Russian Chemical Reviews **78** (5) 465 – 483 (2009) © 2009 Russian Academy of Sciences and Turpion Ltd

DOI 10.1070/RC2009v078n05ABEH004030

Nonlinear mapping techniques for prediction of pharmacological properties of chemical compounds

Ya A Ivanenkov, E V Bovina, K V Balakin

Contents

I.	Introduction	465
II.	Modern technologies and the value of virtual screening	466
III.	Analysis of the chemical space using different mapping methods	468
IV.	The use of the SPE algorithm for analysis of the chemical space	480
V.	Specialized software	481
VI.	Conclusion	482

Abstract. The review concerns the state of the art in the development and application of a special group of data analysis methods called mapping algorithms, which allow one to reduce the dimension of multidimensional spaces of the properties of objects under study (series of chemical compounds) down to corresponding two- or three-dimensional projections (images) with retention of the initial space topology. It is shown how can such maps be used to visually analyze complex structure – property relationships in order to predict the pharmacologically significant properties of chemical compounds. The bibliography includes 86 references.

I. Introduction

Intensive development of methods of organic synthesis in the 20th century led to significant increase in the number of novel chemical compounds and qualitative improvements. The latter manifests itself in the possibility of synthesizing almost all existing compounds, including complex molecules typical of living matter, provided the necessary resources are at hands. In particular, the colossal potential of the available synthetic approaches is indicated by recent syntheses of a huge number of natural compounds.[1] An increase in the number of compounds synthesized is first of all due to the advances in combinatorial chemistry.[2, 3] Recently, considerable progress has been made in this relatively young discipline. Combinatorial chemistry is first of all aimed at discovery of novel drugs. It allows one to generate large series ('libraries') of complex structures in a time-and-material-economic manner. Combinatorial chemistry methods are continuously improved, accumulating the advances of traditional methods of organic synthesis. As a consequence, many compounds with complex structures (*e.g.*, molecules with chiral centres, antibiotics, peptides, oligonucleotides), which have recently been synthesized only by classical methods of organic chemistry, can now successively be obtained using high-throughput combinatorial synthesis schemes.

At present, combinatorial synthesis has become a key method in organic chemistry and related disciplines. When combined with high-throughput screening techniques, it serves as a basis for many research programmes in chemical pharmaceutical and agrochemical industries. However, recently, it has become clear that unrestricted increase in the size of the synthesized and tested 'libraries' of compounds without preliminary rational selection of candidate compounds does not lead to intensification of research programmes. Clearly, the synthesis of complete series of compounds and their subsequent testing are extremely labour-consuming and economically undesired.[4, 5] These processes become even more complicated in the case of multistep syntheses, multicomponent reactions and concurrent combinatorial schemes. Therefore, even in the relatively simple cases, researchers should use methods of rational pre-synthetic selection of candidate compounds. These methods allow one to significantly reduce the sizes of the sets of compounds possessing the desired properties. This approach is necessary for successful realization of many research programmes.

The need in pre-synthetic analysis of 'libraries' of virtual compounds to perform rational selection of best candidate structures for solving particular problems led to emergence of a new branch of science, namely, computer screening (also known as virtual screening). This relatively new discipline becomes increasingly more practically valuable.[4 – 6] Although systematic studies on computer simulation of the properties of organic compounds date back to the 1950s,[7] it is nowadays that this approach plays the role of a key tool in experimental synthesis design.

Ya A Ivanenkov, E V Bovina, K V Balakin Institute of Physiologically Active Compounds, Russian Academy of Sciences, pr. Severnyi 1, 142432 Chernogolovka Moscow Region, Russian Federation.
Fax (7-496) 524 95 08, tel. (7-496) 524 25 63,
e-mail: smart – people@inbox.ru (Ya A Ivanenkov), bovina_e@ipac.ac.ru (E V Bovina), balakin@ipac.ac.ru (K V Balakin)

Received 29 January 2009
Uspekhi Khimii **78** (5) 503 – 523 (2009); translated by A M Raevskiy

Virtual screening of compound libraries involves a number of computational methods and algorithms, which make it possible to predict many fundamental characteristics of molecules, including various pharmacokinetic and physicochemical parameters, and to study various interactions in 'ligand – biological target' systems.

II. Modern technologies and the value of virtual screening

As applied to novel drug design, the practical value of the methods underlying virtual screening is the possibility of assessment of the potential action of low-molecular-mass organic compounds (candidate drugs) on various protein targets involved in the emergence and development of a disease. It is this action that causes a desired therapeutic effect. Computer models described in this review can to the full extent be treated as methods of 'virtual screening' in the sense accepted in chemical pharmaceutical industry.[5,8]

The most important modern approaches to virtual screening are shown in Fig. 1. All of them can to some extent conditionally be divided into two large groups.[9] One of them includes methods of structure-based design, which allow one to design novel molecules using the structures of known active compounds. The other group includes methods of target-based design; here the starting point is the structure of the biological target.

Different sampling algorithms for selection of compounds based on their structural similarity to known drugs include the search using the structural fragment,[10] using the similarity coefficient of small structural fragments of molecules (so-called 'fingerprints'),[11,12] and the search using bioisosteric[13–16] analogues or compounds having similar profiles of 'multilevel neighbourhood of atoms'.[17–19] A fundamental drawback of the methods based on structural similarity is a trend towards known chemical types of compounds.

Modern, more powerful, methods of the first group include various algorithms of multiparametric data analysis, such as the Kohonen self-organizing maps,[20] the Sammon nonlinear maps,[21] artificial neuron networks,[22,23] the support vector machines (SVM),[24] hierarchical clustering algorithms, branching algorithms, genetic algorithms, as well as various regressions and correlations. Unlike the direct methods of structural analysis, the algorithms mentioned above are based on the search for implicit relations in the 'molecular descriptors – property(activity)' system. Such approaches are quite accurate, efficient and powerful.

The novel drug search techniques widely used in modern research projects are based on evaluation of their three-dimensional (3D) structural similarity to particular known, physiologically active molecules.[25] 3D-Pharmacophore simulation is also widely used.[26,27] Here, the search for the structures possessing specific types of activity is based on the 3D-pharmacophore hypothesis, according to which the 3D models are selected by their structural similarity to the known pharmacophore elements, such as hydrogen-bond donors or acceptors, regions with the total positive or negative charge, hydrophobic surfaces, *etc*. These methods allow one to design novel structural chemotypes based on the known active substances, but are computationally expensive.

The second group includes a much smaller number of methods, mainly because their practical implementation requires knowledge of the structure of the biological target under study (specific features of its structure, the shape and chemical composition of the active binding site) and the specific features of the interaction between a known ligand and a particular biological target. In this group, the most widely method is 3D-molecular docking.[28–30] This technique shows no outstanding performance, but often provides a better characterization of the ligand – target interaction and makes it possible to construct 3D computer models which can then be used to predict the physiological activity of organic compounds.

Theoretically, molecular docking is the best tool for assessment of potential target-specific activity of compounds. However, at present its application in solving actual problems of virtual screening faces theoretical and practical difficulties, which makes the technique of very limited use in applied studies. Construction of a 3D model of a ligand complex with a biological target is an extremely expensive task even if one has an access to high-performance computers. In addition, the lack of information on many aspects of the ligand – target interaction (*e.g.*, the dynamics of the interaction, conformational mobility of the ligand and target molecules, the effect of solvent and other substances present in solution on the ligand – target binding, *etc*.) allows one to assume that the quality of the molecular docking-based prediction of the target-specific activity of the ligand will be only slightly better than the quality of the prediction based on the less expensive methods of two-dimensional (2D) structural similarity.[30] One should also keep in mind that at present most computer programmes ignore the conformational mobility of proteins (targets), *i.e.*, proteins are treated as stationary systems. Only minor structural and energy changes in the region of ligand – protein binding are taken into account; of course, they cannot completely characterize all aspects of interactions in such systems. The aforesaid holds also to a great extent for other methods of analysis of 3D structures, which involve preliminary determination of a 3D molecular conformation with the minimum potential energy.

This group of computer methods also includes a relatively young, but very promising branch of science called chemogenomics.[31] Unlike all other methods of virtual screening, chemogenomics first of all uses genomic information. The chemogenomics methods are based on a concept, which, in a simplified formulation, assumes that similar genes encode similar proteins that bind to the same ligand or a group of ligands. Thus, by choosing a known drug and knowing an appropriate biological target (or targets) one can select other biological cell targets, which can probably be affected by the compound under study. As a result, it is possible to elucidate the functional profile of the action of the drug and to determine intracell signal cascades and the related cell responses. This also to the full extent holds for novel active compounds under study.

An important advantage of most virtual screening methods based on the knowledge of the structure of the biological target is that they can be used when only a few ligands active towards a target in question are known, or even when such ligands are unknown at all. In the latter case, rational selection of ligands requires knowledge of the structure of the active binding site of the biological target.

Virtual screening methods allow one to solve a variety of problems. The choice of a particular algorithm of computer analysis equally depends on the availability of necessary information (if the first group of methods is used, information is retrieved from various sources) and on the specific

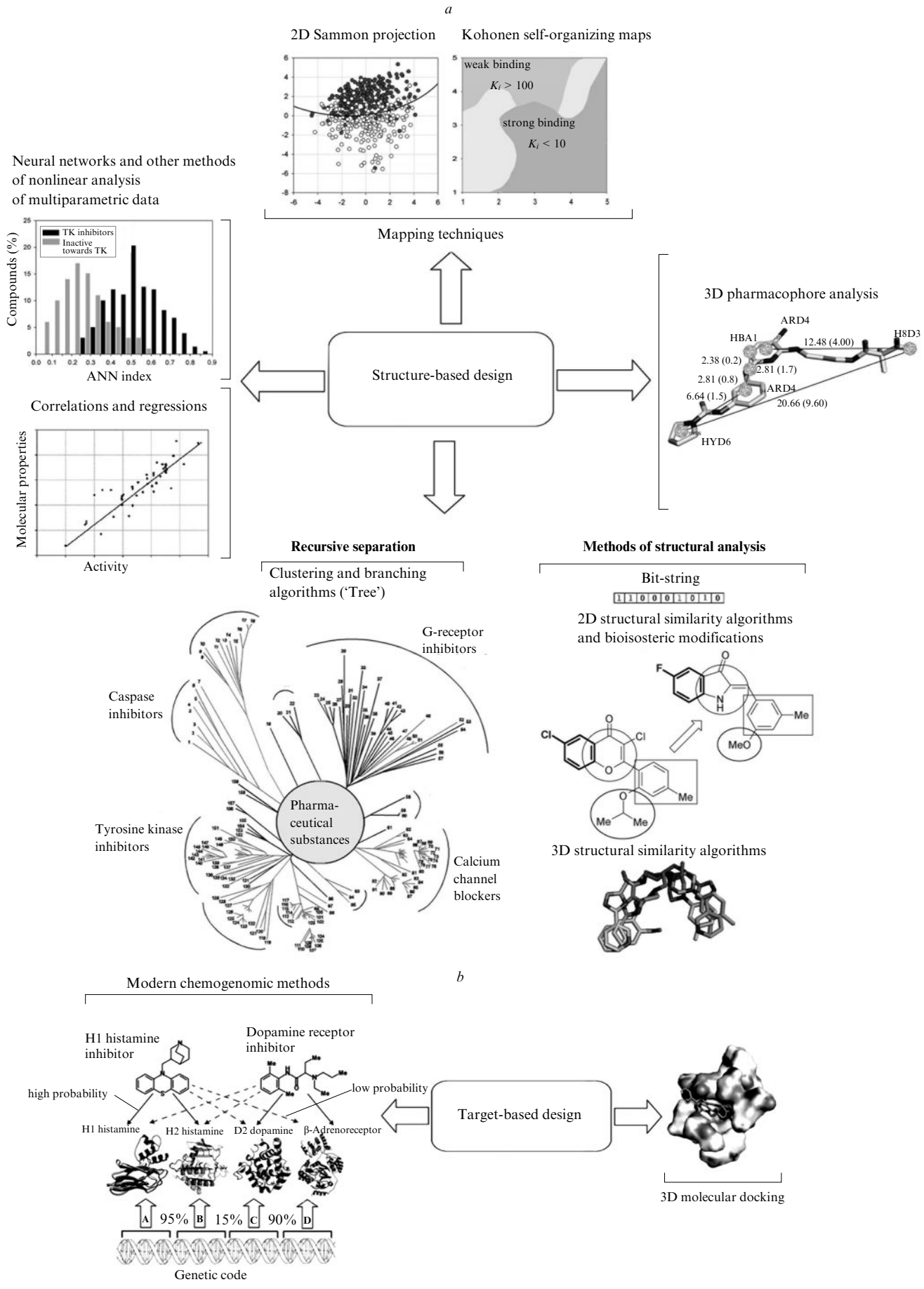
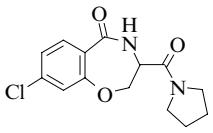
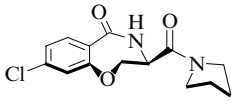


Figure 1. Virtual screening technologies.

Table 1. Classification of molecular descriptors with respect to their 'dimension'.

Category	Initial structure for calculations	Descriptors
One-dimensional (1D)	molecular formula: $C_{14}H_{15}ClN_2O_3$	molecular mass, number of carbon atoms, number of oxygen atoms, *etc.*
Two-dimensional (2D)	2D structural formula 	topological indices, log *P*, structural fragments, *etc.*
Three-dimensional (3D)	3D structure in the conformation having minimum calculated energy 	van der Waals radius, polar surface area, dipole moment, *etc.*

features of the method used. For instance, the neural networks and mapping techniques require at least two necessary factors, *viz.*, a training set and a number of theoretically or experimentally determined parameters. The former group of parameters can include molecular descriptors (Table 1), such as the logarithm of the partition coefficient in octan-1-ol – water system (log *P*), the number of hydrogen-bond donors and acceptors, the number of freely rotatable bonds, the topological and electrotopological indices, the van der Waals atomic radii, *etc*. These parameters are calculated based on the structure of a chemical compound using specialized software, which will be discussed below. Molecular descriptors are used for constructing computer models. Information necessary to form the training set can be retrieved from corresponding commercially available databases (Table 2), which usually include not only the structural formulas, but also certain important characteristics of compounds, *e.g.*, the main biological target, the stage of biological tests, as well as some molecular descriptors and experimental parameters.

Among various methods of computer-assisted screening, construction of models for the description of quantitative structure – activity or structure – property relationships (QSAR and QSPR, respectively) should be emphasized. It is the QSAR/QSPR modelling that was used to construct all prognostic systems described in this review. Generally, QSAR/QSPR modelling includes the following key stages: 1) preparation of a model experiment; 2) constructing a QSAR/QSPR model; and 3) testing and verifying the model constructed. These stages will be discussed in more detail below in the Sections concerning the mapping techniques.

III. Analysis of the chemical space using different mapping methods

Multiparametric data are analyzed using not only artificial neural networks, but also various mapping methods that make it possible to generate 2D or 3D images of multidimensional spaces of the properties of the compound sets under study. Usually, these methods are classified into a more general category of algorithms for reduction of the dimension of the multidimensional space of attributes. At present, the most often used mapping techniques include the Kohonen self-organizing maps,[21,22] the Sammon nonlinear maps[32,33] and the stochastic proximity embedding (SPE) algorithm.[34] There are also other methods suitable for generation of 2D images of the initial space of attributes, *e.g.*, clustering algorithms,[35] diffusion maps[36] and the IsoMap algorithm.[37] Below we will consider some examples of efficient application of the key mapping algorithms in the analysis of the chemical space.

1. The Kohonen self-organizing maps

The Kohonen self-organizing maps are a powerful tool of analysis and visualization of large arrays of high-dimensional multiparametric data, which makes it possible to generate 2D or 3D images of the multidimensional spaces of properties.[21] The method is based on a known learning principle 'winner-take-all' and on a specific learning rule for the Kohonen neurons. Self-organization is attained through iterative learning of each neuron in the network taking account of its activation function (often, its role is played by the Gaussian normal distribution). As a result, the Kohonen neurons are gradually ordered to form a cluster-type structure, which represents a 2D or 3D projection of the initial multidimensional space of attributes and retains, within the limits of the admissible error, its regularities and topology. This method has found increasing application in various fields of organic and medicinal chemistry including prediction of physiological and physicochemical properties of compounds (see, *e.g.*, Refs 22, 38 – 41).

Table 2. Specialized databases used in the present work to form the training sets.

Database name	Information type	Database developer
Prous Ensemble	structures and pharmacologically valuable properties of compounds	Prous Science, Spain (URL: http://www.prous.com)
Beilstein Cross-Fire	structures and physicochemical properties of compounds	Beilstein, USA (URL: http://www.beilstein.com)
MetaDrug	structures and metabolic properties of compounds	GeneGo Inc., USA (URL: http://www.genego.com)
ToxSys	structures and toxic properties of compounds	SciVision Inc. & ToxSys, USA (URL: http://www.toxsys.com)

Table 3. Typical problems in virtual screening.

Problem	Training database	Ref.
Classification of compounds with respect to their ability to inhibit phosphodiesterase V (PDE-V)	PDE-V inhibitors (61 compounds) and 1739 compounds with no activity towards PDE-V	42
Classification of compounds with respect to their ability to inhibit acyl-CoA:cholesterol *O*-acyltransferase (ACAT)	ACAT inhibitors (446 compounds) and 1771 compounds with no activity towards ACAT	42
Classification of compounds with respect to their activity towards estrogen receptors (ER)	estrogen analogues (463 compounds) characterized by different levels of anti-estrogen activity	43
Classification of compounds with respect to their activity towards nine different G-protein-coupled receptors (GPCR)	3365 compounds characterized by activity towards nine different GPCR	44
Classification of compounds with respect to their ability to inhibit carbonic anhydrase II (CA-II)	337 compounds characterized by different levels of CA-II inhibition activity	45
Classification of compounds with respect to their antibacterial activity	661 compounds that possess or do not possess antibacterial activity	46

a. Computer-assisted design of target-specific libraries of organic compounds

A key problem in modern computer-assisted medicinal chemistry is the study of noncovalent interactions in the ligand – receptor (target) system. In particular, a topical problem is the elucidation of a possible mechanism of the action of a drug on a biological target, which causes a therapeutically significant physiological response. *In silico* computer models make it possible to predict the target-specific activity of organic compounds and are, owing to their high performance, an indispensable tool for analysis of large compound libraries. Usually, even in the pre-screening stage of biological tests, virtual compound libraries are 'filtered' using specific QSAR/QSPR models, which predict the target-specific activity of compounds. Examples of some typical classification problems to be solved in the stage of rational selection of compounds possessing a specified target-specific activity are presented in Table 3.

However, it should be noted that successful use of classification statistical QSAR/QSPR models requires a sufficiently large training set including the compounds with a preset type of activity or property. This requirement makes this approach to virtual screening of limited use. A distinctive feature of the Kohonen algorithm is that it allows one, in the framework of a single QSAR/QSPR model, to predict the efficiency of ligand binding to not only one, but also to a number of therapeutically significant biological targets. Theoretical and practical aspects of the use of the Kohonen self-organizing maps have been well documented (see, *e.g.*, Refs 47 – 51 and references cited therein).

A unique computer model [52] based on the algorithm of the Kohonen self-organizing maps was used for efficient separation of drugs characterized by different biological activity profiles into groups. The training set was the Prous Ensemble database (18 703 compounds whose biological activity was established experimentally). It includes both the known drugs and novel compounds that are under clinical or pre-clinical tests. Each compound from this database has a known biological target belonging to one of the five therapeutically significant protein families including (i) tyrosine kinases; (ii) ion channel receptors; (iii) G-protein-coupled receptors (GPCR); (iv) phosphodiesterases; and (v) protease enzymes. Five molecular descriptors were calculated for each compound. These are the molecular mass (MM), the logarithm of the partition coefficient in octan-1-ol – water system ($\log P$), the number of freely rotatable bonds (RB) and the numbers of hydrogen-bond donors (HBD) and acceptors (HBA). They belong to the classical parameters of QSAR/QSPR analysis and are immediately related to such important properties of the molecule as the size, lipophilicity, conformational mobility and ability to form hydrogen bonds.

The preparation and filtration of the training set were followed by generation of a 10×10 Kohonen map. The network was learned independently of the compound class (uncontrollable training); this is the main distinction between this algorithm and conventional 'supervised learning' methods. Having the model constructed, all compounds were classified into groups corresponding to the target-specific activities listed above. All calculations were carried out using the InformaGenesis programme (URL: http://www.Informagenesis.com) (see below). As a result, a 2D projection of the five-dimensional space of properties was obtained; neihbouring objects on the 2D Kohonen map were also neighbours in the five-dimensional space. After training of the Kohonen algorithm, a general map of the distribution of properties of compounds from the complete initial training set was generated (Fig. 2).

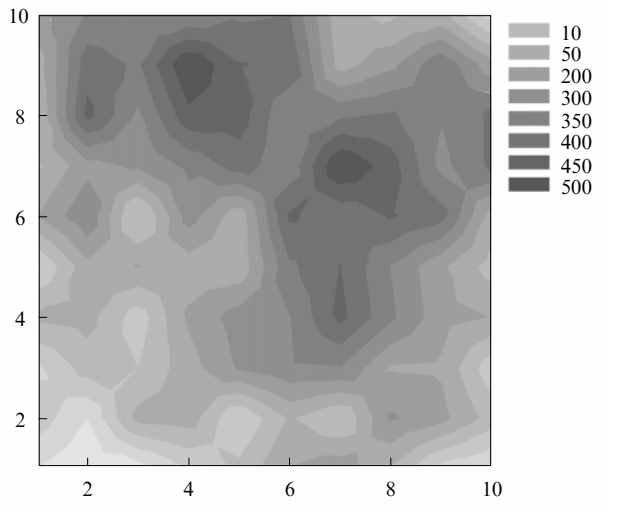


Figure 2. A 10×10 Kohonen map of entire training set (18 703 pharmaceuticals).

As can be seen, the compounds are distributed over the entire map, which clearly shows irregular-shape regions containing a large number of drugs and, *vice versa*, regions containing a few compounds only. The number of compounds in different regions of the map can differ by more than two orders of magnitude.

The common map thus generated exhibits the distribution of compounds that bind to all five targets under study (Fig. 3 *a*–*e*). As can be seen, compounds belonging to different groups are located in different regions of the map. For instance, the regions including the compounds active towards GPCR (Fig. 3 *c*) almost do not overlap with the regions where protease enzyme inhibitors are located (Fig. 3 *e*). For all five groups of compounds, highly populated and unpopulated regions are observed. The average percentage of separation of the target-specific groups in the map was 80%, which is quite admissible from the standpoint of the general concept of rational selection of compounds and mathematical statistics.

The algorithm proposed in Ref. 52 often makes it possible to reveal subtle differences between the properties of different groups of systems (in particular, clustering). Also, its compatibility with high-throughput virtual screening methods enables real-time tests of large virtual compound libraries using the Kohonen map constructed. The model developed can be useful for combinatorial synthesis design because the five target groups studied are the most typical in modern pharmacology, *viz*., they are affected by

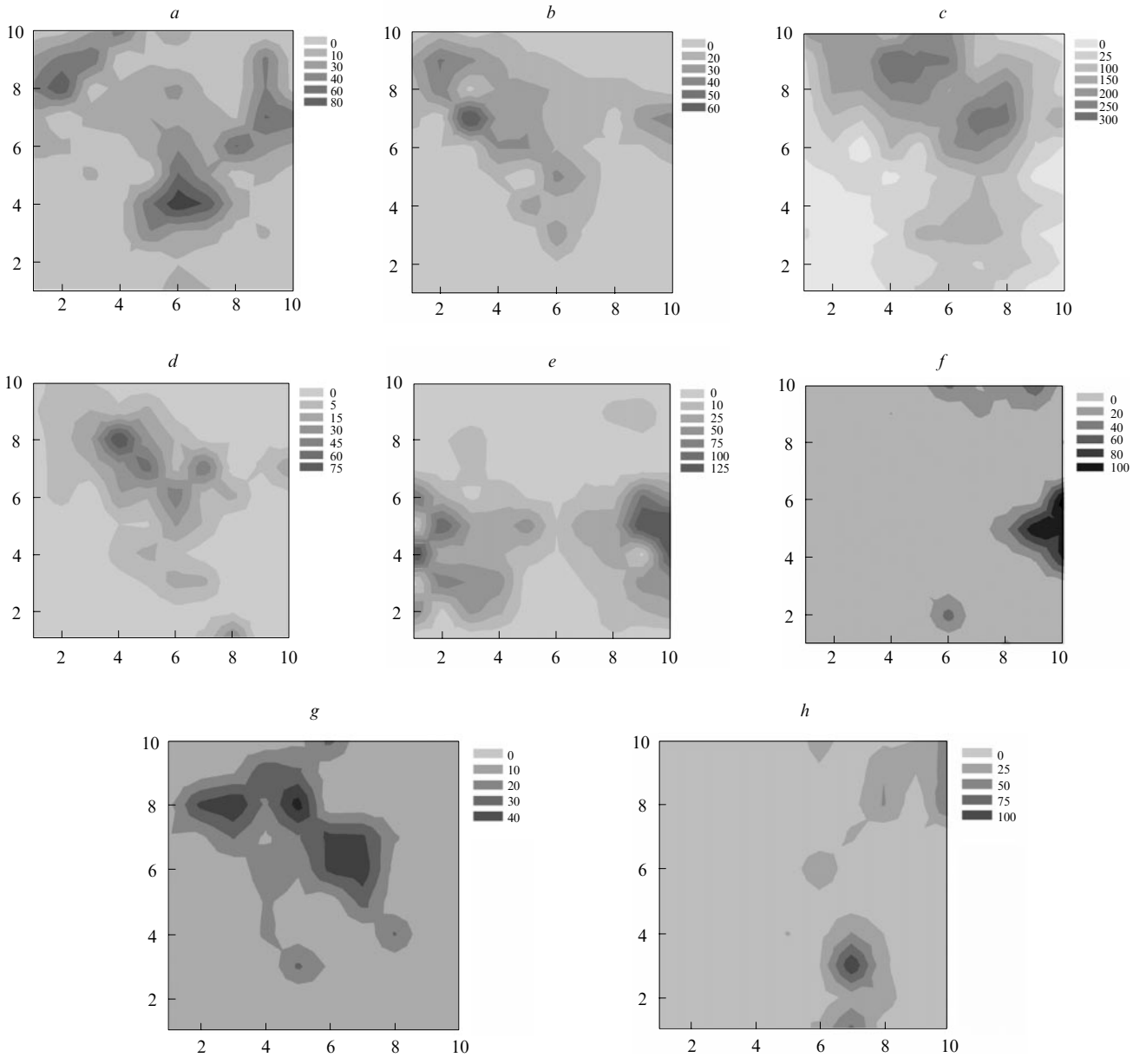


Figure 3. Distribution regions of eight target-specific groups of active compounds in a 10×10 Kohonen map.
(*a*) Tyrosine kinase inhibitors (1400 compounds); (*b*) ion channel receptor modulators (1080 compounds); (*c*) GPCR activity modulators (12 660 compounds); (*d*) phosphodiesterase inhibitors (730 compounds); (*e*) protease enzyme inhibitors (2833 compounds); (*f*) nuclear receptor antagonists/agonists (1021 compounds); (*g*) potassium channel activators (1060 compounds), and (*h*) calcium channel antagonists (1230 compounds). Examples (*a*–*e*) are taken from Ref. 52 and examples (*f*–*h*) are taken from Ref. 50.

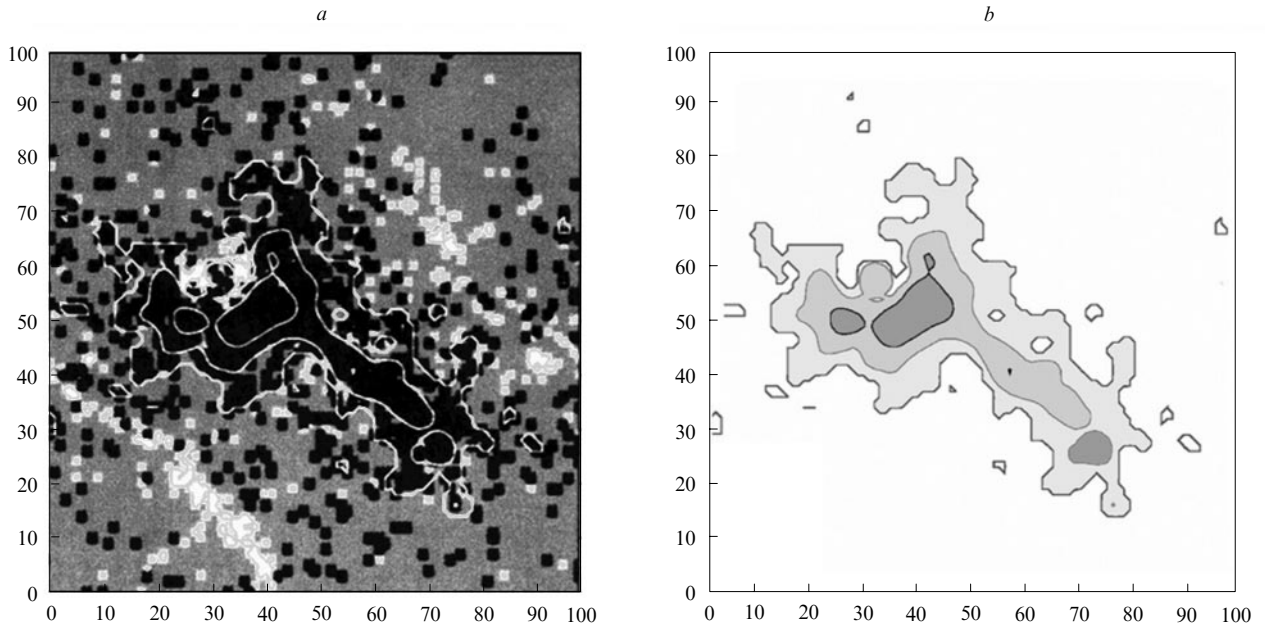


Figure 4. Two 100 × 100 Kohonen maps generated for more than 54 000 organic compounds.
Map (*a*): black and grey regions correspond to pharmaceuticals and non-pharmaceuticals, respectively. Map (*b*): the contour lines show the quantitative distribution of pharmaceuticals (in %); greyscale regions characterize the number of compounds per Kohonen neuron (from a value of 0.2 shown in light-grey to a value of 0.8 shown in dark-grey). The concentrations of pharmaceuticals beyond the outer boundaries of corresponding regions are at most equal to randomly distributed values.

at least 60% of the known commercially available drugs. Using the Kohonen map thus generated, one can select the most efficient compounds (they are located in the highly populated regions) for further screening using analogous protein systems.

The results obtained suggest that the five-dimensional space of properties includes compact regions corresponding to particular types of target-specific activity. The Kohonen map generated is a 2D image of this five-dimensional space and allows one to reveal distinctions between the properties of different groups of active compounds. It is believed [52] that the reasons for this statistical difference between the properties are as follows. Usually, biological targets belonging to the same family of active binding sites of low-molecular-mass ligands are more or less structurally similar. This follows from homologous character of amino acid sequences. Structural similarity of the active binding sites is determined by close values of their geometric parameters, similar lipophilicity, presence of the same positively or negatively charged groups, *etc*. As a consequence, low-molecular-mass ligands that can efficiently bind to these active sites should also be to some extent (in some sense) similar. But this similarity is of probabilistic nature, depends on the specific features of a particular member of a given family and can therefore be disclosed only using rather large training sets.

An alternative version of the Kohonen self-organizing map was also proposed.[41] In this case, a complete training set included the known drugs (2105 'oral drugs') from the World Drug Index database [53] and more than 52 500 non-drugs from the Maybridge database.[54] Each compound was characterized by 26 molecular descriptors; then, the number of descriptors was reduced to three using the principal component analysis (PCA) algorithm. Finally, the descriptors included QsumH (total charge of hydrogen atoms), Espmin (minimum electrostatic potential) and covHBac (covalent hydrogen bond acidity). Using these three descriptors, a common 100 × 100 Kohonen map was generated and regions corresponding to drugs and non-drugs were revealed (Fig. 4 *a*). Learning of the network involved two steps, which enabled a more correct distribution of objects on the map (the second step of learning is often called 'fine tuning' of Kohonen neurons). The learning parameters were as follows: † 10 000 for the number of iterations (learning time) in the first step of learning, 200 for the initial learning radius, 0.03 for the initial learning rate; 100 000 for the number of iterations in the second step of learning, 40 for the learning radius and 0.01 for the initial learning rate.

In the Kohonen map shown in Fig. 4 *a*, the majority of drugs are localized in the central region. This region has the shape of an irregular ellipsis and covers about 25% of the entire map surface area (Fig. 4 *b*). The results obtained suggest that, in the system of molecular descriptors used for simulations. all the compounds selected have quite similar characteristics; this is not surprising in the framework of the fundamental concept of drug similarity.

The Kohonen map was used [41] to study the distribution of each of the three descriptors used in generating the map. As a result, three separate gradient maps were generated (Fig. 5). It follows that different molecular descriptors have different orientations of the data vector distribution; however, all of them contribute to the elliptical regions where drugs are located.

Thus, only three molecular descriptors were necessary to construct the computer model for prediction of the degree

† The learning radius characterizes the number of neurons (except the winner neuron) participating in the learning process (weight correction) in a given iteration; the learning rate is the magnitude of the neuron weight correction in a given iteration.

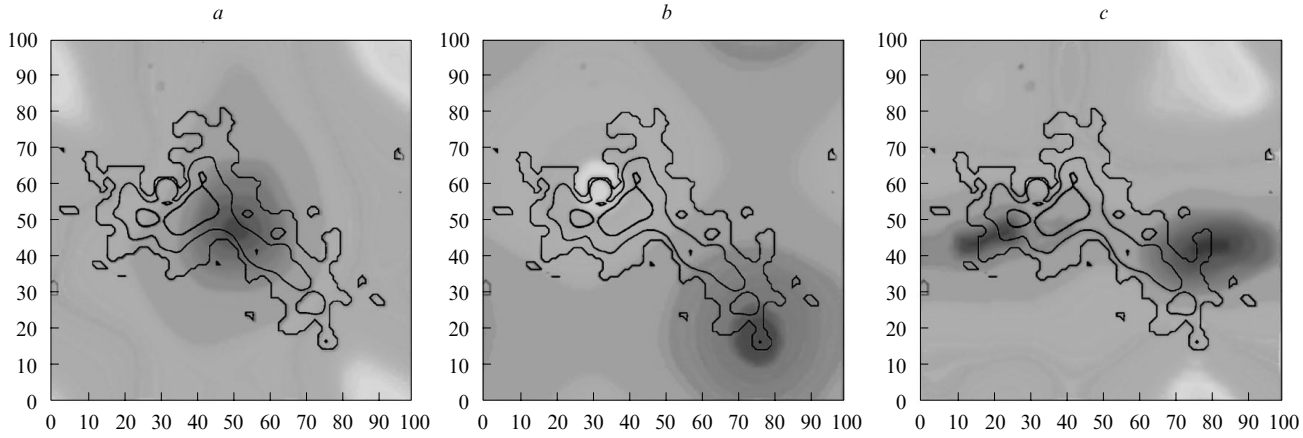


Figure 5. 100×100 gradient Kohonen maps showing the distribution (against the background of the distribution of pharmaceuticals) over three descriptors, *viz*., QsumH (*a*), Espmin (*b*) and covHBac (*c*) used to generate the Kohonen map presented in Fig. 4 *a*.
Light- and dark-grey regions in the greyscale images correspond to minimum and maximum values of the descriptors, respectively.

of similarity between the compounds under study and the known drugs.[41]

b. Computer-assisted assessment of possible toxic effects

Adverse toxic effects are one of the most abundant negative factors influencing the introduction of novel drugs and agrochemicals. As a consequence, the design and practical implementation of the methods of early prediction of toxicity can lead to considerable benefits. At the same time, prediction of potential toxicity of chemicals is one of the most difficult tasks in modern data analysis systems. There is considerable literature on the methods to solution of this problem (see, *e.g*., recently published reviews [55–58]), but up to the present they were of limited use due to either the local character of the effects in question or incomplete information on structural chemical types in the training set.

A number of toxicity prediction algorithms are based on the use of large databases containing information on the structures and action of toxic compounds. In some cases, statistical analysis of correlations between the toxicity of a compound and the presence of particular structural fragments in the molecule allows one to exactly predict the character and intensity of the toxic action. This principle is basic to the most widely used prognostic systems for toxicology, such as DEREK, HazardExpert and TOPKAT.[56,57] But this approach also has fundamental drawbacks. In particular, the prediction is based on information retrieved from a particular database, which contains to some extent incomplete and inexact information, historically determined character of information, different types of data sources and experimental conditions for toxicity determination, the lack of information on new chemotypes of compounds, *etc*. In addition, this type of expert systems usually ignores the influence of physicochemical factors on the toxicokinetic profiles of compounds.

The main objective factor that precludes the development of reliable algorithms for toxicity prediction is the variety of types and mechanisms of toxic action. With few exceptions, neither particular biological target nor particular mechanism of toxic action can be determined in QSAR/QSPR simulation of toxicity. This hampers research activity because unavoidably makes the space of properties of potential toxophores inhomogeneous. To solve practical problems related to, *e.g*., early determination of potential toxicity of low-molecular-mass organic compounds, it is necessary that the predictive model could process various structural chemotypes rather than a particular class of compounds. The lack of relevant models makes toxicity prediction difficult.

In this situation, researchers can only use local models for toxicity prediction, which were developed for particular targets. One can assume that the temporal evolution of the models will lead to accumulation of necessary information and give an impetus to the design of integrated high-performance toxicity prediction systems.

A local model was constructed to predict the toxicity of compounds in aqueous solutions.[59] The model makes it possible to predict toxic effects of a wide range of low-molecular-mass organic compounds towards fathead minnows (*Pimephales promelas*). This is a very useful tool of applied studies on aquatic ecology, in particular, it allows one to predict bioaccumulation of toxic substances in these marine animals. This QSAR/QSPR model was developed using a unique Duluth experimental database,[60] which, in particular, includes information on organic compounds possessing particular toxicity towards *Pimephales promelas*.

The training set included 562 organic compounds from the Duluth database, which were selected using a specialized procedure. A total of 169 individual molecular descriptors were calculated for all compounds from the training set. Rational selection of the most significant descriptors was based on the Kohonen algorithm using a simple rule. Namely, one descriptor whose data vector best characterized the distribution of compounds in the training set was selected from each cluster in the resultant Kohonen map. As a result, it was shown that under the experimental conditions the toxicity can be described using three descriptors, *viz*., the logarithm of the partition coefficient in octan-1-ol–water system, the molecular mass and the Kier–Hall topological index, which characterizes the molecular structure.

Figure 6 *a* shows the distribution of the $\log P$ values in a 30×30 Kohonen map generated after 800 iterations. The $\log P$ distribution almost matches the toxicity distribution of organic compounds towards *Pimephales promelas*, *i.e*., the larger the molecular descriptor ($\log P$) value the higher the toxicity of the compound. This result confirmed an experimentally disclosed interelation between the lipophi-

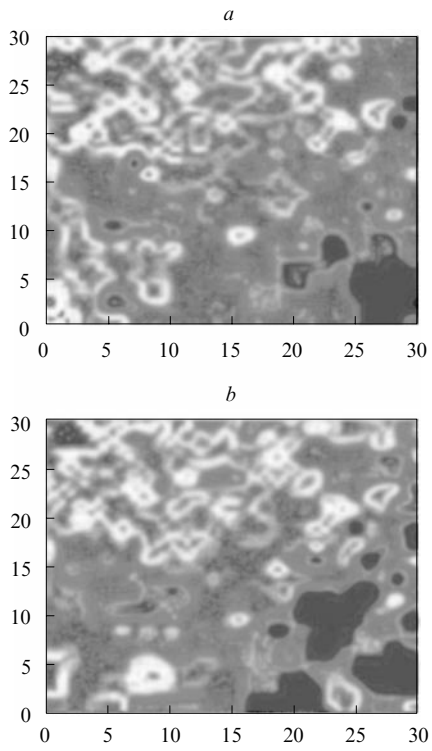


Figure 6. Distribution of $\log P$ values for organic compounds from the Duluth database in a 30×30Kohonen map (*a*) and toxicities of these compounds in aqueous solutions towards fathead minnows (*Pimephales promelas*) (*b*).
White and dark-grey regions correspond to low and high $\log P$ and toxicity values, respectively.

licity of some organic compounds and their toxicity towards *Pimephales promelas*.

The available linear correlations $\log P$ – toxicity do not allow one to predict the toxicity of compounds with good accuracy, whereas the Kohonen map generated has a higher predictive power and can be used for classification of organic structures with respect to their toxicity towards fathead minnows and other marine animals. The QSAR/QSPR model proposed enables automated testing of large virtual libraries of compounds (including up to tens of millions of molecular systems) and selection of compounds possessing the lowest potential toxicity.

c. Analysis of antitumour activity of drugs

Design of antitumour drugs is a key and topical avenue in medicinal chemistry. To date, a large number of organic compounds including natural protein complexes and various low-molecular-mass organic compounds have been in different stages of biological tests for antitumour activity towards various malignant neoplasms. A number of drugs is commercially available on the world drug market. In accordance with the classical pattern of drug design and development, the initial stage of novel antitumour drug design often involves the use of various methods of computer simulation, in particular, the Kohonen self-organizing map algorithm.

A computer model[42] makes it possible to predict some types of antitumour activity of organic compounds. The training set included compounds from the U.S. National Cancer Institute (NCI) database. A total of 362 drugs with experimentally determined activity towards various types of malignant cells were used. Calculations and rational selection of the most significant molecular descriptors were followed by generating a 17×9 rectangular Kohonen map, which revealed regions containing compounds possessing various types of antitumour activity (Fig. 7).

Different classes of antitumour drugs are located in different regions of the map. Dashed lines frame the regions containing various mitotic poisons (region **A**) and compounds affecting tubulin as a therapeutical target (region **B**). The degree of separation of the model compounds exceeded a value of 77%. Thus, the computer model has been constructed,[42] which can be used in the future to predict the antitumour activity of organic compounds.

2. The Sammon nonlinear maps

In addition to Kohonen self-organizing maps, other algorithms for reducing the dimension of the multidimensional spaces of attributes are also available; they generate 2D or 3D maps convenient for visual analysis. An example is provided by 2D mapping of the initial space of attributes

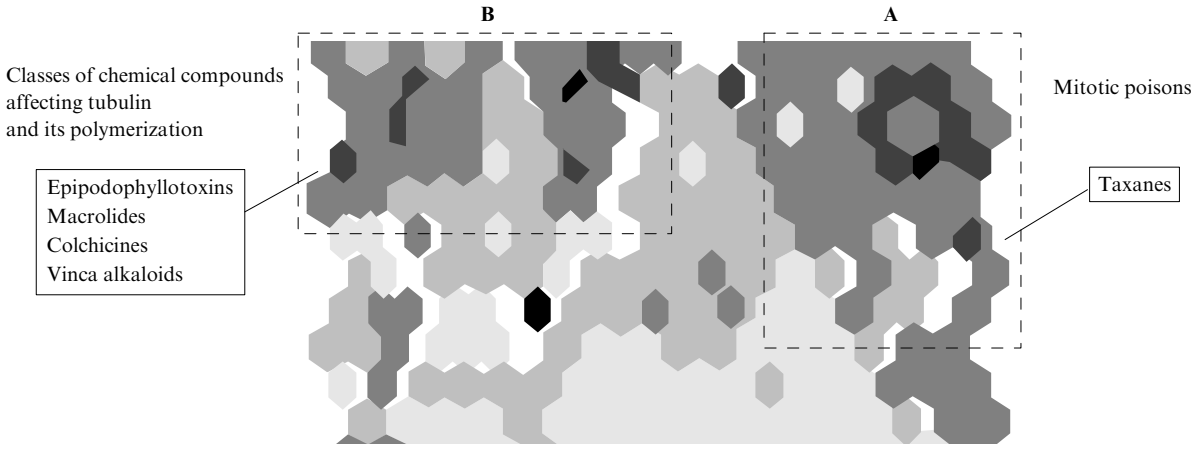


Figure 7. A 17×9 Kohonen map (hexagonal cell) generated for 362 pharmaceuticals possessing antitumour activity.

according to Sammon.[32, 33] As applied to computer-assisted medicinal chemistry, the Sammon algorithm allows one to generate 2D images of the multidimensional space of properties of the biologically active compounds from experimental databases. This gives researchers a convenient tool for visual analysis of the space of properties encoded by molecular descriptors. For instance, it is possible to visually determine the regions of preferred location of compounds belonging to the classes of interest. (In some cases, this is achieved using the SVM algorithm or an algorithm proposed in Ref. 61).

It should be noted that, unlike the Kohonen maps, which correctly map only the cluster topology of the initial space onto the lower-dimension space, the Sammon method permits conservation of the multidimensional space topology, where the distances between objects (vectors mapping these objects) in the lower-dimension projection correspond (within the limits of admissible approximation error) to the initial distances (*e.g.*, Euclidean distances). Sammon's algorithm is used to solve various problems in computer-assisted medicinal chemistry.

a. Permeation of organic compounds through lipid membranes: a computational assessment

Consider how the Sammon mapping technique can be used to predict the permeation of organic compounds through lipid membranes taking the blood–brain barrier (BBB) permeability and human intestinal absorption (HIA) as examples.‡ Recently,[62] a procedure for generating a Sammon map to predict these properties was proposed.

A common Sammon map for two training sets used to study the BBB permeability of compounds and the HIA was generated.[62] The input signal was a set of molecular descriptors including the logarithm of the partition coefficient in octan-1-ol–water system ($\log P$), the number of hydrogen-bond donors and acceptors and the number of freely rotatable C–C bonds. The Sammon map thus generated was used to assess the degree of separation of two categories of compounds characterized by high (denoted by '+') and low (denoted by '−') BBB permeability and HIA by alternative visualization of these categories and drawing a separation line between them using the SVM algorithm.[25] The separation line determines the optimum separation function of objects in the 2D coordinates. Compounds with high and low permeability through lipid membranes are usually located in different regions of the Sammon map (Fig. 8).

For both membranes, the degree of separation approaches a value of 90% (Table 4), the average accuracy of classification being very close in both cases. The degree of separation is quite remarkable with allowance for the fact that generation of the Sammon map is an uncontrollable objective procedure that ignores the experimental classification of molecular systems.

Using the available Sammon maps, additional studies[62] were carried out. In particular, correlations between the permeability of drugs through these lipid membranes and the values of the corresponding molecular descriptors were elucidated. This revealed differences between the properties of the compounds with high BBB permeability and HIA. Thus, the Sammon maps generated showed the possibility

‡ This was also done using the Kohonen approach.

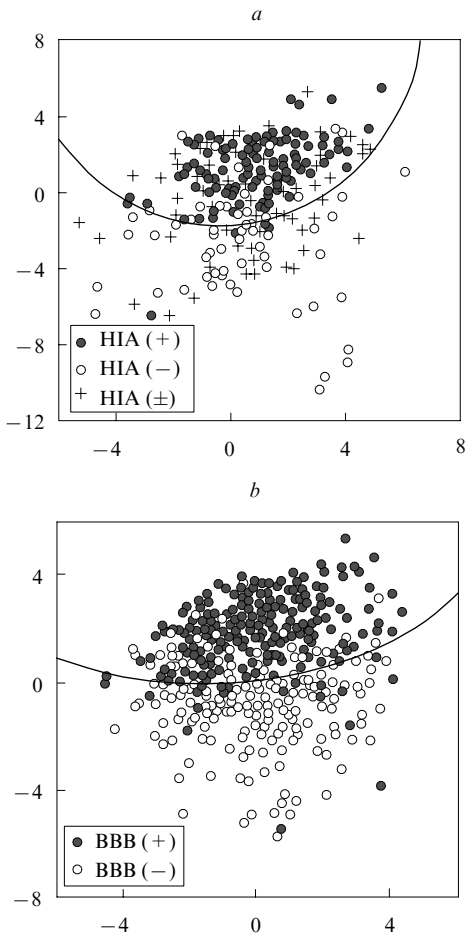


Figure 8. Nonlinear Sammon maps with classification of compounds based on high and low human interstinal absorption (HIA; map *a*) and the ability to cross the blood–brain barrier (BBB; map *b*). Solid and open circles correspond to high and low values of the parameters, respectively.
Hereafter the axes in the Sammon maps are scaled in relative units. The line separating two compound categories (with high and low ability to penetrate through lipid membranes) is drawn using the SVM algorithm, which ensures an optimum (mathematically substantiated) separation.

of classification of compounds using pre-selected descriptors. However, application of the Sammon algorithm offers not only clearly seen advantages, but also some difficulties.

Table 4. Statistical characteristics of classification using Sammon's maps.

Property	Category of compounds	Number of compounds		
		total number of compounds tested	correctly classified compounds	incorrectly classified compounds
Penetration through BBB	BBB (+)	267	237 (88.8%)	30 (11.2%)
	BBB (−)	189	163 (86.2%)	26 (13.8%)
		average:	87.5%	12.5%
Human intestinal absorption	HIA (+)	114	110 (96.5%)	4 (3.5%)
	HIA (−)	61	48 (78.7%)	13 (21.3%)
		average:	87.6%	12.4%

b. Assessment of other pharmacokinetic parameters using the Sammon maps

As mentioned above, pharmacokinetic parameters (absorption, distribution, metabolism and excretion, *i.e.*, the ADME parameters) are the key characteristics of all drugs. In particular, the Directive on Drugs requires that each of them be characterized by a number of attributes including pharmacokinetic ones and possible cytotoxicity. Promising drugs characterized by inadmissible pharmacokinetic parameters are usually excluded from biological tests and undergo chemical modification. Thus, the development of computer methods for prediction of various pharmacokinetic parameters is a topical problem of modern computer-assisted medicinal chemistry.

Unique computer models based on the Sammon maps have been reported.[61] They were used to assess some key ADME parameters of various physiologically active compounds including the HIA (Fig. 8 *a*), BBB permeability (Fig. 8 *b*), plasma protein binding (PPB, Fig. 9 *a*), CYP3A4 binding (Fig. 9 *b*; expressed in units of the Michaelis constants, K_m), P-glycoprotein binding (P-gp, Fig. 9 *c*), as well as the volume distribution of compounds in plasma (V_d, Fig. 9 *d*) and their half-lifes in plasma ($T_{1/2}$, Fig. 9 *e*).

The common training set was an annotated compound library, which included the published data [63–67] and information from private sources (kindly provided by S Ekins and retrieved from specialized databases at the Ivanovo Institute of Solution Chemistry, Russian Academy of Sciences). A total of 1500 different drugs were used to construct the computer model.

Molecular descriptors were calculated using the specialized software Dragon (URL: http://www.talete.mi.it/dragon.htm) and Chemosoft (URL: http://www.chemosoft.com). For each compound, more than 1000 unique molecular descriptors were calculated. The most significant descriptors (reduction of the number of the input attributes) were selected using the PCA algorithm. It was shown that more than 90% of distribution can be described using six to eight principal components whose contributions were estimated from the conventional Kaiser–Guttmann criterion [68] and the Catell criterion (Scree-Test).[69] Based on the results obtained, the molecular descriptors making the greatest contributions to the principal components were selected. The abbreviatons and corresponding definitions of the most widely used descriptors are available from Ref. 70 and from the Dragon software manual.

The Sammon maps for the complete training set were generated using the InformaGenesis software; the positioning of the optimal separation line between the compound classes under study was determined using the SVM algorithm implemented in the LibSVM-2.477 programme (URL: http://www.csie.ntu.edu.tw). The results obtained for all ADME parameters studied [61] are shown in Fig. 9.

As can be seen, in all cases the 'outermost' compound classes are located in different regions of the maps, the overlap being very small. The optimum separation line between two opposite compound categories (no intermedi-

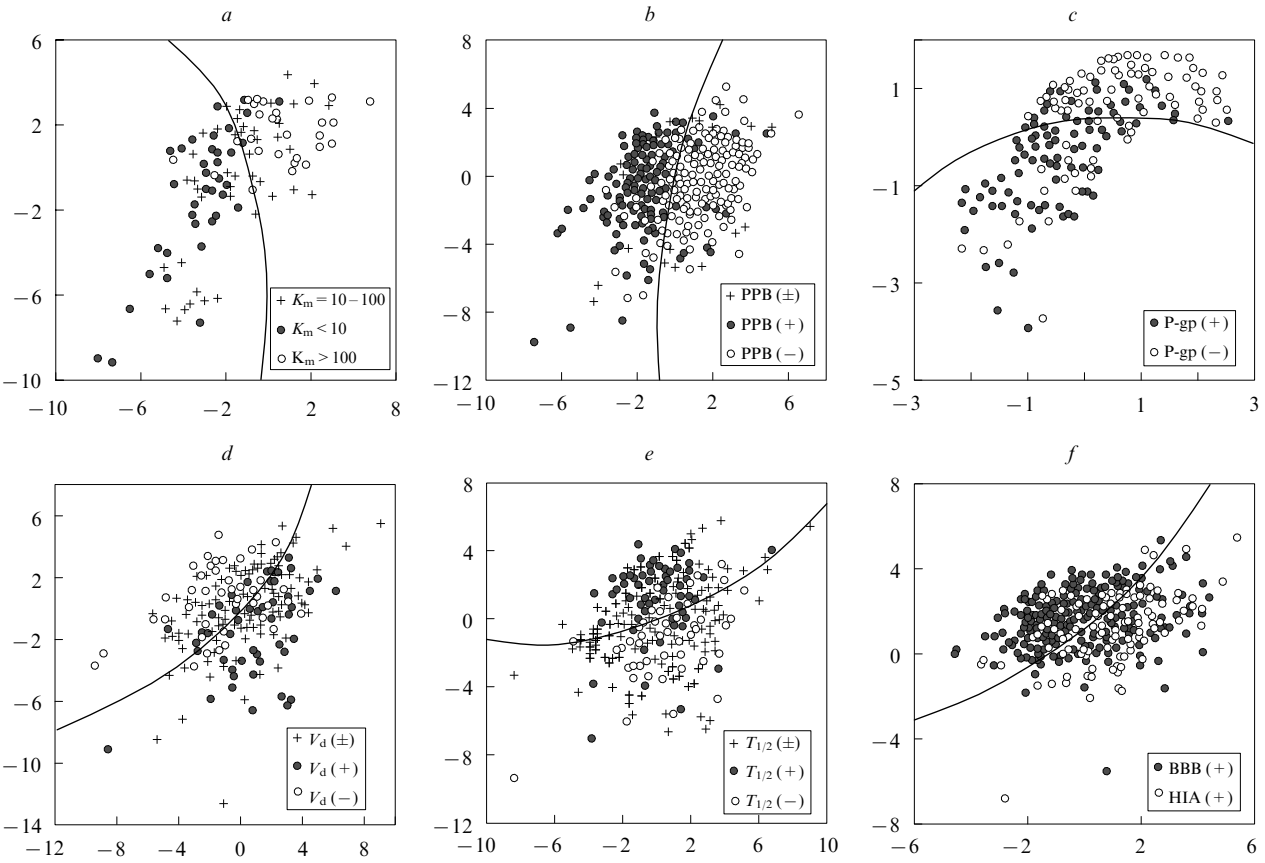


Figure 9. Sammon maps generated for selected key pharmacokinetic parameters of various biologically active compounds. For notations (*a*–*f*), see text.

Table 5. Classification quality of the Sammon maps (optimal positioning of the separation line between opposite compound categories was determined using the SVM algorithm).[61]

Pharmacokinetic parameter	Compound categories	Criterion for classification into a given category	Number of compounds tested	Correctly classified compounds	Percentage of correct predictions (%)
Human intestinal absorption (HIA)	HIA(+)	HIA > 80%	114	110	96.5
	HIA(±)	HIA = 20–80%	145	–	–
	HIA(−)	HIA < 20%	61	48	78.7
Blood–brain barrier permeability (BBB)	BBB(+)	$\log BB \geqslant 0$ [a]	267	237	88.8
	BBB(−)	$\log BB < -1$	189	163	86.2
Plasma protein binding (PPB) (mostly with albumin)	PPB(+)	PPB > 90%	200	169	84.5
	PPB(±)	PPB = 50–90%	172	–	–
	PPB(−)	PPB < 50%	177	143	80.8
CYP3A4 binding ability	CYP(+)	$K_m \leqslant 10$	37	34	91.9
	CYP(−)	$K_m \geqslant 100$	28	26	92.9
P-gp substrates/non-substrates	P-gp(+)	substrates	89	66	74.2
	P-gp(−)	non-substrates	78	62	79.5
Volume of distribution (V_d)	V_d(+)	$V_d < 0.4$ litre kg^{-1}	47	39	83.0
	V_d(±)	$V_d = 0.4$–8 litre kg^{-1}	162	–	–
	V_d(−)	$V_d > 8$ litre kg^{-1}	44	34	77.3
Plasma half-life ($T_{1/2}$)	$T_{1/2}$(+)	$T_{1/2} > 20$ h	63	51	81.0
	$T_{1/2}$(±)	$T_{1/2} = 1$–20 h	347	–	–
	$T_{1/2}$(−)	$T_{1/2} < 1$ h	48	40	83.3

[a] $\log BB = \log([c_{blood}]/[c_{brain}])$, where c is the solute concentration.

ate class was used) was drawn using the SVM algorithm and then the percentage of correct predictions was calculated for all the maps (Table 5).

From the data in Table 5 it follows that the models constructed have a high predictive power (80%–90% of compounds were correctly classified). Other interesting computer experiments (Ref. 62) involved studies on the regularities of location of compounds belonging to different pharmacokinetic classes in the Sammon maps. A number of additional Sammon maps were generated, which characterized the distinctions between different groups of compounds belonging to the same categories (an example is shown in Fig. 9 *f*). As can be seen, compounds with high BBB permeability and HIA are located in different regions of the map, which in turn is an indication of different nature of these biological membranes.

In addition, the authors analyzed the contributions of all individual descriptors used to construct the corresponding model to the classification of the compounds into categories (see Table 5). The gradient Sammon maps showing the distribution over three representative molecular descriptors relative to the separation lines drawn using the SVM algorithm are presented in Fig. 10. As can be seen, the corresponding data sets are characterized by their own gradients. For instance, the smallest values of the descriptor $\log P$ characterizing the lipophilicity of compounds ($\log P < 0$) were observed in the right part of the map, while the largest $\log P$ values ($\log P > 0$) were observed in the left part of the map. Transition from small to large $\log P$ values is relatively smooth and corresponds to a monotonic distribution.

Earlier, it was assumed that the separation of compounds with different BBB permeability first of all depends on their lipophilicity. However, a Sammon map study[41] showed that the topological polar surface area (TPSA)[70] (Fig. 10 *b*) makes at least comparable contribution to separation of these compound classes, although in this case the separation line drawn using the SVM algorithm does not match the boundary between the regions of small and large TPSA values.

As mentioned above, the positioning of the optimum separation line between two opposite compound categories is determined using the SVM algorithm. The separation lines in Fig. 10 were drawn using the LibSVM-2.4 software. They are the best ones from the standpoint of mathematical statistics. Thus, many important pharmacokinetic parameters can be analyzed and predicted using the Sammon algorithm and a few simple molecular descriptors.

c. Selection of the most significant molecular descriptors using the Sammon maps

Rational selection of the most significant attributes, which describe the initial multidimensional space, is a topical problem in modern computer chemistry. In addition to the known statistical algorithms used in this field (PCA algorithm, factor analysis, sensitivity analysis, direct correlations and *t*-statistics), more accurate methods of computer simulation having a higher degree of generalization and descriptive power are often used.

The Sammon maps can also be used to this end.[71] The training set was a digital chromosome database at the Soroka Medical Centre (India). (Earlier, this database has been analyzed using neural network techniques.[72]) The chromosomes from this database were characterized by a specialized set of unique attributes including four descriptors of the structural and geometric features of human chromosomes (centromeric index, length, perimeter length and surface area). The training set included 100 different

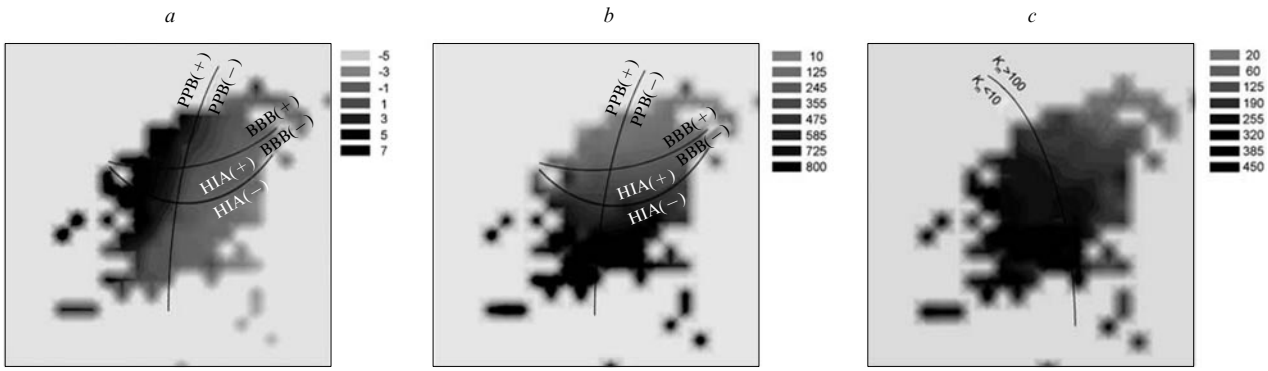


Figure 10. Gradient Sammon maps showing the distribution over three representative descriptors characterizing the lipophilicity of compounds (log *P*; *a*), the total polar surface area (TPSA; *b*) and the polar negative surface area (PNSA, *c*) relative to the separation lines between different compound categories. The separation lines were determined using the SVM algorithm.

chromosomes divided into three main classes, *viz*., the chromosomes No. '13' and '19' and the X-chromosomes. An independent test set included 10 chromosomes, which were randomly selected from the training set. The initial learning rate was 0.9 and the number of training cycles was 30. The Sammon map constructed (2D image of the four-dimensional space of attributes) is shown in Fig. 11.

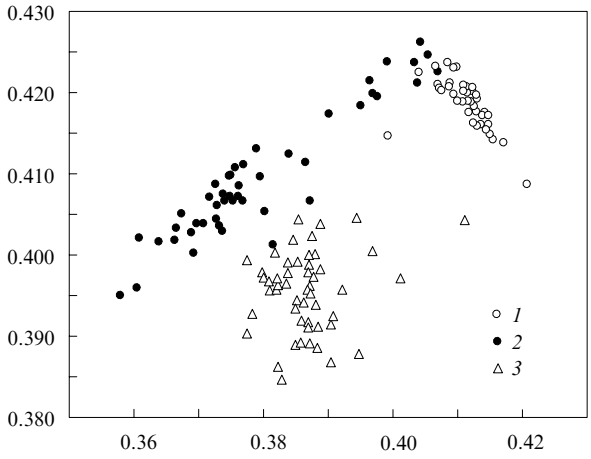


Figure 11. The Sammon map for analysis of human chromosomes '13' (*1*), '19' (*2*) and 'X' (*3*).

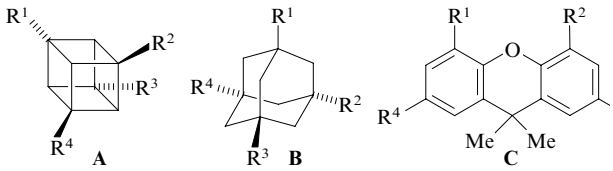
As can be seen, different classes of chromosomes are located in different regions of the map. Their overlap is small. The percentage of correct predictions for the test set was 95% using all four descriptors. When the number of descriptors was reduced to three (the least significant descriptor was excluded), the percentage of correct predictions decreased by 6%. However, a classification quality of 89% is also quite admissible when solving this type of problems.

It should be noted that computer simulation with a too large number of descriptors often leads to overtraining of the model, which thus can not be used for prediction.

d. Studies of various classes of organic compounds

The Sammon maps can be used not only to solve the above-mentioned problems in computer-assisted medicinal chemistry, which are immediately related to investigations of the pharmaceutical properties of organic compounds, but also for systematization of organic compounds with respect to their structural attributes and for optimization and design of organic synthesis, especially combinatorial synthesis. It should be noted that in this case the molecular descriptors often include attributes that describe the qualitative and quantitative structural composition of the molecule, *e.g*., the number of certain peripheral structural fragments. As applied to the solution of such problems, the Sammon maps offer some advantages over the Kohonen self-organizing maps, *e.g*., conservation of the general architecture and topology of the initial space (see above).

The Sammon map for three classes of organic compounds, namely, cubanes (**A**), adamantanes (**B**) and xanthenes (**C**) (Fig. 12) was generated.[73]



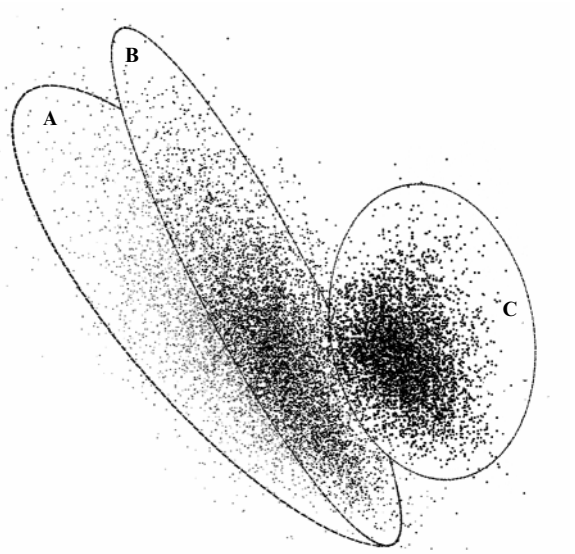


Figure 12. Nonlinear Sammon projection for the distribution of compounds from three libraries of organic compounds including cubanes (**A**), adamantanes (**B**) and xanthenes (**C**).

The training set included a few thousands of individual structures. The Sammon map was generated using twelve autocorrelation descriptors, which provided a complete description of the initial space of attributes. The diversity/similarity of compounds under study was assessed using an Euclidean metric.

The Sammon map allowed one to reveal distinctions between these classes of organic structures. The separation of compounds using the Sammon map was clearly defined and statistically significant, which makes this computer model a promising tool of classification of various combinatorial compound libraries including the cyclic structures considered.

Compared to the Kohonen maps, the Sammon maps give a better image of the general topology of the input space. However, these calculations are more time-consuming because the time taken to generate the Sammon maps increases in proportion to the squared number of objects analyzed. As a result, generation of the Sammon maps for more than 5000 – 7000 compounds is impossible even with powerful personal computers. Usually, this presents no difficulties when analyzing the training databases, but becomes an insurmountable obstacle to testing large virtual libraries, which often include hundreds of thousands and even millions of compounds. Yet another drawback of the Sammon maps is the impossibility of direct testing of new systems using the available map. To this end, the map should be completely re-trained using both the training set and the test set; in this case the classification quality attained by analyzing the training set only is not ensured.

3. The Galaxy™ algorithm

In conclusion, we will consider yet another interesting method of nonlinear mapping, which is in essence a modified Sammon algorithm. The OmniViz computer product (BioWisdom Ltd, United Kingdom, URL: http://www.omniviz.com) includes a Galaxy™ algorithm for generating 2D projections of the multidimensional space of attributes. This algorithm is used in analysis of various types of chemical data. For instance, the Galaxy™ algorithm proved itself [74] as highly efficient tool of analysis of chemical structures in the framework of combinatorial synthesis and related biological tests. The training set included 1107 individual 'products' of high-throughput combinatorial synthesis, which were tested for biological activity under high-throughput biological screening conditions using 19 different biological tests. First, structural descriptors based on various substructural fragments were calculated for all compounds tested. Then, the 2D projection of the multidimensional space of structural attributes (Fig. 13 *a*) was generated using the Galaxy™ algorithm.

Similarly to the reduction of the dimension of the input space according to Sammon, the Galaxy™ algorithm allows one to generate projections in which the distances between all objects in the space of the lower dimension correspond to the distances in the multidimensional space of attributes. Thus, neighbouring compounds in the map are also neighbours in the initial multidimensional space.

The map of biological activity shown in Fig. 13 *b* was generated using an analogous principle, *i.e.*, compounds with similar biological activity are located in neighbouring regions in the map. Interestingly, with arbitrary selection of a group of neighbouring compounds in the biological activity map (framed by the dashed line) the programme also automatically identifies these compounds in the first

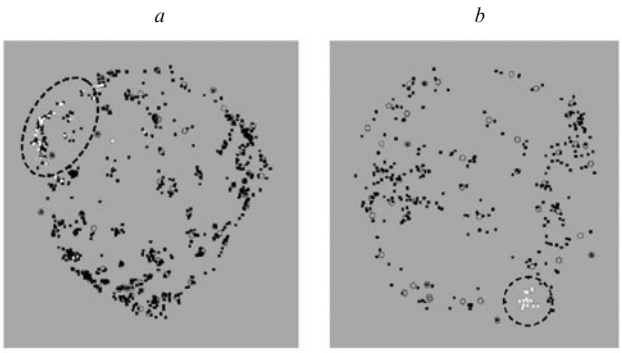


Figure 13. Analysis of various chemical structures using the Galaxy™ algorithm.
A two-dimensional map for the training set of organic compounds (1107 structures) generated using different structural fragments (*a*) and a two-dimensional map characterizing the biological activity of the compounds tested (*b*). Each point in the map corresponds to one compound.

map (Fig. 13 *a*), which characterizes the structural features of compounds, expressed in the number of certain substructural fragments. Analysis of the location of this group of compounds in the map (see Fig. 13 *a*) revealed a high degree of likeness of these compounds; at the same time, a number of compounds has different structural environment. The latter means that structurally dissimilar compounds can have identical biological activity and *vice versa*. Thus, the Galaxy™ algorithm can be considered as promising alternative to the classical Sammon maps and successfully used for multiparametric data analysis in computer-assisted medicinal chemistry.

4. Gene expression studies using a combined mapping algorithm

In some cases, complicated problems (*e.g.*, analysis of gene expression profile) are solved using a combined mapping algorithm based on the Kohonen self-organizing maps and the Sammon maps. The applicability of this hybrid method to analysis of the gene expression profile during a diauxic shift was reported earlier.[75] The training set included the gene expression profile of 6400 yeast genes. The descriptors were the gene expression volumes measured at seven time points.

In the first stage, the data were analyzed using the 16×16 Kohonen maps. Depending on the time point, the number of genes in individual neuron varied from 10 to 64 (Fig. 14 *a*) (another type of visualization of the results obtained using the Kohonen mapping algorithm was used [75]). Analysis of the map constructed showed that genes with the same character of expression are contained in the same Kohonen neuron, genes with similar character of expression are contained in neighbouring neurons, whereas genes with different expression profiles are contained in neurons located far from one another. As an example, Fig. 14 shows four different Kohonen neurons (C1 – C4) highlighted in the common map. The histogram for each of the four neurons characterizes the quantitative character of gene expression (average value) at seven time points separated by 2-h intervals during the overall diauxic shift.

Then, the integrated data thus obtained were analyzed using the Sammon mapping algorithm (Fig. 14 *b*). The

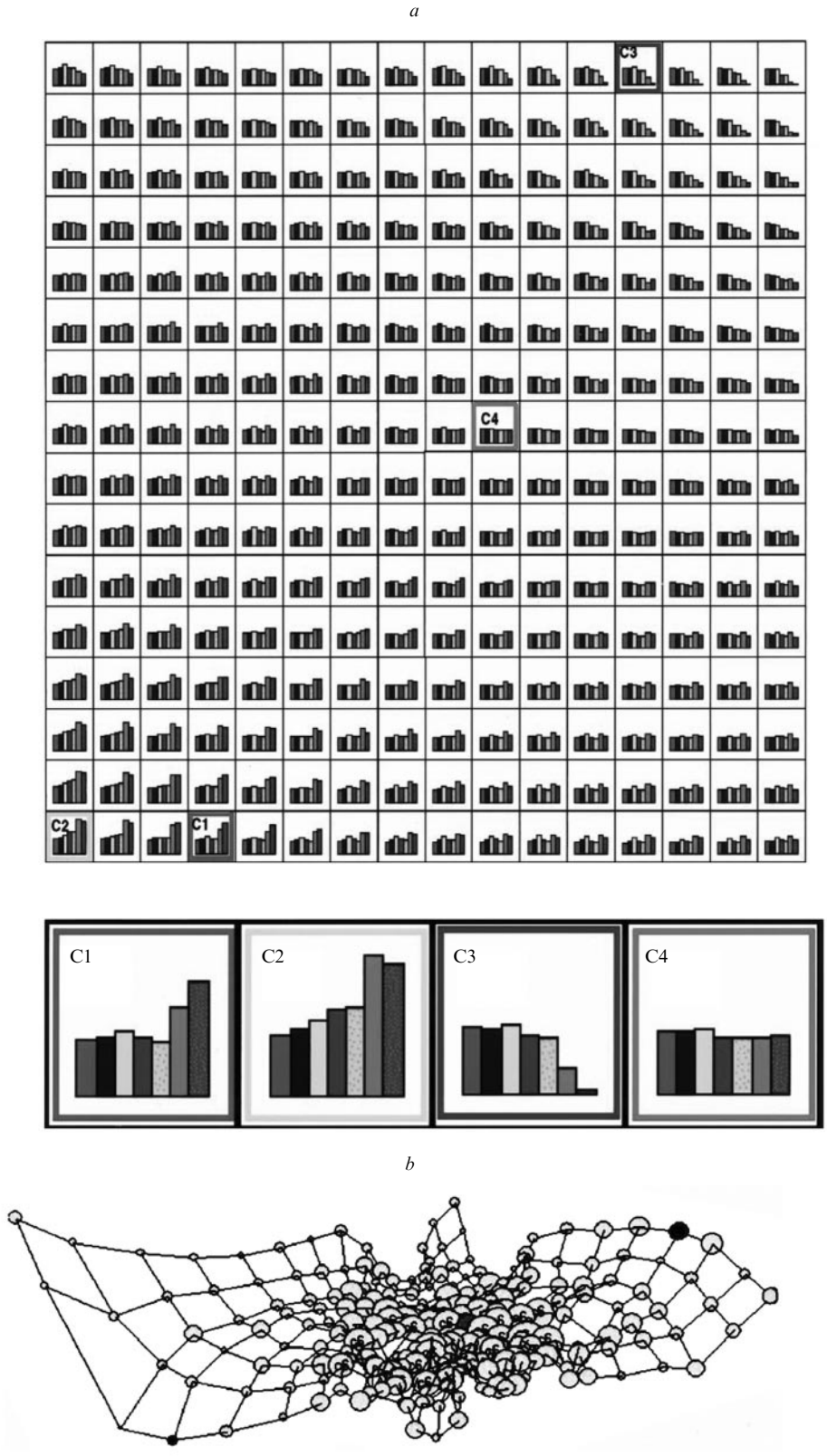


Figure 14. A Kohonen map (*a*) characterizing the expression profile of 6400 yeast genes at different instants (regions shown in greyscale correspond to seven different instants) and the corresponding integrated Sammon map (*b*) (positions corresponding to the selected Kohonen neurons are given in greyscale).

principal axes of the Sammon map quantitatively characterize the gene expression, namely, the genes with the highest integrated activity are located in the left part of the map while the genes with low activity are located in the right part. The genes whose expression was not controlled under the diauxic shift conditions are located in the central part of

the map. Thus, the distances between objects in the Sammon maps characterize differences in the average volume of gene expression and the size of points corresponds to the number of genes in the Kohonen neuron.

This hybrid computer model based on Kohonen self-organizing maps and Sammon nonlinear maps made it possible to mathematically describe the expression of 6400 yeast genes under the diauxic shift conditions.

IV. The use of the SPE algorithm for analysis of the chemical space

To date, the Kohonen self-organizing maps and Sammon nonlinear mapping have proved themselves as efficient and exact methods of computer-assisted medicinal chemistry. They are most widely used for solving many practical virtual screening tasks. However, the demand for even more powerful and exact algorithms, which make it possible to avoid some drawbacks of the methods mentioned above (low performance of the Sammon algorithm or the cluster approximation to the Kohonen algorithm) still persists. An example of such methods is provided by the SPE nonlinear mapping algorithm, which has been recently developed by Agrafiotis and Xu.[76]

Similarly to the Sammon nonlinear mapping, the SPE method is based on the fundamental principle of self-organization, *viz*., generation of the most relevant and complete projection of the structure of the initial space of properties onto the space of the lowest dimension. A basic difference between the SPE and Sammon methods is that the former does not require calculations of all geodesic distances between objects in the new space of low dimension. The SPE algorithm uses a modified pairwise refinement strategy, which does not require calculations of the full-connected similarity matrix. Unlike the Sammon maps, the performance of the SPE algorithm is a linear function of the volume of the input space. In other words, the SPE algorithm allows for the fact that the geodesic distance is always greater than, or equal to, the initial similarity metric. The SPE method is to a greater extent aimed to preserve the principal local distances rather than the global structure of the initial space of attributes. The modified procedure for estimation of the distances between objects used in the SPE algorithm allows one to generate full-connected projections at much lower computational cost and finally generate 2D maps that completely characterize both the global topology of the initial space and the local regularities. The efficiency of the SPE algorithm in virtual screening of organic compounds was proved in a number of studies; some of them will be considered below.

Like the Sammon mapping algorithms, the SPE algorithm can successfully be used to study the chemical space of complex geometry with implicit, hidden interrelations.[34] Figure 15 presents two 2D SPE maps generated to describe the regularities within virtual libraries of organic compounds ('products' of two virtual reactions, a two-component amination reaction (Fig. 15 *a*) and the Ugi multicomponent condensation (Fig. 15 *b*).

In particular, Fig. 15 *a* shows a 2D SPE image of a virtual library including 10 000 organic compounds designed from 100 amines and 100 aldehydes of different structure. Each 'product' of the virtual amination reaction was characterized by 117 different topological descriptors including the molecular connectivity indices, κ-indices of the molecular shape, information theory indices, the Bonchev–Trinajstić indices, *etc*.[34] The dimension of the input space was reduced using the PCA algorithm. It was found that the major contribution to the change in the initial space of attributes ($\sim$99%) comes from 23 principal components, which were used to generate the SPE map. Figure 15 *a* shows that different structural classes of the compounds obtained by the amination reaction occupy their own regions in the map; this indicates a high ability of the SPE algorithm to solve similar classification problems.

The 2D map presented in Fig. 15 *b* characterizes the regularities within the space of a virtual library including 10 000 structures. These compounds are 'products' of a virtual four-component Ugi condensation of carboxylic acids with primary aliphatic or aromatic amines, aldehydes and isocyanides (each group of the reactants included ten different compounds). Each 'product' of the virtual reaction was characterized by 166 different structural descriptors, which encoded the number of certain fragments in the target compound. In this case, the descriptors were selected using the structural similarity/dissimilarity metric based on the classical equation and the Tanimoto coefficients. As can be seen, the maps generated include a few clearly seen compact clusters corresponding to different classes of compounds. (The map generation procedure is described in detail in Ref. 77.)

The SPE algorithm was also successfully used for classification of structurally dissimilar organic compounds often having complex molecular structures (see, *e.g*., Ref. 78). As in the preceding case, computer simulation was carried out using a specially selected universal set of structural descriptors, which characterized the number of certain molecular fragments. The 2D map thus generated is shown in Fig. 15 *c*.

The SPE algorithm also proved itself as an efficient tool for solving other, more complicated tasks, such as conformational analysis of organic molecules,[79] NMR structure elucidation and structure prediction of protein macromolecules.[80] For instance, a combination of the SPE algorithm with the self-organizing superimposition (SOS) algorithm is considered as a promising computer method of analysis of the conformational states of organic molecules and implementation of conformation-based search for compounds.[81–83] In addition, the SPE approach was successfully used for classification and visualization of protein amino acid sequences, where the amino acids were used as descriptors. In particular, Fig. 15 *d* presents a 2D image of the 21-dimensional space of attributes (amino acids), which characterizes the space of kinase enzymes. The SPE algorithm can also be successfully used for reduction of the dimension of the data obtained in the course of genomic and proteomic research.[84] High efficiency of the SPE algorithm in solving problems related to analysis of target-specific activity of organic structures can be well illustrated taking construction of computer models for activity prediction of compounds towards kinases and nuclear receptors as examples.[85]

However, like all currently available methods of computer analysis, the SPE approach has some drawbacks and limitations. They include a relatively complicated algorithm training procedure, which involves the determination of many internal learning parameters and the similarity metric. Similarly to the IsoMap technique, the efficiency of the SPE algorithm depends strongly on the choice of the learning radus. In the case of a too large learning radus, neighbouring objects belonging to different classes will essentially

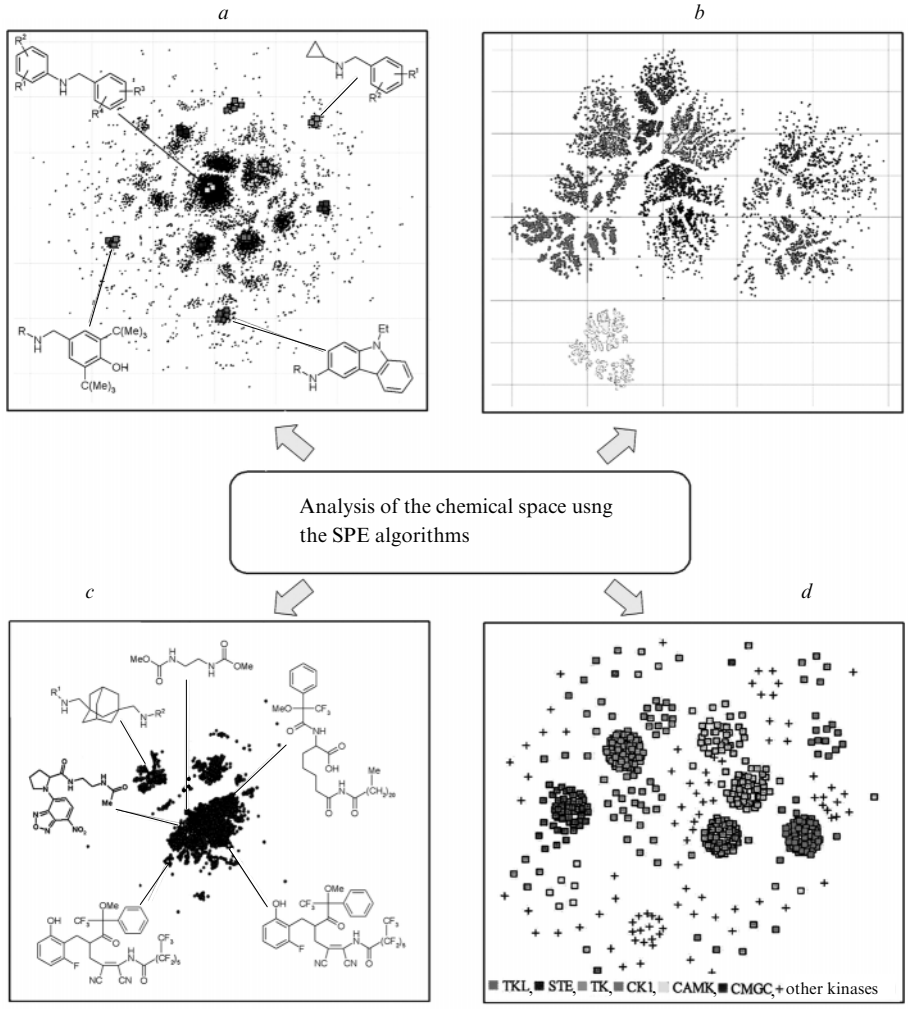


Figure 15. Application of the SPE algorithm to solve various problems in computer-assisted medicinal chemistry: the distribution of different compounds from a combinatorial library of compounds including products of virtual amination reaction (initial learning radus is 0.4) (*a*); the distribution of compounds including products of Ugi classical reaction (*b*); the distribution of different diamines (*c*) and localization of different types of kinases (initial learning radus is 0.89) (*d*).

overlap, thus affecting more distant objects. At the same time, the use of small learning radus can lead to distortion of the global topology and structural organization of the descriptor space of lower dimension, which thus will not correspond to the initial multidimensional space of attributes.

In spite of possible diffuculties, the SPE algorithm, as well as the Kohonen and Sammon mapping techniques mentioned above, is considered as one of the most efficient and exact methods of analysis of the chemical space. Suming up, it should be noted that the Kohonen, Sammon and SPE methods do not complete the list of nonlinear mapping methods. For instance, a nonlinear mapping technique called curvilinear component analysis (CCA) [86] is similar to the SPE algorithm and widely used to solve various problems in mathematical statitics, in particular, to analyze multidimensional data by reducing the dimension of the initial space of attributes.

V. Specialized software

There is a large number of software for computer simulation, which allows one to generate the Kohonen self-organizing maps and Sammon nonlinear maps. A few programmes are listed in Table 6. However, not all of them are efficient and convenient. Many programmes run under old operating system DOS, being console versions with no user interface. Such programmes are most often simple and can solve a variety of simple problems in computer simulation. At present, there is only one specialized programme called SmartMining (URL: http://www.Informagenesis.com), which was developed for computer-assisted medicinal chemistry and virtual screening and is based on the mapping algorithms.

The SmartMining programme is a module of a mathematically more powerful programme InformaGenesis. SmartMining is designed for processing chemical data. It allows one to carry out various types of research in the field of virtual screening and computer simulation. The SmartMining module incorporates many prognostic models, which can automatically predict a wide range of fundamental pharmaceutical characteristics of organic compounds including the target-specific activity profile, permeability through biological membranes, as well as many pharmaco- and toxicokinetic parameters. For instance, the SmartMining programme makes it possible to study the permeation of

Table 6. Commercially available software for constructing QSAR/QSPR models based on the Kohonen and Sammon maps (with separation using the SVM algorithm).

Programme	Algorithm	Software developer
Informagenesis	artificial neural networks, the Kohonen and Sammon maps	Informagenesis company (URL: http://www.informagenesis.com)
NeuroSolution	artificial neural networks and the Kohonen maps	NeuroDimension Inc., USA (URL: http://www.nd.com)
SVMlight	support vector machines	Technische Universität Dortmind, Germany (URL: http://svmlight.joachims.org)
ChemoSoft	artificial neural networks, the Sammon maps	Chemical Diversity Labs Inc., USA; CDRI, Russia (URL: http://www.ChemDiv.com)
OmniViz (Galaxy™)	the modified Sammon algorithm	BioWisdom, Ltd., United Kingdom (URL: http://www.omniviz.com)

compounds through BBB (using the Kohonen self-organizing maps) and HIA (using the Sammon maps).[64]

QSAR/QSPR techniques developed using the SmartMining programme are suitable for automated assessment of the properties of large compound libraries including up to a few millions of structures using conventional Windows-run personal computers. The experience in the use of QSAR/QSPR models points that in most cases computer-assisted prediction is much better than that made by a skilled expert working in a particular field of organic chemistry. The performance of computer methods and human expertise are incommeasurable at all and the gap continuously increases as the available computers become increasingly more powerful.

VI. Conclusion

At present, virtual screening is undoubtfully an essential part of the design of novel pharmaceuticals. Despite a relatively short history of specialized methods of computer-assisted medicinal chemistry, there is a number of efficient *in silico* technologies, which make it possible to perform rational selection of chemical structures to be used in biological tests and to carry out various research activities in this field. Among various QSAR/QSPR methods, the mapping techniques developed for generating two- and three-dimensional projections of the initial multidimensional space of molecular descriptors occupy a special position. Such maps have a great predictive power and are convenient for visualization; they allow researchers to perform a complex and, which is quite important, versatile analysis of the regularities established. The nonlinear projections thus obtained make it possible to elucidate the reasons and key factors responsible for separation of different categories of pharmaceuticals. The main goal of the mapping techniques is to construct relevant computer models suitable for prediction of various therapeutical and physicochemical properties of organic molecules (*e.g.*, target-specific activity profile, various pharmacokinetic parameters, solubility, thermal stability, *etc.*). This gives researchers an efficient and precise computer-assisted tool, which is indeed a must-have in the modern stage of development of organic synthesis and biological screening.

References

1. K C Nicolaou, E J Sorensen *Classic in Total Synthesis* (Weinheim: Wiley-VCH, 1996) p. 798
2. N K Terrett, M Gardner, D W Gordon, R J Kobylecki, J Steele *Tetrahedron* **51** 8135 (1995)
3. W Bannwarth, E Felder *Combinatorial Chemistry: a Practical Approach* (Weinheim: Wiley-VCH, 2000) p. 430
4. H-J Böhm, G Schneider *Virtual Screening for Bioactive Molecules* (Weinheim: Wiley-VCH, 2000)
5. W P Walters, M T Stahl, M A Murcko *Drug Discov. Today* **3** 160 (1998)
6. F L Stahura, J Bajorath *Comb. Chem. High. Throughput Screen.* **7** 259 (2004)
7. H Kubinyi, in *Methods and Principles in Medicinal Chemistry* Vol. 1 (Weinheim: Wiley-VCH, 1993) p. 4
8. J Bajorat *Nature Rev. Drug Discov.* **1** 882 (2002)
9. K V Balakin, Ya A Ivanenkov, A V Skorenko, S N Kovalenko, I A Zhuravel', V P Chernykh *Ukr. Zh. Org. Farm. Khim.* **2** 47 (2004)
10. C Merlot, D Domine, C Cleva, D J Church *Drug Discov. Today* **8** 594 (2003)
11. P Willett *J. Chem. Inf. Comput. Sci.* **38** 983 (1998)
12. R P Sheridan, S K Kearsley *Drug Discov. Today* **7** 903 (2002)
13. G A Patani, E J LaVoie *Chem. Rev.* **96** 3147 (1996)
14. J Krumrine, F Raubacher, N Brooijmans, I Kuntz *Meth. Biochem. Anal.* **44** 443 (2003)
15. K V Balakin, S E Tkachenko, I Okun, A V Skorenko, Y A Ivanenkov, N P Savchuk, A A Ivashchenko, Y Nikolsky *Chem. Today* **22** 15 (2004)
16. N P Savchuk, K V Balakin *Curr. Drug Discov.* **5** 23 (2004)
17. S Anzali, G Barnickel, B Cezanne, M Krug, D Filimonov, V Poroikov *J. Med. Chem.* **44** 2432 (2001)
18. V V Poroikov, D A Filimonov, W-D Ihlenfeldt, T A Gloriozova, A A Lagunin, Y V Borodina, A V Stepanchikova, M C Nicklaus *J. Chem. Inf. Comput. Sci.* **43** 228 (2003)
19. Y Borodina, A Sadym, D Filimonov, V Blinova, A Dmitriev, V Poroikov *J. Chem. Inf. Comput. Sci.* **43** 1636 (2003)
20. T Kohonen *Self-organizing Maps* (3rd Ed.) (New-York: Springer, 2001) p. 528
21. J W Sammon Jr *IEEE Trans. Comput.* **C18** 969 (1969)
22. J P Bigus *Data Mining with Neural Networks* (New York: McGraw-Hill, 1996) p. 220
23. J Devillers *Neural Networks in QSAR and Drug Design* (London: Academic Press, 1996) p. 284
24. V Vapnik *Statistical Learning Theory* (New York: Wiley, 1998) p. 768

25. R D Cramer, M A Poss, M A Hermsmeier, M A Hermsmeier, T J Caulfield, M C Kowala, M T Valentine *J. Med. Chem.* **42** 3919 (1999)
26. A D van Dijk, R Boelens, A M Bonvin *FEBS J.* **272** 293 (2005)
27. D B Kitchen, H Decornez, J R Furr, J Bajorath *Nat. Rev. Drug Discov.* **3** 935 (2004)
28. R D Taylor, P J Jewsbury, J W Essex *J. Comput. Aid. Mol. Des.* **16** 151 (2002)
29. P G Gane, P M Dean *Curr. Opin. Drug Discov. Dev.* **10** 401 (2000)
30. R D Brown, Y C Martin *J. Chem. Inf. Comput. Sci.* **36** 572 (1996)
31. M Bredel, E Jacoby *Nature Rev. Gen.* **5** 262 (2004).
32. S Frickenhaus, S Kannan, M Zacharias *J. Comput. Chem.* **30** 479 (2009)
33. D K Agrafiotis, V S Lobanov *J. Chem. Inf. Comput. Sci.* **40** 1356 (1997)
34. D K Agrafiotis *J. Comput. Chem.* **24** 1215 (2003)
35. M A Johnson, G M Maggiora *Concepts and Applications of Molecular Similarity* (Weinheim: Wiley-VCH, 1990)
36. S Lafon, A B Lee *IEEE Trans. Pattern Anal. Machine Intelligence* **28** 1393 (2006)
37. J B Tenenbaum, in *Advances in Neural Information Processing Systems* Vol. 10 (Cambridge: MIT Press, 1998) p. 682
38. H Bauknecht, A Zell, H Bayer, P Levi, M Wagener, J Sadowski, J Gasteiger *J. Chem. Inf. Comput. Sci.* **36** 1205 (1996)
39. S Anzali, G Barnickel, M Krug, J Sadowski, M Wagener, J Gasteiger, J Polanski *J. Comput. Aid. Mol. Des.* **10** 521 (1996)
40. M Brmstle, B Beck, S T Chindler, W King, T Mitchell, T Clark *J. Med. Chem.* **45** 3345 (2002)
41. A A Rabow, R H Shoemaker, E A Sausville, D G Covell *J. Med. Chem.* **45** 818 (2002)
42. P Labute, S Nilar, C Williams *Comb. Chem. High Throughput Screen.* **5** 135 (2002)
43. H Gao, C Williams, P Labute, J Bajorath *J. Chem. Inf. Comput. Sci.* **39** 164 (1999)
44. K V Balakin, S A Lang, A V Skorenko, S E Tkachenko, A A Ivashchenko, N P Savchuk *J. Chem. Inf. Comput. Sci.* **43** 1553 (2003)
45. V V Zernov, K V Balakin, A A Ivashchenko, N P Savchuk, I V Pletnev *J. Chem. Inf. Comput. Sci.* **43** 2048 (2003)
46. M T Cronin, A O Aptula, J C Dearden, J C Duffy, T I Netzeva, H Patel, P H Rowe, T W Schultz, A P Worth, K Voutzoulidis, G Schüürmann *J. Chem. Inf. Comput. Sci.* **42** 869 (2002)
47. Y Nikolsky, K V Balakin, Y A Ivanenkov, A A Ivashchenko, N P Savchuk *PharmaChem.* **4** 68 (2003)
48. K V Balakin *Modern Drug Discov.* **9** 45 (2003)
49. Y A Ivanenkov, K V Balakin, A V Skorenko, S E Tkachenko, N P Savchuk, A A Ivashchenko, Y Nikolsky *Chem. Today* **21** 72 (2003)
50. N P Savchuk, K V Balakin, in *Virtual Screening in Drug Discovery* (New York: CRC Press, 2005) p. 121
51. N P Savchuk, S E Tkachenko, K V Balakin, in *Cheminformatics in Drug Discovery* (Weinheim: Wiley-VCH, 2004) p. 287
52. K V Balakin, Doctoral Thesis in Chemical Sciences, Ivanovo State University, Ivanovo, 2005
53. *Derwent World Drug Index (Derwent WDI)* (London: Derwent Information, 1997)
54. Maybridge Chemicals Company Ltd., England
55. A G Wilson, A C White, R A Mueller *Curr. Opin. Drug Discov. Devel.* **6** 123 (2003)
56. M T Cronin *Exp. Suppl.* **93** 259 (2003)
57. M T Cronin *Altern. Lab. Anim. Suppl.* **2** 81 (2002)
58. J D McKinney, A Richard, C Waller, M C Newman, F Gerberick *Toxic. Sci.* **56** 8 (2000).
59. P Barbieri, N Piclin, A Szymoszek, M Novic, M Vracko, E Benfenati, in *Meeting of Slovenial Chemical Society 2001, FKKT Maribor, 2001* p. 74
60. C L Russom, S P Brandbury, S J Broderius, D E Hammermeister, D A Drummond *Environ. Toxicol. Chem.* **16** 948 (1997)
61. M Jianchang, A K Jain *IEEE Trans. Neural Netw.* **6** 296 (1995)
62. K V Balakin, Y A Ivanenkov, N P Savchuk, A A Ivashchenko, S Ekins *Curr. Drug Discov. Technol.* **2** 99 (2005)
63. K V Balakin, S Ekins, A Bugrim, Y A Ivanenkov, D Korolev, Y V Nikolsky, A V Skorenko, A A Ivashchenko *Drug Metab. Dispos.* **32** 1183 (2004)
64. U Norinder, M Haeberlein *Adv. Drug Deliv. Rev.* **54** 291 (2002)
65. K Thummel, D D Shen, in *Goodman and Gilman's the Pharmaceutical Basis of Therapeutics* (New York: McGraw-Hill, 2001) p. 1924
66. J E Penzotti, M L Lamb, E Evenson, P D J Grootenhuis *J. Med. Chem.* **45** 1737 (2002)
67. S Ekins, R B Kim, B F Leake, A H Dantzig, E Schuetz, L B Lan, K Yasuda, R L Shepard *Mol. Pharmacol.* **61** 964 (2002)
68. L Guttmann *Psychometrika* **19** 149 (1954)
69. R B Catell *Multi Behav. Res.* **1** 245 (1966)
70. R Todeschini, V Consonni *Handbook of Molecular Descriptors* (Weinheim: Wiley-VCH, 2002)
71. B Lerner, H Guterman, M Aladjem, I Dinstein *Pattern Recognit. Lett.* **20** 7 (1999)
72. B Lerner, H Guterman, I Dinstein, Y Romem *Pattern Recognit.* **28** 1673 (1995)
73. J Sadowski, M Wagener, J Gasteiger *Angew. Chem., Int. Ed. Engl.* **34** 2674 (1995)
74. P Gedeck, P Willett *Curr. Opin. Chem. Biol.* **5** 389 (2001)
75. P Toronen, M Kolehmainen, G Wong, E Castren *FEBS Lett.* **451** 142 (1999)
76. D K Agrafiotis, H Xu *Proc. Natl. Acad. Sci. USA* **99** 15869 (2002)
77. D K Agrafiotis, H Xu *J. Chem. Inf. Comput. Sci.* **43** 475 (2003)
78. D N Rassokhin, D K Agrafiotis *J. Mol. Graph. Model.* **22** 133 (2003)
79. D C Spellmeyer, A K Wong, M J Bower, J M Blaney *J. Mol. Graph. Model.* **15** 18 (1997)
80. T F Havel, K Wuthrich *J. Mol. Biol.* **182** 281 (1985)
81. F Zhu, D K Agrafiotis *J. Comput. Chem.* **28** 1234 (2007)
82. H Xu, S Izrailev, D K Agrafiotis *J. Chem. Inf. Comput. Sci.* **43** 1186 (2003)
83. D K Agrafiotis, D Bandyopadhyay, G Carta, A J S Knox, D G Lloyd *Chem. Biol. Drug Des.* **70** 123 (2007)
84. M A Farnum, H Xu, D K Agrafiotis *Protein Sci.* **12** 1604 (2003)
85. D K Agrafiotis *Protein Sci.* **6** 287 (1997)
86. P Demartines, J Herault *IEEE Trans. Neural Netw.* **8** 148 (1997)