operation and Development (OECD) (Setubal) principles [34]— i.e. they have defined endpoint, an unambiguous algorithm, a defined domain of applicability (DA) and mechanistic interpretation and are good fitted, robust and predictive. Thus, we assume that the proposed strategy allows the user to solve all problems dealing with the virtual screening, modeling of functional (biological) targets, advancing of hypotheses on mechanisms of action and, ultimately, designing of new compounds with desired properties.

Simplex representation of molecular structure (SiRMS)

In the frameworks of SiRMS, any molecule can be represented as a system of different simplexes (tetraatomic fragments of fixed composition, structure, chirality and symmetry) [24, 25, 30, 35] (Fig. 2).

One-dimensional models

At the 1D level, a simplex is a combination of four atoms contained in the molecule (Fig. 2). The simplex descriptor (SD) at this level is a number of quadruples of atoms of the definite composition. For the compound ($A_aB_bC_cD_dE_eF_f...$), the value of SD ($A_1B_1C_1D_m$) is

$$K = f(I) \cdot f(j) \cdot f(l) \cdot f(m),$$

where, for example,

$$f(I) = \frac{a!}{(a-I)! \cdot I!}.$$

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 equated to zero.

Two-dimensional models

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

- atom individuality (nature or more detailed type of atom);
- partial atom charge [36] (see Fig. 2) (reflects the electrostatic properties);
- lipophilicity of atom [37] (reflects the hydrophobic properties);

- atomic refraction [38] (partially reflects the ability of the atom to dispersionic interactions);
- a mark that characterizes the atom as a possible hydrogen donor or acceptor during H-bond formation (A, hydrogen acceptor in H-bond; D, hydrogen donor in H-bond, I, indifferent atom).

For atom characteristics, which have real values (charge, lipophilicity, refraction, etc.), the ranges of values are divided into definite discrete groups during 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 the important part of SiRMS. We consider that the specification of atoms only by their nature (actually reflects atom identity, for example, C, N, O) in many QSAR methods limits the possibilities of pharmacophore fragment selection. For example, if the $-NH-$ group has been selected as the determining activity fragment (pharmacophore) and the ability of H-bond formation is a factor determining its activity, then we shall miss such donors of H-bonds as, for example, an OH-group, among others. The usage of atom differentiation by donor/acceptor of H-bond allows the situation illustrated above to be avoided. Analogical examples can be made for other atom properties (lipophilicity, partial charge, refraction, etc.).

Thus, the SD at 2D level is a number of simplexes of fixed composition and topology. It is necessary to note that for 1D and 2D QSAR analysis along with simplex descriptors, other structural parameters, corresponding to molecular fragments of different size, can be used. The usage of one to four 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.

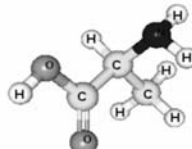
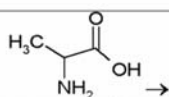
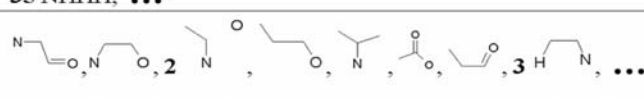
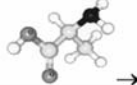
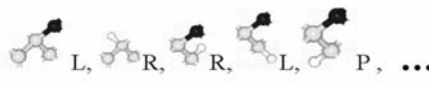
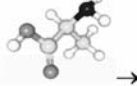

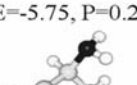
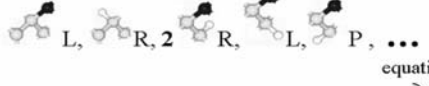
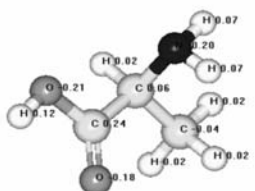
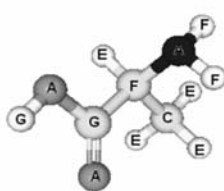
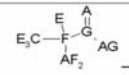
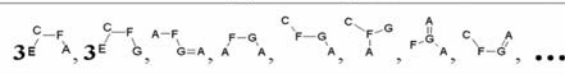
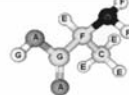
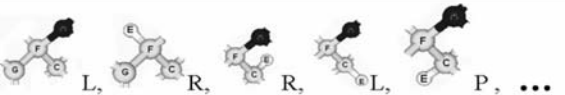
Three-dimensional models

At the 3D level, not only topology but also the stereochemistry of the molecule is taken into account. It is possible to differentiate all of the simplexes as right, left and symmetrical and plane achiral ones. The stereochemical configuration of simplexes is defined by modified Kahn–Ingold–Prelog rules [35]. A SD at this level is a number of simplexes of fixed composition, topology, chirality and symmetry.

Four-dimensional models

For the 4D QSAR models, each SD is calculated by the summation of products of descriptor value for each

Fig. 2 Examples of the generation of simplex descriptors for alanine at one-dimensional to four-dimensional (1D–4D) levels

| Level | Structure | Simplex generation |
|-------------------------|--|--|
| |  | |
| 1D | $C_3H_7O_2N \rightarrow$ | 6 CCNO, 42 CNOH, 63 CNHH, 21 CCNH, 42 NOHH, 7 CCCH, 35 NHHH, ... |
| 2D |  |  2, ... |
| 3D |  |  L, R, R, L, P, ... |
| 4D |  $E=-6.35, P=0.63$  $E=-5.75, P=0.23$  $E=-5.49, P=0.14$ |  L, R, 2 R, L, P, ... equation 3 1 L, 1 R, 2 R, 1.14 L, 0.86 P, ... |
| Division by atom charge | | |
| |  | $A \leq -0.1$ $-0.1 < B \leq -0.05$ $-0.05 < C \leq -0.01$ $-0.01 < D \leq 0.01$ $0.01 < E \leq 0.05$ $0.05 < F \leq 0.1$ $G > 0.1$  |
| 1D | $A_3CE_3F_3G_2 \rightarrow$ | $A_3C, 3A_3E, 9A_2E_2, 3AE_3, 27ACEF, 18CEFG, 9ACF_2, 9CE_2F, 54AEFG, \dots$ |
| 2D |  |  3E, 3E, ... |
| 3D |  |  L, R, R, L, P, ... |

conformer (SD_k) and the probability of realization of the corresponding conformer (P_k).

$$SD = \sum_{k=1}^N (SD_k \cdot P_k), \quad (1)$$

where N is a number of conformers being considered.

As is well known [39], a probability of conformation P_k is defined by its energy:

$$P_k = \left\{ 1 + \sum_{i \neq k} \exp\left(\frac{-(E_i - E_k)}{RT}\right) \right\}^{-1}, \quad \sum_k P_k = 1, \quad (2)$$

where E_i and E_k are the energies of conformations i and k , respectively. The conformers are analyzed within the energy band of 5–7 kcal/mol. Thus, the molecular SD at

the 4D level takes into account the probability of realization of 3D level SD in the set of conformers. On the 4D level, other 3D whole-molecule parameters efficient for description of the spatial form of the conformer (e.g. characteristics of inertia ellipsoid, dipole moment, etc.) can be used along with the SD.

An example of the representation of the molecule as the sets of simplexes on different levels of structure detailing (1D–4D) is depicted in Fig. 2.

Fourier transform and whole-molecule descriptors

Simplex descriptors at all levels of differentiation (1D–4D) are actually the fragmentary parameters that 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-transform [40] for the spectrum of simplex structural parameters.¹ The high-frequency harmonics characterize small fragments, while the low-frequency harmonics correspond to the global molecule properties. The Fourier-transformation of discrete function of parameters $P(I)$ can be presented as

$$P(I) = \frac{a_0}{2} + \sum_{k=1}^{M-1} \left(a_k \cos \frac{2\pi k(I-1)}{N} + b_k \sin \frac{2\pi k(I-1)}{N} \right) + a_{N/2} \cos(\pi(I-1)), \quad (3)$$

where

$$a_k = \frac{2}{N} \cdot \sum_{i=1}^N P_I \cdot \cos \left(\frac{2\pi \cdot k \cdot (I-1)}{N} \right), \quad b_k = \frac{2}{N} \cdot \sum_{i=1}^N P_I \cdot \sin \left(\frac{2\pi \cdot k \cdot (I-1)}{N} \right), \quad (4)$$

or, in alternative form,

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

where amplitudes $q_k = \sqrt{a_k^2 + b_k^2}$, and phase angle

$$\psi_k = \arctan \left(\frac{a_k}{b_k} \right), \quad (6)$$

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 . The values of amplitudes (a_k, b_k, q_k) can be used as the parameters for the solution of QSAR tasks [19]. The 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 (Eq. 4). Amplitudes q_k have a bad mechanistic interpretation because of a more complex dependence from the source structural parameters (Eq. 6). However, all the amplitudes (a_k, b_k, q_k) separately or together allow us to obtain good-fitted, robust and predictive models and, therefore, they can be used as an additional (completely different) tool for the virtual screening.

Such whole-molecule parameters 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, among others have been also used on different levels of representation of the molecular structure.

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

Estimation of the factors determining the interaction with the biological target

Within the framework of SiRMS (as in the CoMFA approach [16]) it is possible to define the relative influence of different physical and chemical factors on the character of the interaction of the molecules with a biological target. For this purpose it is necessary to summarize and compare the normalized contributions (b_j ; Eq. 8, see below) of simplexes in the obtained model separately for each differentiation group. Thus, the relative contribution of simplexes, where the differentiation of vertexes corresponds to the partial charges on atoms, reflects the role of electrostatic factors; the relative contribution of simplexes, where atoms are differentiated by lipophilicity, reflects the role of hydrophobic factors, among others. The relative influence of some physicochemical factors on the variation of anti-HSV-1 activity [3] has been given (see Fig. 3) as an example of the mentioned contributions.

Inverse task solution: molecular design of highly active novel compounds

Using Eq. 8, it is not difficult to make the inverse analysis (interpretation of QSAR models) in the framework 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 the number of atoms (n) in the simplex (or fragment): $C_j = \frac{1}{n} \sum_M b_I$, (for simplex $n = 4$). According to this formula, the atomic contribution depends on the number of simplexes that include this atom. This value

¹ A preliminary step is that the list of structural parameters must be well-organized on a defined principle (for example, lexicographic).

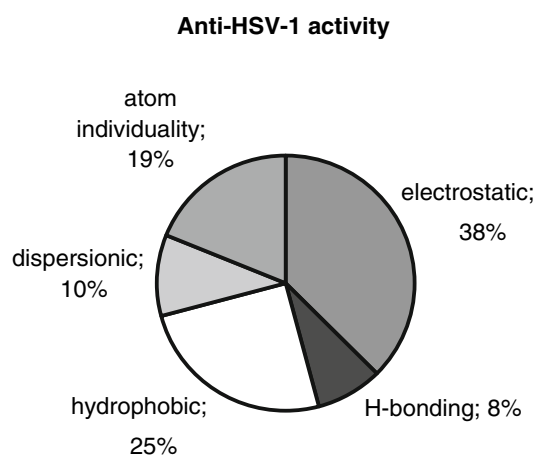


Fig. 3 Relative influence of some physicochemical factors on the variation of anti-herpes simplex virus type 1 (HSV-1) activity [3]

(number of simplexes) is not constant, and it varies in different molecules and depends on other constituents (surroundings); hence, this contribution is non-additive. Atoms that have a positive or negative influence on the studied biological activity of compounds can be colored. This helps to present the results and to determine visually (additionally to the automatic search) the groups of atoms affecting the activity in different directions and with various strength. An example of the representation of the obtained results on the molecule using color-coding according to the contribution of atoms into antiherpetic activity [11] is given in Fig. 4. Atoms and structural fragments reducing antiviral activity are shown in dark gray and those which enhance antiviral activity are shown in light gray and white. Indifferent atoms and fragments are colored in gray.

The procedure of the automatic search of pre-defined fragments from the dataset and their relative effect on activity that is expressed in its (activity) scale has been

realized in the HiT QSAR. The procedure of searching for fragments in the molecule is based on the fast algorithm for solving the maximum clique problem [41]. Some molecular fragments promoting and interfering anti-influenza activity [12, 30, 42] are represented in Table 1 as well as their average relative influence on it.

Based on the SiRMS, it is possible to realize the molecular design of compounds with the desired activity level via the generation of allowed combinations of simplexes determining the property being investigated. The simplest way is the soft drug design [43] that consists of replacing undesired fragments by more active ones, or by the insertion of fragments promoting the activity at positions of indifferent parts of the molecule or hydrogen atoms. The usage of this technique allows the design of new compounds at the same region of structural space as the training set of compounds. The accuracy of prognosis can be estimated using DA techniques (see below). However, the soft drug design usage constrains us within the limits of the initial chemical class of training set compounds. More drastic drug design is, certainly, more risky, but it allows us to obtain much more dramatic results. Almost surely, new structures would lie outside the DA region; however, this does not mean an uncertainty of prediction, but an extrapolation of the model predictivity and a certain lack of any DA procedure. At the same time, however, we can receive compounds of completely different (from initial training set) chemical class as the output of such a design. It has been demonstrated [12, 29, 30] that in searching for a new antiviral and anticancer agents, we started our investigations from macrocyclic pyridinophanes and passed through several convolutions of QSAR analysis to come out to nitrogen analogues of crown ethers in the first and acyclic aromatic structures, with the azomethine fragment in the second case.

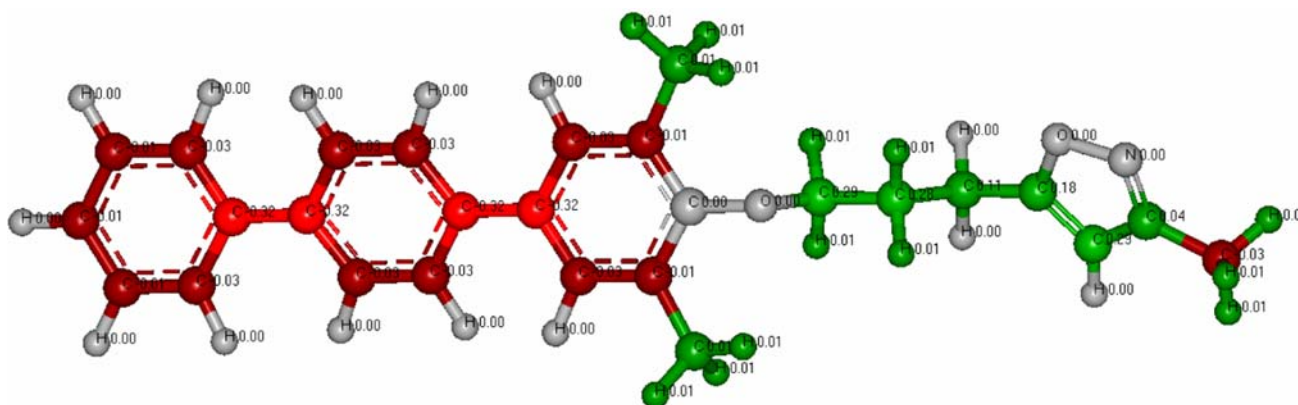
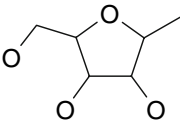
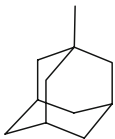
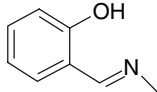
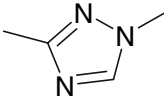
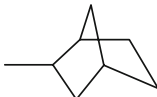
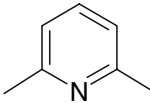


Fig. 4 Color-coded structure according to the contribution of the atoms to activity against HRV-2 [11]. Atoms and structural fragments reducing antiviral activity are colored in *dark gray* and that enhancing antiviral activity in *light grey and white*

Table 1 Molecular fragments governing the change in anti-influenza activity and the average relative influence of these fragments on this activity [12, 30]

| Enhance the activity, $\Delta \lg \text{TID}_{50}$ | | |
|---|--|---|
|  |  |  |
| 3.0 | 2.4 | 1.9 |
|  | $-(\text{CH}_2)_2-\text{O}-$ |  |
| 1.7 | 1.4 | 0.8 |
| Decrease the activity, $\Delta \lg \text{TID}_{50}$ | | |
| $-(\text{CH}_2)_n-\text{NH}-$ $n=2-3$ |  | $-\text{CO}-\text{NH}_2$ |
| -0.3 – -0.4 | -0.2 | -0.2 |

Statistical processing

External validation and test sets formation

To have any practical utility, up-to-date QSAR investigations must be used to make predictions [44]. 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 , etc.). 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 make so-called “test set”. During this procedure a certain proportion of the dataset molecules (10–85%) has been removed into 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 the result can sometimes indicate only the ability of the model to predict the specific test set. It is therefore important that both training and test sets cover the structural space of the complete data set as broadly as possible.

In the HiT QSAR, the following procedure has been used to form the test set: a dissimilarity matrix for all of

the initial training set molecules has been developed based on the relevant structural descriptors. Such a descriptor set can be obtained using different procedures of descriptors selection [for example, trend-vector (TV) procedure; see below) or directly from the model generated for all of the compounds investigated. In our opinion, the usage 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 on its own, but from the point of view of the investigated activity and mentioned descriptors selection will help to avoid some of the distortions caused by the insignificance of the structural parameters from the initial set for this concrete task.

A dissimilarity matrix is based on the estimation of structural dissimilarity between all of the molecules investigated. A measure of the structural dissimilarity for molecules M , M' can be calculated using the Euclidean distance in the multi-dimensional space of the structural parameters S

$$SD(M, M') = \sqrt{\sum_{i=1}^n (S_i - S'_i)^2}, \quad (7)$$

where n is the a number of molecules in the dataset.

Thus, total structural dissimilarity towards the remaining initial training set compounds could be calculated for each molecule like the sum of the corresponding Euclidean distances. At the same time, all of the compounds are divided into groups depending on their activity, where the number of groups equals the number of molecules that one wants to include in the test set. One compound from each group is then chosen into the test set according to its maximal (or minimal) total Euclidean distance towards the other molecules from this group, or by random choice. Most likely, the usage of several (three is an adequate minimum number) test sets constructed by different principles and the subsequent comparison and averaging of the obtained results for the model validation are more preferable than the usage of only one set. In this way, the first test set is constructed to maximize its diversity with the training set, i.e. the compounds with the maximal dissimilarity are chosen. This is the most rigorous estimation, and it can sometimes lead to the elimination of all of the dissimilar compounds from the training set, i.e. such pauperization 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 with the training set, i.e. less dissimilar compounds from each group were removed. The last test set is chosen in random manner, taking into account activity variation only.

Partial least squares method

A large number of simplex descriptors have been generated in the HiT QSAR. The PLS method has proved to be efficient when working with a large number of variables [45–47].

The PLS regression model may be written as [47]

$$Y = b_0 + \sum_{i=1}^N b_i x_i, \quad (8)$$

where Y is an appropriate activity, b_i is the PLS regression coefficient, x_i is an i -th descriptor value and N is the total amount of descriptors. This is apparently not different from that of multiple linear regression (MLR), except that 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) in terms of the number of objects (N) [47].

The removal of highly correlated and constant descriptors, genetic algorithm (GA) [48], the trend-vector method [49, 50] and the automatic variable selection (AVS) strategy that is similar to interactive (IVS) [46] and evolutionary

(EVS) [45] variables selection have all been used to select the descriptors in PLS. The removal of highly correlated descriptors is not necessary for PLS analysis since descriptors are reduced to a series of uncorrelated latent variables. However, this procedure frequently helps to obtain more adequate models and reduce the number of variables used by up to fivefold. During this procedure one descriptor from each pair having a pair correlation coefficient r satisfying $|r| > 0.90$ is eliminated.

Trend-vector procedure

The trend-vector (TV) procedure [19, 49, 50] does not concretize the form of a corresponding dependence and can use many structural parameters. This method can predict the properties of the molecules analyzed only in a rank scale and can be used if biological data are represented in ordinal scale (see Fig. 1). Similarly 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 parameter). Each component of a TV is determined as

$$T_j = \frac{1}{n} \cdot \sum_{i=1}^n (A_i - \bar{A}) \cdot S_{ij}, \quad (9)$$

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

$$\text{rank}(A_i) = \text{rank} \left(\sum_{j=1}^m T_j S_{ij} \right). \quad (10)$$

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

The search of the models using the TV method in HiT QSAR is realized by the methods of exhaustive or partial search after the mutual correlations have been removed. We discovered [10, 24] that descriptors involved in the best TV models (several decades of models with approximately identical quality) form a good subset for their subsequent usage in PLS. The noise elimination can be one of the more probable explanations of the success of the TV procedure.

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 is orthogonal to Y (investigated property). This strategy is similar to that of IVS [46], EVS [45], OSC [51] and O-PLS [52] and has the same objective but uses different means.

The essence of AVS is the following: during the first step of the AVS, the model containing all descriptors is obtained. The variables with the smallest normalized regression coefficients (b_i , Eq. 7) are then excluded from the X -matrix, and in the next step the PLS model is obtained. This procedure is repeated stepwise until the number of variables is equal to 1.

Automatic variable selection strategy can be used either for all structural parameters or after different procedures of variable selection (e.g. removal of highly correlated descriptors, TV procedure, GA). One application of the AVS procedure results in a decrease in model complexity (number of descriptors and latent variables) and an increase in its predictivity and robustness.

Genetic algorithm

The GA imitates such properties of living nature, as natural selection, adaptability, heredity, among others. The usage of the heuristic organized operations of “reproduction”, “crossing” and “mutation” from the casual or selected-by-user started “population” of new “chromosomes” results in the generation of models. The power of the GA lies in its flexibility. By adjusting the small set of algorithm parameters (number of generations, crossover and mutation type, crossover and mutation probability, type of selection), it is possible to displace a balance between the time of search and the quality of decision in one or another side. In the HiT QSAR, GA is used as a tool for the selection of adequate PLS models. Descriptors from the best model obtained by a preliminary AVS procedure are usually used as the starting “population”. The GA is definitely not a tool for the elucidation of the global maximum or minimum, and very often subsequent AVS procedures and different enumerative techniques allow one to increase the quality of the PLS models obtained.

Enumerative techniques

As mentioned above, the usage of the methods of exhaustive or partial search (depending on the number of selected descriptors) after AVS or GA very often allow ones to increase the quality of the obtained models. After

mentioned statistic processing model or models with the best combinations of statistic characteristics (R^2 , Q^2) have been selected from the obtained resulting list for subsequent validation using external test set. General scheme of the PLS models generation and selection applied in the HiT QSAR can be presented by Fig. 5. This procedure can be repeated several times using as input (initial set) SD of different level of the molecular structure representation (usually 2D–4D) and/or with various kinds of atoms 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.

Virtual screening of activity. Estimation of competence regions of PLS models

As mentioned above, QSAR investigations must be used to make predictions for compounds with unknown activity values (so-called “virtual screening”). With the purpose of analyzing the predictivity of the PLS models and according to the QSAR OECD principles [34], different DA procedures are included in the HiT QSAR.

The first one is a DA ellipsoid that we have developed [11]. This model actually represents a line at the 1D level, an ellipse at the 2D level, an ellipsoid at the 3D level and multi-dimensional ellipsoids in more complicated n -dimensional spaces. Its essence consists of the following: the distribution of a training set of molecules in a space of latent variables $T_1 - T_A$ (axes of coordinates) can be obtained from the PLS). For each coordinate axis (T_1 and T_2 in our case), the root-mean-square deviations S_{T1} and S_{T2} are determined. The DA represents an ellipsoid that is built from the

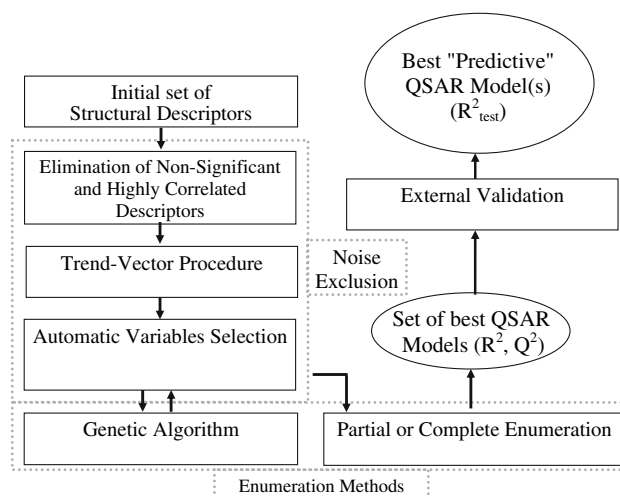


Fig. 5 General scheme of the PLS models generation and selection applied in HiT QSAR

molecules of the center of the training set ($T_1 = 0$; $T_2 = 0$) with the semi-axes length $3S_{T1}$ and $3S_{T2}$ respectively (Fig. 6a [11]). Further, the proper positions in relation to this center are calculated for each molecule (including molecules from the prediction set). If a work set molecule does not correspond to the DA criteria, it is so-called “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 by the corresponding QSAR model is less reliable (model extrapolation). Naturally, the prognoses for molecules approximated to the center of the DA are the most reliable.

The second approach—the DA rectangle—has been also developed by us [11]. Two extreme points (so-called virtual activity and inactivity etalons, respectively) in a space of structural features are determined. The first one has maximal values of descriptors (training set data) promoting activity and minimal-interfering. This point corresponds to a hypothetical molecule—peculiar activity etalon. The second point, analogically, is an inactivity etalon, i.e. it contains maximal values of the interfering activity of

descriptors and minimal-promoting activities. The vector that unites these points (directed from inactive to active one) depicts the tendency of activity change in the space of the variables. This vector is a diagonal for the rectangle that determines DA (Fig. 6a [11]). All of the mentioned trends associated with the “influential” points from the training set and model extrapolation for new molecules from the prediction set remain in the DA rectangle approach.

The third method is based on the estimation of leverage value h_I [53]. It has been visualized as a Williams plot [54] (see Fig. 6b) and is described in detail in [55]. 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 a training set of molecules, high leverage values do not always indicate outliers for the model, i.e. points that are outside of 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 that do not fit the model (i.e. have large residuals) are called “bad high leverage points” or bad influence points. They destabilize the model [55]. The new molecule is situated outside of the DA (model extrapolation) if $h_I > h_{cr} = 3(A + 1)/M$, where A is the number of the PLS latent variables and M is the number of molecules in a work set.

In summary, it is necessary to note that if a new structure is lying inside the DA, it is not a final and definitive proponent of the 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 a rejection of the prediction; it is just an indication of the increased uncertainty of the prediction. Naturally, such compounds can be predicted (by model extrapolation) with the great accuracy, but it will be more of an accident than a regular success. Unfortunately, there is currently no unbiased estimation of prognosis reliability, and the relative character of any DA procedure is reflected in [11, 55]. Thus, one should remember that the DA is not a guide to action but only a probable recommendation.

HiT QSAR: comparison with popular QSAR methods

The HiT QSAR based on SiRMS has proved to be efficient in numerous studies for solving different “structure–activity/property” problems [3, 10–12, 26–33]; as such, it was assessed interesting to compare it with the other successful QSAR approaches and software. The results of such theoretical comparative analysis are shown in Table 1.

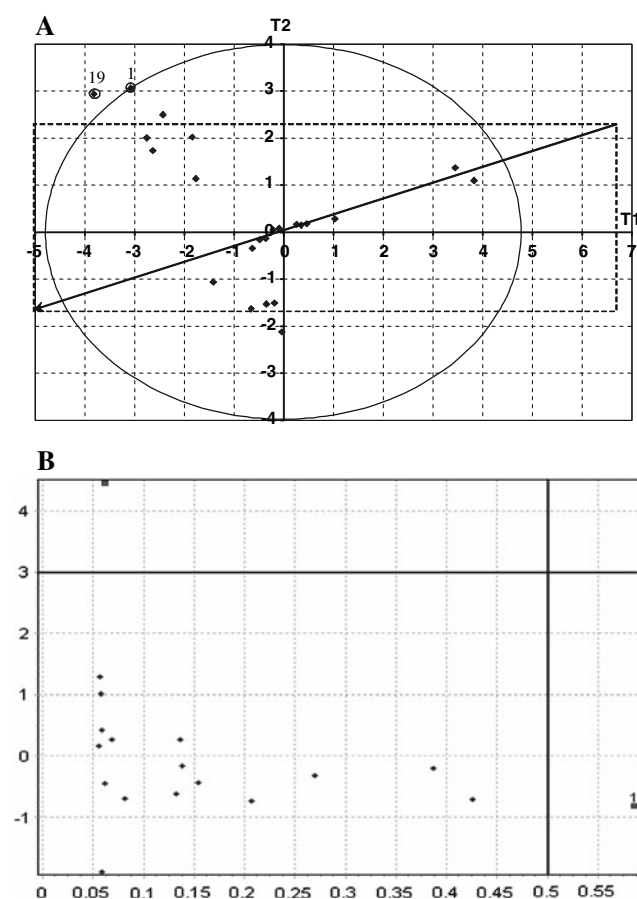


Fig. 6 Domain applicability ellipsoid and rectangle (a) and Williams plot (b) for the HRV-2 QSAR model [11]

Obviously, the HiT QSAR does not have a problem of optimal alignment of the set of the considered molecules that are inherent to the CoMFA and its analogues [16–19]. The SiRMS approach is similar to the HQSAR approach [20] in certain ways but has none of its restrictions (only topological representation of molecular structure) and deficiencies (ambiguity of descriptor formation during the procedure of molecular holograms hashing). In addition, contrary to the HQSAR, different physical and chemical properties of atoms (charge, lipophilicity, etc.) can be taken into account in SiRMS (Table 2).

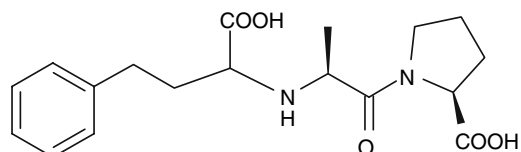
Thus, the main advantages of the HiT QSAR are:

- the usage 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 ways (rules) for molecular design.

Angiotensin converting enzyme (ACE) inhibitors

After such theoretical comparative analysis it was logical to test the efficiency of the proposed HiT QSAR on real

representative sets of compounds. All such sets contain only structurally similar compounds to avoid the “molecular alignment” problem and, therefore, to facilitate the usage of the “lattice” approaches (CoMFA and CoMSIA). A total of 114 ACE inhibitors [56] represent the first set. Different statistic models obtained by HiT QSAR have been compared with ones published by Sutherland et al. [56]. The structure of enalaprat, a representative compound from the ACE dataset, is displayed as follows:



The ability to inhibit ACE (pIC_{50}) has been investigated. The training set consists of 76 compounds, and 38 structures were used in a test set [56]. We have compared the resulting PLS models built with the use of descriptors generated in the following QSAR approaches:

(1) CoMFA, comparative molecular fields analysis [16]; (2) CoMSIA, comparative molecular similarity indexes analysis [18]; (3) EVA, eigenvalue analysis [57]; (4) HQSAR, hologram QSAR [20]; (5) CERUS 2 program package (Accelrys, San Diego, CA), method of

Table 2 The comparison of different QSAR methods

| Criterion | | QSAR methods | | | |
|---|--|--------------|------------------------------|-------------------|--------|
| | | 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 consideration of stereochemistry 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 |

Table 3 Statistical characteristics of the QSAR models obtained for ACE and AchE datasets 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 ^a | 0.80 | 0.88 | 0.68 | 0.52 | 0.49 | 0.47 | 1.04 | 0.41 | 1.54 | 0.95 | 3 | 5 |
| CoMSIA(basic) ^a | 0.76 | 0.86 | 0.65 | 0.45 | 0.52 | 0.44 | 1.15 | 0.45 | 1.48 | 0.98 | 3 | 6 |
| CoMSIA(extra) ^a | 0.73 | 0.86 | 0.66 | 0.46 | 0.49 | 0.44 | 1.22 | 0.45 | 1.53 | 0.98 | 2 | 4 |
| EVA ^a | 0.84 | 0.96 | 0.70 | 0.41 | 0.36 | 0.28 | 0.93 | 0.23 | 1.72 | 1.11 | 4 | 4 |
| HQSAR ^a | 0.84 | 0.72 | 0.72 | 0.33 | 0.30 | 0.37 | 0.95 | 0.64 | 1.80 | 1.01 | 4 | 5 |
| Cerius2 ^a | 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 ^b 2D | 0.86 | 0.81 | 0.80 | 0.69 | 0.75 | 0.74 | 0.9 | 0.53 | 1.07 | 0.67 | 2 | 2 |
| Mix ^b 3D | 0.90 | 0.89 | 0.88 | 0.84 | 0.85 | 0.82 | 0.74 | 0.4 | 0.83 | 0.56 | 2 | 2 |

R^2 , Determination coefficient for training set; Q^2 , cross validation (leave 10% out) determination coefficient; R^2_{test} , determination 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; EVA, eigenvalue analysis QSAR approach; ACE, angiotensin converting enzyme; AchE, acetylcholinesterase

^a Statistic characteristics from Sutherland et al. [56] is shown

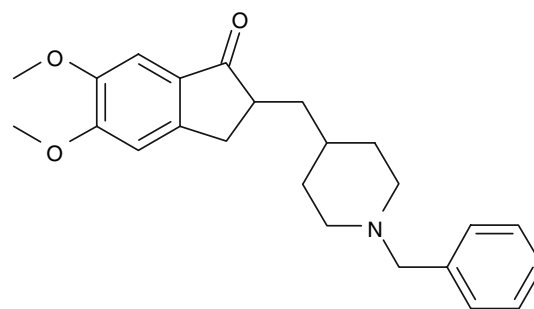
^b Mix, Simplex + Fourier descriptors

traditional integral (whole-molecule) 2D and 2.5D² descriptor generation; (6) HiT QSAR based on SiRMS [3, 11, 24].

Because all of the approaches mentioned used generated parameters of 2D or 3D levels of molecular structure representation for comparison purposes, corresponding SD, the Fourier parameters and united models with mixed (simplex + Fourier) parameters were taken for comparison. The advantage of the HiT QSAR over other methods is revealed by a comparison of such statistical descriptions of the QSAR models as the determination coefficient for training (R^2) and test (R^2_{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 3). For example, the SiRMS $Q^2 = 0.81$ – 0.87 ; for the Fourier models, $Q^2 = 0.73$ – 0.80 ; for the other methods, $Q^2 = 0.65$ – 0.72 . It is necessary to note that the transition to the 3D level allows the quality of the obtained QSAR models to be improved. 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 simplex ones, but because of the presence of integral parameters, the former are different enough to allow the user to cast a glance from the other side on the property being investigated.

Acetylcholinesterase (AchE) inhibitors

The second set used for the comparative analysis consists of 111 AchE inhibitors. The structure of E2020, a representative compound from the AchE dataset, is as follows:



The ability to inhibit AchE (pIC_{50}) is investigated. The training set consists of 74 compounds, and 37 structures were used as a test set [56]. The methods compared and the principles of comparison are similar to the ones described above. The main trends revealed on the ACE set remain correct for the AchE inhibitors. The advantage of the HiT QSAR over other methods are observed on each statistical parameter (see Table 3), but especially on predictivity of the models: for the SiRMS, $R^2_{\text{test}} = 0.74$ – 0.82 ; for the Fourier models, $R^2_{\text{test}} = 0.59$ – 0.61 ; for the other methods, $R^2_{\text{test}} = 0.16$ – 0.47 . As in the previous case, the consideration of the spatial structure of investigated compounds improved the quality of the obtained models.

² This classification is offered by the authors of CERIUS 2.

In summary, it is necessary to note that we understand that the advantage of the SDs generated in HiT QSAR may be partially caused by some differences in the statistical approaches applied (e.g. in addition to GA, TV and AVS procedures have been used by us); however, these mathematical tricks can not cause all of the gap between the approaches investigated. Thus, it is obvious from the obtained results that the HiT QSAR simplex models are well-fitted, robust and, in particular, they are much more predictive than QSAR models developed by other approaches.

Hit QSAR application

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

One-dimensional level

*Nitroaromatics toxicity in vivo*³

One-dimensional QSAR models based on the SiRMS have been applied to predict the oral toxicity in vivo of 28 nitroaromatic compounds, including some well-known explosives. The 50% lethal dose concentration for rats (LD_{50}) was used as the estimation of toxicity in vivo. Despite the fact that SiRMS has proved to be efficient for solving different “structure–activity” problems, 1D models usually have an auxiliary role only, and in this study the first attempt to use them separately as a virtual screening tool has been made. The results of 1D QSAR analysis show that even information on the composition of molecules provides the main tendencies of toxicity changes. Our aim was to describe the molecules investigated on the 1D level in more detail using the reliable ND QSAR model, where $1 < N < 2$ and considering explicitly the presence of different substituents by their representation, such as one pseudoatom being obtained and validated by three different test sets (principles of test set formation have been described above). The necessity of considering the mutual influences of the substituents for the development of adequate QSAR models of nitroaromatics toxicity was demonstrated by the failure of additive QSAR models (including Free–Wilson [58] analogue) of almost the 2D level of differentiation.

³ The authors express their sincere gratitude to Prof. J. Leszczynski, Dr. L. Gorb and Dr. M. Quasim for fruitful cooperation during the development of this task.

Statistic characteristics of the developed QSAR models, with the exception of the additive ones, are quite satisfactory ($R^2 = 0.81–0.92$; $Q^2 = 0.64–0.83$; $R^2_{\text{test}} = 0.84–0.87$). The success of such models has been caused by their non-additivity, i.e. possibility of taking into account the mutual influence of substituents in the benzene ring, which plays the determining role for toxicity change and could be mediated through the different C–H fragments of this ring. The contribution of the nitro group on toxicity substantially depends on the presence/absence of other substituents and varies from highly positive to weakly negative. Hydroxyl and fluorine increase toxicity and the presence of methyl group decreases it. Chlorine has an unambiguous influence on toxicity. Such trends are consistent with the results of 2D QSAR analysis where the structure of the molecule is taken into account. The single QSAR models that were obtained were summarized as the most adequate averaged model that allows improvements in the accuracy of toxicity prediction and shows an ability to be used as a virtual screening tool.

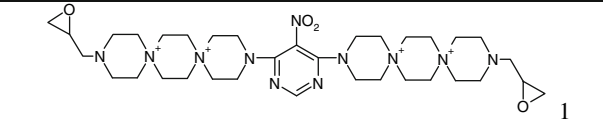
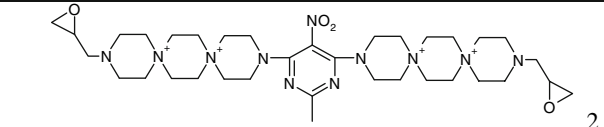
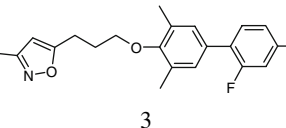
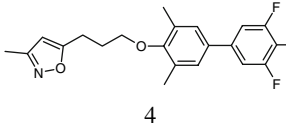
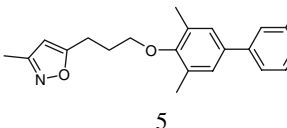
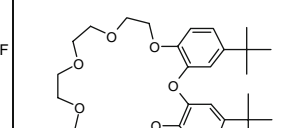
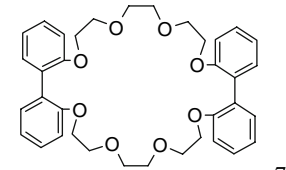
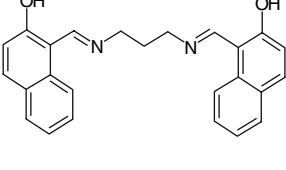
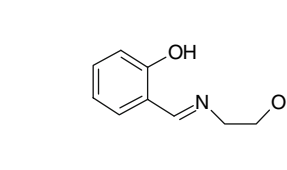
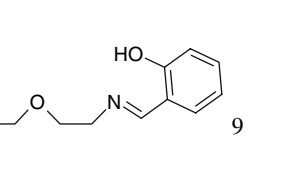
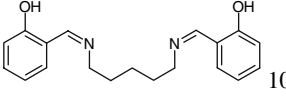
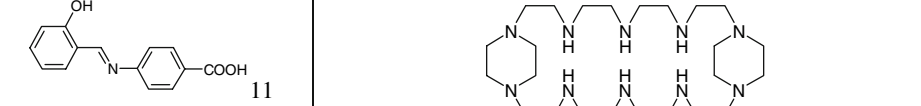
Two-dimensional level

*Antiherpetic activity of N,N'-(bis-5-nitropyrimidyl)dispirotripiperazine derivatives*⁴

HiT QSAR was applied to evaluate the influence of the structure of 48 *N,N'*-(bis-5-nitropyrimidyl)dispirotripiperazines on their anti-herpetic activity, selectivity and cytotoxicity, with the purpose of understanding the chemical–biological interactions governing their activities and designing new compounds with strong anti-viral activity [3]. 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. Spirobro-mine, a medicine with a nitrogen-containing dispiro structure possessing anti-HSV-1 activity, was included in the training set. The statistical characteristics of the 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 the anti-herpetic activity, cytotoxicity and selectivity of the new compounds to be predicted. Electrostatic factors (38%) and hydrophobicity (34%) were the most important determinants of anti-herpetic activity (Fig. 3). The QSAR results demonstrate a high impact of individual structural fragments on anti-viral activity. Molecular fragments that promote and interfere with anti-viral activity were defined on the basis of the

⁴ The authors express their sincere gratitude to Dr. M. Schmidtke, Prof. P. Wutzler, Dr. V. Makarov, Dr. O. Riabova, Mr. N. Kovdienko and Mr. A. Hromov for their most fruitful cooperation that made the development of this task possible.

Table 4 Perspective potent compounds—results of computer-assisted molecular design

| | | | |
|---|--|---|--|
|  1 | |  2 | |
|  3 |  4 |  5 |  6 |
|  7 | |  8 | |
|  9 | |  10 | |
|  11 |  12 | | |

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 dispirotrienamine leads to a complete loss of activity, while the presence of methyloxirane strongly increases it. Using the established results and regularities, several new dispirotrienamine derivatives—potential antiviral agents were computationally designed. Two of these new compounds (**1** and **2**, see Table 4) were synthesized. The results of biological the tests confirm the predicted high values of anti-viral activity and selectivity (they are about two logarithmic units more active and one order more selective than spirobromine) as well as the low toxicity of these compounds.

[(Biphenyloxy)propyl]isoxazole derivatives—human rhinovirus 2 replication inhibitors⁵

The QSAR analysis of anti-viral activity of [(biphenyl-oxo)propyl]isoxazole derivatives was developed using HiT

QSAR on the basis of SiRMS with the purpose to reveal chemico–biological interactions governing their activities as well as their probable mode of action and to design new compounds with strong anti-viral activity [11]. The common logarithms 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 for assessing cytotoxicity, anti-viral activity and selectivity, respectively. The work set consists of 18 compounds, including pleconaril as a reference compound. These have not been divided into training and test sets because of the small amounts (i.e. the structural information contained in each molecule in our case is unique and useful). Statistic characteristics for 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 enable the virtual screening and molecular design of new compounds with strong anti-HRV-2 activity. The results indicate a high influence of atom's individuality on all of the properties investigated (approx. 40%) and of electrostatic factors on selectivity (approx. 50%), where these latter factors along with atom individuality play the determining role together with hydrophobic interactions on the anti-viral activity

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

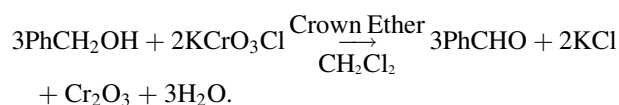
(approx. 40%). The presence of terminal 5-trifluoromethyl-1,2,4-oxadiazole and *p*-fluorophenyl fragments in the molecule leads to a strong enhancement of its useful properties, i.e. increase of activity towards HRV-2 as well as selectivity and decrease of cytotoxicity. An additional terminal aromatic ring—naphthalene or phenyl—strongly decreases activity towards HRV-2 and, to a lesser degree, the SI. The virtual screening and molecular design of new well-tolerated compounds with strong anti-HRV-2 activity have been carried out based on these QSAR results. Three different DA approaches (DA rectangle and ellipsoid as well as leverage) give nearly the same (however, not very correct, it was shown that model extrapolation does not lead to wrong predictions, i.e. real models' DA is wider than calculated) results for each QSAR model and allow us to additionally estimate the quality of the prognosis for all of the compounds designed. A hypothesis to the effect that an external benzene substituent must have negative electrostatic potential and definite length *L* (approximately 5.5–5.6 Å) to possess strong anti-viral activity has been suggested. Most probably, the fluorine atom in the *para*-position of the terminal aromatic ring (compound **3–5**, Table 4) is quite complementary (*L* = 5.59 Å) to receptor cavity for such an interaction. It is necessary to note that the molecule of pleconaril (*L* = 5.54 Å) satisfies completely the indicated criteria. In the case of nitroaromatics, the accumulation of nitro groups in the region of the receptor cavity will lead to a strengthening of the electrostatic interactions with the biological target and, therefore, to an increase in activity.

Several new compounds have been computationally designed and predicted to be highly active and selective. Three of these (**3–5**, Table 4) were synthesized. Experimental testing revealed a strong coincidence between the experimental and predicted anti-HRV-2 activity and SI. Compounds **3–5** are similar based on their cytotoxicity level to pleconaril, but they are more active and selective than the latter compound.

Three-dimensional level

*Catalytic activity of crown ethers*⁶

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:



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

The distinctive feature of this study is the absence of any reliable relationship between the topological structure of crown ethers and their catalytic properties. On the 4D level, a not very robust ($Q^2 = 0.46$) relationship was obtained and only on the 3D level, after the selection of the most acceptable fragment for the formation of complexes with potassium conformations, was the most reliable option for this task model ($R^2 = 0.87$; $Q^2 = 0.66$) developed. The slight preference of “transoid” on *cis*-conformations of crown ethers containing mentioned fragments together with the positive effect of biphenyl and diphenyloxide fragments on the catalytic activity of investigated compounds were shown. The undesirability of the cyclohexyl fragment was determined as well as the certain limits of crown ether dentacy (**4–8**). These facts as well as the predominant role of electrostatic factors in the investigated process (approx. 50%) completely correspond to the known mechanisms of catalytic action of the crown ethers. Two potent catalysts, **6** and **7** (Table 4), were designed and introduced as a result of the realized QSPR analysis.

Fourth-dimensional level

*Anticancer activity of macrocyclic Schiff bases*⁷

Our investigation of the influence of the molecular structure of macrocyclic Schiff bases on their anti-cancer 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. Anti-cancer activity was expressed as the percentage of the corresponding cell growth. The training set is very structurally dissimilar and consists of 30 macrocyclic pyridinophanes, their analogues and a number of other compounds.

The usage of simple topologic models allows the anti-cancer activity of macrocyclic pyridinophanes to be

⁶ The authors express their 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.

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

described for only five cell cultures [59]. The consideration of spatial structure improves the situation, but only on the 4D level were reliable QSAR models ($R^2 = 0.74–0.98$; $Q^2 = 0.54–0.84$), obtained for all of the cells investigated (except leukemia, where $Q^2 < 0.5$; however, even in this case, the designed compound was predicted correctly), and the averaged activity (most of lines and cells are highly correlated) indicates the importance of not the most active or favorable single conformer but of the set of interacting conformers within the limits of the 3 kcal/mol energy gap. The presence of the N^1,N^3 -dimethylenepropane-1,3-diamine fragment strongly promotes the anti-cancer activity. This fragment was used as a linker between two naphthalene-2-oles (compound **8**, Table 4), which leads to the creation of a universal anti-cancer agent active against all mentioned tumors except prostate cancer. It is necessary to note that the usage of SiRMS allows researchers to start from 12 macrocyclic pyridinophanes [59] in searching for anti-cancer agents, ultimately coming to the symmetric open-chained aromatic compounds connected by the linker mentioned above [10].

Two-dimensional–four-dimensional levels

*Anti-influenza activity of macrocyclic pyridinophanes*⁸

All of the advantages of the HiT QSAR were demonstrated during the investigation of anti-influenza activity on the dataset possessing the essential structural variety: different macrocyclic pyridinophanes, their acyclic analogues and well-known anti-viral agents (deiteforin, remantadine, ribavirine, ambenum and other) [12, 30]. Anti-influenza activity (virus A/Hong Kong/1/68 (H3N2)) was expressed in IgTID₅₀ and reflected the suppression of viral replication in “experimental” samples in comparison with “control” ones. Investigated structures were divided between training and test sets (25 and 6 compounds, respectively).

In accordance to hierarchical principles of offered approach, the QSAR task was solved sequentially on the 2D, 4D and 3D levels.⁹ The set of QSAR models with different adequacy levels (2D, 4D, 3D) was obtained as a result of investigations. All of the obtained QSAR models were well fitted, robust and predictive ($R^2 = 0.94–0.98$,

$Q^2 = 0.85–0.95$, $R^2_{\text{test}} = 0.98–0.99$)¹⁰ and have defined DA and clear mechanistic interpretation. For 3D QSAR investigations, the set of “productive” conformers was used. These were determined as the most active ones by the results of 4D QSAR modeling. The results indicate a high impact of atom individuality on the variation of anti-influenza activity (37–50%). Hydrophobic/hydrophilic and electrostatic interactions also play an important role (15–22%). The shape of the molecule (4D and 3D models) also affect the anti-influenza activity, but this property has the smallest influence (11 and 16%, respectively). The cylindrical form of the molecules ($I_X/I_Y \rightarrow 1$) with a small diameter ($I_Y \rightarrow \min$) promotes anti-influenza activity. Molecular fragments governing the change of anti-influenza activity and their average relative influences (Table 1) were determined. For example, the presence of oxyethylene or 2-iminomethylphenol fragments promotes an anti-viral activity, and aminoethylene fragments decreases it.

Table 4) with adjusted activity levels was developed by the obtained results. High levels of all predicted (all of the resulting 2D–4D models show the strong coincidence of predictions) values of anti-influenza activity were confirmed experimentally. Thus, during the QSAR investigations [12, 30], the search for active compounds began with the macrocyclic pyridinophanes and ultimately came to benzene derivatives containing 2-iminomethyl-phenol fragment (**9–11**, Table 4).

Anti-herpetic activity of macrocyclic pyridinophanes (see footnote 7)

The dataset for anti-herpetic research was almost identical to the anti-influenza task and was also characterized by essential structural variety: different macrocyclic pyridinophanes and their acyclic analogues plus well-known anti-viral agents, including acyclovir as a reference compound. Anti-herpetic activity against HSV-1 strain US was expressed in percentage of inhibition of HSV reproduction in the treated cell cultures (Hep-2) in comparison with untreated ones. As in the previous case, our anti-herpetic research has a multi-step cyclic character: synthesis, biological tests, QSAR analysis, virtual screening and computer-assisted drug design, synthesis, etc. [26, 29, 30, 42]. Fourteen compounds (mostly macrocyclic pyridinophanes and their acyclic analogues) were initially investigated for their anti-herpetic activity [30]. At the present stage [26], after several QSAR convolutions, 37

⁸ The anti-influenza and antiherpetic investigations described were carried out as a result of fruitful cooperation with Dr. V.P. Lozitsky, Dr. R.N. Lozyska and 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 their sincere gratitude to all of these colleagues.

⁹ In this and anti-herpetic research 1D modeling was not carried out.

¹⁰ We are aware that these models can approximate not only the variation in activity but also the variation in experimental errors. The high values of the R^2_{test} can be explained by the fact that test compounds are very similar to the training ones, that there are only few compounds in test set, by the high quality of the obtained models, by simple good luck and/or by a combination of all these factors.

compounds have been divided between the training and test sets (26 and 11 compounds, respectively) and the set of QSAR models with the different adequacy levels (2D, 4D, 3D) has been obtained as a result of the investigations. All of the obtained QSAR models were well fitted, robust and predictive ($R^2 = 0.82\text{--}0.90$, $Q^2 = 0.60\text{--}0.65$, $R^2_{\text{test}} = 0.70\text{--}0.78$) and have defined DA and a clear mechanistic interpretation. For 3D QSAR investigations, the set of “productive” conformers was used that were determined to be the most active by the results of 4D QSAR modeling.

All of the developed models (2D–4D) indicate a high impact of hydrophobic (approx. 50%) and electrostatic (approx. 20%) factors on the variation of anti-herpetic activity. The strong promotion of anti-herpetic 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 linked with an aliphatic fragment. There was a tendency for anti-viral activity to increase with the strengthening of the acceptor properties of the compound’s aromatic rings. This information was used for designing drugs with a potent anti-herpetic agent, **12** (Table 4). The usage of the SiRMS allows researchers to start their search for new anti-herpetic agents from macrocyclic pyridinophanes [30], ultimately coming to the symmetric piperazine containing macroheterocycle 1,4,7,10,13,16,19,22,25,28-decaaza-tricyclo[26.2.2.2*13,16*]-tetratriacontane (**12**).

Conclusion

A hierarchical technology for QSAR investigations has been developed based on the Simplex representation of the molecular structure. The main advantages of this technology are:

- absence of “molecular alignment” problem;
- usage of molecular models with different adequacy degrees;
- consideration of the different physical and chemical properties of atoms;
- defined endpoints, unambiguous algorithm, defined applicability domain, easiness of mechanistic interpretation, robustness and predictivity of the developed QSAR models;
- clear ways for molecular design.

The comparative analysis of the HiT QSAR with the most popular modern QSAR approaches reflects its advantages, especially in terms of predictivity.

The efficiency of the HiT QSAR was demonstrated on various QSAR/QSPR tasks at different (1D–4D) levels of molecular modeling.

In summary, we consider the HiT QSAR that we have developed and which is reported here to be an effective instrument for the computer-assisted molecular (drug) design of new compounds and materials with the complexity of desired properties. HiT QSAR has been realized as the complex of computer programs “HiT QSAR” software, which also includes a powerful statistical block and a number of useful utilities.

Acknowledgments This work was partially supported by a grant of the President of Ukraine (President of Ukraine grant for young investigators GP/F11/0115), the Science & Technology Center in Ukraine (STCU project 3147) and INTAS foundation (INTAS Grant 97-31528).

References

1. Ooms F (2000) *Curr Med Chem* 7(2):141
2. Thomas G (2008) *Medicinal chemistry: an introduction*, 2nd edn. Wiley
3. Artemenko AG, Muratov EN, Kuz'min VE, Kovdienko NA, Hromov AI, Makarov VA, Riabova OB, Wutzler P, Schmidtke M (2007) *J Antimicrob Chemother* 60(1):68
4. Bailey TR, Diana GD, Kowalczyk PJ, Akullian V, Eissenstat MA, Cutcliffe D, Mallamo JP, Carabateas PM, Pevear DC (1992) *J Med Chem* 35(24):4628
5. Butina D, Gola JMR (2004) *J Chem Inf Comput Sci* 43:837
6. de Jonge MR, Koymans LM, Vinkers HM, Daeyaert FF, Heeres J, Lewi PJ, Janssen PA (2005) *J Med Chem* 48(6):2176
7. Jenssen H, Gutteberg TJ, Lejon T (2005) *J Pept Sci* 11(2):97
8. Kovatcheva A, Golbraikh A, Oloff S, Xiao Y, Zheng W, Wolschann P, Buchbauer G, Tropsha A (2004) *J Chem Inf Comput Sci* 44:582
9. Kubinyi H (1990) *J Cancer Res Clin Oncol* 116:529
10. Kuz'min VE, Artemenko AG, Lozitska RN, Fedtchouk AS, Lozitsky VP, Muratov EN, Mescheriakov AK (2005) *SAR QSAR Environ Res* 16(3):219
11. Kuz'min VE, Artemenko AG, Muratov EN, Volineckaya IL, Makarov VA, Riabova OB, Wutzler P, Schmidtke M (2007) *J Med Chem* 50:4205
12. Muratov EN, Artemenko AG, Kuz'min VE, Lozitsky VP, Fedchuk AS, Lozitska RN, Boschenko YA, Gridina TL (2005) *Antiviral Res* 65(3):A62
13. Verma RP, Hansch C (2006) *Curr Med Chem* 13(4):423
14. Zhang S, Golbraikh A, Tropsha A (2006) *J Med Chem* 49:2713
15. Selassie CD (2003) In: Abraham DJ (ed) *Burger's medicinal chemistry and drug discovery*, 6th edn, vol 1. Wiley, New York
16. Cramer RD, Patterson DI, Bunce JD (1988) *J Am Chem Soc* 110:5959
17. Doweiko AM (1988) *J Math Chem* 31:1396
18. Klebe G, Abraham U, Mietzner T (1994) *J Med Chem* 37:4130
19. Kuz'min VE, Artemenko AG, Kovdienko NA, Tetko IV, Livingstone DJ (2000) *J Mol Model* 6:517
20. Seel M, Turner DB, Willett P (1999) *QSAR* 18:245
21. Pavan M, Consonni V, Gramatica P, Todeschini R (2006) In: *Partial order in environmental sciences and chemistry*. Springer, Berlin, pp 181–217
22. Baurin N, Mozziconacci JC, Arnoult E, Chavatte P, Marot C, Morin-Allory L (2004) *J Chem Inf Model* 44(1):276
23. Vedani A, Dobler M (2000) *Prog Drug Res* 55:107

24. Kuz'min VE, Artemenko AG, Polischuk PG, Muratov EN, Hromov AI, Liahovskiy AV, Andronati SA, Makan SY (2005) *J Mol Model* 11(6):457
25. Kuz'min VE, Artemenko AG, Muratov EN, Lozitsky VP, Fedchuk AS, Lozitska RN, Boschenko YA, Gridina TL (2005) *Antiviral Res* 65(3):A70
26. Artemenko A, Kuz'min V, Muratov E, Fedchuk A, Lozitsky V, Gridina T, Lozyska R, Basok S, Chikhichin D (2007) *Antiviral Res* 74:A76
27. Artemenko A, Muratov E, Kuz'min V, Fedtchuk A, Mykhaylovska N, Lesyk R, Zimenkovsky B (2006) *Clin Microbiol Infect* 12(4):1557
28. Artemenko A, Muratov E, Kuz'min V, Koroleva L, Silnikov V, Lozitsky V, Fedchuk A (2006) *Antiviral Res* 70:A43
29. Artemenko AG, Kuz'min VE, Muratov EN, Lozitsky VP, Fedchuk AS, Lozitska RN, Boschenko YA, Gridina TL (2005) *Antiviral Res* 65(3):A77
30. Kuz'min VE, Artemenko AG, Lozitsky VP, Muratov EN, Fedtchouk AS, Dyachenko NS, Nosach LN, Gridina TL, Shitikova LI, Mudrik LM, Mescheriakov AK, Chelombitko VA, Zheltvay AI, Vanden Eynde J-J (2002) *Acta Biochim Pol* 49:157
31. Kuz'min VE, Artemenko AG, Muratov EN, Volineckaya IL, Makarov VA, Riabova OB, Wutzler P, Schmidtke M (2007) *Antiviral Res* 74:A49
32. Muratov E, Artemenko A, Kuz'min V, Konup I, Konup L, Kotlyar S, Kamalov G, Fedtchuk A, Mykhaylovska N (2006) *Clin Microbiol Infect* 12(4):1558
33. Muratov EN, Kuz'min VE, Artemenko AG, Makarov VA, Riabova OB, Wutzler P, Schmidtke M, Lozitsky V, Fedchuk A (2006) *Antiviral Res* 70:A77
34. 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.; 49; Organisation for Economic Co-operation and Development, Paris
35. Kuz'min VE (1995) *Zh Strucur Khim* 36:873
36. Jolly WL, Perry WB (1973) *J Am Chem Soc* 95:5442
37. Wang R, Fu Y, Lai L (1997) *J Chem Inf Comput Sci* 37:615
38. Ioffe BV (1983) *Chemistry refractometric methods*, 3rd edn. Himiya, Leningrad
39. Burkert U, Allinger N (1982) *Molecular mechanics*. ACS Publication, Washington D.C.
40. Marple SL Jr (1987) *Digital spectral analysis with applications*. Prentice-Hall, New York
41. Östergard PRJ (2002) *Discrete Appl Math* 120:195
42. Muratov EN (2004) Quantitative evaluation of the structural factors influence on the properties of nitrogen-, oxygen- and sulfur-containing macroheterocycles. A.V. Bogatsky Physical-Chemical Institute, Odessa
43. Bodor N, Buchwald P (2000) *Med Res Rev* 20(1):58
44. Cronin MTD, Schultz TW (2003) *J Mol Struct (Theochem)* 622:39
45. Kubinyi H (1996) *J Chemometr* 10:119
46. Lindgren F, Geladi P, Rannar S, Wold S (1994) *J Chemometr* 8:349
47. Rannar S, Lindgren F, Geladi P, Wold S (1994) *J Chemometr* 8:111
48. Rogers D, Hopfinger AJ (1994) *J Chem Inf Comput Sci* 34(4):854
49. Carhart RE, Smith DH, Venkataraghavan R (1985) *J Chem Inf Comput Sci* 25:64
50. Vitiuk NV, Kuz'min VE (1994) *Zh Anal Khim* 49:165
51. Wold S, Antti H, Lindgren F, Ohman J (1998) *Chemometr Intell Lab Syst* 44:175
52. Trygg J, Wold S (2002) *J Chemometr* 16:119
53. Neter J, Kutner MH, Wasseman W, Nachtsheim C (1996) *Applied linear statistical models*. McGraw-Hill, New York
54. Meloun M, Militku J, Hill M, Brereton MG (2002) *Analyst* 127:433
55. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T (2005) *Altern Lab Anim* 33:445
56. Sutherland JJ, O'Brien LA, Weaver DF (2004) *J Med Chem* 47:5541
57. Heritage TV, Ferguson AM, Turner DB, Willett P (1998) *Perspect Drug Discovery Des* 11:381
58. Free SM, Wilson JW (1964) *J Med Chem* 7:395
59. Kuz'min VE, Lozitsky VP, Kamalov GL, Lozitskaya RN, Zheltvay AI, Fedtchouk AS, Kryzhanovsky DN (2000) *Acta Biochim Pol* 47:867