PERSPECTIVE

Pharmaceutical Perspectives of Nonlinear QSAR Strategies

Lisa Michielan and Stefano Moro*

Molecular Modeling Section (MMS), Dipartimento di Scienze Farmaceutiche, Università di Padova, via Marzolo 5, I-35131 Padova, Italy

Received February 17, 2010

1. INTRODUCTION

1.1. Challenges in Pharmaceutical Research. The drug discovery process is aimed at bringing to market new therapeutic agents with desirable pharmacodynamic profile, favorable ADME (Absorption, Distribution, Metabolism, Elimination) and toxicological properties. The target selectivity is another crucial requirement for drugs to avoid efficacy problems and limiting side-effects, incurring when the compounds do not differentiate between various receptors. During the optimization of drug candidates in terms of potency, the general aim is to design drugs without or with minimum side-effects, while retaining the desired function. Nowadays, the pharmaceutical research has to face many obstacles with the result of a very low success rate, regardless the extreme growth of employed resources.[1] According to recent Tufts Center for the Study of Drug Development data, drug development, starting from the clinical trials to the final approval, is about 8.5 years long with a cost exceeding $40 billion, and only 21.5% of clinical success rate.[2] In fact, pharmaceutical research is actually involved in the study of more complex diseases, while the increasing size and costs of the clinical trials, high attrition rates of candidates, and the late occurrence of failures in the clinical studies are emphasized as the main negative contributions to the economic profile of the research in the pharmaceutical companies. Several aspects have been identified and reported as the causes of the high level of attrition undergone by the compounds during the developmental stages.[3] The reasons for attrition have changed over time, and in 2000, some problems of efficacy, clinical safety, or toxicological effects were identified as highly responsible for the failures, covering more than 50% of the causes for abandoning the candidate, as shown in Figure 1. Clearly, most of the efforts are directed to unproductive clinical trials, since most drug candidates are eliminated late in the clinical development without recovering the starting investment.[3]

In recent decades, the potential of combinatorial chemistry has provided new large databases with unknown compounds. Therefore, at the early stage of drug discovery, suitable computational approaches are needed to shorten the time and increase the success rate by deriving in silico models for the prediction of the desirable properties.[4] The computational tools aim at eliminating lead compounds with undesirable profiles before they enter the costly late phases of drug development and to let compounds to proceed in the optimization step, improving the attrition rate in drug discovery.[5,6] Quantitative structure−activity relationships (QSARs) or quantitative structure−property relationships (QSPRs) approaches represent probably the most robust well-known tools to mathematically analyze the correlation between molecular properties and the corresponding property of interest. Several novel nonlinear machine learning methods have been applied for the prediction of pharmacodynamic and ADMET (Absorption, Distribution, Metabolism, Elimination and Toxicity) properties,[7] and particularly, the regression methods have been recently reviewed.[8] So far, many QSARs have been attempted to correlate molecular descriptors with druglikeness, activity, selectivity, toxicity, and several pharmacokinetic properties.[9−11]

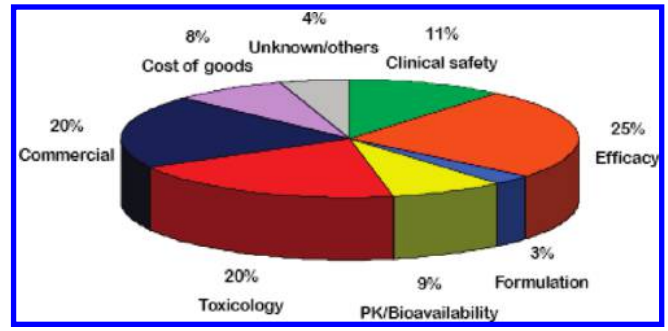

Figure 1. Reasons for attrition (2000). PK: pharmacokinetics.

1.2. Predictive Toxicology in Drug Discovery. The regulatory frame is considered an additional obstacle in the drug discovery process, since a very accurate risk evaluation is required to assess the safety of the drug once on the marketplace.[1] The poorly efficient risk assessment process and the limited information on hazard properties of chemicals has driven the need for new regulatory dispositions that were introduced in European Community on June 1, 2007 with the chemical management system REACH (Registration, Evaluation, Authorization, and Restriction of Chemicals).[12,13] The immediate objective of REACH, in a relatively short time period (11 years), is to characterize the toxicological

* To whom correspondence should be addressed. Tel: +39 049 8275704. Fax: +39 049 8275366. E-mail: stefano.moro@unipd.it.

10.1021/ci100072z
Published on Web 06/07/2010

properties of a large group of substances, manufactured or imported in quantities in excess of 1 ton per year. The attempt of this regulation is to increase the production of useful data for the decisions involving the protection of human health and environment through a better identification and understanding of the chemical properties hazardous to safety. Moreover, diverse, expensive animal testing experiments are usually expected for in vivo toxicological data requirements.

Very recently, a paradigm shift has been suggested in toxicology with a specific reference to the computational methods as reliable support in the toxicity assessment.[14] In more detail, the predictive toxicology represents an attractive tool to investigate the effects on human health and the potential ecotoxicological risk of chemical substances in the drug discovery process, as well as in the environmental hazard assessment.[15] In this context, pharmaceuticals, personal health care products, nutritional ingredients and products of the chemical industries are all potentially dangerous and need to be assessed. Then, the aim of the computational toxicology is to accelerate their assessment through in silico models and a brief overview of both tools and models in computational toxicology have been considered.[16] Moreover, a recent interesting review about toxicity databases available and in silico toxicology tools, together with their advantages and limitations, has been published.[17]

The same introduction of REACH should speed up the risk assessment process by prioritizing compounds for traditional toxicity testing and providing information on the exposure scenarios (ESs) concerning the chemical safety profile.[18] In fact, REACH promotes alternative tools to collect extensive information on hazards of chemicals to reduce animal use in toxicology. As a consequence, several intelligent or integrated testing strategies (ITS) have been proposed as rapid, efficient approaches to obtain exposure and effects data and identify different modes of toxic action.[19,20] In more detail, in vitro or computational methods, optimized in vivo studies, chemical categories, read-across analysis, and thresholds of toxicological concern (TTCs) are admitted nontesting strategies to replace missing data or endpoints and profitably reduce costly animal experiments.[18] So far, powerful computational toxicology prediction systems have been developed for the exposure and hazard assessment to satisfy the new regulatory requests.[21]

The in silico approaches, among these several machine learning methodologies, for the toxicity prediction of safety-relevant endpoints are precious contributions to early discovery of adverse drug reactions.[22,23] In the last years, several "data driven systems" and "expert systems" have become available for the assessment of toxicological endpoints. Data-driven programs generate statistically valuable structure−activity relationships (SARs) by processing large groups of unrelated chemicals, without user bias or prior organization, to find associations based on similar chemical structures, known as structural alerts, that most probably correspond to the same toxicological mechanism. Unfortunately, the ease of prediction in these techniques is penalized by the accurate statistical validation needed. For this reason, they are better suggested to detect general alerting properties. On the other hand, the expert systems embody a series of knowledge based rules, which are the result of a previous toxicity assessment of compounds by human experts. They consider small classes of similar-acting chemicals or groups of compounds with similar structure to build classes of potential toxicity. Regardless of their more limited application in comparison with the data driven systems, the expert systems offer more easily interpretable results.[22] In fact, if data driven systems require experimental data to carry out predictive models, expert systems provide a direct and easy access to toxicological information. However, a careful evaluation is needed as a consequence of the moderate sensitivity obtained in the prediction of positive samples by the expert systems.

In toxicology QSAR are widely used strategies to infer the toxicological properties of compounds from their molecular structure.[24] Several studies have focused on the prediction of the environmental toxicity properties of drugs.[25] In particular, aquatic toxicity of chemical substances is ultimately investigated as basic information in the hazard and environmental risk assessment.[26−30]

In this paper, we present the most recent machine learning applications in computational toxicology, leaving out the statistically weak modeling examples to concentrate on the most robust models, in terms of number of compounds in the training set and performance measures. In limited cases, we have considered less-performing models with the only objective to exemplify novel methodologies.

2. QSAR

The quantitative structure−activity relationships are based on the principle that the compounds can be mathematically codified as distributions of molecular properties or molecular descriptors. Therefore, a statistical modeling method is able to achieve the correlation between the molecular descriptors and the defined target property to predict the corresponding property of new compounds. If the considered endpoint is represented by continuous data, the QSAR model is referred to as regression; on the other hand, if binary data are introduced to represent the property in the input matrix for the model generation, discrete classification models are derived with the aim to separate compounds into different classes.[31] The present review will consider the recent development of QSAR approaches to predict both quantitative and categorical properties.

Several key steps are involved in any QSAR modeling approach: data collection, calculation of suitable molecular descriptors from chemical structures and, if necessary, their selection, model generation, and finally, internal and external validation. Furthermore, the evaluation of the applicability of the model for a particular endpoint, the optimization of the modeling parameters, and the update of the model once new data are available are also required steps. The whole process is summarized in Figure 2.

In drug discovery, QSAR methodologies have demonstrated to be powerful in the prediction of simple chemical-physical properties as well as complex pharmacodynamic, pharmacokinetic and toxicological profiles. The selection of the modeling technique represents a crucial point in the model development. As reported by Winkler, the classical linear QSARs have problems of overfitting and they are not able to handle nonlinear relationships.[32] In recent QSAR studies, the machine learning methodologies have shown promising potentialities with respect to the more classical linear techniques.[31,32] Moreover, due to the increasing influence of the regulatory aspect in QSAR, the predictive

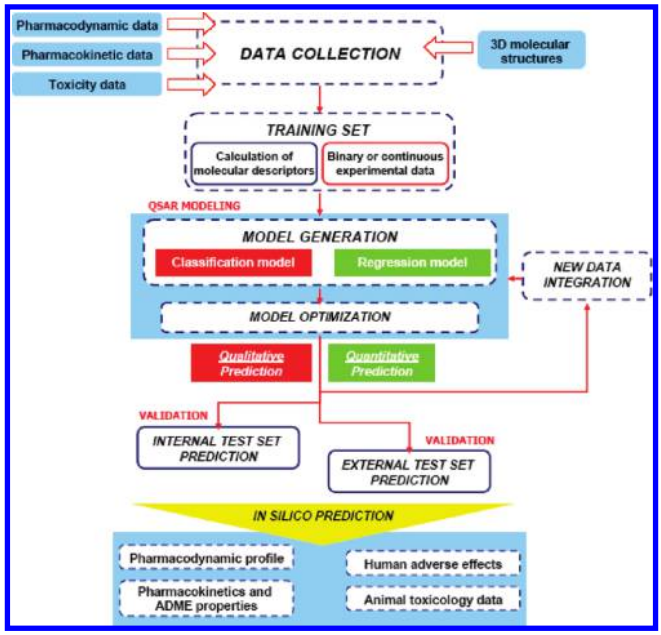

Figure 2. Workflow for the use of pharmacodynamic, pharmacokinetic, and toxicity databases and models: from the data collection to the in silico prediction of the properties for new compounds.

capability of the models has been discussed.[33] The predictivity defines the practical use and the validity of the models, so local QSARs for congeneric series of chemicals have demonstrated better reliability by external validation in comparison with global models, considering large diverse sets of chemicals. However, the imbalanced distribution of the chemical classes and the unknown mechanistic interpretation are important limitations of local models.

For regulatory purposes, some reference principles, introduced by Organization for Economic Cooperation and Development (OECD), have been recommended in the QSAR model development.[34] The guideline document has underlined the following requirements: (a) a defined endpoint, (b) an unambiguous algorithm, (c) a defined domain of applicability, (d) appropriate measures of goodness-of-fit, robustness, and predictivity, and (e) a mechanistic interpretation, if possible. Some of the principles and their role in the regulatory context have been clarified.[35] In particular, QSAR models should be reproducible and based on specific molecular descriptors, while the external validation should be performed as additional evaluation of the model stability and predictivity. The modeling strategies have been briefly discussed by Zvinavashe et al. in the context of the prediction of metabolism and toxicity, focusing on the influence of mechanistic interpretation in the selection of the molecular descriptors.[36]

3. MODELING METHODS

3.1. Molecular Descriptors. Various molecular descriptors are provided by different softwares for the generation of robust QSARs. The model quality is highly dependent on both type and number of molecular descriptors; then a feature selection procedure is commonly applied to find the best descriptor set, as underlined by Li et al.[6] Molecular descriptors are quantitative representations of chemical structures and structural or physicochemical properties and their complexity is based on the structure dimensionality (1D, 2D, or 3D). Recently, the most popular available computer programs for descriptors calculation have been reported and they are easily accessible.[6,37,38] The classical molecular descriptors are classified into different groups: constitutional, that is, molecular weight, geometrical, that is, surface area, electronic, topological, that is, the number of rotatable bonds, RDF descriptors, that refer to interatomic distances, molecular walk counts, 3D-MoRSE, BCUT, GRIND, and WHIM descriptors, Galvez topological charge indices and charge descriptors, GETAWAY descriptors, 2D and 3D autocorrelations, functional groups, randic molecular profiles, electrotopological state descriptors, and other physicochemical properties. 1D and 2D molecular descriptors, based on molecular formula or the connectivity of the considered compound, are easy to interpret. 3D descriptors introduce the conformational problem, and this aspect influence their computation. The choice of the combination of descriptor types depend on the aim of the model development: in the optimal case the obtained model is reliable, it is based on easily computed molecular descriptors, it has good prediction capability and mechanistic interpretation. In general, the selection of the descriptor set is driven by their influence on the target property and the model achieved robustness.

3.2. Machine Learning Methodologies. Machine learning represents a well-known family of algorithmic techniques based on a solid statistical theory and able to handle various experimental data sets. In the last years, novel nonlinear modeling methods, suitable for solving both regression and classification problems, have been developed as robust alternative tools to the traditional linear statistical techniques, such as the well-known Partial Least Squares (PLS) analysis.[39−41] Very recently, Liu et al. published an exhaustive paper reporting the current mathematical techniques applied in QSAR studies.[42] Promising strategies have been newly developed and, furthermore, the existing linear techniques have been upgraded by combining them with other methods. In the present review we briefly present the basic concepts related to some machine learning methodologies: Artificial Neural Networks (ANNs), Support Vector Machine (SVM), Decision Trees (DTs), and Random Forest (RF).

The innovative potentialities and applicability of the neural network methodology in drug discovery have been recently described.[32] Artificial neural networks algorithm has been introduced as a more flexible class of modeling techniques naturally able to deal with complex nonlinear systems both in classification and regression problems. Its architecture is particularly suitable in the studies involving a large number of observations. ANNs are based on layers of processing neurodes, interconnected to form a network to simulate the behavior of neuronal brain cells.[43] The single input information of each neuron is weighted, summed to the other inputs, and modified to get the output result. During several training cycles in the learning process, the weights are adapted to minimize the difference between the target and the resulting output for each data. Different factors influence the final output; among these the architecture of the network, the parameters controlling the learning process and the weights that connect the neurodes. Regardless their good performance as nonlinear modeling technique, ANNs have some disadvantages, since they tend to overfit the data and it is difficult to evaluate the contribution of the different descriptors. Among ANNs, self-organizing maps (SOM) represent an

unsupervised learning methodology, able to classify data basing only on the input vectors. In k nearest neighbor (k-NN), the compounds with similar physicochemical properties and activity belong to the same class. They are represented by vectors defined in the physicochemical space, then, the Euclidean distance between an unclassified vector and each remaining individual vector is calculated to assign the class to the input vector, according to the majority vote. Finally, radial basis function neural network (RBFNN) is currently one of the most frequently used techniques among ANN ones.[42]

Support vector machines (SVMs) were developed by Vapnik and represent a group of supervised learning techniques, characterized by novel attractive features and optimal generalization performance.[44,45] SVM was first applied in pattern recognition but now is very utilized to solve both classification and regression problems.[46,47] The last several classification problems have been solved, in fact, using SVM approach, such as the discrimination between active and nonactive compounds.[48−50] Moreover, support vector regression (SVR) has been widely applied as nonlinear methodology to derive quantitative structure−activity relationships for the prediction of different chemical and biological properties.[51,52] In general, SVMs are used in supervised learning problems, where the available data set is represented as a set of pairs $T = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, called examples, where x_i is an input sample and y_i is the corresponding desired value. Usually $x_i \in \boldsymbol{R}^n$, while if $y_i \in \{-1, +1\}$, the learning problem is a binary classification problem, and if $y_i \in \boldsymbol{R}$, the learning problem is a regression problem. In both cases, SVMs select the best function approximation in the given hypothesis space by minimizing some risk functional. In the classification studies, the SVM method builds an hyperplane, able to discriminate data points in two distinct classes, such that the maximization of the margin is achieved. On the other hand, the support vector regression is able to select the hyperplane that approximates the desired responses y_i over the data set in an optimal fashion, that is, the function $f(x)$ should be a reasonable estimate of the functional relation between unseen input−output pairs. Both position and orientation of the hyperplane are defined by a subset of training points, known as support vectors. Moreover, for the nonlinear decision surfaces, SVM projects the feature vectors into a high dimensional feature space by introducing a kernel function.

Decision tree (DT) learning provides informative models for the classification tasks by grouping data according to similar features, as described in the literature.[53−56] The input information is propagated along the branches of the tree in the learning process as long as the leaves of DT architecture contain data that belongs to the same class. Among the possible tree structures, the learning process selects the simplest one to achieve a given accuracy. However, despite various methods and implementations of DT learning system exist, this technique is not an extensively used classification approach.

The random forest (RF) method is based on an ensemble of classification or regression trees, such that each tree grows on the value of an input random vector, independently introduced and with the same distribution for all trees in the forest. RF has been presented as powerful technique in the previous literature.[57,58]

4. APPLICATIONS IN PHARMACODYNAMICS

4.1. Regression Strategy. The quantitative prediction of the activity of compounds can be achieved by applying powerful regression approaches. We have reviewed in details the most recent papers and high confidence models in this field, as reported in Table 1.

In most studies, the nonlinear methods, especially SVM, have showed better prediction capability with respect to the more traditional linear regression approaches. In fact, the linear regression method is not always able to accurately describe a complex correlation between the molecular descriptors and the activity. Neural networks performances were investigated by Zheng et al. to predict the affinity of lobeline and tetrabenazine analogs binding at the vesicular monoamine transporter-2.[59] In more detail, the application of ANN method in combination with 11 molecular descriptors has resulted in the values of correlation coefficient r^2 of 0.91 and 0.93 on the training set and the test set, respectively. Further examples that demonstrate the applicability of ANN as QSAR tool have been reported by Worachartcheewan et al.,[60] while Mandal et al. emphasized the predictive power of both ANN and genetic function approximation (GFA).[61] Interestingly, ANN has been also applied to the prediction of the antibacterial activity of peptides, with the aim to perform high-throughput in silico screening for novel antibacterial drug candidates.[62] Moreover, the activity of carbonic anhydrase II inhibitors has been investigated by ANN in combination with nonlinear feature selection technique: in this case the values of correlation coefficients r^2 for the training, test and validation sets were 0.891, 0.850, and 0.845, respectively.[63]

In several papers, the comparison between SVM and ANN prediction capabilities was performed. Fatemi et al. considered the apoptosis-inducing activity of several 4-aryl-4-H-chromenes by using SVM and ANN.[64] The results demonstrated that the SVM model, combined to nonlinear variable selection, is more accurate in the evaluation of $\log(1/EC_{50})$ activity values. ANN and SVM analysis were also compared in the prediction of glycogen synthase kinase-3β (GSK-3β) inhibitory activity by Goodarzi et al., showing the best SVM prediction results.[65] The inhibition of histone deacetylases (HDAC) is ultimately considered a key strategy in the treatment of human cancer and other diseases, since this enzyme is involved in the regulation of cell cycle. In recent years, the linear QSAR strategies have contributed to model the HDAC inhibitory activity. Several inhibitors of histone deacetylases (HDACIs) have been used to develop k-NN and SVM models in combination with MolConnZ descriptors and good values of LOO cross-validated r^2 were obtained (0.81 and 0.93, respectively).[66] The authors emphasized their potential in the virtual screening to identify novel hits. The best results were obtained by SVM in combination with a feature selection procedure, showing the alternative useful applicability to describe nonlinear relationships.

Other papers have highlighted the better SVM performances with respect to linear strategies. Li et al. applied least-squares SVM (LS-SVM) in the prediction of the inhibitory activity of a small set of acyl ureas on human glycogen phosphorylase (hlGPa), demonstrating the potential use of SVM in combination with linear PLS in QSAR analysis.[67] Regarding the central benzodiazepine receptor (BzR), a SVM

Table 1. Performance of Regression-Based Nonlinear Machine Learning Models for the Prediction of Pharmacodynamic Properties

property/target	experimental data	methodology	molecular descriptors	validation method	reported statistics	refs
VMAT2	K_i	ANN	11 (DRAGON descriptors)	training set (89) LOO cross-validation test set (15)	r^2 = 0.91, rmsd = 0.225 r^2_{cv} = 0.82, rmsd = 0.316 r^2 = 0.93, rmsd = 0.282	59
furin-dependent processing of antrax PA	K_i	ANN	11 (DRAGON descriptors) 6 (RECON descriptors)	training set (12) training set (11) LOO cross-validation	r_{cv} = 0.807, rms = 0.666 r_{cv} = 0.923, rms = 0.304	60
HIV reverse transcriptase	IC_{50}	ANN GFA	Cerius2 descriptors	training set (52) test set (18)	(ANN) r^2 = 0.731 (GFA) r^2 = 0.612	61
carbonic anhydrase II	IC_{50}	GA-KPLS-ANN	8 (DRAGON descriptors)	training set (76) LOO cross-validation test set (19) validation set (19)	r^2 = 0.891 r^2_{cv} = 0.899, RMSE = 0.176 r^2 = 0.850, RMSE = 0.214 r^2 = 0.845, RMSE = 0.205	63
GSK- 3β	IC_{50}	ANN SVM	16 (DRAGON descriptors)	training set (100) validation set (23) test set (29) 	(ANN) r^2 = 0.932, RMSEP = 0.229 (SVM) r^2 = 0.960, RMSEP = 0.179 (ANN) r^2 = 0.928, RMSEP = 0.279 (ANN) r^2 = 0.915, RMSEP = 0.289 (SVM) r^2 = 0.927, RMSEP = 0.240	65
HDAC	IC_{50}	*k*-NN SVM	262 (MolConnZ descriptors)	training set (34) LOO cross-validation test set (16) external test set (9) 	(*k*-NN) r^2_{cv} = 0.81 (SVM) r^2_{cv} = 0.93 (*k*-NN) r^2 = 0.80, RMSE = 0.38 (SVM) r^2 = 0.87, RMSE = 0.36 (*k*-NN) r^2 = 0.76 (SVM) r^2 = 0.62	66
hIGPa	IC_{50}	LS-SVM	6 (CODESSA descriptors)	training set (33) LOO cross-validation test set (11)	r^2 = 0.957 r^2_{cv} = 0.917, MSE = 0.048 r^2 = 0.889, MSE = 0.148	67
BzR	IC_{50}	SVM	7 (CODESSA descriptors)	training set (48) test set (15)	r^2 = 0.93, MSE = 0.26 r^2 = 0.96, MSE = 0.10	68
BzR	IC_{50}	RBFNN	618 (CODESSA descriptors)	training set (45) test set (13)	r = 0.911, RMS = 0.570 r = 0.903, RMS = 0.724	69
CCR5	IC_{50}	SVM PPR	8 (CODESSA constitutional, topological, geometrical and quantum chemical descriptors)	training set (59) test set (20) 	(SVM) r^2 = 0.867 (PPR) r^2 = 0.837 (SVM) r^2 = 0.732 (PPR) r^2 = 0.726	70
anti-HIV-1 activity	IC_{50}	RBFNN GRNN PPR SVM	600 (CODESSA descriptors)	training set (48) test set (16) 	(RBFNN) r^2 = 0.791, MSE = 0.19 (GRNN) r^2 = 0.814, MSE = 0.18 (PPR) r^2 = 0.890, MSE = 0.10 (SVM) r^2 = 0.831, MSE = 0.16 (RBFNN) r^2 = 0.833, MSE = 0.18 (GRNN) r^2 = 0.686, MSE = 0.32	71
HIV-1 reverse transcriptase	IC_{50}			training set (38) test set (13) 	(PPR) r^2 = 0.882, MSE = 0.15 (SVM) r^2 = 0.850, MSE = 0.21 (RBFNN) r^2 = 0.825, MSE = 0.18 (GRNN) r^2 = 0.808, MSE = 0.23 (PPR) r^2 = 0.843, MSE = 0.16 (SVM) r^2 = 0.811, MSE = 0.20 (RBFNN) r^2 = 0.651, MSE = 0.29 (GRNN) r^2 = 0.694, MSE = 0.29 (PPR) r^2 = 0.843, MSE = 0.15 (SVM) r^2 = 0.802, MSE = 0.16	
HIV-1 protease	IC_{50}	PLS SVR LS-SVM	14 (1D, 2D, 3D descriptors)	training set (32) LOO cross-validation test set (16) 	(PLS) r^2 = 0.836 (SVR) r^2 = 0.935 (LS-SVM) r^2 = 0.938 (PLS) r^2_{cv} = 0.641 (SVR) r^2_{cv} = 0.844 (LS-SVM) r^2_{cv} = 0.819 (PLS) r^2 = 0.841 (SVR) r^2 = 0.891 (LS-SVM) r^2 = 0.899	72
hERG	IC_{50}	*k*-NN	340 (2D descriptors) 88 (3D descriptors)	training set (86) test set I (18) test set II (61)	r^2 = 0.7035	73
H3	K_i	SVM	8 (VolSurf and CPSA electrostatic, quantum chemical descriptors)	training set (118) test set (26)	r^2 = 0.858, SEE = 0.298	74
BzR binding site (BZDbs) of the γ-aminobutyric acid type A (GABA(A)) receptor complex	K_i	SVM BPN hga-SVM	DRAGON descriptors	training set (50) test set (28) 	(SVM) r^2 = 0.9042, RMSEP = 0.3943 (BPN) r^2 = 0.9208, RMSEP = 0.2983 (hga-SVM) r^2 = 0.9809, RMSEP = 0.1513 (SVM) r^2 = 0.8846, RMSEP = 0.3938 (BPN) r^2 = 0.9097, RMSEP = 0.2785 (hga-SVM) r^2 = 0.9637, RMSEP = 0.1731	75
HIV-1 reverse transcriptase	IC_{50}	RBFN-SVM	structural, spatial, topological, electronic descriptors	training set (35) test set (16)	r = 0.954, RSS = 2.627 r = 0.940, RSS = 2.729	76
PDGFR	IC_{50}			training set (55) test set (20)	r = 0.911, RSS = 6.535 r = 0.919, RSS = 3.186	

Table 1. Continued

property/target	experimental data	methodology	molecular descriptors	validation method	reported statistics	refs
PDGFR	IC_{50}			training set (55) test set (20)	$r = 0.911$, RSS = 6.535 $r = 0.919$, RSS = 3.186	
HIV-1 reverse transcriptase	EC_{50}	GEP SVM	CODESSA descriptors	training set (36) test set (12) 	(GEP) $r = 0.91$ (SVM) $r = 0.86$ (GEP) $r = 0.63$ (SVM) $r = 0.52$	77
anticoccidial activity	IC_{50}	GEP	473 (CODESSA descriptors)	training set (26) test set (12)	$r = 0.96$, ME = 0.24 $r = 0.91$, ME = 0.52	78
5HT1E 5HT1F	K_i	*Regression:* *k*-NN	MolConnZ descriptors (5HT1E)	training set (33) LOO cross-validation	$r^2 = 0.92$, RMSE = 0.22 $r^2 = 0.69$	79
			DRAGON descriptors (5HT1F)	training set (31) LOO cross-validation	$r^2 = 0.89$, RMSE = 0.68 $r^2 = 0.64$	
		Classification: *k*-NN	MolConnZ descriptors (selectivity 1E/1F)	training set (15) test set (7)	CCR = 0.88, SE = 1.00, SP = 1.00 CCR = 1.00	
$hA_{2A}R$ hA_3R	K_i	*Classification:* SVM	12 (3D descriptors)	training set (104) LOO cross-validation test set (51)	 $accuracy_{cv}$ = 93.3%, SE = 92.0%, SP = 94.4% accuracy = 78.4%, SE = 71.9%, SP = 89.5%	80
		Regression: SVM	12 (3D descriptors) ($hA_{2A}R$)	training set (104) LOO cross-validation test set (51)	$r = 0.83$ $r_{cv} = 0.78$, RMSR = 0.050 $r = 0.82$	
			12 (3D descriptors) (hA_3R)	training set (104) LOO cross-validation test set (51)	$r = 0.95$ $r_{cv} = 0.85$, RMSR = 0.046 $r = 0.85$	
Akt/protein kinase B	IC_{50}	*Classification:* SVM	17 (DRAGON descriptors)	training set (104) LOO cross-validation test set (44)	accuracy = 95.2% accuracy = 86.8% accuracy = 88.6%	81
		Regression: SVM		training set (104) test set (44)	$r^2 = 0.882$ $r^2 = 0.762$	
VEGFR2 kinase	IC_{50}	*Regression:* SVM	351 (DRAGON descriptors)	training set (60) LOO cross-validation 5-fold cross-validation test set (14)	 $r^2_{cv} = 0.737$ $r^2 = 0.688$ $r^2 = 0.743$	82
		Classification: LDA SVM		training set (60) LOO cross-validation training set (60) LOO cross-validation test set (14)	 accuracy = 0.797 accuracy = 0.838 accuracy = 0.857	

model was carried out by Qin et al. for a small set of pyrazolo[4,3-c]quinolin-3-ones as BzR ligands, while RBFNN methodology has been applied to other 58 structurally different non-benzodiazepine derivatives active to the same target.[68,69] SVM is also very useful for the optimization of a particular scaffold. Yuan et al. developed a QSAR model by using SVM to analyze the binding affinity of substituted 1-(3,3-diphenylpropyl)-piperidinyl amides and ureas derivatives to the chemokine receptor CCR5.[70] The statistical parameters confirm the slight superiority of SVM over projection pursuit regression (PPR) method with good values of the correlation coefficients ($r^2 = 0.867$ and $r^2 = 0.732$ on the training set and the test set, respectively). Hu et al. compared the performances of six different QSAR regression methods in the analysis of the anti-HIV-1 activity and HIV-1 reverse transcriptase binding affinity of a data set of 2-amino-6-arylsulfonylbenzonitriles.[71] In both studies, PPR and SVM techniques have demonstrated the best predictive power. So, the authors have suggested the application of these models for the assessment of activity and binding affinity of molecules with structures similar to the compounds in the training set. An analysis on HIV-1 protease inhibitors have emphasized that SVM performs better than linear PLS method, but the validation performance of SVM has been more deeply discussed in the paper by Hernández et al.[72] Fourteen molecular descriptors were combined with SVR and LS-SVM techniques to model 32 compounds, achieving a value of correlation coefficient r^2 of 0.935 and 0.938, respectively. Despite the limited number of compounds used to train the model, the results after LOO cross-validation ($r^2_{cv} = 0.641$ in PLS models and $r^2_{cv} > 0.80$ in LS-SVM model) suggest better SVM performances. However, SVM seems to give better results in the prediction step, LOO cross-validation and external validation but shows a lower prediction capability in *n*-fold cross-validation and *y*-randomization processes after the comparison with the described linear models. Alternatively, Gunturi et al. carried out a different *k*-NN regression model to predict hERG inhibition.[73] SVM can also be used to select the most suitable descriptor set to improve the results of additional 3D-QSAR models, as reported by Chen.[74] A new-hybrid genetic-based SVR approach has been presented by Goodarzi et al. as improving tool both in calibration and validation processes with respect to linear and nonlinear strategies.[75] The SVM algorithm has been modified in various ways to improve the modeling results: a more flexible radial basis function network-based transform optimized by particle swarm optimization was proposed in the analysis of HIV-1 reverse transcriptase inhibitors and a set of ATP-site inhibitors of the platelet derived growth factor receptor (PDGFR).[76]

Among nonlinear strategies, a new algorithm has been suggested as good alternative to SVM: the gene expression programming (GEP) was applied to predict the EC_{50} values of nucleoside reverse transcriptase inhibitors (NRTIs), known as specific agents against HIV-1.[77] Similar performances of GEP were obtained in comparison with SVM ($r = 0.91$ versus $r = 0.86$ on the training set, including 36 compounds). GEP technique has also been profitably used for the prediction of IC_{50} values of 38 imidazopyridine anticoccidial compounds ($r = 0.96$ and $r = 0.91$ on the training set and the test set, respectively).[78]

Both regression and classification analysis can be applied to the same problem to obtain different information on binding affinity and selectivity of the compounds. This approach has been utilized to predict the specificity and subtype selectivity of 5HT1E and 5HT1F serotonin receptors ligands.[79] In particular, *k*-NN method has been combined to different sets of molecular descriptors for 5HT1E and 5HT1F ligands, obtaining values of the correlation coefficient r^2 of 0.92 and 0.89, respectively. In a second step, once calculated a selectivity index for each compound, based on the corresponding binding affinity K_i values, a classification model has been developed by combining *k*-NN with MolConnZ descriptors, achieving a prediction accuracy of 0.88 and 1.00 on the training set and the test set, respectively. A similar scheme has been followed by Michielan et al. to analyze both binding affinity and selectivity of antagonists to human A_{2A} and A_3 adenosine receptor (hAR) subtypes.[80] In this case, SVM in combination with simple autocorrelated molecular descriptors encoding for the molecular electrostatic potential (*auto*MEP) were used to build a first binary classifier able to discriminate A_{2A} versus A_3 antagonists. Then, two parallel SVR models for A_{2A} and A_3 receptor subtypes ($r = 0.78$ and $r = 0.85$, respectively, after LOO cross-validation) have been derived to predict the binding affinity to the corresponding AR subtype. Moreover, the combined application of SVM classification and regression has been performed as reliable tool in the exploration of the activity of Akt/protein kinase B inhibitors.[81] The recent publication by Sun et al. has shown that SVM technique can give more satisfactory results in the classification approach rather than in the regression analysis in comparison with linear strategies.[82] In more detail, they have applied both regression and classification SVM-based methods to discriminate between highly active and moderate active diaryl ureas against vascular endothelial growth factor receptor-2 (VEGFR2) kinase. The best performances were achieved by the SVM classification model, with values of prediction accuracies of 0.838 and 0.857 after LOO cross-validation and external validation, respectively.

Most of the case studies discussed in this paragraph represent a proof that the prediction of intervals of activity instead of the precise activity value is a valid alternative to the traditional regression-based QSAR analysis.

4.2. Classification Strategy. The classification approach has generally shown some advantages over the regression technique. In fact, some nonlinear effects can be easily incorporated by the classification models that are also able to discriminate different biological targets. Finally, the precise experimental values are not needed in the classification method; therefore, it is more suitable to study multiple mechanisms and structural classes. However, in the modeling of biological activity, the classification approach requires the selection of a threshold value, according to the experimental parameter: as examples, a value of the diameter of inhibition zone (DIZ) for the prediction of the antimicrobial activity, as proposed for the classification analysis of some 3-nitrocoumarins in the work by Debeljak et al.;[83] IC_{50} or K_i values in the paper by Bruce et al., classifying different families of drugs: angiotensine converting enzyme (ACE), acetylcholinesterase (AchE) inhibitors, benzodiazepine receptor (BzR) ligands, cyclooxygenase-2 (COX2) inhibitors, dihydrofolate reductase (DHFR) inhibitors, glycogen phoshorylase b (GPB) inhibitors, termolysis (THER), and thrombin (THR) inhibitors.[84] These applications and further examples of nonlinear classification approaches to solve pharmacodynamic problems are summarized in Table 2.

In many cases, SVM or modifications of SVM techniques have confirmed their superiority over the linear or the other nonlinear classification-based QSAR approaches applied to different targets. The activities of 45 inhibitors of cyclin-dependent kinases (CDK1 and CDK1) with an oxindole scaffold have been modeled by Li et al., yielding more than 90% correct predictions on both training and test sets.[85] 1098 factor Xa (FXa) inhibitors of diverse structures have been investigated by using several nonlinear methods (DT, probabilistic neural network, *k*-NN and SVM combined with feature selection procedure): the introduction of a weighting function to select the descriptor set improved the prediction accuracy.[86] A similar approach was used to predict the endocrine disruptor activity of chemicals by Liu et al.[87] In this study they have collected the estrogen receptor (ER) binding affinity data of chemicals to derive least-squares SVM (LS-SVM), CP-ANN and *k*-NN models in combination with a large monodimensional and bidimensional descriptor set. The best performance was given by the LS-SVM model with percentages (%) of correct predictions of 89.66% and 83.91% on the training set and test set, respectively. SVM has been applied also in the prediction of the vasodilator activity: 228 compounds were selected as training set, while a 115-compounds test set was used to validate the final model with the aim to predict the activity of some prenylated flavonoids derivatives.[88] Dong et al. obtained a percentage (%) of correct predictions of 93% and 82.6% on the training and test set, respectively. A well performing SVM classification model has been carried out also for estrogen receptor-β ligands (105 diphenolic azoles) by Luan et al.[89] In this case, the accuracy of the model prediction on the test set, comprising 26 compounds, was 91.4%, a percentage (%) value higher than the accuracy achieved by the linear discriminant analysis (LDA), based on the same descriptor set. Yang et al. have focused their study on the prediction of the antibacterial activity by evaluating the performances of SVM classification with respect to DT and *k*-NN.[90] After feature selection process and 5-fold cross-validation, they obtained a percentage (%) of accuracy of 98.40% and a Matthews correlation coefficient of 0.96. The selection of an IC_{50} value as threshold has been used to define two classes of antagonist activity to the melanocortin-4 (MC-4) receptor and derive an accurate SVM model with only five descriptors.[91] The prediction capabilities of SVM, back-propagation neural networks, *k*-NN and DT were compared in modeling selective TNF-α converting enzyme (TACE) inhibitors by Cong et al.[92] The application of the relative models on a

Table 2. Performance of Classification-Based Nonlinear Machine Learning Models for the Prediction of Pharmacodynamic Properties

property/target	experimental data	methodology	molecular descriptors	validation method	reported statistics	refs
ACE	IC_{50}	SVM	2.5D and linear fragments descriptors	training set (114)	accuracy = 78.9%	84
				10-fold cross-validation	accuracy = 82.2%	
AchE	IC_{50}			training set (111)	accuracy = 69.4%	
				10-fold cross-validation	accuracy = 77.1%	
BzR	IC_{50}			training set (163)	accuracy = 74.0%	
				10-fold cross-validation	accuracy = 75.8%	
COX2	IC_{50}			training set (322)	accuracy = 72.6%	
				10-fold cross-validation	accuracy = 71.1%	
DHFR	IC_{50}			training set (397)	accuracy = 83.5%	
				10-fold cross-validation	accuracy = 86.5%	
GPB	K_i			training set (66)	accuracy = 77.4%	
				10-fold cross-validation	accuracy = 76.7%	
THER	K_i			training set (76)	accuracy = 75.3%	
				10-fold cross-validation	accuracy = 73.4%	
THR	K_i			training set (88)	accuracy = 71.1%	
				10-fold cross-validation	accuracy = 69.8%	
CDK1 CDK2	classes: active: $IC_{50} \leq 100$ nM inactive: $IC_{50} > 100$ nM	LDA SVM	8 (CODESSA descriptors) 6 (CODESSA descriptors)	training set (45)	(LDA) CDK1 accuracy = 91.1%, SE = 0.89, SP = 1.00 CDK2 accuracy = 88.89%, SE = 1.00, SP = 0.73 (SVM) CDK1 accuracy = 93.3%, SE = 1.00, SP =1.00 CDK2 accuracy = 100%, SE = 0.93, SP = 0.88	85
				test set (22)	(LDA) CDK1 accuracy = 95.45% CDK2 accuracy = 86.36% (SVM) CDK1 accuracy = 100% CDK2 accuracy = 90.91%	
FXa	K_i	DT PNN *k*-NN SVM+RFE	199 (connectivity, shape, quantum chemical, electro-topological state and geometrical descriptors)	training set (715) 5-fold cross-validation	(DT) accuracy = 92.3%, SE = 89.1%, SP =93.8% (PNN) accuracy = 94.0%, SE = 97.5%, SP = 92.3% (*k*-NN) accuracy = 93.5%, SE = 94.1%, SP = 93.2% (SVM) accuracy = 97.0%, SE = 94.6%, SP = 98.1%	86
				test set (383)	(DT) accuracy = 93.5%, SE =91.0%, SP =94.1% (PNN) accuracy = 95.3%, SE = 94.9%, SP = 95.4% (*k*-NN) accuracy = 95.6%, SE = 98.7%, SP = 94.8% (SVM) accuracy = 98.2%, SE = 98.7%, SP = 98.0%	
ER	logRBA	*k*-NN CP-ANN LS-SVM	1D and 2D descriptors	training set (232)	(*k*-NN) accuracy = 89.22%, SE = 83%, SP = 93.9% (CP-ANN) accuracy = 87.50%, SE = 76%, SP = 96.2% (LS-SVM) accuracy = 89.66% SE =80%, SP = 96.9%	87
				test set (87)	(*k*-NN) accuracy = 83.91%, SE = 70%, 91.2% (CP-ANN) accuracy = 82.76%, SE = 76.6%, 91.2% (LS-SVM) accuracy = 83.91%, SE = 70%, SP = 91.2%	
vasodilator activity	EC_{50}	SVM	9 (DRAGON descriptors)	training set (228) test set (115)	accuracy = 93.0%, SE = 87.1%, SP = 95.5% accuracy = 82.6%, SE = 78.0%, SP = 85.1%	88
ER-β	Classes: active: $IC_{50} < 100$ nM inactive: $IC_{50} \geq 100$ nM	LDA SVM	CODESSA descriptors	training set (79) test set (26)	(LDA) accuracy = 81.9%, SE = 88.8%, SP = 66.7% (SVM) accuracy = 91.4%, SE = 94.4%, SP = 84.8%	89
Antibacterial activity	EC_{50}	SVM DT *k*-NN	36 (simple, molecular connectivity and shape, electrotopological state, quantum chemical and geometrical descriptors)	training set (285) test set (171) external validation set (155)	(SVM) accuracy = 98.06%, SE = 100%, SP = 100% (DT) accuracy = 90.97%, SE = 91.5%, SP = 92.8% (*k*-NN) accuracy = 96.13%, SE = 98.3%, SP = 97.3% (SVM) SE = 98.15%, SP = 98.02% (DT) SE = 92.59% , SP = 90.10% (*k*-NN) SE = 98.15% , SP = 95.05%	90
TACE	IC_{50}	SVM BPNN *k*-NN DT	23 (simple, connectivity and shape, electrotopological state, quantum chemical and geometrical descriptors)	training set (531) test set (348) validation set (323)	(SVM) accuracy = 98.45% , SE = 96.64%, SP = 99.51%, MCC = 0.967 (BPNN) accuracy = 97.52%, SE = 94.96%, SP = 99.02%, MCC = 0.947 (*k*-NN) accuracy = 98.45%, SE =98.32%, SP = 98.53%, MCC = 0.967 (DT) accuracy = 97.83%, SE = 98.32%, SP = 97.55%, MCC = 0.954	92
PXR	Classes: active: $EC_{50} \leq 100$ μM inactive: $EC_{50} > 100$ μM	RP RF SVM	68 (VolSurf descriptors)	training set (177) 10-fold cross-validation test set 1 (14) test set 2 (145)	(RP) accuracy = 87.57%, SE = 95.92%, SP = 77.21% (RF) accuracy = 73.45% , SE = 82.65%, SP = 62.02% (SVM) accuracy = 94.35%, SE = 98.98%, SP = 88.61% (RP) accuracy = 63.45%, SE = 64.63%, SP = 61.9% (RF) accuracy = 65.52%, SE = 64.63%, SP = 66.67% (SVM) accuracy = 66.9%, SE = 68.29%, SP = 65.08%	93
hERG 5-HT_{2B}	Classes: inhibitors: $IC_{50} < 1$μM non-inhibitors: $IC_{50} > 10$μM	SVM *k*-NN SVM	shape signatures	training set (83) 10-fold cross-validation leave-20-out	(SVM) accuracy = 78% (*k*-NN) accuracy = 69% (SVM) accuracy = 74%, SE = 73%, SP = 74% (*k*-NN) accuracy = 67%, SE = 79%, SP = 56%	94
	Classes: active: $K_i \leq 100$ nM inactive: $K_i \geq 1$ μM	SOM *k*-NN		training set (182) 10-fold cross-validation leave-42-out	(SVM) accuracy = 87% (SOM) accuracy = 86% (*k*-NN) accuracy = 74% (SVM) accuracy = 83%, SE = 91%, SP = 69% (SOM) accuracy = 70%, SE = 78%, SP = 54% (*k*-NN) accuracy = 79%, SE = 93%, SP = 53%	
PRX	Classes: active: $EC_{50} \leq 100$ μM inactive: $EC_{50} > 100$ μM	SVM	20 (MOE descriptors) shape signatures	training set (168) 10-fold cross-validation test set (130) 10-fold cross-validation	(1DSS+MOE) accuracy = 81%, MCC = 0.531 (2DSS+MOE) accuracy = 75%, MCC = 0.431 (1DSS+MOE) accuracy = 72%, MCC = 0.289 (2DSS+MOE) accuracy = 73%, MCC = 0.370	95

Table 2. Continued

property/target	experimental data	methodology	molecular descriptors	validation method	reported statistics	refs
Serotonin release NF-κB	Classes: 1: $IC_{50} < 5$ μM 2: 5 μM ≤ $IC_{50} < 10$ μM 3: 10 μM ≤ $IC_{50} < 40$ μM 4: 40 μM ≤ $IC_{50} < 100$ μM 5: 100 μM ≤ $IC_{50} < 300$ μM 6: $IC_{50} \geq 300$ μM	CPGNN	global, atomic, surface potentials descriptors	training set (54) 10-fold cross-validation	Serotonine release accuracy = 85.2% NF-κB accuracy = 80.6%	98
MRP2	Classes: inhibitors non-inhibitors	SVM	16 (PCLIENT descriptors)	training set (257) 5-fold cross-validation test set (61)	accuracy = 82.9%, SE = 83.3%, SP = 82.4% accuracy = 77.1%, SE = 78.6%, SP = 75.8%	99
AchE BzR DHFR COX2 CYPC17	Classes: active: Low IC_{50} inactive: High IC_{50}	(MILP) based hyper-boxes	DRAGON descriptors	10-fold cross-validation training set	AchE accuracy = 100% (7 attributes) BzR accuracy = 96.36% (7 attributes) DHFR_TG accuracy = 97.74% (7 attributes) DHFR_RL accuracy = 96.99% (7 attributes) DHFR_PC accuracy = 97.62% (7 attributes) COX2 accuracy = 98.13% (7 attributes) CYPC17 accuracy = 100% (7 attributes)	100
Protein-protein interaction	Classes: inhibitors non-inhibitors	DT	1664 (constitutional, molecular profile, functional group count descriptors)	training set (25) 10-fold cross-validation	SE = 88%, SP = 98% SE = 84%, SP = 98%	103
Lck	Classes: inhibitors: $IC_{50} \leq 10$ μM non-inhibitors: $IC_{50} > 500$ μM	SVM	100 (MODEL descriptors)	training set (810) 5-fold cross-validation test set (80)	accuracy = 99.7%, SE = 87.8%, SP = 99.9%, MCC = 0.788 SE = 83.8%	105
Abl	Classes: inhibitors: $IC_{50} < 50$ μM non-inhibitors: $IC_{50} > 50$ μM	SVM	98 (simple, chemical properties, electro-topological state, molecular connectivity and shape descriptors)	training set (708) 5-fold cross-validation	accuracy = 99.86%, SE = 89.12%, SP = 99.97%, MCC = 0.865	106

training set (531 compounds), a test set (348 compounds) and an external validation set (323 compounds), after feature selection procedure, has demonstrated the good performances of machine learning methods, with particular reference to SVM (a percentage (%) of accuracy of 98.45% on the validation set). Khandelwal et al. demonstrated the good prediction capability of machine learning methods in the prediction of pregnane X receptor (PXR) agonists also in comparison with the docking technique.[93]

Novel tools were introduced to represent the compounds in QSAR studies. In particular, the shape signatures are codification of the molecular shape and polarity, calculated by an algorithm able to explore the volume enclosed by the surface of a molecule. Chekmarev et al. have used these molecular descriptors in combination with nonlinear classification methods to model two series of 5-HT_{2B} ligands and hERG potassium channel inhibitors.[94] The percentages (%) values of accuracy of 87% and 78% on the 5-HT_{2B} and hERG training sets, respectively, after 10-fold cross validation procedure, denote that the molecular shape and polarity are crucial aspects to describe the molecular activity. Shape signatures, in addition to docking experiments, have also been used in a hybrid method for the prediction of pregnane X receptor activators.[95] Moreover, the topological-fragment-spectra (TFS) representation considers the occurrence of all possible substructures in the molecule and was introduced to transform the drugs of 100 different therapeutic classes into pattern vectors in the work by Kawai et al.[96] They were able to build multilabel SVM models by using "one against all" method for all classes, yielding almost all sensitivity and specificity percentages (%) higher than 90%. In this case, the multilabel task is decomposed into several binary classifiers, one for each class and each SVM classification model was trained to separate the drugs that belong to one class from the drugs that do not belong to it.

In the traditional single-label classification, classes are considered mutually exclusive, but in the multilabel tasks, some samples might belong to multiple classes, that result to be overlapping. Very recently, a multilabel SVM approach was applied for the classification of human adenosine receptor (hAR) antagonists in order to predict both their potency and selectivity.[97] This paper has reported one of the very few models able to predict the selectivity toward a target. The combination of *auto*MEP vectors with ct-SVM analysis has been introduced as a novel strategy to discover potent and selective hAR antagonists with xanthine and pyrazolo-triazolo-pyrimidine scaffolds. In more detail, a large collection of hAR antagonists has been utilized to carry out and validate three classification models. They have been applied *in series* as quantitative sieves, based on decreasing thresholds of potency (500 nM, 250 nM and 100 nM), corresponding to different binding affinity K_i values. The prediction results on internal and external test sets confirm the *auto*MEP/ct-SVM strategy as valuable tool for this multilabel problem.[97] The quantitative use of the classification approaches is also shown in a work considering 54 sesquiterpene lactones, known to inhibit NF-κB and the serotonin release.[98] CPGNN analysis has been applied to classify these derivatives into six increasing activity classes by using 3D global, atomic and surface properties descriptors. The prediction accuracies for the NF-κB and serotonin release models (80.6% and 85.2%, respectively) confirm the reliability of both selected descriptors and neural networks technique in the treatment of these particular structural data. Some SVM models in combination with both molecular descriptors and pharmacophore analysis have been built for MRP2 inhibitors.[99] In this study, satisfactory values of accuracy were obtained in the validation step: 82.9% after 5-fold cross-validation on the training set and 77.1% after the test set prediction.

A novel mixed-integer programming (MILP) based hyperboxes method has been described by Armutlu et al.[100] This strategy was successfully applied, giving superior results over several nonlinear techniques, in the evaluation of the activity of inhibitor data sets of molecules targeting acetylcholinesterase (AchE), benzodiazepine receptor (BzR), dihydrofolate reductase (DHFR), cyclooxygenase-2 (COX2), and cytochrome P450 C17 (CYPC17). Eighteen different selectivity sets have been considered to investigate the performances of a new SVM-based approach with the aim to search for target-selective compounds and to discriminate selective and nonselective compounds.[101] In particular, four ranking strategies have been combined to different kernel functions and 2D fingerprints as molecular descriptors for SVM classification to solve a three-classes problem. The capability of this methodology to distinguish selective, nonselective but active and nonactive compounds as potential tool in the preliminary study of the structural determinants of selectivity has been discussed. A further interesting application of SVM is the prediction of ligands for orphan targets by using target-ligand kernel functions to incorporate the information on protein and small molecules.[102]

The decision tree method can be applied in the prediction of protein−protein interaction inhibitory activity of molecules.[103] Furthermore, the in silico screening of drug-like compounds has been performed by using decision trees and support vector machine, demonstrating the suitability of these methods in such application.[104] In more detail, 90.2% of the drugs in the external test set were successively retained after the application of different classification schemes based on various descriptor sets. Ultimately, further applications of SVM models in the virtual screening process have been shown for both Lck and Abl inhibitors.[105,106]

5. EVALUATION OF ADME PROPERTIES

Nowadays, the investigation of the properties related to absorption, distribution, metabolism and excretion of drug candidates is needed as well as their pharmacodynamic profile evaluation in the initial stage of drug discovery.[107,108] The computational strategies are considered valuable tools for the study of the various aspects of the pharmacokinetic profile and for the successful selection of compounds for the synthesis prioritization. Recent prediction examples of ADME properties (aqueous solubility, Caco-2 and MDCK permeability, blood-brain barrier, human intestinal absorption, plasma protein binding, as well as oral bioavailability) have been reported by Wang et al.[109] Recently, several computational models for the prediction of ADME properties have been reviewed by Norinder et al.[110] In particular, they analyzed the practical application of both linear and nonlinear strategies for the prediction of the parameters influencing the intestinal drug absorption (solubility, permeability and the fraction absorbed). They have also underlined the need for a large number of chemically diverse high-quality data and, once identified the applicability domain, the development of mechanism-based models have been suggested to improve the predictive accuracy. Some additional publications regarding in silico tools for toxicokinetic studies have been reported by Ruiz-Garcia et al.[111] The prediction of drug absorption and permeability represents a difficult task, since these phenomena are influenced by multiple physiological processes. However, the availability of large data set and the possibility to combine multiple models might help the judgment of the predictions reliability, as reported by Hou et al.[112]

Several in silico methods have focused on the metabolic endpoints, and the prediction of drug metabolism directly from structure represents an advanced approach integrated into expert systems.[113,114] Moreover, the computational tools are suggested to successfully assist the in vitro methods for human drug metabolism to compensate the limitations of the use of each of these approaches alone.[115] In the past years, ANN and SVM techniques have gained interesting progresses in both classification and quantitative prediction of different metabolic endpoints, as summarized by Fox et al.[116]

The metabolic profile of a drug candidate is an important aspect to be considered in the selection of potential new drugs and several problems related to stability, toxicity of xenobiotics and drug−drug interactions might represent serious adverse effects. In fact, in the case of coadministration of therapeutic agents, the pharmacological profile of each drug might be modified by the presence of other drugs in the human body. In particular, cytochrome P450 (CYP450) class of enzymes is responsible for phase 1 metabolism. This detoxification system is highly complex, since it includes many different CYP450 isoforms characterized by multiple binding sites, polymorphism and enzyme induction or modulation phenomena. Many drugs are known to inhibit specific CYP450 isoforms and to be substrates for others. The early detection of potential drug−drug interactions is highly desirable to avoid costly failures in drug development. Some valid machine learning applications in the prediction of the CYP450 specificity and drug−drug interactions have been summarized by Arimoto,[117] while Yap et al. have discussed the performances of support vector machine to predict CYP450 substrates and inhibitors.[118] Furthermore, the ultimately used QSAR approaches to study the binding processes to CYP450 or UDP-glucuronosyltransferase metabolizing enzymes have been reviewed by Chohan et al.[119] Drug metabolism comprises several aspects: different metabolic site, types of metabolic enzymes, interactions and sites of interactions. Moreover, the potentialities of QSAR strategies in the analysis of both recognition and inhibition processes in the CYP450-mediated metabolism have been considered by Li et al.[120] Finally, the potential application of modeling the metabolism properties in the screening approach has been shown by Crivori et al.[121]

Moreover, the computational approach in the analysis of the blood-brain partitioning process of drugs has been applied to predict the brain permeability.[122] In this field, novel nonlinear methods were used to improve the robustness, and the accuracy of the predictions in comparison with the traditional linear approaches. Both neural networks and SVM have been introduced as more promising strategies to model larger collections of logBB data in comparison with the previously attempted linear methods.

In recent years, the machine learning methods have been applied to predict different pharmacokinetic properties and each of the references reported in Table 3 focuses on a particular pharmacokinetic aspect.

Regarding the analysis of the passive drug absorption process, Hou et al. have carried out both correlation and classification approaches, by using genetic function approximation (GFA) and recursive partitioning (RP) tech-

niques, respectively.[123] They have collected the fraction absorption (% FA) of 579 drug and drug-like molecules, known to be transported by passive diffusion. In the classification approach, a threshold % FA was introduced to discriminate the training set (481 compounds) into poor (%FA ≤ 30%) and good (%FA > 30%) intestinal absorption classes. The RP method has been combined with 45 different molecular descriptors to develop a decision tree, and a 98 compounds-test set was used to assess the performance of the classification model. The selected molecular properties in the final decision tree have shown the importance of hydrophobicity, hydrogen-bonding potential, and molecular weight in the human intestinal absorption process. A good modeling performance was achieved for the training set on the poor intestinal absorption class (percentage of accuracy of 95.9%) and the good intestinal absorption class (percentage of accuracy of 96.1%). Moreover, all five compounds in the test set with %FA ≤ 30% were correctly classified, while the remaining 93 compounds were predicted with an accuracy of 96.8%. The same data set has been considered in a further study by using SVM analysis in combination with seven molecular descriptors.[124] The satisfactory results, demonstrated by the prediction accuracies for the training and the test set, and the reduced descriptor set, support the higher robustness of SVM in comparison with the RP technique. Wan et al. have focused on the prediction of the in vitro brain tissue binding (fraction unbound, f_u) by applying various linear and nonlinear regression techniques.[125] This experimental parameter is crucial to evaluate the brain penetration of drugs targeting CNS and, consequently, their pharmacological effects. Even in this case, SVM has given the best results (a value of correlation coefficient r^2 of 0.871 after prediction of the test set).

Table 3. Performance of Nonlinear Machine Learning Models for the Prediction of ADME Properties

property/target	experimental data	methodology	molecular descriptors	validation method	reported statistics	refs
human intestinal absorption	fraction absorption (FA%) classes: 1: FA% ≤ 30% 2: FA% > 30%	*Classification:* RP	45 (ACDLAB and Cerius2 descriptors)	training set (481) test set (98)	accuracy(class 1) = 95.9% accuracy(class 2) = 96.1% accuracy(class 1) = 100% accuracy(class 2) = 96.8%	123
human intestinal absorption	fraction absorption (FA%) classes: HIA- : FA% ≤ 30% HIA+ : FA% > 30%	*Classification:* SVM	7 (ACDLAB and Cerius2 descriptors)	training set (480) test set (98)	accuracy(Class HIA-) = 97.8% accuracy(Class HIA+) = 94.5% accuracy(Class HIA-) = 100% accuracy(Class HIA+) = 97.8%	124
in vitro brain tissue binding	fraction unbound (f_u)	*Regression:* NN SVM	1D and 2D descriptors	training set (56) test set (24)	(NN) r^2 = 0.819, RMSE = 0.41 (SVM) r^2 = 0.871, RMSE = 0.36	125
blood-to-plasma concentration ratio	R_b (C_{Blood}/C_{Plasma})	*Regression:* ANN	30 (ALOGPS and DRAGON descriptors)	training set (93) test set (7)	r^2 = 0.927, RMSE = 0.17, %outside$_{1.25\text{-fold error}}$ = 14 r^2 = 0.871, RMSE = 0.20, %outside$_{1.25\text{-fold error}}$ = 16	128
metabolic stability	Cl_{int}	*Classification:* RF SVM	193 (MOE descriptors)	training set (1952) 10 –fold cross-validation test set (487)	(RF) SP = 0.64 (SVM) SP = 0.56 (RF) accuracy > 0.8, SE > 0.9, SP > 0.6 kappa = 0.71, MCC = 0.72 (SVM) accuracy > 0.8, SE > 0.9, SP > 0.6 kappa = 0.70, MCC = 0.71	129
metabolic stability	% recovery after 30 min classes: stable: recovery ≥ 50% unstable: recovery < 50%	*Classification:* GP SVM *Regression:* GP SVM	1664 (constitutional, topological, walk and path counts, eigenvalue-based indeces, functional group counts, atom-centered fragments descriptors)	training set (1931) 2-fold cross-validation test set (700)	(GP) AUC = 85.0 (GP) AUC = 86.2 (SVM) AUC = 83.5 (SVR) AUC = 85.5 (GP) AUC = 71.8	130
CYP450 isoform specificity	Classes: 3A4 2D6 2C9	*Classification:* SVM	12 (ADRIANA.*Code* descriptors)	training set (146) LOO cross-validation test set (233)	accuracy = 89%, SE (3A4) = 88.7%, SE (2D6) = 95.5%, SE (2C9) = 76.2%, SP (3A4) = 91.0%, SP (2D6) = 84.3%, SP (2C9) = 94.1% accuracy = 83%, SE (3A4) = 79.9%, SE (2D6) = 89.8%, SE (2C9) = 80.0%, SP (3A4) = 92.0%, SP (2D6) = 74.7%, SP (2C9) = 64.0%	131
CYP450 isoform specificity	Classes: 1A2 2C9 2D6 2E1 3A4	*Classification:* ct-SVM SVM	27 (ADRIANA.*Code* descriptors) 18 (ADRIANA.*Code* descriptors)	training set (345) test set (209) training set (293) test set (191)	accuracy = 70%, SE (1A2) = 52%, SE (2C9) = 61%, SE (2D6) = 70%, SE (2E1) = 92%, SE (3A4) = 72%, SP (1A2) = 85%, SP (2C9) = 93%, SP (2D6) = 88%, SP (2E1) = 97%, SP (3A4) = 79% accuracy = 78%, SE (1A2) = 64%, SE (2C9) = 73%, SE (2D6) = 77%, SE (2E1) = 100%, SE (3A4) = 78%, SP (1A2) = 97%, SP (2C9) = 98%, SP (2D6) = 91%, SP (2E1) = 97%, SP (3A4) = 85%	132
CYP450 1A2, 2D6 and 3A4 interaction	classes: active, inactive (1A2) active, inactive (2D6) active, inactive (3A4)	*Classification:* RF *k*-NN ANN DT SVM	118 (ChemAxon and Chemistry Development Kit descriptors)	training set (353) 10-fold cross-validation	(RF) accuracy(1A2) = 66.7% accuracy(2D6) = 78.1% accuracy(3A4) = 67.5% (*k*-NN) accuracy(1A2) = 69.7% accuracy(2D6) = 79.0% accuracy(3A4) = 62.3% (ANN) accuracy(1A2) = 67.4% accuracy(2D6) = 79.6% accuracy(3A4) = 61.9% (DT) accuracy(1A2) = 91.5%	133

Table 3. Continued

property/target	experimental data	methodology	molecular descriptors	validation method	reported statistics	refs
					accuracy(2D6) = 89.2% accuracy(3A4) = 81.4% (SVM) accuracy(1A2) = 71.2% accuracy(2D6) = 83.1% accuracy(3A4) = 67.2%	
CYP450 2D6 and 3A4 inhibition	classes: non-inhibitors: inhibition < 50% inhibitors: inhibition ≥ 50%	*Classification:* Gaussian kernel weighted *k*-NN	extended connectivity fingerprints (ECFP) functional class fingerprints (FCFP)	training set (865) test set (288) training set (1037) test set (345)	accuracy = 82%, SE = 88.2%, SP = 94.0% accuracy = 90%, SE = 84.1%, SP = 97.5% accuracy = 90%, SE = 77.9%, SP = 96.5% accuracy = 88%, SE = 80.0%, SP = 94.7%	134
CYP450 2D6 inhibition	classes: inhibitors: $IC_{50} \leq 50$ μM non-inhibitors: $IC_{50} > 50$ μM	*Classification:* SVM	557 (atom-count, constitutional, topological, autocorrelations, aromaticity descriptors)	training set (185) test set (78)	hit rate = 0.92, false alarm rate = 0.23	135
CYP450 3A4 inhibition	classes: inhibitors non-inhibitors	*Classification:* RP	2D (constitutional, topological, geometric descriptors)	training set (741) test set (186)	accuracy = 77.33%, SE = 75.82%, SP = 70.3% accuracy = 72.58%, SE = 82.64%, SP = 53.85%	136
CYP450 1A2 inhibition	classes: inhibitors non-inhibitors	*Classification:* SVM RF *k*-NN DT	324 (MOE and VolSurf+ descriptors)	training set (400) test set (7000) external test set (89)	(SVM) accuracy = 82%, TP = 79%, TN = 84% (RF) accuracy = 100%, TP =100%, TN = 100 % (*k*-NN) accuracy = 83%, TP = 86%, TN = 80% (DT) accuracy = 97%, TP = 98%, TN = 97% (SVM) accuracy = 75% , TP = 73%, TN = 78% (RF) accuracy = 76%, TP = 78%, TN = 74% (*k*-NN) accuracy = 74% , TP = 79%, TN = 68% (DT) accuracy = 71% , TP = 71%, TN = 70% (SVM) accuracy = 67%, TP = 62%, TN = 72% (RF) accuracy = 58%, TP = 56%, TN = 70% (*k*-NN) accuracy = 58%, TP = 77%, TN = 44% (DT) accuracy = 53%, TP = 54%, TN = 52%	137
CYP450 2D6, 1A2, 3A4, 2A6, 2C9, 2C8, 2C19 and 17 inhibition	pIC_{50} BFE classes: low active high active	*Classification:* MILP hyper-boxes	6 (DRAGON descriptors)	10-fold cross-validation	pIC_{50} 2D6 accuracy = 97.05% 1A2 accuracy = 91.94% 3A4 accuracy = 89.27% 2A6 accuracy = 88.50% 2C9 accuracy = 83.00% 2C8 accuracy = 81.67% 2C19 accuracy = 83.17% 17 accuracy = 82.80% BFE 2D6 accuracy = 92.90% 1A2 accuracy = 92.44% 3A4 accuracy = 87.99% 2A6 accuracy = 86.25% 2C9 accuracy = 87.86% 2C8 accuracy = 79.83% 2C19 accuracy = 84.83% 17 accuracy = 92.60%	138

Table 4. Performance of Nonlinear Machine Learning Models for the Toxicity Prediction

property/target	experimental data	methodology	molecular descriptors	validation method	reported statistics	refs
skin sensitization	potency categories	*Classification:* ANN	CODESSA and DRAGON descriptors	training set 1 (358) training set 2 (307) training set 3 (251)	accuracy = 90%, SE = 92%, SP = 88% accuracy = 95%, SE = 95%, SP = 95% accuracy = 90%, SE = 82%, SP = 96%	140
toxicity	EC_{50} relative binding affinity	*Regression:* RBFNN SVM	CODESSA descriptors	training set (61) LOO cross-validation test set (15) training set (118) LOO cross-validation test set (28)	(RBFNN) $r = 0.91$ (SVM) $r = 0.96$ (RBFNN) $r^2 = 0.81$ (SVM) $r^2 = 0.92$ (RBFNN) $r = 0.80$ (SVM) $r = 0.93$ (RBFNN) $r = 0.86$ (SVM) $r = 0.88$ (RBFNN) $r^2 = 0.72$ (SVM) $r^2 = 0.76$ (RBFNN) $r = 0.75$ (SVM) $r = 0.93$	141
genotoxicity	maximal SOS induction factor (IMAX)	*Classification:* Grid search SVM	7 (CODESSA descriptors)	training set (123) test set (27) external test set (11)	accuracy = 92.9%, TP = 87.5%, TN = 95.1% accuracy = 92.6%, TP = 100%, TN = 90% accuracy = 90.9%, TP = 66.7%, TN = 100%	145
hepatocarcinogenic toxicity	TD_{50} (mg/Kg)	*Regression:* SVM	11 (CODESSA descriptors)	training set (46) LOO cross-validation 10-fold cross-validation test set (9)	$r^2 = 0.919$ $r^2 = 0.58$ $r^2 = 0.60$ $r^2 = 0.707$	146

The steady-state volume of distribution (VD_{SS}) in human is a crucial pharmacokinetic parameter that determines, together with clearance, the half-life of drugs. Half-life regulates their free plasma concentration, which influences the pharmacological and the potential side effects. Based on the evaluation of this pharmacokinetic parameter, the dosage regimen can be established. The last progress in the field of VD prediction has been reported by Sui et al.[126] Very recently, a collection of 669 drugs has been utilized to model and predict VD_{SS} by using the regression random forest (RF)

approach in combination with 280 MOE and VolSurf+ descriptors.[127]

A further parameter related to the volume of distribution and the clearance able to describe the drug distribution within blood is the blood-to-plasma concentration ratio (R_b), below 1 if a drug concentrates in plasma or above 1 if drug binds more in blood cells than in plasma. A QSAR model by using ANN has demonstrated the best prediction capability in comparison with the PLS analysis, yielding a value of correlation coefficient r^2 of 0.927 and of 0.871 for the training and the test set, respectively.[128] In particular, ANN model has given the highest number of drugs predicted within a 1.25-fold error, showing its superior performance over a linear method. Several machine learning methods have been used for the prediction of the metabolic stability. In the paper by Sakiyama et al., a particular value of in vitro intrinsic clearance (Cl_{int}) was selected to distinguish the data set into two groups of stable and unstable compounds.[129] The RF, SVM, logistic regression, and RP methods were applied to perform the classification analysis on a large training set, and their performance has been evaluated in the prediction of the test set (487 compounds). By considering the values of Matthews correlation coefficient on the test set (0.72 and 0.71, respectively), the prediction accuracy, sensitivity, and specificity, the RF as well as SVM have shown a slightly more predictive capability than the other classification strategies. Schwaighofer et al. have considered the outcome of experiments with liver microsomes of different species to measure the in vitro metabolic stability.[130] A probabilistic approach was used by applying the Gaussian Process (GP) and SVM analysis to generate both regression and classification models. The selected GP classification model was evaluated by 2-fold cross-validation procedure on the training set and by predicting a blind test set, with the resulting area under the ROC curve 85.0 and 71.8, respectively. Moreover, the model has output the probability of being metabolically stable and the performance of GP model has been demonstrated to be competitive with respect to the previously published works.

As anticipated drug metabolism is highly related to the CYP450 superfamily of hemoprotein enzymes. They play an important role in the degradation of drugs through oxidation reactions and various CYP450 isoforms are differently responsible for the metabolism process. Recently, Terfloth et al. have considered the isoform specificity of cytochrome P450 3A4, 2D6, and 2C9 substrates.[131] Different model building methods (*k*-NN, DT, multilayer perceptron, RBFNN, logistic regression, and SVM) have been applied to a 146-drugs training set. The best model was obtained by performing the automatic variable selection, leading to 12 variables, in combination with SVM. The results achieved after an extensive cross-validation process have assessed the model robustness, with a percentage (%) of 89% correctly predicted drugs in the training set after LOO cross-validation procedure. Moreover, 83% correct predictions were obtained for the external validation set (233 compounds). The values of accuracy have shown an improvement in comparison with the previous literature. Moreover, a novel multilabel classification analysis by using SVM has been applied to predict the isoform specificity of CYP1A2, CYP2C9, CYP2D6, CYP2E1, and CYP3A4 substrates. It was demonstrated that the multilabel classification approach is more suitable than the traditional single-label one to study the CYP450 metabolism, with drugs metabolized by multiple isoforms.[132] CYP450 substrates have also been classified according to their interaction with CYP1A2, CYP2D6, and CYP3A4 isoforms by applying different machine learning techniques.[133] In this application, a decision tree algorithm has outperformed the remaining techniques with accuracies on the global training set after 10-fold cross-validation procedure of 91%, 89.2%, and 81.4% for CYP1A2, CYP2D6, and CYP3A4 compounds, respectively. Regarding the inhibition of CYP450, both CYP2D6 and CYP3A4 inhibition was studied by Jensen et al.[134] Classification models for CYP2D6 and CYP3A4 inhibitors were built by using Gaussian kernel weighted *k*-NN based on connectivity and functional class fingerprints. This modeling method has improved the performances obtained by applying the classical *k*-NN technique, with percentages (%) of accuracy close to 85% for both data sets after LOO cross-validation and the test set prediction. These results suggest that the decomposition of compounds into ring fragments and functional groups is a useful strategy to represent CYP2D6 and CYP3A4 inhibitors. Eitrich et al. have focused on the 2D6 isoform and showed how the classification model can be used to model unbalanced data sets.[135] They have classified the compounds in the training set (185 drugs) and in the test set (78) into 2D6 inhibitors and noninhibitors. Then, they have applied the SVM method in combination with feature selection, oversampling and threshold moving techniques to solve the problem of unbalanced classification. In fact, regarding the unbalanced data of CYP2D6 inhibitors, the authors have concluded that the application of these strategies improves the accuracy and the sensitivity of the SVM model in the test set prediction. CYP3A4 is the predominant isoform, responsible for the metabolism of about 50% of drugs. Therefore, the prediction of the inhibitory activity of drug candidates on this isoform is important to know the potential accumulation of the therapeutic agents in the case of coadministration. Choi et al. have carried out a classification model to discriminate CYP3A4 inhibitors and noninhibitors by using RP and RF techniques in combination with constitutional, electrostatic and geometric descriptors.[136] In particular, the RP model achieved the percentages (%) of 72.58 and 82.64 as accuracy and sensitivity values, respectively, but a low specificity, on the test set prediction. The results have shown that most of the noninhibitors are predicted with more uncertainty in both training and test set, while most of the inhibitors are correctly predicted. Very recently, the inhibitory activity on CYP1A2 has been investigated by Vasanthanathan et al.[137] SVM, RF, *k*-NN, and DT methods in combination with the BestFirst variable selection procedure were compared in the classification of CYP1A2 inhibitors and noninhibitors. Nonlinear SVM and RF models provided the best results, with 75% and 76% correctly predicted compounds in the test set, respectively. The advantage of the RF method, is the prediction of the probability to belong to one of the classes: in fact, if the probability is lower than 50%, then the compounds are classified as noninhibitor, while the compounds are inhibitors if the predicted probability is higher than 50%. In this paper, the predictive capability was assessed on an external 89-drugs test set, showing the superiority of the nonlinear SVM method, able to correctly classify 60 out of 89 compounds.

Finally, a novel mixed-integer programming (MILP) based hyper-boxes technique in combination with six molecular descriptors, as mentioned in the previous paragraph, has been used for the prediction of different classes of activity, based on pIC_{50} and binding free energy (BFE) values, to eight different CYP450 isoforms.[138] Interestingly, docking and molecular dynamic simulation experiments have selected the conformations used in this study. Here, the best results were given by MILP-hyperboxes method with respect to other nonlinear strategies, as shown in the previous application.

All models for the prediction of toxicokinetic properties reviewed here are potentially applicable for virtual screening, with the aim of preventing any potential drug−drug interaction at the early stage of the drug discovery process.

6. QSAR IN PREDICTIVE TOXICOLOGY

The need for various toxicological endpoints data of chemicals in limited time and animal experiments requires the application of alternative computational solutions. The computational toxicology offers a valuable tool to speed up the costly screening of high numbers of compounds. In fact, computer-based identification of molecular structure properties qualitatively (SAR) or quantitatively (QSAR) related to biological activity represents the main useful application in predictive toxicology.[21,22,24]

In Table 4 we have reported the most recent examples of the use of machine learning in the prediction of toxicological properties. Recently, Saliner et al. have given an overview on the development of QSAR approaches to model irritation and corrosion properties of chemicals.[139] The limitation of the current models, despite their promising performances, is their debatable applicability for regulatory purposes. Consequently, according to REACH mentality, QSAR models should be investigated further to assess their regulatory use. Golla et al. carried out model for the prediction of skin sensitization effects of chemicals.[140] In this paper, three different data sets have been considered and the toxicity scores were assigned to classify chemicals into levels of toxicity. ANN in combination with several literature and structural descriptors was used to derive QSAR models able to predict skin sensitization with percentages (%) of accuracy of 90%, 95%, and 90%, respectively. SVM technique was compared with other linear and nonlinear methods in the prediction of toxicity for different training sets (toxicity EC_{50} values and binding to the androgen receptor) by Zhao et al.[141] The best performances were given by the SVM model, achieving values of the correlation coefficient r^2 of 0.92 and 0.76, after LOO cross-validation on the first and second training set, respectively. The prediction results on the test set have confirmed the good generalization capability of SVM ($r = 0.93$ for both test sets).

Several QSAR models have been carried out to predict cytotoxicity.[142,143] Kohonen self-organizing maps (SOM) analysis is able to group samples into two-dimensional maps according to similar features. An application of SOM to discriminate 55 sesquiterpene lactones into classes of cytotoxic activity has been shown by Fernandes et al.[144]

Interestingly, a classification approach has been applied to the prediction of genotoxicity of thiophene derivatives: the selection of a threshold is needed to discriminate between genotoxic and nongenotoxic compounds.[145] Then, grid search SVM has been combined to seven molecular descriptors to derive the classification model, achieving the percentages (%) of prediction accuracy of 92.9%, 92.6%, and 90.9% on the training set, test set and external test set, respectively.

Recently, the hepatocarcinogenic toxicity was investigated by developing a SVM regression model in combination with the feature selection procedure applied to 55 chemicals.[146] The prediction results, corresponding to a value of correlation coefficient r^2 of 0.919, have confirmed the SVM technique as a reliable methodology to infer the potential hepatocarcinogenicity of new chemicals. However, the reliability of this model could be improved by integrating new compounds in the training set.

In the toxicity assessment process, the grouping of compounds according to structural similarity may represent a valid approach to generate reliable QSAR local models. In this way, the predictive performances of local regressions models can improve the statistical quality of global models built from diverse data sets. For this purpose, Yuan et al. have compared the local with the global approach in the prediction of acute aquatic toxicity of two different data sets.[147] In a first step, they performed a clustering analysis for the training sets and the classification of both validation and test sets by using *k*-NN technique in combination with 169 molecular descriptors. Then, a regression analysis by using linear PLS was carried out. The results obtained in the following validation step suggest the extension of this scheme for the prediction of other toxicological and pharmacodynamic properties.

Despite the modeling examples reported here, most of the literature have considered the traditional linear techniques to predict toxicological properties. Moreover, the regression-based approach is still more widely applied than the classification one with the aim to achieve a precise quantitative estimation of the toxicological data.

7. CONCLUSIONS AND PERSPECTIVES

In this article, we have focused on the development of QSAR models by machine learning methods as an attractive and helpful strategy in drug discovery. The promising predictive power underlined by the reported case studies in the field of pharmacodynamics, pharmacokinetics, and toxicology has been obtained by combining several types of molecular descriptors with powerful nonlinear techniques, to properly describe the relationship existing between the structural features and the desired property. Especially SVM has demonstrated the good prediction and generalization capabilities in a large number of regression and classification problems. The classification-based approaches aim at predicting classes of activity (high and low) or mechanisms of activity, while the regression-based methods offer the precise prediction of activity data. Further algorithms and filtering approaches are being developing to improve the results in the virtual screening process. Regarding the prediction of the toxicological profile of chemicals, in silico approaches are more and more applied as alternative methods to animal testing to refine and reduce animal experiments. Consequently, in computational toxicology investigations, the QSAR models should be in agreement with the recent regulatory system for the evaluation of the environmental hazards.

ACKNOWLEDGMENT

The molecular modeling work coordinated by S.M. was carried out with financial support from the University of Padova, Italy, and the Italian Ministry for University and Research (MIUR), Rome, Italy. We thank the Molecular Networks GmbH (Erlangen, Germany; http://www.molecular-networks.com) for the assistance in using the ct-SVM classification method and Adriana modeling suite. S.M. is also very grateful to the Chemical Computing Group for the scientific and technical partnership.

REFERENCES AND NOTES

(1) Kaitin, K. I. Obstacles and Opportunities in New Drug Development. *Nat. Clin. Pharm. Ther.* **2008**, *83*, 210–212.
(2) Tufts Center for the Study of Drug Development. Optimizing Protocol Design Strategies to Improve Clinical Research Performance Tufts University School of Medicine, Boston, 2008. http://csdd.tufts.edu/reports/description/rd_single_issues (accessed November 20, 2009).
(3) Kola, I.; Landis, J. Can the Pharmaceutical Industry Reduce the Attrition Rates. *Nat. Rev. Drug Discovery* **2004**, *3*, 711–715.
(4) Hutter, M. C. In Silico Prediction of Drug Properties. *Curr. Med. Chem.* **2009**, *16*, 189–202.
(5) Chadwick, A.; Hajek, M. Learning to Improve the Decision-Making Process in Research. *Drug Discovery Today* **2004**, *9*, 251–257.
(6) Li, H.; Yap, C. W.; Xue, Y.; Li, Z. R.; Ung, C. Y.; Han, L. Y.; Chen, Y. Z. Statistical Learning Approach for Predicting Pharmacodynamic, Pharmacokinetic, or Toxicological Properties of Pharmaceutical Agents. *Drug Dev. Res.* **2006**, *66*, 245–259.
(7) Duch, W.; Swaminathan, K.; Meller, J. Artificial Intelligence Approaches for Rational Drug Design and Discovery. *Curr. Pharm. Des.* **2007**, *13*, 1497–1508.
(8) Yap, C. W.; Li, H.; Ji, Z. L.; Chen, Y. Z. Regression Methods for Developing QSAR and QSPR Models to Predict Compounds of Specific Pharmacodynamic, Pharmacokinetic and Toxicological Properties. *Mini-Rev. Med. Chem.* **2007**, *7*, 1097–1107.
(9) Mager, D. E. Quantitative Structure-Pharmacokinetic/Pharmacodynamic Relationships. *Adv. Drug Delivery Rev.* **2006**, *58*, 1326–1356.
(10) Khan, M. T.; Sylte, I. Predictive QSAR Modeling for the Successful Predictions of the ADMET Properties of Candidate Drug Molecules. *Curr. Drug Discovery Technol.* **2007**, *4*, 141–149.
(11) Yap, C. W.; Xue, Y.; Li, H.; Li, Z. R.; Ung, C. Y.; Han, L. Y.; Zheng, C. J.; Cao, Z. W.; Chen, Y. Z. Prediction of Compounds with Specific Pharmacodynamic, Pharmacokinetic or Toxicological Property by Statistical Learning Methods. *Mini-Rev. Med. Chem.* **2006**, *6*, 449–459.
(12) European Union. Corrigendum to Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorization and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC (OJ L 396, 30.12.2006). *Off. J. Eur. Union* **2007**, *L136*, 50.
(13) Registration, Evaluation, Authorization and Restriction of Chemicals (REACH). http://ecb.jrc.it/reach/reach-legislation/ (accessed September 20, 2009).
(14) Collins, F. S.; Gray, G. M.; Bucher, J. R. Transforming Environmental Health Protection. *Science Tox.* **2009**, *319*, 906–907.
(15) Kavlock, R. J.; Ankley, G.; Blancato, B.; Breen, M.; Conolly, R.; Dix, D.; Houck, K.; Hubal, E.; Judson, R.; Rabinowitz, J.; Richard, A.; Setzer, R. W.; Shah, I.; Villeneuve, D.; Weber, E. Computational Toxicology—A State of the Science Mini Review. *Toxicol. Sci.* **2008**, *103*, 14–27.
(16) Nigsch, F.; Macaluso, N. J. M.; Mitchell, J. B.; Zmuidinavicius, D. Computational Toxicology: An Overview of the Sources of Data and of Modelling Methods. *Exp. Opin. Drug Metab. Toxicol.* **2009**, *5*, 1–14.
(17) Valerio, L. G., Jr. In Silico Toxicology for the Pharmaceutical Sciences. *Toxicol. Appl. Pharmacol.* **2009**, *241*, 356–370.
(18) Schaafsma, G.; Kroese, E. D.; Tielemanns, E. L.; Van de Sandt, J. J.; Van Leeuwen, C. J. REACH, Non-Testing Approaches and the Urgent Need for a Change in Mind Set. *Regul. Toxicol. Pharmacol.* **2009**, *53*, 70–80.
(19) Bradbury, S. P.; Feijtel, T. C.; Van Leeuwen, C. J. Meeting the Scientific Needs of Ecological Risk Assessment in a Regulatory Context. *Environ. Sci. Technol.* **2004**, *38*, 463–470.
(20) Nendza, M.; Wenzel, A. Discriminating Toxicant Classes by Mode of Action. 1.(Eco)toxicity Profiles. *Environ. Sci. Pollut. Res.* **2006**, *13*, 192–203.
(21) Benfenati, E. Predicting Toxicity through Computers: a Changing World. *Chem. Cent. J.* **2007**, *32*, 1–7.
(22) Muster, W.; Breidenbach, A.; Fischer, H.; Kirchner, S.; Müller, L.; Pähler, A. Computational Toxicology in Drug Development. *Drug Discovery Today* **2008**, *13*, 303–310.
(23) Ma, X. H.; Wang, R.; Xue, Y.; Li, Z. R.; Yang, S. Y.; Wei, Y. Q.; Chen, Y. Z. Advances in Machine Learning Prediction of Toxicological Properties and Adverse Drug Reactions of Pharmaceutical Agents. *Curr. Drug Saf.* **2008**, *3*, 100–114.
(24) Helma, C. In Silico Predictive Toxicology: the State-of-the-art Strategies to Predict Human Health Effects. *Curr. Opin. Drug Discovery* **2005**, *8*, 27–31.
(25) Fent, K.; Weston, A. A.; Caminada, D. Ecotoxicology of Human Pharmaceuticals. *Aquat. Toxicol.* **2006**, *76*, 122–159.
(26) Papa, E.; Villa, F.; Gramatica, P. Statistically Validated QSARs, based on Theoretical Descriptors, for Modeling Aquatic Toxicity of Organic Chemicals in Phimephales Promelas (Fathead Minnow). *J. Chem. Inf. Model.* **2005**, *45*, 1256–1266.
(27) Mazzatorta, P.; Smiesko, M.; Lo Piparo, E.; Benfenati, E. QSAR Models fore Predicting Pesticide Aquatic Toxicity. *J. Chem. Inf. Model.* **2006**, *45*, 1767–1774.
(28) Netzeva, T. I.; Pavan, M.; Worth, A. P. Review of (Quantitative) Structure−Activity Relationships for acute Aquatic Toxicity. *QSAR Comb. Sci.* **2008**, *27*, 77–90.
(29) Castillo-Garit, J. A.; Marrero-Ponce, Y.; Escobar, J.; Torrens, F.; Rotondo, R. A Novel Approach to Predict Aquatic from Molecular Structure. *Chemosphere* **2008**, *73*, 415–427.
(30) Zhu, H.; Tropsha, A.; Fourches, D.; Varnek, A.; Papa, E.; Gramatica, P.; Oberg, T.; Dao, P.; Cherkasov, A.; Tetko, I. V. Combinatorial QSAR Modeling of Chemical Toxicants tested against Tetrahymena Pyriformis. *J. Chem. Inf. Model.* **2008**, *48*, 766–784.
(31) Burbidge, R.; Trotter, M.; Buxton, B.; Holdenm, S. Drug Design by Machine Learning: Support Vector Machines for Pharmaceutical Data Analysis. *Comput. Chem.* **2001**, *26*, 5–14.
(32) Winkler, D. A. Neural Networks as Robust Tools in Drug Lead Discovery and Development. *Mol. Biotechnol.* **2004**, *27*, 139–167.
(33) Benigni, R.; Brossa, C. Predictivity of QSAR. *J. Chem. Inf. Model.* **2008**, *48*, 971–980.
(34) OECD Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure−Activity Relationship Models. http://www.oecd.org/dataoecd/33/37/37849783.pdf (accessed September 20, 2009).
(35) Gramatica, P. Principles of QSAR Models Validation: Internal and External. *QSAR Comb. Sci.* **2007**, *26*, 694–701.
(36) Zvinavashe, E.; Murk, A. J.; Rietjens, I. M. C. M. Promises and Pitfalls of Quantitative Structure−Activity Relationship Approaches for Predicting Metabolism and Toxicity. *Chem. Res. Toxicol.* **2008**, *21*, 2229–2236.
(37) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Wiley-VCH: Weinheim, Germany, 2009; Vol. II.
(38) Molecular Descriptors. http://www.moleculardescriptors.eu/softwares/softwares.htm. (accessed September 20, 2009).
(39) Wold, H. *Research Papers in Statistics*; Wiley: New York, 1966; pp 411−444.
(40) Wold, H. *Partial Least Squares*; Wiley: New York, 1985; Vol. 6, pp 581−591.
(41) Jores-Kong, H.; Wold, H. *Systems under Indirect Observation: Causality, Structure, Prediction*; North-Holland: Amsterdam, The Netherlands, 1982; Vol. 2, pp 1−54.
(42) Liu, P.; Long, W. Current Mathematical Methods used in QSAR/QSPR Studies. *Int. J. Mol. Sci.* **2009**, *10*, 1978–1998.
(43) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*; Wiley-VHC: Weinheim, Germany, 1999; pp 9−154.
(44) Vapnik, V. *the Nature of Statistical Learning Theory*; Springer: New York, 1995.
(45) Vapnik, V. *Statistical Learning Theory*; Wiley-WHC: New York, 1998.
(46) Burges, C. J. C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Min. Knowl. Discovery* **1998**, *2*, 121–167.
(47) Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines*; Cambridge University Press: Cambridge, U.K., 2000; pp 93−121.
(48) Czermiński, R.; Tyasri, A.; Hartsough, D. Use of Support Vector Machine in Pattern Classification: Application to QSAR Studies. *QSAR* **2001**, *20*, 227–240.
(49) Warmuth, M. K.; Liao, J.; Rätsch, G.; Mathieson, M.; Putta, S. Active Learning with Support Vector Machine in the Drug Discovery Process. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 667–673.
(50) Jorissen, R. N.; Gilson, M. K. Virtual Screening of Molecular Databases Using a Support Vector Machine. *J. Chem. Inf. Model.* **2005**, *45*, 549–561.

(51) Norinder, U. Support Vector Machine Models in Drug Design: Applications to Drug Transport Processes and QSAR Using Simplex Optimisations and Variable Selection. *Neurocomputing* **2003**, *55*, 337–346.
(52) Xue, C. X.; Zhang, R. S.; Liu, H. X.; Yao, X. J.; Liu, M. C. QSAR Models for the Prediction of Binding Affinities to Human Serum Albumin Using the Heuristic Method and a Support Vector Machine. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1693–1700.
(53) Quinlan, J. R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann: San Mateo, CA, 1993.
(54) Mitchell, T. Decision Tree Learning. In *Machine Learning*; The McGraw-Hill Companies: New York, 1997; pp 52−78.
(55) Winston, P. Learning by building Identification Trees. In *Artificial Intelligence*; Addison-Wesley: London, U.K., 1992; pp 423−442.
(56) Myles, A. J.; Feudale, R. N.; Liu, Y.; Woody, N. A.; Brown, S. D. An Introduction to Decision Tree Modeling. *J. Chemom.* **2004**, *18*, 275–285.
(57) Breiman, L. Random Forests. *Machine Learning* **2001**, *45* (1), 5–32.
(58) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.
(59) Zheng, F.; Zheng, G.; Deaciuc, A. G.; Zhan, C.-G.; Dwoskin, L. P.; Crooks, P. A. Computational Neural Network Analysis of the Affinity of Lobeline and Tetrabenazine Analogs for the Vesicular Monoamine Transporter-2. *Bioorg. Med. Chem.* **2007**, *15*, 2975–2992.
(60) Worachartcheewan, A.; Nantasenamat, C.; Naenna, T.; Isarankura-Na-Ayudhya, C.; Prachayasittikul, V. Modeling the Activity of Furin Inhibitors Using Artificial Neural Network. *Eur. J. Med. Chem.* **2008**, *44*, 1664–1673.
(61) Mandal, A. S.; Roy, K. Predictive QSAR Modeling of HIV Reverse Transcriptase Inhibitor TIBO Derivatives. *Eur. J. Med. Chem.* **2009**, *44*, 1509–1524.
(62) Fjell, C. D.; Jenssen, H.; Hilpert, K.; Cheung, W. A.; Panté, N.; Hancock, R. E. W.; Cherkasov, A. Identification of Novel Antibacterial Peptides by Chemoinformatics and Machine Learning. *J. Med. Chem.* **2009**, *52*, 2006–2015.
(63) Jalali-Heravi, M.; Kyani, A. Application of Genetic Algorithm-Kernel Partial Least Square as a Novel Nonlinear Feature Selection Method: Activity of Carbonic Anhydrase II Inhibitors. *Eur. J. Med. Chem.* **2007**, *42*, 649–659.
(64) Fatemi, M. H.; Gharaghani, S. A Novel QSAR Model for Prediction of Apoptosis-Inducing Activity of 4-aryl-4-H-chromenes Based on Support Vector Machine. *Bioorg. Med. Chem.* **2007**, *15*, 7746–7754.
(65) Goodarzi, M.; Freitas, M. P.; Jensen, R. Feature Selection and Linear/Nonlinear Regression Methods for the Accurate Prediction of Glycogen Synthase Kinase-3 Inhibitory Activities. *J. Chem. Inf. Model.* **2009**, *49*, 824–832.
(66) Tang, H.; Wang, X. S.; Huang, X.-P.; Roth, B. L.; Butler, K. V.; Kozikowski, A. P.; Jung, M.; Tropsha, A. Novel Inhibitors of Human Histone Deacetylase (HDAC) Identified by QSAR Modeling of Known Inhibitors, Virtual Screening, and Experimental Validation. *J. Chem. Inf. Model.* **2009**, *49*, 461–476.
(67) Li, J.; Liu, H.; Yao, X.; Liu, M.; Hu, Z.; Fan, B. Quantitative Structure−Activity Relationship Study of Acyl Ureas as Inhibitors of Human Liver Glycogen Phosphorylase Using Least Squares Support Vector Machines. *Chemom. Intell. Lab. Syst.* **2007**, *87*, 139–146.
(68) Qin, S.; Liu, H.; Wang, J.; Yao, X.; Liu, M.; Hu, Z.; Fan, B. Quantitative Structure−Activity Relationship Study on a Series of Novel Ligands Binding to Central Benzodiazepine Receptor by Using the Combination of Heuristic Method and Support Vector Machines. *QSAR Comb. Sci.* **2007**, *26*, 443–451.
(69) Xia, B.; Ma, W.; Zheng, B.; Zhang, X.; Fan, B. Quantitative Structure−Activity Relationship Studies of a Series of Non-Benzodiazepine Structural Ligands Binding to Benzodiazepine Receptor. *Eur. J. Med. Chem.* **2008**, *43*, 1489–1498.
(70) Yuan, Y.; Zhang, R.; Hu, R.; Ruan, X. Prediction of CCR5 Receptor Binding Affinity of Substituted 1-(3,3-Diphenylpropyl)-piperidinyl Amides and Ureas Based on the Heuristic Method, Support Vector Machine and Projection Pursuit Regression. *Eur. J. Med. Chem.* **2009**, *44*, 25–34.
(71) Hu, R.; Doucet, J.-P.; Delamar, M.; Zhang, R. QSAR Models for 2-Amino-6-arylsulfonylbenzonitriles and Congeners HIV-1 Reverse Transcriptase Inhibitors Based on Linear and Nonlinear Regression Methods. *Eur. J. Med. Chem.* **2009**, *44*, 2158–2171.
(72) Hernández, N.; Kiralj, R.; Ferreira, M. M. C.; Talavera, I. Critical Comparative Analysis, Validation and Interpretation of SVM and PLS Regression Models in a QSAR Study on HIV-1 Protease Inhibitors. *Chemom. Intell. Lab. Syst.* **2009**, *98*, 65–77.
(73) Gunturi, S. B.; Archana, K.; Khandelwal, A.; Narayanan, R. Prediction of hERG Potassium Channel Blockade Using kNN-QSAR and Local Lazy Regression Methods. *QSAR Comb. Sci.* **2008**, *27*, 1305–1317.
(74) Chen, H.-F. Computational Study of Histamine H3-Receptor Antagonist with Support Vector Machines and Three Dimension Quantitative Structure Activity Relationship Methods. *Anal. Chim. Acta* **2008**, *624*, 203–209.
(75) Goodarzi, M.; Duchowicz, P. R.; Wu, C. H.; Fernández, F. M.; Castro, E. A. New Hybrid Genetic Based Support Vector Regression as QSAR Approach for Analyzing Flavonoids-GABA(A) Complexes. *J. Chem. Inf. Model.* **2009**, *49*, 1475–1485.
(76) Tang, L.-J.; Zhou, Y.-P.; Jiang, J.-H.; Zou, H.-Y.; Wu, H.-L.; Shen, G.-L.; Yu, R.-Q. Radial Basis Function Network-Based Transform for a Nonlinear Support Vector Machine as Optimized by a Particle Swarm Optimization Algorithm with Application to QSAR Studies. *J. Chem. Inf. Model.* **2007**, *47*, 1438–1445.
(77) Si, H.; Yuan, S.; Zhang, K.; Fu, A.; Duan, Y.-B.; Hu, Z. Quantitative Structure−Activity Relationship Study on EC50 of Anti-HIV Drugs. *Chem. Intel. Lab. Systems* **2008**, *90*, 15–24.
(78) Si, H.; Lian, N.; Yuan, S.; Fu, A.; Duan, Y.-B.; Zhang, K.; Yao, X. Predicting the Activity of Drugs for a Group of Imidazopyridine Anticoccidial Compounds. *Eur. J. Med. Chem.* **2009**, *44*, 4044–4050.
(79) Wang, X. S.; Tang, H; Golbraikh, A.; Tropsha, A. Combinatorial QSAR Modeling of Specificity and Subtype Selectivity of Ligands Binding to Serotonin Receptors 5HT1E and 5HT1F. *J. Chem. Inf. Model* **2008**, *48*, 997–1013.
(80) Michielan, L.; Bolcato, C.; Federico, S.; Cacciari, B.; Bacilieri, M.; Klotz, K. N.; Kachler, S.; Pastorin, G.; Cardin, R.; Sperduti, A.; Spalluto, G.; Moro, S. Combining Selectivity and Affinity Predictions Using an Integrated Support Vector Machine (SVM) Approach: An Alternative Tool to Discriminate Between the Human Adenosine A_{2A} and A_3 Receptor Pyrazolo-triazolo-pyrimidine Antagonists Binding Sites. *Bioorg. Med. Chem.* **2009**, *17*, 5259–5274.
(81) Dong, X.; Jiang, C.; Hu, H.; Yan, J.; Chen, J.; Hu, Y. QSAR Study of Akt/Protein Kinase B (PKB) Inhibitors Using Support Vector Machine. *Eur. J. Med. Chem.* **2009**, *44*, 4090–4097.
(82) Sun, M.; Chen, J.; Wei, H.; Yin, S.; Yang, Y.; Ji, M. Quantitative Structure−Activity Relationship and Classification Analysis of Diaryl Ureas Against Vascular Endothelial Growth Factor Receptor-2 Kinase Using Linear and Non-Linear Models. *Chem. Biol. Drug Des.* **2009**, *73*, 644–654.
(83) Debeljak, Z.; Skrbo, A.; Jasprica, I.; Mornar, A.; Plecko, V.; Banjanac, M.; Medić-Sarić, M. QSAR Study of Antimicrobial Activity of Some 3-Nitrocoumarins and Related Compounds. *J. Chem. Inf. Model.* **2007**, *47*, 918–926.
(84) Bruce, C. L.; Melville, J. L.; Pickett, S. D.; Hirst, J. D. Contemporary QSAR Classifiers Compared. *J. Chem. Inf. Model.* **2007**, *47*, 219–227.
(85) Li, J.; Liu, H.; Yao, X.; Liu, M.; Hu, Z.; Fan, B. Structure−Activity Relationship Study of Oxindole-Based Inhibitors of Cyclin-Dependent Kinases Based on Least-Squares Support Vector Machines. *Anal. Chim. Acta* **2007**, *581*, 333–342.
(86) Lin, H. H.; Han, L. Y.; Yap, C. W.; Xue, Y.; Liu, X. H.; Zhu, F.; Chen, Y. Z. Prediction of Factor Xa Inhibitors by Machine Learning Methods. *J. Mol. Graphics Modell.* **2007**, *26*, 505–518.
(87) Liu, H.; Papa, E.; Walker, J. D.; Gramatica, P. In Silico Screening of Estrogen-Like Chemicals Based on Different Nonlinear Classification Models. *J. Mol. Graphics Modell.* **2007**, *26*, 135–144.
(88) Dong, X.; Liu, Y.; Yan, J.; Jiang, C.; Chen, J.; Liu, T.; Hu, Y. Identification of SVM-Based Classification Model, Synthesis and Evaluation of Prenylated Flavonoids as Vasorelaxant Agents. *Bioorg. Med. Chem.* **2008**, *16*, 8151–8160.
(89) Luan, F.; Liu, H. T.; Ma, W. P.; Fan, B. T. Classification of Estrogen Receptor-β Ligands on the Basis of their Binding Affinities Using Support Vector Machine and Linear Discriminant Analysis. *Eur. J. Med. Chem.* **2008**, *43*, 43–52.
(90) Yang, X.-G.; Chen, D.; Wang, M.; Xue, Y.; Chen, Y. Z. Prediction of Antibacterial Compounds by Machine Learning Approaches. *J. Comput. Chem.* **2009**, *30*, 1202–1211.
(91) Yuan, Y.; Zhang, R.; Luo, L. Classification Study of Novel Piperazines as Antagonists for the Melanocortin-4 Receptor Based on Least-Squares Support Vector Machines. *Chemom. Intell. Lab. Syst.* **2009**, *96*, 144–148.
(92) Cong, Y.; Yang, X.-G.; Lv, W.; Xue, Y. Prediction of Novel and Selective TNF-α Converting Enzyme (TACE) Inhibitors and Characterization of Correlative Molecular Descriptors by Machine Learning Approaches. *J. Mol. Graphics Modell.* **2009**, *28*, 236–244.
(93) Khandelwal, A.; Krasowski, M. D.; Reschly, E. J.; Sinz, M. W.; Swaan, P. W.; Ekins, S. Machine Learning Methods and Docking for Predicting Human Pregnane X Receptor Activation. *Chem. Res. Toxicol.* **2008**, *21*, 1457–1467.
(94) Chekmarev, D. S.; Kholodovych, V.; Balakin, K. V.; Ivanenkov, Y.; Ekins, S.; Welsh, W. J. Shape Signatures: New Descriptors for Predicting Cardiotoxicity In Silico. *Chem. Res. Toxicol.* **2008**, *21*, 1304–1314.

(95) Kortagere, S.; Chekmarev, D.; Welsh, W. J.; Ekins, S. Hybrid Scoring and Classification Approaches to Predict Human Pregnane X Receptor Activators. *Pharm. Res.* **2009**, *26*, 1001–1011.
(96) Kawai, K.; Fujishima, S.; Takahashi, Y. Predictive Activity Profiling of Drugs by Topological-Fragment-Spectra-Based Support Vector Machines. *J. Chem. Inf. Model.* **2008**, *48*, 1152–1160.
(97) Michielan, L.; Terfloth, L.; Gasteiger, J.; Moro, S. Exploring Potency and Selectivity Receptor Antagonist Profiles using a Multilabel Classification Approach: the Human Adenosine Receptors as a Key Study. *J. Chem. Inf. Model.* **2009**, *49*, 2820–2836.
(98) Wagner, S.; Arce, R.; Murillo, R.; Terfloth, L.; Gasteiger, J.; Merfort, I. Neural Networks as Valuable Tools To Differentiate between Sesquiterpene Lactones' Inhibitory Activity on Serotonin Release and on NF-κB. *J. Med. Chem.* **2008**, *51*, 1324–1332.
(99) Zhang, H.; Xiang, M.-L.; Zhao, Y.-L.; Wei, Y.-Q.; Yang, S.-Y. Support Vector Machine and Pharmacophore-based Prediction Models of Multidrug-Resistance Protein 2 (MRP2) Inhibitors. *Eur. J. Pharm. Sci.* **2009**, *36*, 451–457.
(100) Armutlu, P.; Ozdemir, M. E.; Uney-Yuksektepe, F.; Kavakli, I. H.; Turkay, M. Classification of Drug Molecules Considering their IC_{50} Values Using Mixed-Integer Linear Programming Based Hyper-Boxes Method. *BMC Bioinf.* **2008**, *9*, 411.
(101) Wassermann, A. M.; Geppert, H.; Bajorath, J. Searching for Target-Selective Compounds Using Different Combinations of Multiclass Support Vector Machine Ranking Methods, Kernel Functions, And Fingerprint Descriptors. *J. Chem. Inf. Model.* **2009**, *49*, 582.
(102) Wassermann, A. M.; Geppert, H.; Bajorath, J. Ligand Prediction for Orphan Targets Using Support Vector Machines and Various Target-Ligand Kernels is Dominated by Nearest Neighbor Effects. *J. Chem. Inf. Model.* **2009**, *49*, 2155–2167.
(103) Neugebauer, A.; Hartmann, R. W.; Klein, C. D. Prediction of Protein−Protein Interaction Inhibitors by Chemoinformatics and Machine Learning Methods. *J. Med. Chem.* **2007**, *50*, 4665–4668.
(104) Schneider, N.; Jäckels, C.; Andres, C.; Hutter, M. C. Gradual in Silico Filtering for Druglike Substances. *J. Chem. Inf. Model.* **2008**, *48*, 613–628.
(105) Liew, C. Y.; Ma, X. H.; Liu, X.; Yap, C. W. SVM Model for Virtual Screening of Lck Inhibitors. *J. Chem. Inf. Model.* **2009**, *49*, 877–885.
(106) Liu, X. H.; Ma, X. H.; Tan, C. Y.; Jiang, Y. Y.; Go, M. L.; Low, B. C.; Chen, Y. Z. Virtual Screening of Abl Inhibitors from Large Compound Libraries by Support Vector Machines. *J. Chem. Inf. Model.* **2009**, *49*, 2101–2110.
(107) Tang, W.; Lu, A. Y. Drug Metabolism and Pharmacokinetics in Support of Drug Design. *Curr. Pharm. Des* **2009**, *15*, 2170–2183.
(108) Korfmacher, W. A. Advances in the Integration of Drug Metabolism into the Lead Optimization Paradigm. *Mini-Rev. Med. Chem.* **2009**, *9*, 703–716.
(109) Wang, J.; Hou, T. Recent Advances on in Silico ADME Modeling. *Ann. Rep. Comput. Chem.* **2009**, *5*, 101–127.
(110) Norinder, U.; Bergström, C. A. S. Prediction of ADMET Properties. *Chem. Med. Chem.* **2006**, *1*, 920–937.
(111) Ruiz-Garcia, A.; Bermejo, M.; Moss, A.; Casabo, G. V. Pharmacokinetics in Drug Discovery. *J. Pharm. Sci.* **2006**, *97*, 654–690.
(112) Hou, T.; Wang, J.; Zhang, W.; Wang, W.; Xu, X. Recent Advances in Computational Prediction of Drug Absorption and Permeability in Drug Discovery. *Curr. Med. Chem.* **2006**, *13*, 2653–2667.
(113) Czodrowski, P.; Kriegl, J. M.; Scheuerer, S.; Fox, T. Computational Approaches to Predict Drug Metabolism. *Exp. Opin. Drug Metab. Toxicol.* **2009**, *5*, 15–27.
(114) Madden, J. C.; Cronin, M. T. D. Structure-Based Methods for the Prediction of Drug Metabolism. *Exp. Opin. Drug Metab. Toxicol.* **2006**, *2*, 545–557.
(115) Jolivette, L. J.; Ekins, S. Methods for Predicting Human Drug Metabolism. *Adv. Clin. Chem.* **2007**, *43*, 131–176.
(116) Fox, T.; Kriegl, J. M. Machine Learning Techniques for In Silico Modeling of Drug Metabolism. *Curr. Top. Med. Chem.* **2006**, *6*, 1579–1591.
(117) Arimoto, R. Computational Models for Predicting Interactions with Cytochrome P450 Enzyme. *Curr. Top. Med. Chem.* **2006**, *6*, 1609–1618.
(118) Yap, C. W.; Xue, Y.; Chen, Y. Z. Application of Support Vector Machines to In Silico Prediction of Cytochrome P450 Enzyme Substrates and Inhibitors. *Curr. Top. Med. Chem.* **2006**, *6*, 1593–1607.
(119) Chohan, K. K.; Paine, S. W.; Waters, N. J. Quantitative Structure−Activity Relationships in Drug Metabolism. *Curr. Top. Med. Chem.* **2006**, *6*, 1569–1578.
(120) Li, H.; Sun, J.; Fan, S.; Sui, X.; Zhang, L.; Wang, Y.; He, Z. Considerations and Recent Advances in QSAR Models for Cytochrome P450 Mediated Drug Metabolism Prediction. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 843–855.
(121) Crivori, P.; Poggesi, I. Computational Approaches for Predicting CYP-related Metabolism Properties in the Screening of New Drugs. *Eur. J. Med. Chem.* **2006**, *41*, 795–808.
(122) Mehdipour, A. R.; Hamidi, M. Brain Drug Targeting: a Computational Approach for Overcoming Blood-Brain Barrier. *Drug Discovery Today* **2009**, *14*, 1030–1036.
(123) Hou, T.; Wang, J.; Zhang, W.; Xu, X. ADME Evaluation in Drug Discovery. 7. Prediction of Oral Absorption by Correlation and Classification. *J. Chem. Inf. Model.* **2007**, *47*, 208–218.
(124) Hou, T. ADME Evaluation in Drug Discovery. 8. The Prediction of Human Intestinal Absorption by a Support Vector Machine. *J. Chem. Inf. Model.* **2007**, *47*, 2408–2415.
(125) Wan, H.; Rehegren, M.; Giordanetto, F.; Bergström, F.; Tunek, A. High-Throughput Screening of Drug-Brain Tissue Binding and in Silico Prediction for Assessment of Central Nervous System Drug Delivery. *J. Med. Chem.* **2007**, *50*, 4606–4615.
(126) Sui, X.; Sun, J.; Wu, X.; Li, H.; Liu, J.; He, Z. Predicting the Volume of Distribution of Drugs in Humans. *Curr. Drug Metab.* **2008**, *9*, 574–580.
(127) Berellini, G.; Springer, C.; Waters, N. J.; Lombardo, F. In Silico Prediction of Volume of Distribution in Human Using Linear and Nonlinear Models on a 669 Compound Data Set. *J. Med. Chem.* **2009**, *52*, 4488–4495.
(128) Paixão, P.; Gouveia, L. F.; Morais, J. A. G. Prediction of Drug Distribution within Blood. *Eur. J. Pharm. Sci.* **2009**, *36*, 544–554.
(129) Sakiyama, Y.; Yuki, H.; Moriya, T.; Hattori, K.; Suzuki, M.; Shimada, K.; Honma, T. Predicting Human Liver Microsomal Stability with Machine Learning Techniques. *J. Mol. Graphics Modell.* **2008**, *26*, 907–915.
(130) Schwaighofer, A.; Schroeter, T.; Mika, S.; Hansen, k.; Ter Laak, A.; Lienau, P.; Reichel, A.; Heinrich, N.; Müller, K.-R. A Probabilistic Approach to Classifying Metabolic Stability. *J. Chem. Inf. Model.* **2008**, *48*, 785–796.
(131) Terfloth, L.; Bienfait, B.; Gasteiger, J. Ligand-based Models of the Isoform Specificity of Cytochrome P450 3A4, 2D6 and 2C9 Substrates. *J. Chem. Inf. Model.* **2007**, *47*, 1688–1701.
(132) Michielan, L.; Terfloth, L.; Gasteiger, J.; Moro, S. Comparison of Multilabel and Single-Label Classification Applied to the Prediction of the Isoform Specificity of Cytochrome P450 Substrates. *J. Chem. Inf. Model.* **2009**, *49*, 2588–2605.
(133) Hammann, F.; Gutmann, H.; Baumann, U.; Helma, C.; Drewe, J. Classification of Cytochrome P450 Activities Using Machine Learning Methods. *Mol. Pharmaceutics* **2009**, *6*, 1920–1926.
(134) Jensen, B. F.; Vind, C.; Padkjaer, S. B.; Brockoff, P. B.; Refgaard, H. H. F. In Silico Prediction of Cytochrome P450 2D6 and 3A4 Inhibition Using Gaussian Kernel Weighted k-nearest Neighbor and Extended Connectivity Fingerprints, Including Structural Fragments Analysis of Inhibitors versus Noninhibitors. *J. Med. Chem.* **2007**, *50*, 501–511.
(135) Eitrich, T.; Kless, A.; Druska, C.; Meyer, W.; Grotendorst, J. Classification of Highly Unbalanced CYP450 Data of Drugs Using Cost Sensitive Machine Learning Techniques. *J. Chem. Inf. Model.* **2007**, *47*, 92–103.
(136) Choi, I.; Kim, S. Y.; Kim, H.; Kang, N. S.; Bae, M. A.; Yoo, S.-E.; Jung, J.; No, K. T. Classification Models for CYP450 3A4 Inhibitors and Non-inhibitors. *Eur. J. Med. Chem.* **2009**, *44*, 2354–2360.
(137) Vasanthanathan, P.; Taboureau, O.; Oostenbrink, C.; Vermeulen, N. P. E.; Olsen, L.; Jørgensen, F. S. Classification of Cytochrome P450 1A2 Inhibitors and Noninhibitors by Machine Learning Techniques. *Drug Metab. Dispos.* **2009**, *37*, 658–664.
(138) Dagliyan, O.; Kavakli, I. H.; Turkay, M. Classification of Cytochrome P450 Inhibitors with Respect to Binding Free Energy and pIC_{50} Using Common Molecular Descriptors. *J. Chem. Inf. Model.* **2009**, *49*, 2403–2411.
(139) Saliner, A. G.; Patlewicz, G.; Worth, A. P. A Review of (Q)SAR Models for Skin and Eye Irritation and Corrosion. *QSAR Comb. Sci.* **2007**, *27*, 49–59.
(140) Golla, S.; Madihally, S.; Robinson, R. L.; Gasem, K. A. M. Quantitative Structure−Property Relationship Modeling of Skin Sensitization: A Quantitative Prediction. *Toxicol. In Vitro* **2009**, *23*, 454–465.
(141) Zhao, C. Y.; Zhang, H. X.; Zhang, X. Y.; Liu, M. C.; Hu, Z. D.; Fan, B. T. Application of Support Vector Machine (SVM) for Prediction Toxic Activity of Different Data Sets. *Toxicology* **2006**, *217*, 105–119.
(142) Chamjangali, M. A.; Beglari, M.; Bagherian, G. Prediction of Cytotoxicity Data (CC_{50}) of Anti-HIV 5-phenyl-1-phenylamino-1H-imidazole Derivatives by Artificial Neural Network Trained with Levenberg-Marquardt Algorithm. *J. Mol. Graphics Modell.* **2007**, *26*, 360–367.
(143) Chamjangali, M. A. Modelling of Cytotoxicity Data (CC_{50}) of Anti-HIV 1-[5-chlorophenyl)sulfonyl]-1H-pyrrole Derivatives Using Calcu-

lated Molecular Descriptors and Levenberg-Marquardt Artificial Neural Network. *Chem. Biol. Drug Des.* **2009**, *73*, 456–465.
(144) Fernandes, M. B.; Scotti, M. T.; Ferreira, M. J. P.; Emerenciano, V. P. Use of Self-Organizing Maps and Molecular Descriptors to Predict the Cytotoxic Activity of Sesquiterpene Lactones. *Eur. J. Med. Chem.* **2008**, *43*, 2197–2205.
(145) Du, H.; Wang, J.; Watzl, J.; Zhang, X. Y.; Hu, Z. Classification Structure−Activity Relationship (CSAR) Studies for Prediction of Genotoxicity of Thiophene Derivatives. *Toxicol. Lett.* **2008**, *177*, 10–19.
(146) Massarelli, I.; Imbriani, M.; Coi, A.; Saraceno, M.; Carli, N.; Bianucci, A. M. Development of QSAR Models for Predicting Hepatocarcinogenic Toxicity of Chemicals. *Eur. J. Med. Chem.* **2009**, *44*, 3658–3664.
(147) Yuan, H.; Wang, Y.; Cheng, Y. Local and Global Quantitative Structure−Activity Relationship Modeling and Prediction for the Baseline Toxicity. *J. Chem. Inf. Model.* **2007**, *47*, 159–169.

CI100072Z