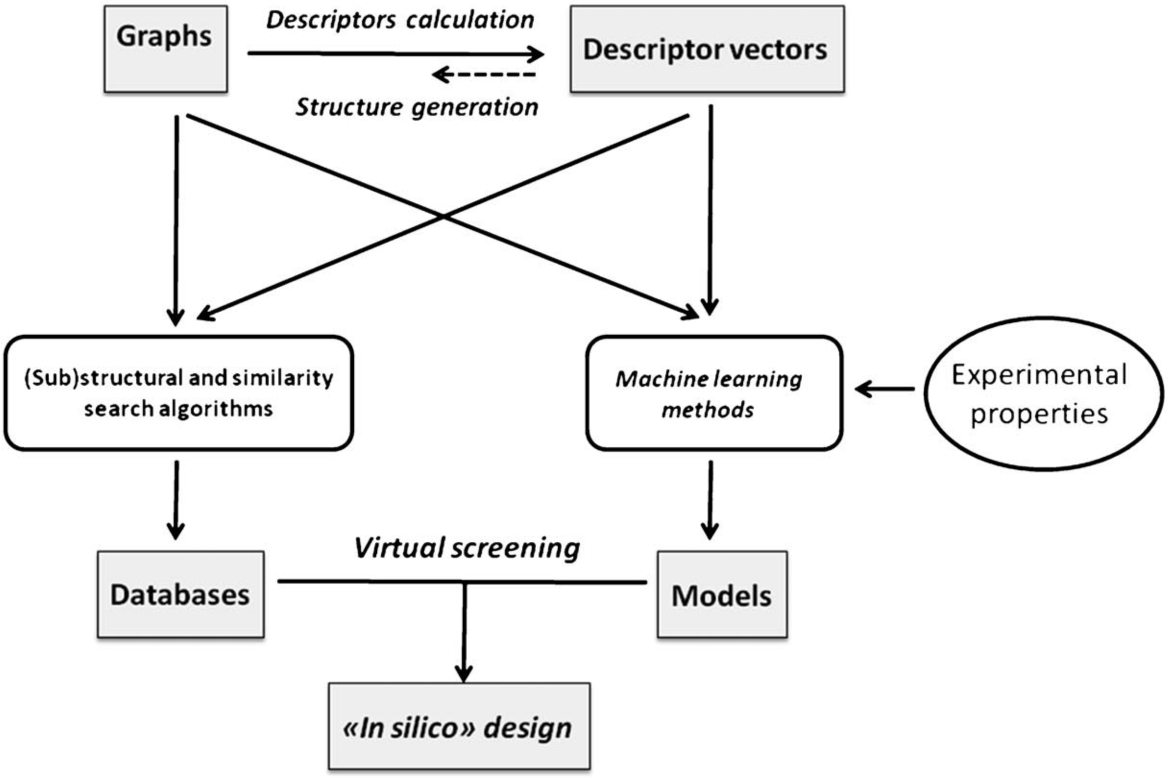
DOI: 10.1002/minf.201000100

Chemoinformatics as a Theoretical Chemistry Discipline

Alexandre Varnek*[a] and Igor I. Baskin[b]

Contribution to the 2nd Strasbourg Summer School on Chemoinformatics, VVF Obernai, France, 20–24 June 2010



Abstract: Here, chemoinformatics is considered as a theoretical chemistry discipline complementary to quantum chemistry and force-field molecular modeling. These three fields are compared with respect to molecular representation, inference mechanisms, basic concepts and application areas. A chemical space, a fundamental concept of chemoinformatics, is considered with respect to complex relations between chemical objects (graphs or descriptor vectors). Statistical Learning Theory, one of the main mathematical approaches in structure-property modeling, is briefly reviewed. Links between chemoinformatics and its "sister" fields – machine learning, chemometrics and bioinformatics are discussed.

Keywords: Chemoinformatics · Chemical space · Similarity · Computational learning theory

1 Introduction

Chemoinformatics, a young field incorporating several "old" fields (QSAR and chemical databases development),[1] is approaching maturity.[2–10] Indeed, it is widely applied in academia and industry (especially in the drug design area), it is taught in many universities at the undergraduate and graduate level, and there are several specialized international journals, as well as many international meetings being held every year. At the same time, it has not still been recognized as an individual scientific discipline, but mostly considered as an interface between chemistry and informatics, or as a collection of methods and tools specifically oriented toward drug design. This is clearly seen from the early definitions of chemoinformatics suggested by Brown, Paris, Gasteiger, and Faulon (Table 1). In fact, any scientific discipline should satisfy some obvious requirements: it should be based on its own concepts and approaches, and its differences from and complementarity to related disciplines must be clearly identified.

One of the ultimate applications of chemoinformatics is the development of models linking chemical structure and various molecular properties. This logically relates chemoinformatics with two other modeling approaches – quantum chemistry and force-field simulations. These three complementary fields differ with respect to the form of their molecular models, their basic concepts, inference mechanisms and domains of application (Table 2). Unlike the molecular models used in quantum mechanics (ensembles of nuclei and electrons) and force field molecular modeling (ensembles of "classical" atoms and bonds), chemoinformatics treats molecules as molecular graphs or related descriptor vectors with associated features (physicochemical properties, biological activity, 3D geometry, etc.) (Figure 1). The ensemble of graphs or descriptor vectors forms a *chemical space* in which some relations between the objects must be defined. Unlike real physical space, a chemical space is not unique: each ensemble of graphs and descriptors defines its own chemical space. *Thus, chemoinformatics could be defined as a scientific field based on the representation of molecules as objects (graphs or vectors) in a chemical space.*

Here, we attempt to define chemoinformatics as a theoretical chemistry discipline by characterizing its fundamental concepts and underlining its links with some "sister" disciplines. First, we present theoretical chemistry as an ensemble of three complementary disciplines: chemoinformatics, quantum chemistry and force field simulations. Then, we discuss two fundamental concepts of chemoinformatics: chemical space and statistical learning theory. Finally, some relations of chemoinformatics with machine learning, bioinformatics and chemometrics are discussed.

2 Complementarities of the Chemoinformatics, Quantum Chemistry and Force Field Approaches

2.1 Basic Molecular Models

Differences and complementarities of three theoretical chemistry disciplines – Chemoinformatics, Quantum Chemistry and Force Field approach – are directly related to the way they represent molecular structures, i.e., their basic molecular models (Table 2). Quantum chemistry (QC) explicitly considers ensembles of electrons and nuclei which are described by the Schrödinger wave equation. Since this equation can only be solved analytically for atoms with one electron, in practice various approximate methods (commonly Hartree–Fock or Density Functional Theory) are used. Since even such calculations are time-consuming, they are usually performed on single molecules or reactions in the gas phase, or on relatively small ensembles of molecules. The Force Field (FF) approach considers "classical" atoms and bonds and it uses empirical equations to calculate the molecular potential energy as a sum of terms corresponding to both bonding and nonbonding interactions. This approach can be easily coupled with classical mechanics, allowing one to calculate molecular trajectories (Molecular Dynamics simulations), or with statistical mechanics in order to generate Boltzmann ensembles (Monte-Carlo simulations), or, simply, with optimization techniques (Molecular Mechanics).[11] Due to the simplicity

[a] A. Varnek
Laboratoire d'Infochimie, UMR 7177 CNRS, Université de Strasbourg
4, rue B. Pascal, Strasbourg 67000, France
*e-mail: varnek@unistra.fr

[b] I. I. Baskin
Department of Chemistry, Moscow State University
Moscow 119991, Russia

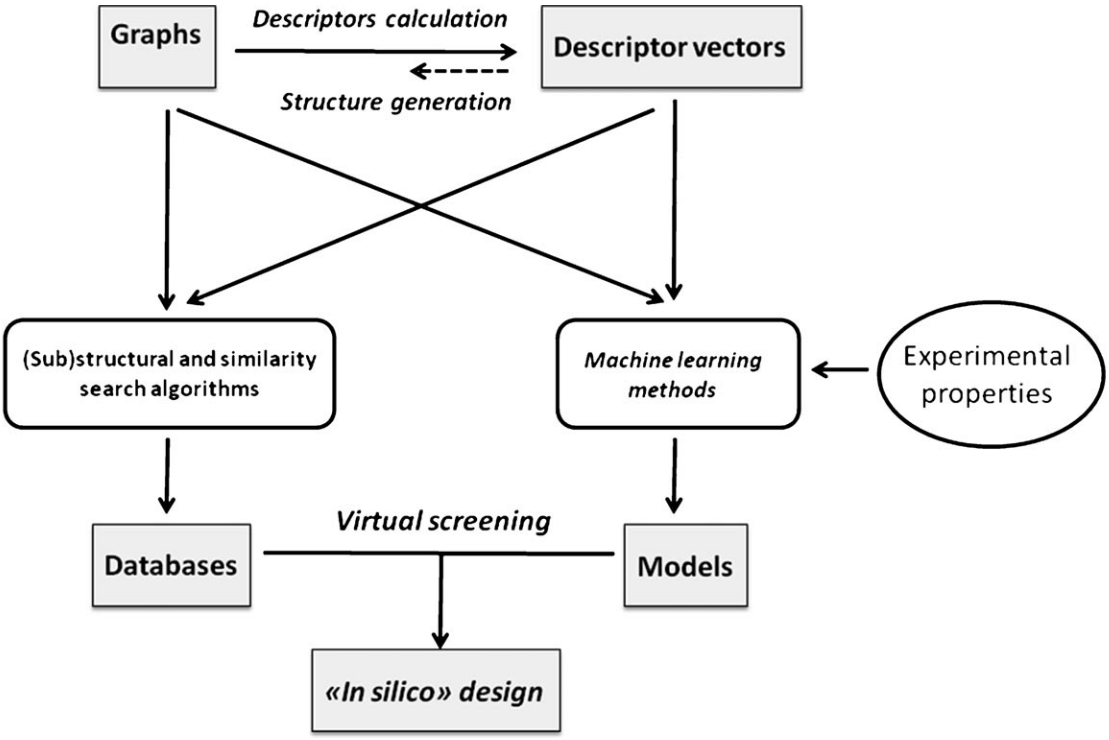
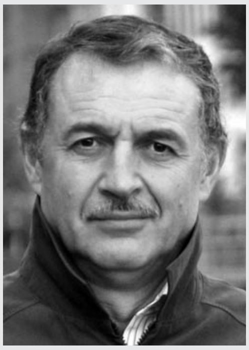


Figure 1. Chemoinformatics: from objects to major applications. Notice that for each Chemoinformatics Object (graph, descriptor vector in the input or in the feature space) there exist associated machine-learning approaches: graph based, vector-based or kernel-based methods, respectively.

of the basic molecular model and potential energy equations, Force Field methods can be applied to rather large systems containing many thousands of atoms (proteins, solutions, etc.). Chemoinformatics considers a molecule as a graph or an ensemble of descriptors generated from this graph. A set of molecules forms a chemical space for which the relationships between the objects themselves, on one hand, and between their chemical structures and related properties, on the other hand, are established using two main mathematical approaches: graph theory and statistical learning. Due to the rapidity of such calculations, these structure-property relationships can be applied to fast



Alexandre Varnek got his PhD in physical chemistry from the Institute of Inorganic and General Chemistry of the Russian Academy of Sciences, Moscow. In 1988–1995, he was Associate Professor in theoretical chemistry at the Moscow Mendeleyev University of Chemical Technology. In 1995, Alexandre joined the University of Strasbourg, France, where he holds the position of a Professor in theoretical chemistry, head of the laboratory on chemoinformatics and the director of the master courses on chemoinformatics. His research interests focus on the development of new approaches and tools for virtual screening and "in silico" design of new compounds and chemical reactions.

Igor Baskin received his PhD in organic chemistry (1990) and habilitation in mathematical and quantum chemistry (2010) from Lomonosov Moscow State University, Russia. After holding several positions at the Semenov Institute of Chemical Physics and Zelinsky Institute of Organic Chemistry of the Russian Academy of Sciences, Moscow, he joined in 2001 the Chemistry Department of Lomonosov Moscow State University, where since 2005 he holds the position of a Leading Scientist. He is regularly engaged as Visiting Scientist and Invited Professor at the University of Strasbourg, France. He has published more than 100 articles related to SAR/QSAR/QSPR methodology, artificial neural networks, medicinal chemistry, as well as molecular modeling of biological receptors and supramolecular systems. Igor Baskin is a member of the International Academy of Mathematical Chemistry since 2009. His current work focuses on the application of advances machine learning approaches in chemoinformatics.

Table 1. Different definitions of chemoinformatics as a field.

Frank Brown[5]	The use of information technology and management has become a critical part of the drug discovery process. Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organization.
Greg Paris[45]	Chemoinformatics is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization, and use of chemical information.
Johann Gasteiger[2]	Chemoinformatics is the application of informatics methods to solve chemical problems.
Jean-Loup Faulon and Andreas Bender[8]	Chemoinformatics is the field of handling chemical information
This work	Chemoinformatics is a field based on the representation of molecules as objects (graphs or vectors) in a chemical space.

Table 2. Interrelations between three branches of theoretical chemistry.

	Quantum chemistry	Force field based molecular modeling	Chemoinformatics
Molecular model	Electrons and Nuclei	Atoms and bonds	Graphs and descriptor vectors
Inference mechanism	Deductive$\gg$inductive	Deductive$\cong$inductive	Deductive$\ll$inductive
Typically applied to	Individual species or ensemble of a few species	Individual species, complex system representing an ensemble of many species	Ensemble of species (both for knowledge extraction and predictions), individual species (for predictions only)
Basic concept	Wave/particle dualism	Classical mechanics	Chemical space
Basic mathematical approaches	Schrödinger equation and approximate methods (HF, DFT, …)	Force field method and its implementation in molecular mechanics, molecular dynamics, Monte-Carlo and free energy perturbation techniques	Statistical learning, graph theory

screening of large databases. Any property for which a sufficient number of experimental data is available can be modeled in chemoinformatics, whereas this is not always so for QC and FF approaches.

Thus, Chemoinformatics, Quantum Chemistry and Force Field approaches are interrelated areas. Indeed, QC influenced the development of many popular molecular connectivity indices such as E-state, whereas molecular mechanics is indispensible part of 3D shape descriptors generation. On the other hand, various machine-learning methods can be used to fit the parameters in some QC and FF approaches.

Nonetheless, QC, FF and chemoinformatics are different, if highly complementary approaches. Each has its own application area, its advantages and problems. A good knowledge of the all these areas is beneficial for a theoretical chemist to enable selection of the most suitable tools for a particular task.

2.2 Inference in Chemoinformatics

One of the main distinctions of chemoinformatics from QC and FF concerns the inference (learning) mechanism. Quantum chemical studies are a typical example of deductive inference, where a general physical model is applied to particular molecules. In chemoinformatics, the logic of inference is different, because it is generally not based on existing physical theories. Chemoinformatics considers the world too complex to be a priori described by any set of rules. The incompleteness of our knowledge changes the inference paradigm: instead of searching for exact solutions, chemoinformatics applies *plausible* reasoning quantified by probability theory.[13] The rules (models) in chemoinformatics are not explicitly taken from rigorous physical models, but learned inductively from the data. Thus, in inductive learning, the models are the result of generalization of patterns in the data. More general models have a greater chance to be predictive. Various approaches to assess the generalization ability of models have been suggested in the statistical learning theory[14–15] that is the mathematical basis of modeling in chemoinformatics.

It should be noted that the inductive learning approach is also used to some extent in QC and FF methods. In quantum chemistry, the parameterization of the electron density functional[16] and pseudopotentials[17–18] is often based on empirical parameters fitted to experimental data, as is the case in numerous semi-empirical methods.[8,19] The number of these parameters is sometimes so great that some quantum chemical methods, like DFT with the functional M06[20] or B97D,[21] can be considered to define a sort of "Schrödinger force field".[16] In Force-Field simulations, inductive learning is at least as important as deductive, since potential energy calculations involve many empirical parameters.

3 Fundamentals of Chemoinformatics

For the objects in chemical space, chemoinformatics builds its models using two main mathematical approaches: graph theory and statistical learning. While these mathematical methods can be applied to other fields, the chemical space is a particular concept of chemoinformatics describing a way to handle ensembles of chemical structures.

3.1 Chemical Space Paradigm

As pointed out by C. Lipinski and A. Hopkins, "chemical space can be viewed as being analogous to the cosmological universe in its vastness, with chemical compounds populating space instead of stars".[20] Any attempt even to count the number of chemical compounds which potentially could be synthesized leads to combinatorial explosion and yields an absolutely unrealistic number estimated as more than 10^{60},[22] which exceeds the number of elemental particles in the cosmological universe. Clearly that this number is so huge that it is impossible not only to synthesize these molecules but even to generate computationally their structures. The goal of chemoinformatics is to find a rational way of representing this literally infinite chemical space and to navigate in this space. Efficient strategies for navigating chemical space are crucially important for the development of new biologically active compounds and the design of new drugs for medicine.[20] This is due to the fact that biologically active compounds of a certain type are not distributed evenly over the whole chemical space, but form very compact regions in it, like galaxies in the cosmological universe.[20] This is certainly true for any other chemical property. A special term, *chemography*, analogous to geography, has even been suggested for the art of navigating in chemical space.[23]

Although the expression "*Chemical space*" is widely used in the chemoinformatics literature, it is not still well defined. Generally speaking, the notion of "space" stands for a set of objects with some particular properties and some relationships between them (metric). Below, we consider two types of chemical objects (graphs and descriptor vectors), different metrics, and related chemical spaces.

3.1.1 Representation of Chemical Objects in Chemoinformatics

In chemoinformatics, the molecules are treated as informational objects, identifying their structure and properties. Generally, two main types of objects are used: graphs and descriptor vectors. In a vertex- and edge-labeled undirected graph, the vertices and edges correspond to atoms and chemical bonds, respectively. The vertex labels identify symbols of chemical elements, whereas the edge labels characterize the bond type. The label corresponds either to the bond order in molecules or to some special bond types in more complex systems. For instance, different types of "coordination" bonds can be defined for supramolecular systems, whereas "dynamic" bonds corresponding to chemical transformations can be used to encode chemical reactions.[24] More complex chemical systems, like polymers or mixtures can be described by ensembles of graphs.

For several practical purposes, more generalized representations of chemical structures are needed. For example, for pharmacophore analysis, the graph vertexes can be labeled as pharmacophoric centers (H-donors, H-acceptors, cation, anion, aliphatic, aromatic), while the separation of two centers can be depicted by an edge labeled by the value of the 2D or 3D distance.[25] In Markush structures used for patent searches, a graph vertex can stand for several types of either individual atoms or whole substructures (e.g., substituents).[26] The same is true for substructure queries used for searching chemical databases.[27]

Consideration of some complex chemical objects reveals, however, some limitations of graph theory to code chemical structures and their ensembles. Instead, hypergraphs[28] have been suggested as a more adequate mathematical model to encode stereochemical information and multicenter bonds. However, hypergraphs are much more difficult objects to operate compared to graphs, and, therefore, their use is still very limited.

Another popular representation of molecular structure is based on molecular descriptors defined by Todeschini and Consonni as "…the final result of a logical and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment."[29] This molecular representation is extremely popular in chemoinformatics because: (a) various descriptors can be generated from one and the same molecular graph, thus describing different facets of the information hidden in the graph; (b) it is invariant to any renumbering of graph vertices; (c) most of the descriptors are easy interpretable; (d) inductive transfer of knowledge can be performed via descriptors;[30] and, (e) descriptors define a vector space which is mathematically much easier to handle compared to the graph-based space. Descriptor vectors can be prepared not only for individual molecules but for more complex systems like chemical reactions[24] or multicomponent mixtures.[31] Nowadays, more than 5000 types of descriptors of different types have been reported.[29] They are used for database processing (as screens or fingerprints), for building SAR/QSAR/QSPR models, in similarity searching, clustering, etc.

At the same time, several weak points of molecular descriptors should be mentioned: (a) If descriptors are not well selected, in the resulting chemical space two different molecules can be superposed on one point; (b) The number of existing descriptors is very large and despite numerous variables selection techniques reported in the literature,[32] there is always a risk of selecting irrelevant and redundant descriptors; (c) A serious drawback of molecular descriptors is the loss of reciprocity with the molecular structure. Indeed, the reverse reconstruction of molecular graphs from descriptors is a very difficult and, in some cases, impossible task known in QSAR as the "inverse" problem.[33–34] From the practical point of view, it concerns generation of molecular structures possessing desired property values. Attempts to solve this problem have been reported by Gordeeva et al.,[35] Skvortsova et al.,[36] and Faulon et al.[37] who observed some degeneracy of solutions, when several chemical structures corresponded to one set of molecular descriptor values. As pointed out in,[38] this prevents a reverse engineering of chemical structures from molecular descriptors, but, on the other hand, can be useful to safely exchange chemical information in the form of molecular descriptors.

3.1.2 Chemical Similarity as a Metric of Chemical Space

By definition, a metric is a function which defines a distance between the elements of a set. For all *x*, *y*, *z*, this function must satisfy the following conditions: (i) $d(x, y) \geq 0$ (nonnegativity); (ii) $d(x, y) = d(y, x)$ (symmetry) and, (*iii*) $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality). Strictly speaking, the distance $d(x, z)$ is a dissimilarity measure which is zero for identical elements and increases with the decrease of similarity between them. Thus, it can be defined as *distance* = 1 − *similarity*. Some similarity measures are briefly considered below.

Molecular similarity (or *chemical similarity*) is one of the most basic concepts in chemoinformatics.[39–40] It is widely used in virtual screening and *in silico* design of new compounds. Such studies are based on the *similar property principle* which states that similar compounds have similar properties.[39] In application to classification problems this means that similar chemical compounds tend to belong to the same class (e.g., possessing similar biological activity), whereas as applied to regression problems it means that the approximating function should be as smooth as possible. It should also be pointed out that molecular similarity always depends on the choice of descriptors and methods to compare molecular graphs.

Chemical similarity measures described in the literature can be calculated from (a) molecular graphs; (b) descriptor vectors; (c) molecular fields; they can also be assessed from (d) kernels, and (e) unsupervised or (f) supervised modeling studies. This classification is rather fuzzy, and some similarity measures belong simultaneously to several classes. Some details are given below.

A similarity measure based on the size of the maximum common subgraph (MCS) for a pair of graphs is perhaps the most well-known graph-based similarity measure. Due to the relative complexity and inefficiency of computational algorithms to search for an MCS,[41] this approach, however, is rarely used to perform a similarity search[42] or to cluster chemical databases.[43]

Another type of graph-based similarity measure is that of graph kernels which assign to each pair of graphs a positive real number characterizing similarity.[44–45] They are used to map a graph-based chemical space to a vector (feature) space in which the structure–property model is built. This approach has been successfully used in SAR and QSAR.[46]

The most popular similarity measures are based on fixed-sized descriptor vectors. These are various types of distances (Euclidean, Manhattan, Mahalanobis, Minkowski) measuring molecular dissimilarity or some indices (Tanimoto, Dice, cosine, Tversky, etc.) measuring similarity. These measures are widely discussed in the literature, e.g., see the review paper by Willett[47] and references therein.

Several approaches have been developed to compare molecular fields. The Carbo index is computed by integrating overlaps of electronic densities of two molecules assessed using quantum-chemical approaches.[48–51] The SEAL index[52] is used to assess an alignment of steric and electrostatic fields of the molecules. Since any molecular field could be represented as a descriptor vector based on the field value on the grid points, a similarity measure can be simply calculated as the product of two vectors.

Similarity measures for which all matrices of values are semipositive definite (the determinant is larger or equal to zero) are called "Mercer kernels", or simply "kernels". Generally, kernels are used to project the objects (graphs or vectors) into a Hilbert "feature space", in which a similarity measure between these objects is equal to dot-product of their projections. A dot product of vectors, which can be viewed as the cosine similarity measure for normalized vectors, is the simplest type of kernel.

Unsupervised machine-learning methods of nonlinear neighborhood-preserving projections of data can also be used to assess similarity. A typical example is mapping to Self-Organizing (Kohonen) Maps, SOM,[53] where the similarity is measured as a distance between different cells. This offers the possibility to use SOMs for property predictions[54] and in virtual screening.[55]

If several QSAR models are simultaneously applied to predict a property for a series of compounds, the similarity can be assessed in the "models' space". Indeed, for each compound, one can form a vector based on the prediction results. A dot product of these vectors can be considered as a measure of the similarity of two molecules. This approach has been used by Tetko in the ASNN (Associative Neural Networks) method.[56]

Generally, similarity measures could be used both for similarity-based predictions and similarity searching.[39] Similarity-based prediction approaches in the initial descriptor space are based on the *k* nearest neighbors method (*k*NN). However, kernel similarity measures implemented in kernel-based machine learning methods lead generally to more computationally efficient and predictive models.[44] Both in similarity-based prediction methods and in querying large chemical databases, the computational efficiency largely depends on whether a given similarity measure defines a metric in chemical space.[57]

For most of similarity measures, the metric axioms (i)–(iii) are valid, and, therefore, they can be perceived as distances in chemical space.

3.1.3 Navigation in Graph-based Chemical Space

In principle, each ensemble of molecular graphs forms a discrete metric topological space. Its topology is defined by a set of all its possible subsets, where the simplest discrete metric gives the distance 0 if two chemical objects are equivalent, i.e. corresponding chemical graphs are isomorphic to each other and 1 otherwise. This simplest metric is however not useful in practical applications, because in such space all distinct objects are equally similar to each other. More flexible relationship between graphs can be expressed as a degree of their mutual similarity/dissimilarity. In particular, this relationship can be established by mapping an ensemble of graphs onto a descriptor vector space followed by an assessment of standard similarity measures.

The three main approaches used to describe a set of molecular graphs and to navigate in this space are: (a) substructure-based, (b) superstructure-based, and (c) mutation-based.

In the *substructure-based* approach a special "navigation" graph is usually constructed. It can be used for the visualization of chemical databases, exploring relations between compounds and discovering unexplored regions in the chemical space. In the navigation graph, the nodes correspond to individual molecular graphs and edges correspond to some transition rules. Bemis and Mursko have considered transitions between an unlabelled graph (framework) to a labeled graph (full chemical structure).[58–59] They invented the concept of molecular frameworks,[58–59] used to organize the structural data by grouping the atoms of each drug molecule into ring, linker, framework, and side chain atoms. Thus, a huge database can be described by a limited number of frameworks. In the "scaffold tree" graph approach of Schuffenhauer et al.,[60–61] transitions are allowed between a molecular graph and its subgraph. It has been demonstrated that this type of navigation graphs allows one to perform an efficient and intuitive activity mapping, visualization and navigation of the chemical space defined by a given library, which in turn leads to building correlations with bioactivity and further compound design.[62] Thus, the hierarchical scaffold classification proposed in[61] helps to chart biologically relevant chemical space using data on natural products. The idea of a "scaffold tree" is implemented in the open source "Scaffold Hunter" software,[63] an interactive tool for navigation in chemical space, which facilitates recognition of complex structural relationships associated with bioactivity.

To represent relationships in analogous series of compounds having the same scaffold and different substitution patterns, multilayer-rooted "combinatorial analogue graphs" (CAGs) have been proposed by Peltason et al.[19] These graphical representations hierarchically organize compounds according to substitution patterns and are annotated with SARI discontinuity scores[64] in order to account for SAR discontinuity at the level of functional groups. The approach makes it possible to identify undersampled regions and highlight key substitution patterns which determine the SAR of a compound series. An alternative way to visualize SARs in analogous series with a common scaffold is offered by the "SAR maps" invented by Agrafiotis et al.[65] In a "SAR map", each series is rendered as a rectangular matrix of cells, each representing a unique combination of substituents (i.e., a unique compound). Color-coding the cells by their potency easily identifies SAR patterns.

Pollock et al.[66] introduced the scaffold topology approach, which represents a connected graph with the minimum number of nodes and edges required to fully describe its ring structure. An algorithm for systematic generation of scaffold topologies allows one to analyze systematically all scaffold topologies for up to eight-ring molecules and four-valence atoms, thus providing coverage of the lower portion of the chemical space of small molecules.[66] Scaffold topology distributions were analyzed for several of the most popular chemical structure databases with huge number of compounds, both real and virtual, and many interesting features were found.[67] It is claimed that "scaffold topologies can be the first step toward an efficient coarse-grained classification scheme of the molecules found in chemical databases".[67]

In the *superstructure-based* approach, each individual molecular graph is considered as a subgraph of a common supergraph corresponding to the ensemble of individual graphs.[68] Although this approach is limited to relatively small congeneric sets of compounds, it has been found very suitable to build QSAR models, as demonstrated in the positional analysis by Magee,[69–70] the DARC/CALPHI system by Mercier et al.,[71] the MTD-PLS approach of Kurunczi et al.,[72–74] and the MFTA approach by Palyulin et al.[68,75] For each individual chemical structure, the occupancies of supergraph nodes or local physicochemical descriptors of atoms matching these nodes, form a fixed-size descriptor vector used in machine-learning methods as an input.

An alternative *mutation-based* approach to travel in graph-based chemical space has been suggested by van Deursen et al.[76] They represent a chemical space as a graph in which vertices correspond to individual molecules and edges correspond to structural mutations: change of atom type; inversion of stereochemical configuration at chiral centers, removal and addition of atom; saturation and unsaturation of bond; bond rearrangement; and aromatic ring addition. Traveling in such space from one active molecule to another one, one can discover along the trajectory a certain number of novel structures which can be further analyzed in the context of lead optimization. A similar approach has been reported by Bishop et al.[77] who suggested the use of chemical reactions as structural mutations connecting in the chemical space known organic compounds taken from the Beilstein database. The supergraph created in such a way enabled the authors to select a set of the "most useful compounds" from which the majority of chemical compounds can be synthesized.

3.1.4 Navigation in Descriptor-Based Chemical Space

Descriptor-based chemical space is a multidimensional space in which molecules are represented as vectors. Two main approaches – dimensionality reduction and clustering -are used to facilitate the navigation in this space.

Dimensionality reduction is achieved in classical multivariate data analysis by the Principal Component Analysis (PCA) procedure.[78–79] In PCA, several features (called "principal components") corresponding to the principal inertia axes of the "cloud" of data points in the initial descriptor space are used as axes of a new low-dimensional space, onto which the initial data points are projected. Such projection occurs with the minimal loss of information and, therefore, maximal conservation of the neighborhood relationships between data points. Thus, representation of the data points in the resulting low-dimensional space can be considered as a "navigation map" of the descriptor space. This idea has been implemented in the ChemGPS (chemical global positioning system) technique[23] which positions chemical structures in drug-like chemical space (drug space). This makes this approach as well as the related ChemGPS-NP[80–81] tool a well-suited reference system to compare multiple libraries and to keep track of previously explored regions of the chemical descriptor space.[23]

Although the axes of the PCA "navigation map" are orthogonal, corresponding latent variables are statistically independent only for a Gaussian distribution of data points. Since this distribution in the descriptor space is usually strongly non-Gaussian, this can hamper the chemical interpretability of particular latent variables and reduce the usefulness of the whole "navigation map". To solve this problem, Independent Component Analysis (ICA) has been suggested.[82–85] It has been demonstrated that the application of ICA instead of PCA yields chemically more readily interpretable latent variables.[86]

Hierarchical cluster analysis represents an alternative approach to navigate in the descriptor space. The resulting dendrogram gives a clear picture of the neighborhood relations between chemical objects, although for a large number of compounds it becomes too burdensome.

The combined application of dimensionality reduction and clustering methods is realized in Kohonen Self-Organizing Maps (SOM).[53] In SOMs, the dimensionality reduction is achieved by embedding a net of neurons onto a 2D surface. The SOMs provide more efficient solutions than PCA, because the former are more suitable to analyze complex topological structures of the descriptor space. The ability of SOMs to build "navigation maps" for visualizing chemical space has been demonstrated on GPCR ligands,[54] toxic compounds,[87] inhibitors of P-glycoprotein[88] and different organic reactions.[89]

A set of chemical structures can be presented as a graph in which the vertices correspond to individual molecules and the edges connecting them correspond to certain neighborhood relations.[90] This technique has been used to represent relationships between different classes of drug molecules,[91] to elucidate similarity relationships within the sets of active compounds,[92] and to explore structure-selectivity relationships.[93]

Hierarchical clustering techniques using some similarity measures also offer the possibility of analyzing large chemical data sets. Thus, Agrafiotis et al.[94] have used radial clusterograms, different segments of which are color-coded by biological activity or any other user-defined property.

To characterize structure–activity landscapes in the descriptor-based chemical space, SARI and SALI indices have been suggested. The SARI index[64] globally characterizes structure-activity landscapes. It consists of two terms: the continuity score which measures the potency-weighted structural diversity, and the discontinuity score calculated as the average potency difference among similar pairs of molecules. The SALI index[95] is local, considering two related molecules, and it is often used to quantify "activity cliffs".[96]

3.2 Modeling Background

The two main mathematical approaches used in chemoinformatics are graph theory and computational learning theory. Whilst the chemical applications of graphs are described in numerous books and review articles (e.g., see Bonchev[97]), the latter is described mostly in the data-mining literature. Here, we give some general information about some basic concepts of computational learning theory.

3.2.1 Computational Learning Theory

In recent years, in statistical modeling there has been a shift from the classical statistical paradigm of "model parameterization" to a new paradigm of "predictive flexible modeling". The first paradigm supposes that the functional dependence between the input and output data is established from some external knowledge and the goal of the statistical study is to find a few independent free parameters by fitting to experimental data. This usually requires a certain number of experimental observations per each free parameter. Unfortunately, this requirement can be met only in very few cases, e.g., within the classical Hansch-Fujita approach based on three descriptors only.[98] The aim of the second paradigm is to build models with maximal predictive performance by fitting to experimental data rather flexible families of functions involving large numbers of intercorrelated parameters. Such a setup is evidently much more appropriate for most chemoinformatics studies. The first attempts to implement the second paradigm in the framework of so-called nonparametric statistical analysis failed because of the "curse of dimensionality" (which required a huge number of observations exponentially growing with the number of free parameters).[99] Nonetheless, early works on predictive modeling were successfully carried out using completely heuristic methodologies of artificial neural networks[100–101] and decision trees.[102] For the first time, a strong theoretical background to build statistical models using finite (even small) data sets was developed by Vapnik in his Statistical Learning Theory (SLT).[14] This approach, together with that developed later as the PAC (Probably Approximately Correct) theory by Valiant[15] and the MDL (Minimum Description Length) concept by Rissanen[103] constitute the basis of modern computational learning theory.

According to SLT, the goal of statistical study is to choose from a given set of functions $f(x, \theta)$ the "best" one $f(x, \theta^*)$ with the minimum value of the risk functional $R[f]$, which is defined as an expected prediction error on new data taken from the same distribution as the training set (i.e., the mean prediction performance on all possible test sets). Here x denotes the variables (descriptors in QSAR studies) and θ the adjustable parameters. Another important characteristic is the empirical risk functional R_{emp} $[f]$, which is defined as an error on the training set (fitting error). For regression tasks, R_{emp} is usually calculated as:

$$R_{\mathrm{emp}}[f] = \frac{1}{N}\sum_{i=1}^{N} (y_i - f(x_i, \theta))^2 \tag{1}$$

Here i denotes the observations (compounds in QSAR studies) in the training set, N is the size of the training set, y_i is the response value in i-th observation (the property value of i-th compound in QSAR studies). According to Vapnik,[14,99] for the classification tasks the risk can be estimated as

$$R[f] \leq R_{\mathrm{emp}}[f] + c(h,N) \tag{2}$$

Here, the complexity term $c(h,N)$ characterizes the flexibility of the set of functions $f(x, \theta)$ to fit experimental data. It increases with the Vapnik-Chervonenkis (VC) dimension h and decreases with the number of data N. It follows from Equation 2 that in order to obtain a predictive_model, one should minimize both the empirical risk R_{emp} (i.e., fitting error) and the complexity term.

In fact, the notion of complexity is related to the smoothness of functions for regression tasks. If f is not flexible enough, the complexity term is small, but R_{emp} could be large (underfitting). Too complex (flexible) f perfectly fits the data, thus reducing R_{emp}. On the other hand, f could fit not only a trend but also noise in the data, thus increasing the complexity term (overfitting). Thus, to minimize the risk $R[f]$, one should find a compromise between R_{emp} and the complexity term in (2). This can be achieved by introduction of some trade-off parameters depending on particular machine learning method. For example, these include the number of descriptors in multiple linear regression models with variables selection; the ridge value in the ridge regression; the number of "leaves" in decision trees; the number of iterations in neural networks; the parameter k in kNN and the C and ν parameters in SVM calculations. These parameters should be optimized in order to achieve the best prediction performance of the model.

One of the most interesting conclusions of SLT is that the value of the complexity term does not directly depend on the number of free parameters θ in the function class f, the flexibility (capacity, complexity) of which is measured by the VC dimension h. The value of h can be considered as an "effective" number of free parameters. (Note that h is equal to the number of free parameters in classical multiple linear regression without descriptor selection).

According to SLT, h is controlled by the trade-off parameter used to simultaneously minimize both terms in Equation 2. This offers an opportunity to build models with any (even very huge) number of variables using kernel approaches, which approximate nonlinear functional dependencies of any form by projecting descriptors onto a feature space of any (even infinite) dimensionality and build linear models in this feature space.[44]

Nowadays, computation learning theory represents a quickly developing area. Thus recently, a Bayesian learning approach to predictive flexible modeling has been described.[104] Instead of one single model (as in STL), it considers the whole statistical distributions of models weighted by their ability to fit data, thus allowing one to make probabilistic predictions by averaging these distributions. This approach has come to be rather popular in chemoinformatics: its implementations in Bayesian Neural Networks,[105] Gaussian Processes,[106] and Bayesian Networks[107] have been recently published.

3.2.2 Different Facets of Statistical Modeling

It should be pointed out that the range of application of different statistical (machine learning) methods in chemoinformatics is currently very wide (Figure 2). Most of the existing machine learning approaches can provisionally be divided into two large families: supervised and unsupervised machine learning. (Some other approaches – semisupervised, active and multi-instant learning – are very rarely used in chemistry so far).

The goal of the supervised learning in chemistry is to predict physicochemical properties and biological activities of chemical compounds. The quantitative prediction of real-valued properties is performed by regression models,

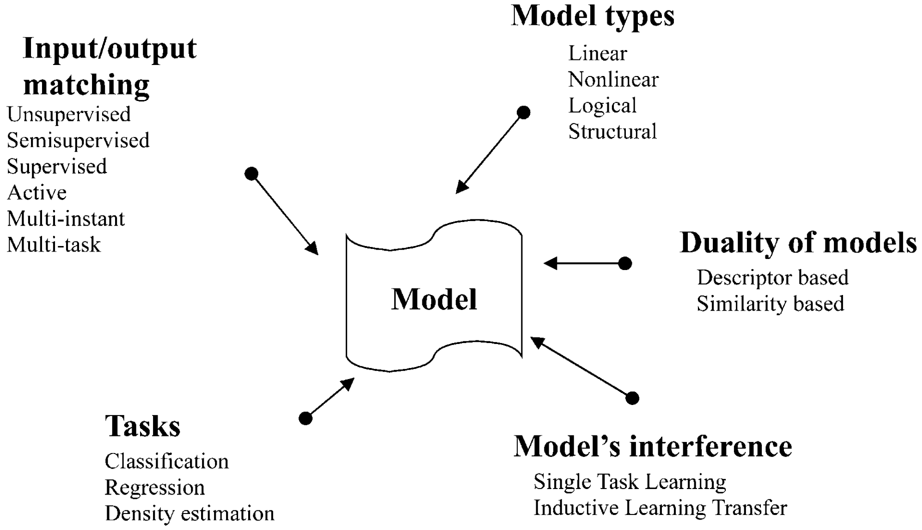


Figure 2. Different approaches to model description.

whereas qualitative predictions ("active" or "inactive"?) are assessed in classification models. The most popular regression methods currently used in chemoinformatics applications are multiple linear regression (MLR), partial least squares (PLS), neural networks, support vector regression (SVR), and *k*NN, whereas the naïve Bayes, support vector machines (SVM), neural networks and classification trees (especially the Random Forest method[108]) are widely used for classification. There are also ranking models,[109] in which ranking order instead of property values are predicted, and models with structured output,[110] in which predicted values belong to classes of any complexity. Models of the latter two types can be built using some special modifications of SVM.

Unsupervised learning describes the data and reveals their hidden patterns. The most important tasks treated by unsupervised modeling approaches are: (a) cluster analysis (data reduction); (b) dimensionality reduction; (c) novelty (outlier) detection. All these tasks can be perceived as particular cases of data density estimation. Many standard algorithms for both nonhierarchical (e.g., *k*-means) and hierarchical clustering algorithms are used. The most popular algorithms for dimensionality reduction are PCA (Principal Component Analysis) and ICA (Independent Component Analysis). Tasks (a) and (b) are solved simultaneously in the Kohonen Self-Organizing Maps (SOMs),[53] which are intensively used for the purposes of visualization and analysis of the chemical space. The ability of several machine learning methods, such as one-class SVM,[44] to tackle the problem of novelty detection is currently used to define the applicability domains of QSAR/QSPR models[111] as well as in virtual screening experiments.[112]

With respect to data description, two types of models – primal and dual – can be identified. Primal models are based on the direct use of descriptors, whereas dual models are based on measures describing similarity relationships between chemical structures. Kernels represent the most useful types of such measures; they can be computed both from molecular descriptors and by direct comparison of chemical structures. Both primal and dual approaches can be used within supervised and unsupervised modeling tasks.

Finally, statistical models can be built for a net of mutually related models, in which their predictive performance can be leveraged due to Inductive Learning Transfer phenomenon,[30,113] in the framework of the Multi-Task Learning and Feature Net approaches.[30]

4 Relations of Chemoinformatics with the "Sister" Disciplines

4.1 Chemoinformatics and Machine Learning

Although machine learning is widely used for structure-property modeling, chemoinformatics can be considered as a very specific area of its application. The specificity of chemoinformatics results from (i) the nature of chemical objects, (ii) the complexity of the chemical universe and (iii) a possibility to take into account an extra-knowledge.

The basic chemical object is a graph (or hypergraph), rather than simple fixed-sized vector of numbers as in the typical applications in mathematical statistics and machine learning. This dictates the need to apply graph theory, to develop novel descriptors and structured graph kernels, and to apply machine learning methods capable of dealing with structured discrete data.

The second important distinction comes from the fact that the chemical data result from an explorative process in a huge chemical space rather than from specially organized sampling. Hence, they cannot be considered as representative, independent and identically distributed sampling from a well defined distribution. Thus, special approaches are needed to treat this problem: various strategies to explore chemical space, the "applicability domain" concept, the active learning approach, etc.

Finally, one can use the relationships between different properties issued from physicochemical theory. (For example, the Arrhenius law could be particularly useful upon the modeling the rate constants). These relationships could be integrated into chemoinformatics workflow as an external knowledge.

4.2 Chemoinformatics and Chemometrics

Massart[114] has defined c*hemometrics as "a chemical discipline that applies mathematics, statistics and formal logic (a) to design and select optimal experimental procedures; (b) to provide maximum relevant chemical information by analyzing chemical data; and (c) to obtain knowledge about chemical systems"*. Generally, chemometrics requires no information about chemical structure and, therefore it overlaps with chemoinformatics only in the area of application of machine learning methods. It is widely used in experiment design, chemical engineering, analytical chemistry and treatment of spectra – fields where an exhaustive treatment of multivariate data is needed.

4.3 Chemoinformatics and Bioinformatics

Unlike chemoinformatics dealing with "chemical size" molecules, bioinformatics uses computational tools to study the structure and function of biomolecules (proteins, nucleic acids). This is a broad field mostly involving 3D (force field and quantum mechanics calculations) and 1D (sequence alignment) modeling. In the latter, a biomolecule is represented as a string of characters (building blocks). Graph and fixed size vector models used in chemoinformatics are very rarely used in bioinformatics. In this sense, chemo- and bioinformatics are "complementary". On the other hand, there are many examples of interpenetration of these fields. Thus, in docking calculations, protein structures could be generated by bioinformatics tools, whereas some scoring functions involve vector representation of ligands.

Another way to combine bio- and chemoinformatic approaches is related to the construction of protein-ligand descriptors or fingerprints based on available 3D information about protein-ligand complexes. Thus, Tropsha et al.[115] developed CoLiBRI descriptors calculated for a pseudomolecule constructed from interacting atoms of the protein and the ligand. Marcou and Rognan[116] have developed "interaction fingerprints" accounting for eight interaction types per each protein atom interacting with the ligand: hydrophobic; aromatic (face to face); aromatic (edge to face); H-bond (protein donor atom); H-bond (protein acceptor atom); ionic (positively charged protein atom); ionic (negatively charged protein atom); metal complexation., Langer et al.[117] have reported a technique to build pharmacophoric ligand models based on the analysis of 3D protein-ligand structures.

A promising way to describe ligand–receptor complexes concerns construction of protein-ligand kernels (PLK) as products of "chemical" ligand–ligand (LLK) and "biological" protein–protein kernels (PPK). The resulting feature space for PLK is a tensor product of the features spaces corresponding to LLK and PPK. Machine learning models involving PLK are based on the idea that similar ligands bind to similar proteins. Using these kernels, one can predict binding potency of both different ligands with respect to a given protein, and different proteins with respect to a given ligand. Several articles describing PPK have been published. Erhan et al. combined "chemical" kernels based on MOE descriptors and "biological" kernels based on protein-ligand "interaction fingerprints".[118] Faulon et al.[119] used the signature molecular descriptors to calculate "chemical" and "biological" Tanimoto kernels. Jacob and Vert[120] combined a Tanimoto kernel for the ligands and several types of kernels for the proteins. In particular, for PPK they compared either protein sequences or EC numbers. Bajorath et al.[121] used a linear kernel for the ligands and protein-protein kernels calculated from sequence identity matrix.

5 Conclusions

Here, chemoinformatics has been described as a fundamental theoretical chemistry discipline complementary to quantum chemistry and force-field molecular modeling. Chemoinformatics represents molecules as graphs or descriptor vectors whose ensembles form, respectively, graph-based or descriptor-based spaces. Chemical similarity measures or hierarchical relationships between graphs are used as metrics in the chemical space. Chemoinformatics uses two main mathematical approaches – graph theory and statistical learning theory; the latter is briefly described here.

In this paper, we have not aimed to describe all facets of chemoinformatics, but have attempted to delineate some important points identifying this field as an independent scientific discipline. This view is probably incomplete. However, we hope it will initiate a discussion which in any case could be useful for the chemoinformatics community.

Acknowledgements

We thank Prof. *L. Morin-Allory*, Dr. *P. Vayer* and Dr. *G. Marcou* for fruitful discussion and Prof. *J. Harrowfield* for his help and advice.

References

[1] J. Gasteiger, *Anal. Bioanal. Chem.* **2006**, *384*, 57 – 64.
[2] J. Gasteiger, T. Engel, *Chemoinformatics: A Textbook*, Wiley-VCH, Weinheim, **2003**.

[3] J. Gasteiger, *Handbook of Chemoinformatics: From Data to Knowledge,* Wiley-VCH, Weinheim, **2003**.

[4] J. Bajorath, *Mol. Divers.* **2002**, *5*, 305 – 313.

[5] N. Brown, *Computing Surveys* **2006**.

[6] W. L. Chen, *J. Chem. Inf. Model.* **2006**, *46*, 2230 – 2255.

[7] T. Engel, *J. Chem. Inf. Model.* **2006**, *46*, 2267 – 2277.

[8] J.-L. Faulon, A. Bender, *Handbook of Chemoinformatics Algorithms*, CRC Press, Boca Raton, **2010**.

[9] B. Nathan, *ACM Comput. Surv.* **2009**, *41*, 1 – 38.

[10] I. Baskin, A. Varnek, in *Chemoinformatics Approaches to Virtual Screening* (Eds: A. Varnek, A. Tropsha), RSC Publisher, Cambridge, **2008**, pp. 1 – 43.

[11] A. R. Leach, *Molecular Modelling Principles and Applications*, 2nd ed, Prentice Hall, Upper Saddle River, **2001**.

[12] N. Brown, *ACM Comput. Surv.* **2009**, *41*, 1 – 38.

[13] E. T. Jaynes, *Probability Theory. The Logic of Science*, Cambridge University Press, Cambridge, **2003**.

[14] V. Vapnik, *Statistical Learning Theory*, Wiley-Interscience, New York, **1998**.

[15] L. G. Valiant, *Commun. ACM* **1984**, *27*, 1134 – 1142.

[16] A. Nicholls, in *ACS Fall 2009 National Meeting & Exposition*, **2009**, p. CINF55.

[17] T. R. Cundari, M. T. Benson, M. L. Lutz, S. O. Sommerer, *Rev. Comput. Chem.* **1996**, *8*, 145 – 202.

[18] G. Frenking, I. Antes, M. Bahme, S. Dapprich, A. W. Ehlers, V. Jonas, A. Neuhaus, M. Otto, R. Stegmann, A. Veldkamp, S. F. Vyboishchikov, *Rev. Comput. Chem.* **1996**, *8*, 63 – 144.

[19] L. Peltason, N. Weskamp, A. Teckentrup, J. Bajorath, *J. Med. Chem.* **2009**, *52*, 3212 – 3224.

[20] Y. Zhao, D. G. Truhlar, *Theor. Chem. Acc.* **2008**, *120*, 215 – 241.

[21] S. Grimme, *J. Comput. Chem.* **2006**, *27*, 1787 – 1799.

[22] C. M. Dobson, *Nature* **2004**, *432*, 824 – 828.

[23] T. I. Oprea, J. Gottfries, *J. Comb. Chem.* **2001**, *3*, 157 – 166.

[24] A. Varnek, D. Fourches, F. Hoonakker, V. P. Solov'ev, *J. Comput. Aided Mol. Des.* **2005**, *19*, 693 – 703.

[25] T. Langer, R. D. Hoffman, *Pharmacophores and Pharmacophore Searches*, Wiley-VCH, Weinheim, **2000**.

[26] J. M. Barnard, *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 64 – 68.

[27] U. Schoch-Grübler, *Online Inform. Rev.* **1990**, *14*, 95 – 108.

[28] C. Berge, *Hypergraphs*, Elsevier, Amsterdam, **1989**.

[29] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, **2000**.

[30] A. Varnek, C. Gaudin, G. Marcou, I. Baskin, A. K. Pandey, I. V. Tetko, *J. Chem. Inf. Model.* **2009**, *49*, 133 – 144.

[31] N. M. Halberstam, I. I. Baskin, V. A. Palyulin, N. S. Zefirov, *Dokl. Chem. (Engl. Transl.)* **2002**, *384*, 140 – 143.

[32] D. J. Livingstone, D. W. Salt, *Rev. Comput. Chem.* **2005**, *21*, 287 – 348.

[33] I. I. Baskin, E. V. Gordeeva, R. O. Devdariani, N. S. Zefirov, V. A. Palyulin, M. I. Stankevich, *Dokl. Akad. Nauk. SSSR* **1989**, *307*, 613 – 617 [Chem].

[34] M. I. Skvortsova, I. I. Baskin, V. A. Palyulin, O. L. Slovokhotova, N. S. Zefirov, in *AIP Conf. Proc. 330. E.C.C.C.1 Comput. Chem. F.E.C.S. Conf., Nancy, France* (Eds: F. Bernardi, J.-L. Rivail), AIP Press, Woodbury, New York, **1995**, pp. 486 – 499.

[35] E. V. Gordeeva, M. S. Molchanova, N. S. Zefirov, *Tetrahedron Comput. Methodol.* **1990**, *3*, 389 – 415.

[36] M. I. Skvortsova, I. I. Baskin, O. L. Slovokhotova, V. A. Palyulin, N. S. Zefirov, *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 630 – 634.

[37] J.-L. Faulon, C. J. Churchwell, D. P. Visco, Jr., *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 721 – 734.

[38] J. L. Faulon, W. M. Brown, S. Martin, *J. Comput. Aided Mol. Des.* **2005**, *19*, 637 – 650.

[39] A. M. Johnson, G. M. Maggiora, *Concepts and Applications of Molecular Similarity*, Wiley, New York, **1990**.

[40] N. Nikolova, J. Jaworska, *QSAR Comb. Sci.* **2003**, *22*, 1006 – 1026.

[41] J. W. Raymond, P. Willett, *J. Comput. Aided Mol. Des.* **2002**, *16*, 521 – 533.

[42] T. R. Hagadone, *J. Chem. Inf. Model.* **1992**, *32*, 515 – 521.

[43] I. L. Ruiz, C. G. Garcia, M. A. Gomez-Nieto, *J. Chem. Inf. Model.* **2005**, *45*, 1178 – 1194.

[44] B. Schölkopf, A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, MA, USA, **2002**.

[45] G. Paris, (August 1999 Meeting of the American Chemical Society), quoted by W. Warr at http://www.warr.com/warrzone.htm.

[46] M. Rupp, G. Schneider, *Mol. Inf.* **2010**, *29*, 266 – 273.

[47] P. Willett, J. M. Barnard, G. M. Downs, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983 – 996.

[48] E. Besalu, X. Girones, L. Amat, R. Carbo-Dorca, *Acc. Chem. Res.* **2002**, *35*, 289 – 295.

[49] X. Fradera, L. Amat, E. Besalu, R. Carbo-Dorca, *Quant. Struct.-Act. Rel.* **1997**, *16*, 25 – 32.

[50] A. Gallegos, R. Carbo-Dorca, R. Ponec, K. Waisser, *Int. J. Pharm.* **2004**, *269*, 51 – 60.

[51] X. Girones, L. Amat, R. Carbo-Dorca, *SAR QSAR Environ. Res.* **1999**, *10*, 545 – 556.

[52] S. K. Kearsley, G. M. Smith, *Tetrahedron Comput. Methodol.* **1990**, *3*, 615 – 633.

[53] T. Kohonen, *Self-Organizing Maps*, Springer, **2001**.

[54] M. von Korff, M. Steger, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1137 – 1147.

[55] D. Hristozov, T. I. Oprea, J. Gasteiger, *J. Chem. Inf. Model.* **2007**, *47*, 2044 – 2062.

[56] I. V. Tetko, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 717 – 728.

[57] T. G. Kristensen, *J. Math. Chem.* **2010**, *48*, 287 – 289.

[58] G. W. Bemis, M. A. Murcko, *J. Med. Chem.* **1996**, *39*, 2887 – 2893.

[59] G. W. Bemis, M. A. Murcko, *J. Med. Chem.* **1999**, *42*, 5095 – 5099.

[60] A. Schuffenhauer, P. Ertl, S. Roggo, S. Wetzel, M. A. Koch, H. Waldmann, *J. Chem. Inf. Model.* **2007**, *47*, 47 – 58.

[61] M. A. Koch, A. Schuffenhauer, M. Scheck, S. Wetzel, M. Casaulta, A. Odermatt, P. Ertl, H. Waldmann, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 17272 – 17277.

[62] S. Renner, W. A. L. Van Otterlo, M. Dominguez Seoane, S. Möcklinghoff, B. Hofmann, S. Wetzel, A. Schuffenhauer, P. Ertl, T. I. Oprea, D. Steinhilber, L. Brunsveld, D. Rauh, H. Waldmann, *Nature Chem. Biol.* **2009**, *5*, 585 – 592.

[63] S. Wetzel, K. Klein, S. Renner, D. Rauh, T. I. Oprea, P. Mutzel, H. Waldmann, *Nature Chem. Biol.* **2009**, *5*, 581 – 583.

[64] L. Peltason, J. Bajorath, *J. Med. Chem.* **2007**, *50*, 5571 – 5578.

[65] D. K. Agrafiotis, M. Shemanarev, P. J. Connolly, M. Farnum, V. S. Lobanov, *J. Med. Chem.* **2007**, *50*, 5926 – 5937.

[66] S. N. Pollock, E. A. Coutsias, M. J. Wester, T. I. Oprea, *J. Chem. Inf. Model.* **2008**, *48*, 1304 – 1310.

[67] M. J. Wester, S. N. Pollock, E. A. Coutsias, T. K. Allu, S. Muresan, T. I. Oprea, *J. Chem. Inf. Model.* **2008**, *48*, 1311 – 1324.

[68] E. V. Radchenko, V. A. Palyulin, N. S. Zefirov, in *Chemoinformatics Approaches to Virtual Screening* (Eds: A. Varnek, A. Tropsha), RSC, **2008**, pp. 150 – 181.

[69] P. S. Magee, *Quant. Struct.-Act. Rel.* **1990**, *9*, 202 – 215.

[70] P. S. Magee, in *QSAR: Rational Approaches to the Design of Bioactive Compounds* (Eds: C. Silipo, A. Vittoria), Elsevier, Amsterdam, **1991**.

[71] C. Mercier, V. Fabart, Y. Sobel, J. E. Dubois, *J. Med. Chem.* **1991**, *34*, 934–942.
[72] L. Kurunczi, E. Seclaman, T. I. Oprea, L. Crisan, Z. Simon, *J. Chem. Inf. Model.* **2005**, *45*, 1275–1281.
[73] L. Kurunczi, M. Olah, T. I. Oprea, C. Bologa, Z. Simon, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 841–846.
[74] T. I. Oprea, L. Kurunczi, M. Olah, Z. Simon, *SAR QSAR Environ. Res.* **2001**, *12*, 75–92.
[75] V. A. Palyulin, E. V. Radchenko, N. S. Zefirov, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 659–667.
[76] R. Van Deursen, J. L. Reymond, *ChemMedChem* **2007**, *2*, 636–640.
[77] K. J. M. Bishop, R. Klajn, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2006**, *45*, 5348–5354.
[78] K. Varmuza, in *Handbook of Chemoinformatics. From Data to Knowledge* (Ed: J. Gasteiger), Wiley-VCH, Weinheim, **2003**, pp. 1098–1133.
[79] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed., Springer, Heidelberg **2002**.
[80] J. Larsson, J. Gottfries, L. Bohlin, A. Backlund, *J. Nat. Prod.* **2005**, *68*, 985–991.
[81] J. Larsson, J. Gottfries, S. Muresan, A. Backlund, *J. Nat. Prod.* **2007**, *70*, 789–794.
[82] A. Hyvarinen, *Acta Polytech. Sc. Ma.* **1997**, *88*.
[83] A. Hyvarinen, E. Oja, *Neural Comput.* **1997**, *9*, 1483–1492.
[84] A. Hyvarinen, E. Oja, *Neural Networks* **2000**, *13*, 411–430.
[85] J. Chen, X. Z. Wang, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 992–1001.
[86] M. G. Gustafsson, *J. Chem. Inf. Model.* **2005**, *45*, 1244–1255.
[87] P. Mazzatorta, M. Vracko, A. Jezierska, E. Benfenati, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 485–492.
[88] Y.-H. Wang, Y. Li, S.-L. Yang, L. Yang, *J. Chem. Inf. Model.* **2005**, *45*, 750–757.
[89] H. Satoh, O. Sacher, T. Nakata, L. Chen, J. Gasteiger, K. Funatsu, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 210–219.
[90] A. Tropsha, D. Fourches, *Chem. Central J.* **2009**, *3*.
[91] J. Hert, M. J. Keiser, J. J. Irwin, T. I. Oprea, B. K. Shoichet, *J. Chem. Inf. Model.* **2008**, *48*, 755–765.
[92] M. Wawer, L. Peltason, N. Weskamp, A. Teckentrup, J. Bajorath, *J. Med. Chem.* **2008**, *51*, 6075–6084.
[93] L. Peltason, Y. Hu, J. Bajorath, *ChemMedChem* **2009**, *4*, 1864–1873.
[94] D. K. Agrafiotis, D. Bandyopadhyay, M. Farnum, *J. Chem. Inf. Model.* **2007**, *47*, 69–75.
[95] R. Guha, J. H. Van Drie, *J. Chem. Inf. Model.* **2008**, *48*, 646–658.
[96] G. M. Maggiora, *J. Chem. Inf. Model.* **2006**, *46*, 1535–1535.
[97] D. Bonchev, D. H. Rouvray, *Chemical Graph Theory. Introduction and Fundamentals*, Gordon and Breach, New York, **1991**, p. 300.
[98] C. Hansch, T. Fujita, *J. Am. Chem. Soc.* **1964**, *86*, 1616–1626.
[99] V. Cherkassky, F. Mulier, *Learning from Data: Concept, Theory and Methods.*, 2nd ed., Wiley, Hoboken, New Jersey, **2007**.
[100] J. Zupan, J. Gasteiger, *Neural Networks in Chemistry*, Wiley-VCH, Weinheim, **1999**.
[101] I. I. Baskin, V. A. Palyulin, N. S. Zefirov, *Methods Mol. Biol.* **2008**, *458*, 137–158.
[102] L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen, *Classification and Regression Trees*, Chapman & Hall/CRC, Wadsworth, CA **1984**.
[103] J. Rissanen, *Ann. Stat.* **1983**, *11*, 416–431.
[104] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, **2006**.
[105] F. R. Burden, D. A. Winkler, *J. Med. Chem.* **1999**, *42*, 3183–3187.
[106] O. Obrezanova, G. Csanyi, J. M. R. Gola, M. D. Segall, *J. Chem. Inf. Model.* **2007**, *47*, 1847–1857.
[107] A. Abdo, B. Chen, C. Mueller, N. Salim, P. Willett, *J. Chem. Inf. Model.* **2010**, *50*, 1012–1020.
[108] L. Breiman, *Mach. Learn.* **2001**, *45*, 5–32.
[109] S. Agarwal, D. Dugar, S. Sengupta, *J. Chem. Inf. Model.* **2010**, *50*, 716–731.
[110] T. Joachims, T. Hofmann, Y. Yue, C. N. Yu, *Commun. ACM* **2009**, *52*, 97–104.
[111] I. I. Baskin, N. Kireeva, A. Varnek, *Mol. Inf.* **2010**, *29*, 581–587.
[112] N. Fechner, A. Jahn, G. Hinselmann, A. Zell, *J. Cheminformatics* **2010**, *2*.
[113] I. I. Baskin, N. I. Zhokhova, V. A. Palyulin, A. N. Zefirov, N. S. Zefirov, *Dokl. Chem. (Engl. Transl.)* **2009**, *427*, 172–175.
[114] D. L. Massart, *Handbook of Chemometrics and Qualimetrics*, Elsevier, New York, **1998**.
[115] S. Oloff, S. Zhang, N. Sukumar, C. Breneman, A. Tropsha, *J. Chem. Inf. Model.* **2006**, *46*, 844–851.
[116] G. Marcou, D. Rognan, *J. Chem. Inf. Model.* **2007**, *47*, 195–207.
[117] C. Laggner, G. Wolber, J. Kirchmair, D. Schuster, T. Langer, in *Chemoinformatics Approaches to Virtual Screening* (Eds: A. Varnek, A. Tropsha), RSC Publisher, Cambridge, **2008**, pp. 76–101.
[118] D. Erhan, P.-J. L'Heureux, S. Y. Yue, Y. Bengio, *J. Chem. Inf. Model.* **2006**, *46*, 626–635.
[119] J. L. Faulon, M. Misra, S. Martin, K. Sale, R. Sapra, *Bioinformatics* **2008**, *24*, 225–233.
[120] L. Jacob, J. P. Vert, *Bioinformatics* **2008**, *24*, 2149–2156.
[121] H. Geppert, J. Humrich, D. Stumpfe, T. Gaertner, J. Bajorath, *J. Chem. Inf. Model.* **2009**, *49*, 767–779.

Received: August 31, 2010
Accepted: January 14, 2011
Published online: January 24, 2011