

Comprehensive Computational Assessment of ADME Properties Using Mapping Techniques

Konstantin V. Balakin, Yan A. Ivanenkov, Nikolay P. Savchuk, Andrey A. Ivashchenko and Sean Ekins*

Chemical Diversity, Inc., 11558 Sorrento Valley Road, San Diego, CA 92121, USA; *GeneGo, Inc., 500 Renaissance Drive Suite 106, St Joseph, MI 49085, USA

Abstract: One strategy to potentially improve the success of drug discovery is to apply computational approaches early in the process to select molecules and scaffolds with ideal binding and physicochemical properties. Numerous algorithms and different molecular descriptors have been used for modeling ligand-protein interactions as well as absorption, distribution, metabolism and excretion (ADME) properties. In most cases a single data set has been evaluated with one approach or multiple algorithms that have been compared for a single dataset. These models have been primarily evaluated by leave-one out analysis or boot strapping with groups representing 25-50% of the training set left out of the final model. In a very few examples a test set of molecules not included in the model has been used for an external evaluation. In the present study we have applied Sammon non-linear maps, Support Vector Machines and Kohonen Self Organizing Maps to modeling numerous datasets for ADME properties including human intestinal absorption, blood brain barrier permeability, cytochrome P450 binding, plasma protein binding, P-gp inhibition, volume of distribution and plasma half life.

Keywords: Sammon non-linear maps, support vector machines, self-organizing maps, ADME, absorption, cytochrome P450, volume of distribution, plasma half-life, P-glycoprotein.

INTRODUCTION

Poor pharmacokinetics and toxicity are important causes of costly late-stage failures in drug development, and it is widely recognized that these issues should be addressed as early as possible in the drug discovery process. Despite the development of a variety of medium and high-throughput *in vitro* screens in recent years, the modern high-throughput synthesis and screening technologies have enormously increased the number of compounds for which early absorption, distribution, metabolism and excretion (ADME) data are needed. We are currently witnessing a movement towards the wider utilization of various computational technologies at all stages of the process of drug discovery. For instance various computational scoring methods for predicting ligand-protein interactions have been applied to docking or *de novo* growth in the binding site of therapeutic proteins [1-5]. However binding to the target protein is only part of the process of drug discovery which requires molecules that are readily synthesizable with favorable molecular properties or what has been termed 'drug likeness' [6-9]. Drug likeness studies are a clear attempt to understand the chemical properties that make molecules either successful or possibly expensive clinical failures. Similarly the contribution of molecular properties which influence ADME are recognized alongside therapeutic potency as key determinants of whether a molecule can be successfully developed as a drug [10-12]. Many of these molecular

properties are used as later lead selection criteria that can be predicted before molecules are synthesized, purchased or even tested in order to improve overall lead quality. Some research groups have used relatively simple filters like the rule of 5 [13] and others [14] to limit the types of molecules evaluated with *in silico* methods and to focus libraries for high throughput screening. However, the proactive use of computational models for ADME properties has been less widely described in conjunction with lead discovery.

We now describe several *in silico* approaches, which increase our ability to predict some important pharmacokinetic, metabolic (and possibly also toxicity) endpoints. The approaches are based on advanced algorithms of dimensionality reduction and data visualization, Sammon non-linear maps, Support Vector Machines and Kohonen Self Organizing Maps. The developed models are useful for virtual ADME profiling of combinatorial libraries and selecting compounds for *in vitro* and *in vivo* testing. This type of approach will aid in improving the decision making process of compound selection in drug discovery.

MATERIALS AND METHODS

ADME Data

We used an annotated library consisting of approximately 1500 drugs and pharmaceutical leads with experimentally determined ADME properties. Each compound in the studied database was characterized by at least one ADME-related property: CYP3A4 binding affinity (CYP), plasma protein binding affinity (PPB), human intestinal absorption (by passive diffusion mechanisms) (HIA), blood-brain barrier permeability (by passive diffusion mechanisms) (BBB), P-

*Address correspondence to this author at the Vice President, Computational Biology, GeneGo, Inc. 500 Renaissance Drive, Suite 106, St. Joseph, MI 49085, USA; Tel: 269-930-0974; Fax: 269-983-7654; E-mail: sean@gene.go.com, ekinssean@yahoo.com

gp binding affinity (P-gp), volume of distribution (VD), and plasma half-life ($T_{1/2}$). The data were collected from the literature (for example, [15-19]).

Molecular Descriptors

A wide range of molecular descriptors of different types were calculated for all compounds with the Dragon (University of Milano, 2000) and ChemoSoft™ (Chemical Diversity Labs, Inc., 2004) software tools. These descriptors included electronic, topological, spatial, structural, and thermodynamic descriptors. A total of more than 1000 initial

descriptors were calculated for each compound. To reduce the number of descriptors that could contain redundant information, principal component analysis (PCA) was performed. Usually, about 90% of the variance could be explained by the first 6-10 PCs. The significance of the PCs was tested using standard Kaiser-Guttman and Scree tests [20, 21]. Descriptors maximally contributing to the first significant PCs, were selected based on these results as the most relevant and were used as input parameters in all further computational experiments (Table 1). The same descriptors were used for generation of Sammon and Kohonen maps.

Table 1. Descriptors used for generation of *in silico* models.

	Descriptor	Definition	CYP	PPB	HIA	BBB	P-gp	V _d	T _{1/2}
1	MLOGP	Moriguchi octanol-water partition coeff. (logP)	+	+	+	+	-	+	+
2	PSA	Fragment-based polar surface area	-	-	+	+	-	+	+
3	ATS7v	Broto-Moreau autocorrelation of a topological structure/weighted by atomic van der Waals volumes	-	+	-	-	-	-	-
4	ATS4e		-	-	-	-	-	+	-
5	BELp3	Eigenvalues of Burden matrix/weighted by atomic polarizabilities	-	-	-	-	+	-	-
6	BELv1		-	-	-	-	-	-	+
7	BEHp2		-	+	-	-	-	-	-
8	HATS7p	Leverage-weighted autocorrelation of lag 7 / weighted by atomic polarizabilities	-	-	-	-	-	-	+
9	H3u	H autocorrelation/ unweighted	-	-	-	-	+	-	-
10	CIC0	Complementary information content (neighborhood symmetry of 0-order)	-	-	-	-	+	-	-
11	RDF025v	Radial distribution function/ weighted by atomic Van der Waals volumes	-	-	-	-	+	-	-
12	HTu	H total index/unweighted	-	-	-	-	+	-	-
13	IVDM	Mean information content vertex degree magnitude	-	+	-	-	-	-	-
14	Ui	Unsaturation index	-	+	-	-	-	-	-
15	nBM	Number of multiple bonds	-	+	-	-	-	-	-
16	Ms	Mean electrotopological state	-	-	-	-	-	+	-
17	SIC1	Structural information content (neighborhood symmetry)	-	-	-	-	-	+	-
18	GATS2e	Geary autocorrelation/ weighted by atomic sanderson electronegativities	-	-	-	-	-	+	-
19	IC1	Information content index (neighborhood symmetry)	-	-	-	-	-	+	-
20	AAC	Mean information index on atomic composition	-	-	-	-	-	+	-
21	nHAcc	Number of acceptor atoms for H-bonds (N O F)	+	-	+	+	-	-	-
22	nHDon	Number of donor atoms for H-bonds (with N and O)	+	-	+	+	-	-	-
23	ZM1V	Zagreb index by valence vertex degrees	+	-	+	+	-	-	-
24	RBN	Number of rotatable bonds	+	-	+	+	-	-	-

Sammon Non-Linear Mapping and Support Vector Machines

The Sammon map [22] generation was conducted using a program developed internally at Chemical Diversity Labs as part of the ChemoSoft™ software suite. The non-linear map was built based on the following parameters: maximal number of iterations 300, optimization step 0.3; Euclidean distance was used as a similarity measure. The Sammon NLM procedure allows the creation of a 2-D image of the studied multi-dimensional property space.

For visual discrimination of the studied drug categories on the map we used the separation lines. The positioning of the separation line was determined using the nonlinear Support Vector Machine (SVM) algorithm [23] as implemented in the LibSVM-2.4 program (URL: <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>). The separation line represents the largest margin separating the studied compounds classes which is defined as the sum of the shortest distances from the decision line to the closest points for both classes, and thus can serve as an optimal discriminator between the two studied compound categories. All the calculations were performed using the ChemoSoft software suite.

Kohonen Self-Organizing Maps

The generation of the Kohonen self-organizing maps (SOMs) [24] was also conducted using the ChemoSoft software. The training parameters for the SOM were as follows: the number of interactions for the training runs was 2000, the starting adjustment radius for the training runs was 0.1, and the decay factor was 0.001. After the SOM was generated, we studied the distribution of various compound groups (such as strong or poor binders etc.) as separate maps.

RESULTS

Data Visualization and Non-Linear Maps

The visual analysis of multivariate data sets has established itself as a powerful means in data mining to detect non-obvious and relevant information for further exploitation. The human data analyst, with domain-specific experience (as well as perceptive and associative capabilities in the search process) are able to ascertain structure and correlations in data to ultimately provide potentially useful and exploitable information. Visual analysis thus provides qualitative information for efficient extraction. With the increasing importance of high-throughput synthesis and bioassay technologies, the resulting increase in data volume requires effective methods for visualization and interpretation of the obtained information. In the case of compounds represented by a set of descriptor values, the available data are multidimensional. To explore such information, it is necessary to map the data points into two- or three-dimensional space. The aim of the mapping procedure is to preserve the topology of the multidimensional space, so that data points that are close together in the multidimensional space will be close together in the low-dimensional space.

Numerous options and methods for projection computation based on neural and statistical approaches, have been introduced in the last three decades. Specifically,

topology and distance preserving mappings, e.g. using the self-organizing feature map (SOM) of Kohonen [24], or the distance preserving nonlinear mapping (NLM) of Sammon [22], are well suited for data visualization and data mining purposes as well as for analysis of ADME properties. NLM is an advanced multivariate statistical technique, which approximates local geometric relationships on a two- or three-dimensional plot. In theory, the NLM is especially attractive for data visualization and data mining, as the resulting mapping gives an insight into the presence and the structure of clusters in the data, and each projection point corresponds with a data entry. By contrast to SOM, non-linear maps represent all relative distances between all pairs of compounds in the descriptor space in a two-dimensional map. The distance between two points on the map directly reflects the similarity of the compounds. NLMs have previously been used for the visualization of protein sequence relationships in two dimensions, and comparisons between large compound collections, represented by a set of molecular descriptors [25, 26]. The purpose of this series of experiments is to show the possibility of discrimination between the principal categories of compounds present in the studied data sets (e.g. poor and strong binders, highly- and poorly permeable compounds, etc.).

Human Intestinal Absorption

Human intestinal absorption (HIA) is a major issue in the design, optimization, and selection of candidates for development of orally active pharmaceuticals (for example, [27-30]). The combination of several molecular descriptors believed to be important for HIA has been used in various multivariate analysis studies. The physicochemical properties of drugs, such as lipophilicity, molecular weight, ionization profile, H-bonding capacity, determine the extent to which drugs can cross the cellular barriers using passive diffusion mechanisms. In general, the molecular properties affecting HIA via passive diffusion mechanisms are well understood, and the reported models adequately describe this phenomenon [27-30]. Nevertheless, while much effort continues to be expended in this field with some success on existing datasets, perhaps the most pressing need at this time is for considerably larger, high-quality sets of experimental data and for an effective data mining algorithm to provide a sound basis for model building.

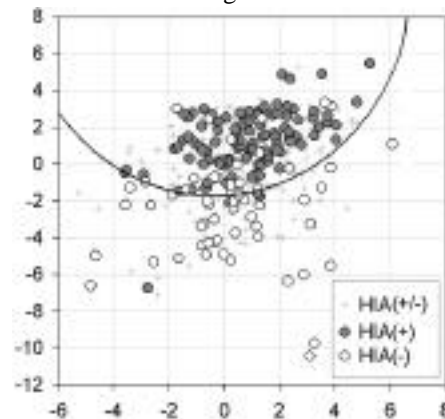


Fig. (1). Sammon NLM with SVM classification of compounds based on their experimental human intestinal absorption.

We studied the possibility of *in silico* recognition of compounds with poor intestinal absorption. The model is based on a relatively large training set consisting of 320 drugs with known values of HIA and six calculated molecular descriptors (Table 1) encoding some properties crucial for effective penetration through biological membranes. After the non-linear map was generated, we observed statistically significant differences in the molecular properties of highly-absorbed (HIA >80%) and poorly absorbed (HIA <20%) drugs (Fig. 1). Obviously, these categories of compounds occupied distinctly different areas on the map, and these differences can be used for assessment of HIA profile for novel compounds. Compounds belonging to the intermediate category with HIA =20-80% (shown as grey crosses) occupy a wide area on the map overlapping with the sites of location of HIA(+) and HIA(-) compounds.

Blood-Brain Barrier Permeability

The optimization of the distribution of therapeutic compounds between brain and blood is a very important issue in the design of CNS-active drugs. For drugs targeting the CNS, blood-brain barrier (BBB) penetration is a necessary attribute unless invasive or carrier-based strategies are being considered. On the other hand, for drugs aimed at other sites of action, BBB permeation would be undesirable as it can produce unwanted side-effects. Considering the great importance of the problem, the development of a reliable method of effective pre-synthetic assessment of BBB permeability is a requirement in the discovery of CNS-active agents. There have been many published computational models for this property using a variety of techniques (for example, [31-34]).

We have created a qualitative model for the early assessment of the BBB permeability of therapeutically relevant molecules. A comprehensive set of experimental BBB-permeability data on 456 compounds was collected. It was assumed that only passive diffusion mechanisms are involved in the BBB transport of these compounds. Statistical analysis enabled the selection of an optimal set of molecular descriptors for the effective prediction of BBB penetration. The projection of the combined data set of well-

and poorly permeable compounds onto a Sammon map was generated (Fig. 2). The data set of BBB-permeable compounds occupies a distinct area on the map that is substantially different from the regions of localization of BBB poorly permeable agents. Therefore, the sites of compound's localization on the map can be used for the assessment of its BBB-permeability with reasonable confidence.

Cytochrome P450 Binding

The majority of drugs undergo metabolism via the cytochrome P450 (CYP) enzymes [35] to result in either inactivate or active metabolites mediated by fewer than a dozen unique enzymes in human liver [36]. CYPs often have distinct roles in xenobiotic metabolism with active sites that enable broad and overlapping substrate specificity, complicated by ligand binding promiscuity [37]. Various *in vitro* systems are now widely used to study metabolism [38, 39] and characterize the potential for drug-drug interactions mediated by CYP enzymes. Lipophilicity expressed as log P or molecular refractivity was one of the first important properties found to be important for enzyme substrate binding [40, 41]. Later steric, electronic and molecular shape properties were also considered important for enzyme binding and transformation while metabolite release likely requires the opposite properties to binding [42]. Lewis and co-workers have provided many QSAR studies that enabled them to suggest a simple decision tree for human CYP substrates [43]. Computational pharmacophore type models have also been used to describe key molecular features of ligands for human CYPs [44-46] providing insight into the important features for interaction of ligands and the proteins. Within this family of enzymes, the CYP3A enzymes are the most important in terms of human drug metabolism because they have very broad substrate specificity [47, 48]. Recently more complex QSAR methods such as Kohonen maps have been useful for differentiating high and low affinity CYP3A4 substrates [15] and has also been applied to the differentiation of CYP substrates from non-substrates [49, 50]. Neural networks have been used to predict N-dealkylation rates for CYP3A4 and CYP2D6 substrates [51]. A k-nearest neighbour

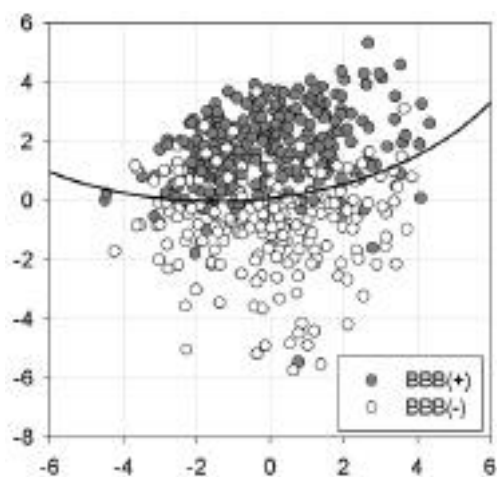


Fig. (2). Sammon map with SVM classification of compounds based on their ability to cross the blood-brain barrier.

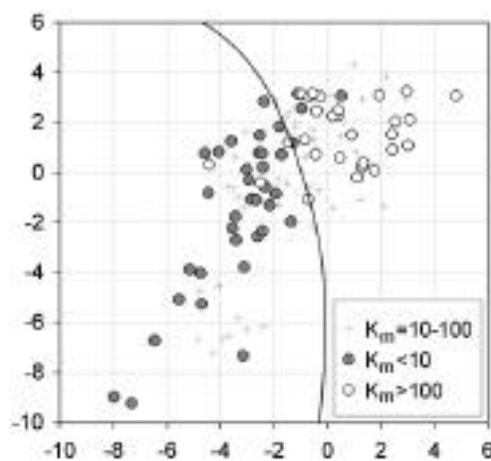


Fig. (3). Sammon map with SVM classification of compounds based on strong and weak binders to the active site of human CYP3A4.

statistical model using metabolic stability data from human S9 homogenate for 631 diverse molecules was able to adequately classify metabolism of a further set of over 100 molecules [52].

We have used non-linear maps for the assessment of ability of drugs and drug-like compounds to bind to the cytochrome P450s (CYPs). This approach was applied for CYP specific classification of nearly 500 drug compounds. We observed statistically significant differences in the molecular properties of strong ($K_m < 10$) and weak ($K_m > 100$) binders for various CYPs (K_m – Michaelis constant). The weak and strong binders occupied distinctly different areas of the map for all the groups. For illustration, we show the Sammon map of strong and weak binders for the CYP3A4 isozyme only (Fig. 3), which represents the largest isozyme-specific group in the dataset studied.

There exists a strong correlation between P450 binding affinity and the reversible competitive P450 inhibition of drugs. CYP inhibition is thought to be the most common cause of drug-drug interactions, and several prominent drugs have been withdrawn from the market due to such undesirable effects. Although inhibition of CYP enzymes *in vitro* is not necessarily associated with drug-drug interactions in clinical studies, lead compounds with weak CYP inhibition are favored in drug development based on these considerations. Reliable *in silico* methods for assessing CYP inhibition or substrate affinity can provide a valuable complement to the early stage selection of lead compounds.

Plasma Protein Binding

Plasma proteins are the major vehicle for transport and buffering of drug compounds. Understanding of drug-target and drug-plasma protein binding characteristics throughout the course of the drug development process is essential in the ADME evaluation of novel drug candidates. The clinical potential of drug compounds is greatly affected by the nature of their interactions with circulating plasma proteins, such as human serum albumin (HSA) and α_1 -acid glycoprotein (AGP) [53, 54]. Plasma protein binding varies greatly and affects the free concentration of the drug in the circulation, as well as transport and distribution in the body, and the duration of drug action.

A definite relationship was observed between plasma-protein binding (in percentage of drug bound or unbound) and $\log D$ at pH 7.4 [55]. Another approach was based on 107 descriptors and the genetic function approximation (GFA) algorithm [53]. For a set of 80 compounds, a QSAR with 12 descriptors and a correlation coefficient $r = 0.91$ between measured and predicted serum-albumin binding data was obtained. Using the multiple computer-automated structure evaluation (M-CASE) program and protein-affinity data for 154 drugs, models were generated that correctly predicted the percentage of drug bound in plasma for ~80% of the test compounds [54]. A pharmacophoric similarity concept was also used to predict drug-association constants to human serum albumin (HSA) for 138 drugs [56].

We have developed an approach to *in silico* classification of drugs and drug-like compounds according to their binding affinity to plasma proteins (in relation to multiple-protein binding). Using non-linear Sammon maps, we completed a

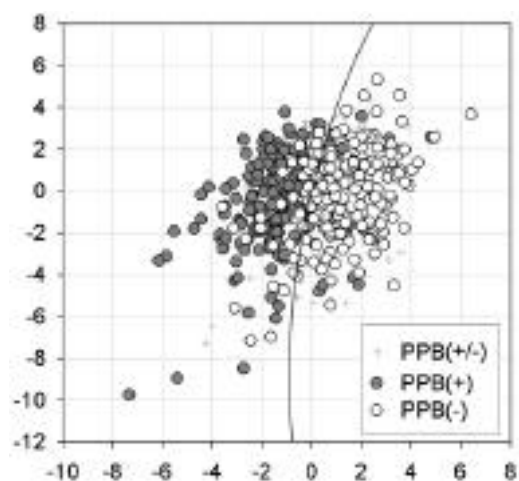


Fig. (4). Sammon map with SVM classification of drugs based on their plasma protein binding affinity.

knowledge-based classification analysis of 549 drugs with experimental % protein bound values. Fig. 4 demonstrates that strong and poor binders occupy different sites on the map with minimal overlap.

P-gp Substrates

P-glycoprotein is a large membrane-bound protein present in the canalicular domain of hepatocytes, brush border of proximal tubule cells and capillary endothelial cells in the central nervous system, which acts as a barrier to limit the exposure to xenobiotics [57]. There have been many computational models for inhibitors of P-gp which have been reviewed previously [58, 59] and more recently pharmacophore-based models have appeared for substrates [60, 61] and inhibitors [62-64]. These methods suggest the positions and importance of hydrogen bond donors, hydrophobic and aromatic ring features. We have used the published dataset of 191 P-gp substrates and non-substrates [18] after excluding Pt-complexes and compounds with very high MW (>800). The remaining 167 compounds were used for the generation of a non-linear Sammon map. Fig. 5

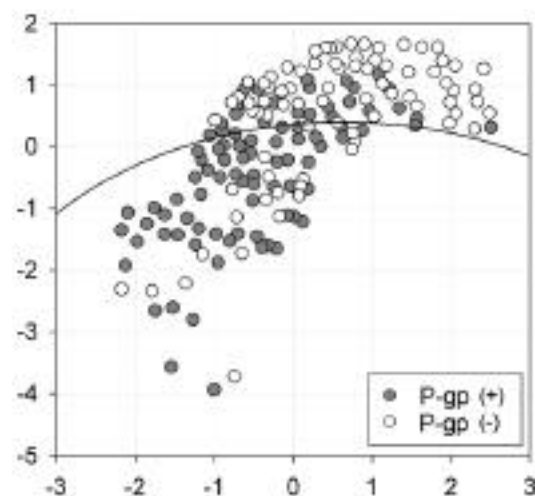


Fig. (5). Sammon map with SVM classification of drugs based on their P-gp substrate efficacy.

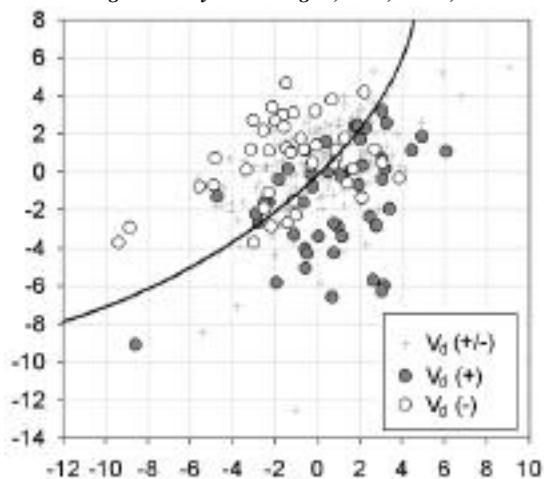


Fig. (6). Sammon map with SVM classification of drugs based on their volume of distribution.

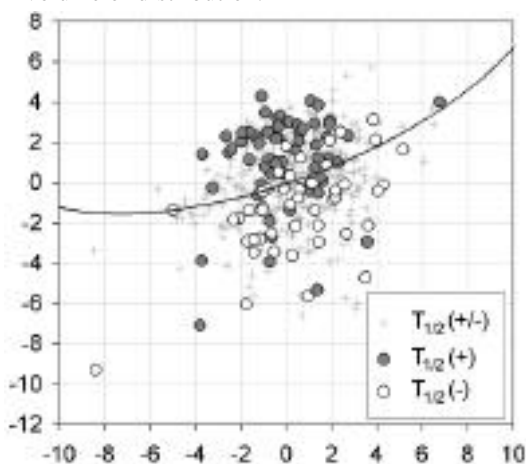


Fig. (7). Sammon map with SVM classification of drugs based on their plasma half-life.

shows the differences in distribution of P-gp(+) and P-gp(-) compounds on the map.

Volume of Distribution and Plasma Half-Life

The volume of distribution, together with the half-life, determine the dose regimen of a drug, and so the early prediction of both properties would be of great benefit. The logarithm of the volume of distribution corrected for plasma-protein binding and plotted against $\log D$ revealed a clear linear trend, with $\log V_d$ increasing at higher [55]. Recently, a more sophisticated approach for predicting volume of distribution values has been presented that used experimental distribution coefficients at pH 7.4 in octanol/water, the ionization constant (pK_a) of the compounds and measured plasma-protein-binding data [65]. In principle, this approach could be fully computational, as predictive models are available for $\log P$ and pK_a , and models for plasma-protein binding can be effectively generated as described above.

In this work we demonstrate that such complex properties as V_d and $T_{1/2}$ can be effectively modeled using the non-linear mapping techniques (Fig. 6, 7). The achieved level of discrimination can therefore be used as a guide in modifying and optimizing these important pharmacokinetic properties.

Discrimination Efficacy

Quantitative data on the irregularity of distribution of opposite compound categories on the generated NLMs described above are summarized in Table 2. In general, the discrimination efficacy provided by the algorithm used is good: on average, 80-90% of compounds from the training sets (categories 1 and 3) could be correctly assigned using the NLMs.

In the experiments described above, we demonstrated the possibility of differentiation between the principal compound categories in the studied ADME data sets. Specifically, in all the cases, we obtained visual evidence of statistically significant differences in molecular properties of the different compound groups. However, for large data sets, the NLM computation is increasingly intractable. The approach may also generate a 2D mapping that poorly approximates the original distances when the number of compounds is large [66]. Also, if additional data points or data sets are to be included in the projection, a complete recomputation of the mapping is required for all data points. The latter feature makes it difficult to properly validate and analyze the generated NLMs with new molecules. To overcome these difficulties, we performed an additional series of experiments, where self-organizing Kohonen maps were generated for the same ADME data sets and molecular descriptors.

Kohonen Self-Organizing Maps

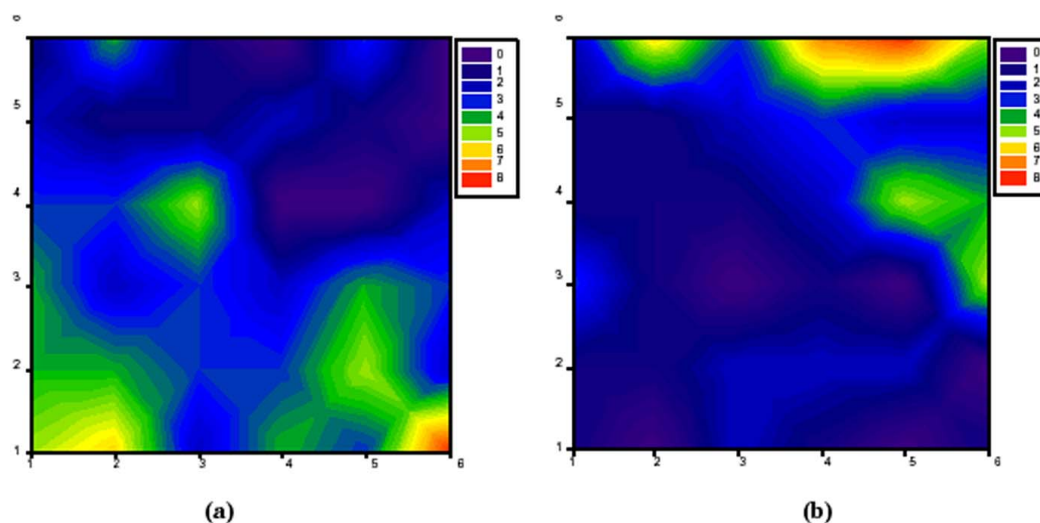
Model generation and internal validation. Self-organizing maps (SOM) or Kohonen networks [24] were originally designed in an attempt to model intelligent information processing, i.e., the ability of the brain to form reduced representations of the most relevant facts without any loss of information about their interrelationships. The general idea behind this method is to map a set of vectorial samples onto a two-dimensional lattice in a way that preserves the topology of the original space. SOMs belong to a class of neural networks known as competitive learning or self-organizing networks. The SOM consists of artificial neurons that are characterized by weight vectors with the same dimensionality as the descriptor set. The neurons are connected by a distance dependent function. In an unsupervised training algorithm the neurons self-organize until their pairwise neighborhoods represent the correct topology of the original data set.

As an example of our approach, we describe here the results of SOM generation for P-gp substrates and non substrates. The same training set and molecular descriptors were used for training and internal validation of the model. The Kohonen map was generated for the entire P-gp training dataset including 96 P-gp substrates and 79 non-substrates. Fig. 8 separately shows the sites of distribution of these compound categories. Obviously, their positions on the map are quite different. The results of a leave-10%-out experiment indicate that the developed model is general enough and can be used for prediction purposes. On average, 77.4% of P-gp substrates (+) and 80.6% of P-gp non-substrates (-) compounds were correctly classified with this method.

Table 2. Classification quality using Sammon nonlinear maps for ADME properties. Categories were separated using the SVM (support vector machine) algorithm.

ADME parameter	compound categories		cmpds	correctly classified	total cmpds
human intestinal absorption (HIA)	1	HIA(+), HIA>80%	114 (100%)	110 (96.5%)	320
	2	HIA(+/-), HIA=20-80%	145	-	
	3	HIA(-), HIA<20%	61 (100%)	48 (78.7%)	
blood brain barrier permeability (BBB)	1	BBB(+), logBB* 0	267 (100%)	237 (88.8%)	456
	3	BBB(-), logBB<-1	189 (100%)	163 (86.2%)	
plasma protein binding (PPB)	1	PPB(+), PPB>90%	200 (100%)	169 (84.5%)	549
	2	PPB(+/-), PPB=50-90%	172	-	
	3	PPB(-), PPB<50%	177 (100%)	143 (80.8%)	
CYP3A4 binding affinity	1	CYP(+), K_m 10	37 (100%)	34 (91.9%)	65
	3	CYP(-), K_m 100	28 (100%)	26 (92.9%)	
P-gp substrates/non-substrates	1	P-gp(+), substrates	89 (100%)	66 (74.2%)	167
	3	P-gp(-), non-substrates	78 (100%)	62 (79.5%)	
volume of distribution (V_d)	1	V_d (+), V_d <0.4 l/kg	47 (100%)	39 (83.0%)	253
	2	V_d (+/-), V_d =0.4-8 l/kg	162	-	
	3	V_d (-), V_d >8 l/kg	44 (100%)	34 (77.3%)	
plasma half-life ($T_{1/2}$)	1	$T_{1/2}$ (+), $T_{1/2}$ 20 h	63 (100%)	51 (81.0%)	458
	2	$T_{1/2}$ (+/-), $T_{1/2}$ =1-20 h	347	-	
	3	$T_{1/2}$ (-), $T_{1/2}$ 1 h	48 (100%)	40 (83.3%)	

* $\log BB = \log([C_{\text{blood}}]/[C_{\text{brain}}])$.

**Fig. (8).** Kohonen map generated for the entire P-gp training set (167 cmpds). The areas of substrates (a) and non-substrates (b) are shown separately.

Using the same procedure, we generated Kohonen maps for all the remaining ADME datasets studied in this work. Table 3 shows the results of leave-10%-out internal validation experiments, which demonstrate good generalization ability of *in silico* models based on the SOMs.

DISCUSSION

Considerable research efforts have focused on novel machine learning algorithms that can be used for drug discovery for predicting these molecular properties. Such calculations can be performed with very large numbers of molecules and act as a form of multidimensional selection

filter. Comparative molecular fields analysis (CoMFA) and pharmacophore approaches (for review, see [44, 45]) have been used to model cytochrome P450 (CYP) enzymes involved in drug metabolism as well as transporters such as P-glycoprotein [19], nuclear hormone receptors [68, 69] and ion channels [70, 71] important for drug-drug interactions. These approaches have been rarely used to model more complex processes such as absorption, bioavailability and clearance processes which have required other more robust algorithms capable of dealing with larger data matrices and hundreds to thousands of descriptors. Recursive partitioning methods have been used extensively with these large sets of molecules and either continuous [72, 73] or binary data, for therapeutic target end points as well as CYP inhibition [74]

Table 3. Classification quality using Kohonen maps for ADME properties.*

Model	Number of compounds				Average percentage of correctly classified compounds in leave-10%-out experiment			
					training set		test set	
	cat. 1, (+)	cat. 2, (+/-)	cat. 3, (-)	entire training set	cat. 1	cat. 3	cat. 1	cat. 3
HIA	114	145	61	320	90.3	78.5	86.2	76.9
BBB	267	-	189	456	89.6	88.3	87.2	85.6
PPB	200	172	177	549	90.4	86.7	89.4	85.4
CYP3A4	37	-	28	65	92.1	93.8	90.5	92.3
P-gp	89	-	78	167	80.4	76.9	77.4	80.6
V _d	47	162	44	253	73.5	77.5	83.3	81.8
T _{1/2}	63	347	48	458	80.4	83.8	75.4	79.0

* for the category definition, see Table 2.

and toxicity properties such as AMES mutagenicity status [75]. Smaller QSAR datasets have also been used with inductive logic programming as a pharmacophore approach [76] but this method does not appear to have been used with ADME data sets to date. Kohonen self organizing maps have only recently been applied to model cytochrome P450 mediated drug metabolism [15, 49], while *k*-nearest neighbors has been used to predict metabolic stability [52]. Genetic programming has also been applied to QSAR studies in which model complexity is controlled by a penalty function [77], although there are no published applications with this technology for modeling other ADME properties to our knowledge.

At present, the computational method garnering the most interest for potential wider drug discovery applications is support vector machines that have been adopted in the biosciences for pattern classification tasks (for example, [78-80]). In contrast to SVM, partial least squares has been widely applied for QSAR studies in which there are few data points but many descriptors [81, 82]. Rosipal and Trejo have applied PLS with nonlinear regression and a kernel function [83] to produce K-PLS. Recently a number of machine learning approaches including PLS, SVM, K-PLS and Kernel ridge regression have all been implemented in a single piece of software and used with several benchmark datasets [84]. One of the datasets used was a protein binding regression model [53] in which 94 molecules had 511 MOE and wavelet descriptors calculated. The K-PLS based model produced similar q^2 statistics and had faster execution times than the SVM models used for comparison [84]. These results with K-PLS indicate that it could be favorably applied to other datasets to enable QSAR model construction and aid drug discovery research as an alternative to SVM in future.

In this work, we demonstrate two effective data mining approaches, which extract information from knowledge databases of compounds with experimentally determined ADME properties. Molecular features encoding the relevant physicochemical and topological properties of compounds were calculated from 2D molecular representations. After the

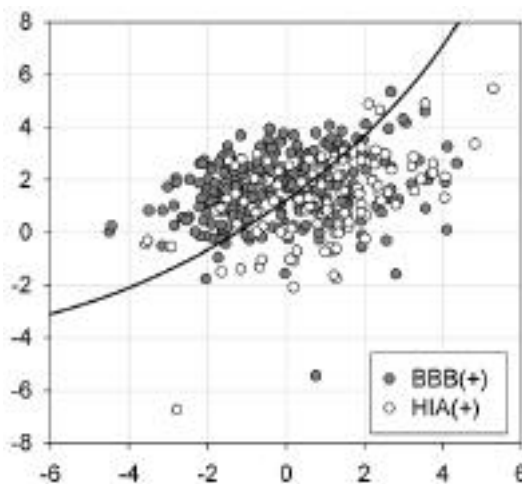


Fig. (9). Differences in distribution of HIA(+) and BBB(+) compounds on the Sammon map.

relevant molecular descriptors were calculated and selected, non-linear Sammon maps with SVM or Kohonen self-organizing maps were generated and analyzed for each ADME dataset. These mapping methods are multivariate statistical techniques, which approximate local geometric relationships of a multidimensional property space on a two-dimensional plot. In other words, the methodology used allows the creation of two-dimensional images of multidimensional property space for the studied datasets. The generated Sammon maps represent all relative distances between all pairs of compounds, and the distance of two points on the map directly reflects the similarity of the compounds. Among the other dimensionality reduction techniques that have appeared in the statistical literature, Sammon non-linear maps are unique due to their conceptual simplicity and ability to reproduce the topology and structure of the data space in a faithful and unbiased manner. This method has a practical value and can be recommended for analysis of small-sized combinatorial libraries (up to several thousands of compounds) aimed at the selection of subsets with enhanced knowledge-based informational content.

The generated Sammon maps are useful qualitative tools for analysis of various phenomena associated with the studied ADME properties. As an illustration, we can demonstrate a Sammon map, which shows that the sites of preferable localization of HIA(+) compounds are somewhat different from those of BBB(+) compounds (Fig. 9). These observations are consistent with the fact that the phenomena of BBB and HIA permeability are different in their nature [85], requiring different molecular properties. The differences illustrated in this way are valuable in the design of orally active drugs, which should not cross the blood-brain barrier (for example, to avoid their CNS-toxicity).

To overcome some drawbacks inherent to the non-linear mapping technique, we performed an additional series of experiments in which the studied ADME datasets were used to generate self-organizing Kohonen maps. The internal validation results demonstrate the good generalization ability

of these *in silico* models based on SOMs. In addition, the developed models can be used for high-throughput operations with larger virtual data sets.

Predictive computational models of value for practical use should be used after external validation. In this current work, five of the developed models based on Kohonen maps were validated using small external test sets, which were not used for training. These data were collected from additional literature sources. After all the necessary molecular descriptors were calculated, we tested the models for P-gp, plasma protein binding, CYP3A4 inhibition, volume of distribution and plasma half-life. For example, a small set of P-gp inhibitors with log IC₅₀ data [19] was used for external validation of the Kohonen map generated with the entire training set. Identical compounds present in the training set were removed from this test set to leave a set of 18 compounds. The results of prediction are shown in Table 4.

Table 4. Individual data for each compound from P-glycoprotein external test set.

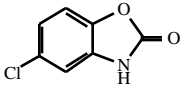
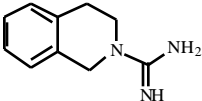
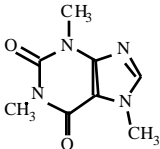
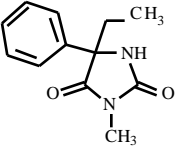
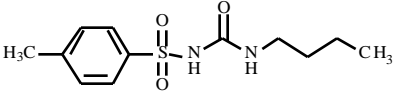
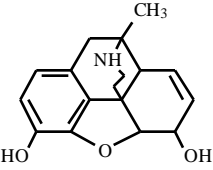
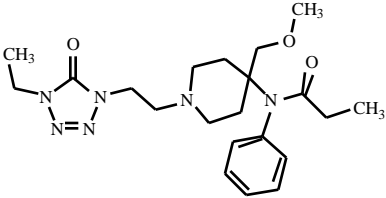
No.	Structure	logIC ₅₀	category	
			exp.	predict.
1		2	3	3
2		2	3	3
3		2	3	3
4		2	3	3
5		2	3	3
6		2	3	3
7		2.0	3	1

Table 4. Contd.....

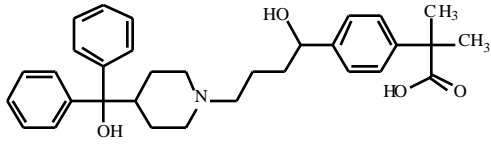
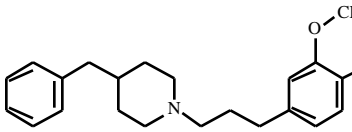
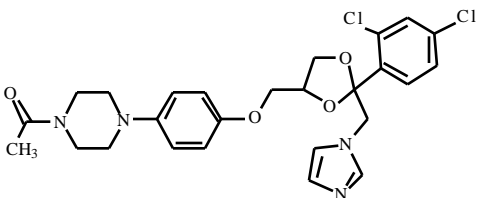
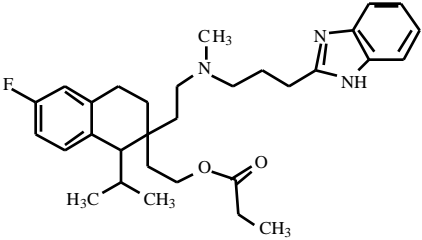
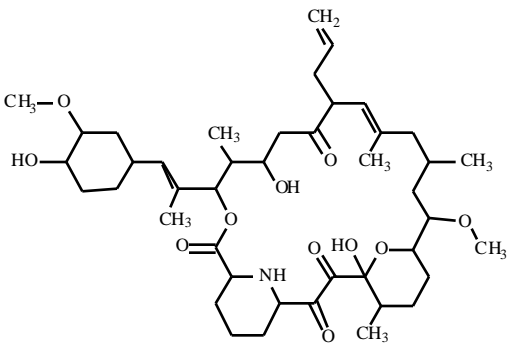
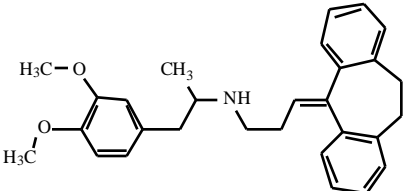
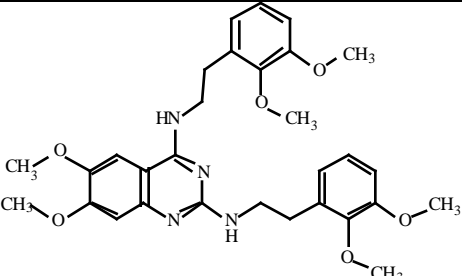
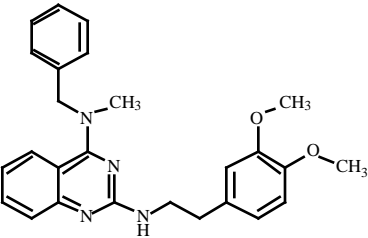
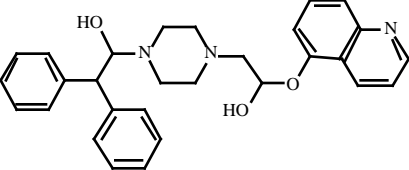
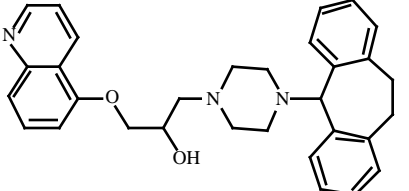
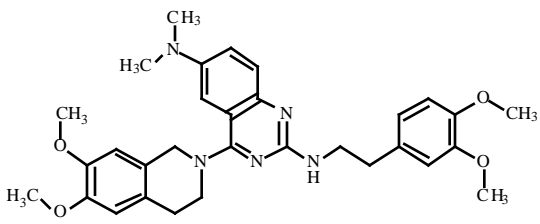
No.	Structure	logIC ₅₀	category	
			exp.	predict.
8		2.0	3	Undefined
9		0.36	1	1
10		0.08	1	1
11		0.08	1	1
12		-0.13	1	1
13		-0.15	1	1

Table 4. Contd.....

No.	Structure	logIC ₅₀	category	
			exp.	predict.
14		-0.22	1	1
15		-0.52	1	1
16		-0.85	1	1
17		-1.13	1	1
18		-1.44	1	1

* for the category definition, see Table 2.

The results of calculations and experimental data for all the external test sets are summarized in Table 5. Overall they demonstrate a good predictive power of the developed models with up to 80-90% correctly classified compounds.

It should be noted that the Kohonen SOMs demonstrate a significant speed gain compared to NLM. In addition, they allow the instant inclusion of new individual or multiple data points on the map without the need of recomputing the entire dataset. These features are favorable for visualization and analysis of larger databases or virtual libraries than

NLM. On the other hand, NLMs provide better distance and topology preservation as compared with Kohonen maps and the latter often contain gaps or undefined regions. Even small gaps in Kohonen maps can include huge parts of chemical space that are not covered by the mapped dataset. Therefore, there must be a tradeoff between mapping quality (in terms of topology and distance preservation) and the computational time for these methods which should be assessed in further experiments with similar or larger sized datasets relevant to drug discovery.

Table 5. External testing of the classification quality of the Kohonen maps using published data.

model	categories (exp.)	compounds predicted		Total cmpds
		cat. 1	cat. 3	
P-gp	cat. 1, PGP(+)	10 (100%)	0/0	10
	cat. 3, PGP(-)	2 (25%)	6 (75%)	8
CYP3A4	cat. 1, CYP(+)	31 (94%)	2 (6%)	33
PPB	cat. 1, PPB(+)	17 (77.3%)	5 (22.6%)	22
	cat. 3, PPB(-)	1 (4.3%)	22 (95.7%)	23
T _{1/2}	cat. 1, T _{1/2} (+)	20 (74.1%)	7 (25.9%)	27
	cat. 3, T _{1/2} (-)	3 (18.7%)	13 (81.3%)	16
V _d	cat. 1, V _d (+)	9 (75.0%)	3 (25%)	12
	cat. 3, V _d (-)	4 (18.2%)	18 (81.8%)	22

CONCLUSION

In this work, we have studied seven ADME datasets of drug compounds using Sammon non-linear maps (with SVM to differentiate between compound classes) and Kohonen self-organizing maps. Our results combined with literature data demonstrate that these data mining techniques are efficient clustering, classification and visualization tools.

It should be stressed that these statistics-oriented data mining approaches are sensitive to the properties of the molecules to be tested. In those cases when the molecular parameters are too far from the parameters of the training set (for example, large molecular weight, high lipophilicity, increased number of polar groups, etc.), the predictions can be incorrect. Therefore, some preprocessing of the tested structural datasets is usually required to ensure their compatibility with the training set. Simple threshold filtering criteria can be used for such preprocessing, as well as the more sophisticated methods of removal of outliers using special statistical techniques (as an example, see our recent article [51]).

Combinations of different computational models for ADME are applicable to the selection of molecules during early drug discovery and represent an approach to filtering large libraries alongside other predicted properties. With the addition of further data it is likely that even more significant models can be generated. At present the different methods we have used could be combined and used in parallel as a consensus-modeling approach to perhaps improve the predictions for external molecules. Certainly there are also many other ADME related datasets that could be modeled similarly, including hERG, and many general toxicity properties.

Pharmaceutical lead discovery and optimization has now become an integrated process where *in vitro*, *in vivo* and *in silico* methods should be considered simultaneously. The described models are applicable as general *in silico* screening tools in the early discovery stages, for example, for the design and selection stage of diverse and focused library construction. The models can be used as efficient 'pre-screen' tools to limit the need for future high-capacity ADME *in vitro* screening studies. For example, these models can be used as filters for compounds prior to further pharmacological testing, where the capacity for such testing is unable to cope with the number of hits in primary

screening (e.g. transporters). The obvious next step is the use of these *in silico* models for assisting library design by virtual screening of proposed structures to ensure that the multidimensional ADME properties are favorable before carrying out the synthesis. We can speculate perhaps optimistically that the further evolution of such computational visualization and predictive technologies will result in the development of truly integrated cheminformatics platforms. These future tools will account for all the issues related to ADME alongside "drug-or lead likeness" which will be solved with maximal accuracy, efficiency and cost-effectiveness.

REFERENCES

- [1] Welch W., Ruppert J., Jain A.N.: *Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites*. Chem. Biol. 3(6), 449, (1996).
- [2] Stahl M., Todorov N.P., James T., Mauser H., Boehm H.-J., Dean P.M.: *A validation study on the practical use of automated de novo design*. J. Comput. Aided Mol. Des. 16(7), 459, (2002).
- [3] Schneider G., Bohm H.-J.: *Virtual screening and fast automated docking methods*. Drug Discov. Today 7(1), 64, (2002).
- [4] Ishchenko A.V., Shakhnovich E.I.: *SMoG2001: An improved knowledge-based scoring function for protein-ligand interactions*. J. Med. Chem. 45(13), 2770, (2002).
- [5] Shimada J., Ekins S., Elkin C., Shakhnovich E.I., Wery J.-P.: *Integrating computer-based de novo drug design and multidimensional filtering for desirable drugs*. Targets 1, 196, (2002).
- [6] Lipinski C.A.: *Drug-like properties and the causes of poor solubility and poor permeability*. J. Pharm. Toxicol. Methods 44(1), 235, (2000).
- [7] Walters W.P., Murcko M.A.: *Prediction of 'drug-likeness'*. Adv. Drug Del. Rev. 54(3), 255, (2002).
- [8] Oprea T.I.: *Current trends in lead discovery: are we looking for the appropriate properties?* J. Comput. Aided Mol. Des. 16(5-6), 325, (2002).
- [9] Wenlock M.C., Austin R.P., Barton P., Davis A.M., Leeson P.D.: *A comparison of physicochemical*

- property profiles of development and marketed oral drugs. *J. Med. Chem.* 46(7), 1250, (2003).
- [10] Ekins S., Boulanger B., Swaan P.W., Hupcey M.A.Z.: *Towards a new age of virtual ADME/TOX and multidimensional drug discovery.* *J. Comput. Aided Mol. Des.* 16(5-6), 381, (2002).
- [11] Ekins S., Rose, J.P.: *In silico ADME/Tox: the state of the art.* *J. Mol. Graph.* 20(4), 305, (2002).
- [12] Van de Waterbeemd H., Gifford E.: *ADMET in silico modelling: towards prediction paradise?* *Nat. Rev. Drug Discov.* 2(3), 192, (2003).
- [13] Lipinski C.A., Lombardo F., Dominy B.W., Feeney P.J.: *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.* *Adv. Drug Del. Rev.* 46(1-3), 3, (1997).
- [14] Veber D.F., Johnson S.R., Cheng H.-Y., Smith B.R., Ward K.W., Kopple K.D.: *Molecular properties that influence the oral bioavailability of drug candidates.* *J. Med. Chem.* 45(12), 2615, (2002).
- [15] Balakin K.V., Ekins S., Bugrim A., Ivanenkov Y.A., Korolev D., Nikolsky Y.V., Skorenko A.V., Ivashchenko A.A., et al. *Kohonen maps for prediction of binding to human cytochrome P450 3A4.* *Drug Metab. Dispos.* 32(10), 1183, (2004).
- [16] Norinder U., Haerberlein M.: *Computational approaches to the prediction of the blood-brain distribution.* *Adv. Drug Del. Rev.* 54(3), 291, (2002).
- [17] Thummel K., Shen D.D.: *In Goodman & Gilman's the pharmaceutical basis of therapeutics*; Tenth K.E., Ed.; McGraw-Hill: New York, pp. 1924-2023, (2001).
- [18] Penzotti J.E., Lamb M.L., Evenson E., Grootenhuis P.D.J.: *A computational ensemble pharmacophore model for identifying substrates of P-glycoprotein.* *J. Med. Chem.* 45(9), 1737, (2002).
- [19] Ekins S., Kim R.B., Leake B.F., Dantzig A.H., Schuetz E., Lan L.B., Yasuda K., Shepard R.L., et al. *Three-dimensional quantitative structure-activity relationships of inhibitors of P-glycoprotein.* *Mol. Pharmacol.* 61(5), 964, (2002).
- [20] Guttman L.: *Some necessary conditions for common factor analysis.* *Psychometrika* 19, 149, (1954).
- [21] Cattell R.B.: *The scree test for the number of factors.* *Multi Behav. Res.* 1, 245, (1966).
- [22] Sammon J.W.: *A nonlinear mapping for data structure analysis.* *IEEE Trans. Comp. C-18(5)*, 401, (1969).
- [23] Vapnik V.: *Statistical learning theory*, Wiley: New York, (1998).
- [24] Kohonen T.: *Self-Organizing Maps*, 3rd edn, Springer Verlag: New York, (2000).
- [25] Agrafiotis D.K., Myslik J.C., Salemme F.R.: *Advances in diversity profiling and combinatorial series design.* *Mol. Divers.* 4, 1, (1999).
- [26] Agrafiotis D.K., Lobanov V.S.: *Nonlinear mapping networks.* *J. Chem. Inf. Comp. Sci.* 40(3), 1356, (1997).
- [27] Yoshida F., Topliss J.G.: *QSAR model for drug human oral bioavailability.* *J. Med. Chem.* 43(13), 2575, (2000).
- [28] Clark D.E.: *Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption.* *J. Pharm. Sci.* 88(8), 807, (1999).
- [29] Palm K., Stenberg P., Luthman K., Artursson P.: *Polar molecular surface properties predict the intestinal absorption of drugs in humans.* *Pharm. Res.* 14(5), 568, (1997).
- [30] Wessel M.D., Jurs P.C., Tolan J.W., Muskal S.M.: *Prediction of human intestinal absorption of drug compounds from molecular structure.* *J. Chem. Inf. Comput. Sci.* 38(4), 726, (1998).
- [31] Kelder J., Grootenhuis P.D.J., Bayada D.M., Delbressine L.P.C., Ploeman J.-P.: *Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs.* *Pharm. Res.* 16(10), 1514, (1999).
- [32] Luco J.M.: *Prediction of the brain-blood distribution of a large set of drugs from structurally derived descriptors using partial least-squares (PLS) modeling.* *J. Chem. Inf. Comput. Sci.* 39(2), 396, (1999).
- [33] Lombardo F., Blake J.F., Curatolo W.J.: *Computation of brain-blood partitioning of organic solutes via free energy calculations.* *J. Med. Chem.* 39(24), 4750, (1996).
- [34] Crivori P., Cruciani G., Carrupt P.A., Testa B.: *Predicting blood-brain barrier permeation from three-dimensional molecular structure.* *J. Med. Chem.* 43(11), 2204, (2000).
- [35] Williams J.A., Ring B.J., Cantrell V.E., Jones D.R., Eckstein J., Ruterbories K., Hamman M.A., Hall S.D., et al. *Comparative metabolic capabilities of CYP3A4, CYP3A5, and CYP3A7.* *Drug Metab. Dispos.* 30(8), 883, (2002).
- [36] Nelson D.R.: *Introductory remarks on human CYPs.* *Drug Metab. Rev.* 34(1-2), 1, (2002).
- [37] Ekins S.: *Predicting undesirable drug interactions with promiscuous proteins in silico.* *Drug Discov. Today* 9(10), 276, (2004).
- [38] VandenBranden M., Wrighton S.A., Ekins S., Gillespie J.S., Binkley S.N., Binkley S.N., Ring B.J., Gadberry M.G., et al. *Alterations of the catalytic activities of drug-metabolizing enzymes in cultures of human liver slices.* *Drug Metab. Dispos.* 26(11), 1063, (1998).

- [39] Ekins S., Maenpaa J., Wrighton S.A.: In *Handbook of drug metabolism*; Marcel Dekker: New York, pp. 363-399, (1999).
- [40] Hansch C.: *The QSAR paradigm in the design of less toxic molecules*. Drug Metab. Rev. 15(7), 1279, (1984).
- [41] Hansch C., Zhang L.: *Quantitative structure-activity relationships of cytochrome P-450* Drug Metab. Rev. 25(1-2), 1, (1993).
- [42] Austel V.A.: *Quantitative structure-activity relationships of drugs*, Academic Press, pp. 437, (1983).
- [43] Lewis D.F.V.: *On the recognition of mammalian microsomal cytochrome P450 substrates and their characteristics: towards the prediction of human p450 substrate specificity and metabolism*. Biochem. Pharmacol. 60(3), 293, (2000).
- [44] de Groot M.J., Ekins S.: *Pharmacophore modeling of cytochromes P450*. Adv. Drug Del. Rev. 54(3), 367, (2002).
- [45] Ekins S., de Groot M., Jones J.P.: *Pharmacophore and three-dimensional quantitative structure activity relationship methods for modeling cytochrome p450 active sites*. Drug Metab. Dispos. 29(7), 936, (2001).
- [46] Jones J.P., He M., Trager W.F., Rettie A.E.: *Three-dimensional quantitative structure-activity relationship for inhibitors of cytochrome P450C9*. Drug Metabol. Dispos. 24(1), 1, (1996).
- [47] Wrighton S.A., Schuetz E.G., Thummel K.E., Shen D.D., Korzekwa K.R., Watkins P.B.: *The human CYP3A subfamily: practical considerations*. Drug Metab. Rev. 32(3-4), 339, (2000).
- [48] Ekins S., Stresser D.M., Williams J.A.: *In vitro and pharmacophore insights into CYP3A enzymes*. Trends Pharmacol. Sci. 24(4), 191, (2003).
- [49] Korolev D., Balakin K.V., Nikolsky Y., Kirillov E., Ivanenkov Y.A., Savchuk N.P., Ivashchenko A.A., Nikolskaya T.: *Modeling of human cytochrome p450-mediated drug metabolism using unsupervised machine learning approach*. J. Med. Chem. 46(17), 3631, (2003).
- [50] Bugrim A., Nikolskaya T., Nikolsky Y.: *Early prediction of drug metabolism and toxicity: systems biology approach and modeling*. Drug Discov. Today 9(3), 127, (2004).
- [51] Balakin K.V., Ekins S., Bugrim A., Ivanenkov Y.A., Korolev D., Nikolsky Y.V., Ivashchenko A.A., Savchuk N.P., et al. *Quantitative structure-metabolism relationship modeling of metabolic N-dealkylation reaction rates*. Drug Metab. Dispos. 32(10), 1111, (2004).
- [52] Shen M., Xiao Y., Golbraikh A., Gombar V.K., Tropsha A.: *Development and validation of k-nearest-neighbor QSPR models of metabolic stability of drug candidates*. J. Med. Chem. 46(14), 3013, (2003).
- [53] Colmenarejo G., Alvarez-Pedraglio A., Lavandera J.-L.: *Cheminformatic models to predict binding affinities to human serum albumin*. J. Med. Chem. 44(25), 4370, (2001).
- [54] Saiakhov R., Stefan L.R., Klopman G.: *Multiple computer-automated structure evaluation model of the plasma protein binding affinity of diverse drugs*. Persp. Drug Disc. Design 19, 133, (2000).
- [55] van de Waterbeemd H., Smith D.A., Jones B.C.: *Lipophilicity in PK design: methyl, ethyl, futile*. J. Comput. Aid. Mol. Des. 15(3), 273, (2001).
- [56] Kratochwil N.A., Huber W., Müller F., Kansy M., Gerber D.: *Predicting plasma protein binding of drugs: a new approach*. Biochem. Pharmacol. 64(9), 1355, (2002).
- [57] Andrews C.W., Bennett L., Yu L.X.: *Predicting human oral bioavailability of a compound: development of a novel quantitative structure-bioavailability relationship*. Pharm. Res. 17(6), 639, (2000).
- [58] Zhang E.Y., Phelps M.A., Cheng C., Ekins S., Swaan P.W.: Adv. Drug Del. Rev. 54, 354, (2002).
- [59] Stouch T.R., Gudmundsson O.: *Progress in understanding the structure-activity relationships of P-glycoprotein*. Adv. Drug Del. Rev. 54(3), 315, (2002).
- [60] Garrigues A., Loiseau N., Delaforge M., Ferte J., Garrigos M., Andre F., Orłowski S.: *Characterization of two pharmacophores on the multidrug transporter P-glycoprotein*. Mol. Pharmacol. 62(6), 1288, (2002).
- [61] Ekins S., Kim R.B., Leake B.F., Dantzig A.H., Schuetz E., Wikel J.H., Wrighton S.A.: *Application of three-dimensional quantitative structure-activity relationships of P-glycoprotein inhibitors and substrates*. Mol. Pharmacol. 61(5), 974, (2002).
- [62] Pajeva I.K., Wiese M.: *Pharmacophore model of drugs involved in P-glycoprotein multidrug resistance: explanation of structural variety (hypothesis)*. J. Med. Chem. 45(26), 5671, (2002).
- [63] Langer T., Eder M., Hoffmann R.D., Chiba P., Ecker G.F.: *Lead identification for modulators of multidrug resistance based on in silico screening with a pharmacophoric feature model*. Arch. Pharm. (Weinheim) 337(6), 317, (2004).
- [64] Pajeva I.K., Globisch C., Wiese M.: *Structure-function relationships of multidrug resistance P-glycoprotein*. J. Med. Chem. 47(10), 2523, (2004).
- [65] Lombardo F., Obach R.S., Shalaeva M.Y., Gao F.: *Prediction of volume of distribution values in humans for neutral and basic drugs using physicochemical measurements and plasma protein binding data*. J. Med. Chem. 45(13), 2867, (2002).
- [66] Clark R.D., Patterson D.E., Soltanshahi F., Blake J.F., Matthew J.B.: *Visualizing substructural*

- fingerprints*. J. Mol. Graph. Model 18(4-5), 404, (2000).
- [67] Ekins S., de Groot M., Jones J.P.: *Pharmacophore and three-dimensional quantitative structure activity relationship methods for modeling cytochrome p450 active sites*. Drug Metab. Dispos. 29(7), 936, (2001).
- [68] Ekins S., Mirny L., Schuetz E.G.: *A ligand-based approach to understanding selectivity of nuclear hormone receptors PXR, CAR, FXR, LXRA, and LXRB*. Pharm. Res. 19(12), 1788, (2002).
- [69] Ekins S., Erickson J.A.: *A pharmacophore for human pregnane X receptor ligands*. Drug Metab. Dispos. 30(1), 96, (2002).
- [70] Cavalli A., Poluzzi E., De Ponti F., Recanatini M.: *Toward a pharmacophore for drugs inducing the long QT syndrome: insights from a CoMFA study of HERG K(+) channel blockers*. J. Med. Chem. 45(18), 3844, (2002).
- [71] Aronov A.M., Goldman B.B.: *A model for identifying HERG K+ channel blockers*. Bioorg. Med. Chem. 12(9), 2307, (2004).
- [72] Chen X., Rusinko III A., Young S.S.: *Recursive partitioning analysis of a large structure-activity data set using three-dimensional descriptors*. J. Chem. Inf. Comput. Sci. 38(6), 1054, (1998).
- [73] Chen X., Rusinko I.A., Tropsha A., Young S.S.: *Automated pharmacophore identification for large chemical data sets*. J. Chem. Inf. Comput. Sci. 39(5), 887, (1999).
- [74] Ekins S., Berbaum J., Harrison R.K.: *Generation and validation of rapid computational filters for Cyp2D6 and Cyp3A4*. Drug Metab. Dispos. 31(9), 1077, (2003).
- [75] Young S.S., Gombar V.K., Emptage M.R., Cariello N.F., Lambert C.: *Mixture deconvolution and analysis of Ames mutagenicity data*. Chemo. Intell. Lab. Sys. 60(), 5, (2002).
- [76] Marchand-Geneste N., Watson K.A., Alsberg B.K., King R.D.: *New approach to pharmacophore mapping and QSAR analysis using inductive logic programming. Application to thermolysin inhibitors and glycogen phosphorylase B inhibitors*. J. Med. Chem. 45(2), 399, (2002).
- [77] Nicolotti O., Gillet V.J., Fleming P.J., Green D.V.S.: *Multiobjective optimization in quantitative structure-activity relationships: deriving accurate and interpretable QSARs*. J. Med. Chem. 45(23), 5069, (2002).
- [78] Furey T.S., Christianini N., Duffy N., Bednarski D.W., Schummer M., Haussler D.: *Support vector machine classification and validation of cancer tissue samples using microarray expression data*. Bioinformatics 16(10), 906, (2000).
- [79] Brown M.P.S., Grundy W.N., Lin D., Christianini N., Sugnet C.W., Ares M., Haussler D. Jr.: *Knowledge-based analysis of microarray gene expression data by using support vector machines*. Proc. Natl. Acad. Sci. USA 97, 262, (2000).
- [80] Zernov V.V., Balakin K.V., Ivashchenko A.A., Savchuk N.P., Pletnev I.V.: *Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions*. J. Chem. Inf. Comp. Sci. 43(6), 2048, (2003).
- [81] Wold S., Johansson E., Cocchi M.: In *3D-QSAR in drug design; Theory, methods and applications*, ESCOM: Leiden, pp. 523-550, (1993).
- [82] Afzelius L., Masimirembwa C.M., Karlen A., Andersson T.B., Zamora I.: *Discriminant and quantitative PLS analysis of competitive CYP2C9 inhibitors versus non-inhibitors using alignment independent GRIND descriptors*. J. Comput. Aided Mol. Des. 16(7), 443, (2002).
- [83] Rosipal R., Trejo L.J.: *Kernel partial least squares regression in reproducing kernel hilbert space*. J. Machine Learning Research. 2(12), 97, (2001).
- [84] Bennett K.P., Embrechts M.J.: In *Advances in learning theory; methods, models and applications*, IOS Press: Amsterdam, pp. 227-250, (2003).
- [85] Lundquist S., Renftel M., Brillault J., Fenart L., Cecchelli R., Dehouck M.P.: *Prediction of drug transport through the blood-brain barrier in vivo: a comparison between two in vitro cell models*. Pharm. Res. 19(7), 976, (2002).