# Predictive QSAR Modeling: Methods and Applications in Drug Discovery and Chemical Risk Assessment

57

Alexander Golbraikh, Xiang Simon Wang, Hao Zhu, and Alexander Tropsha

## Contents



A. Golbraikh (✉) • X.S. Wang • H. Zhu
Laboratory for Molecular Modeling and Carolina Center for Exploratory Cheminformatics Research, Division of Medicinal Chemistry and Natural Products, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, USA
e-mail: golbraik@email.unc.edu; xiang.wang@Howard.edu; haozhu@email.unc.edu

A. Tropsha
Division of Medicinal Chemistry and Natural Products, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, USA
e-mail: alex_tropsha@unc.edu



J. Leszczynski et al. (eds.), *Handbook of Computational Chemistry*,
DOI 10.1007/978-3-319-27282-5_37

**Abstract**

Quantitative structure–activity relationship (QSAR) modeling is the major chemin- formatics approach to exploring and exploiting the dependency of chemical, biological, toxicological, or other types of activities or properties on their molecular features. QSAR modeling has been traditionally used as a lead optimization approach in drug discovery research. However, in recent years QSAR modeling found broader applications in hit and lead discovery by the means of virtual screening as well as in the area of drug-like property prediction, and chemical risk assessment. These developments have been enabled by the improved protocols for model development and most importantly, model validation that focus on developing models with independently validated external prediction power. This chapter reviews the predictive QSAR modeling workflow developed in this laboratory that incorporates rigorous procedures for QSAR model development, validation, and application to virtual screening. It also provides several examples of the workflow application to the identification of experimentally confirmed hit compounds as well as to chemical toxicity modeling. We believe that methods and applications considered in this chapter will be of interest and value to researchers working in the field of computational drug discovery and environmental chemical risk assessment.

## QSAR Methodology: Summary of Approaches for Model Building and Validation

In order to find new leads in the process of drug design and discovery, there is a need for efficient and robust computational procedures that can be used to screen chemical databases and virtual libraries against molecules with known activities or properties. For this purpose, quantitative structure–activity relationship (QSAR) analysis is widely used. QSAR modeling provides an effective way for establishing and exploiting the relationship between chemical structures and their biological actions toward the development of novel drug candidates. Theoretically, QSAR analysis is the application of mathematical and statistical methods for the development of models for the prediction of biological activities or properties of compounds. Formally, a QSAR model can be expressed in the following generic format:

$$\text{Predicted Biological Activity} = \text{Function (Chemical Structure)} \quad (1)$$

A QSAR procedure tries to minimize the error of prediction, for example, in the form of the sum of squares between predicted and observed activities.

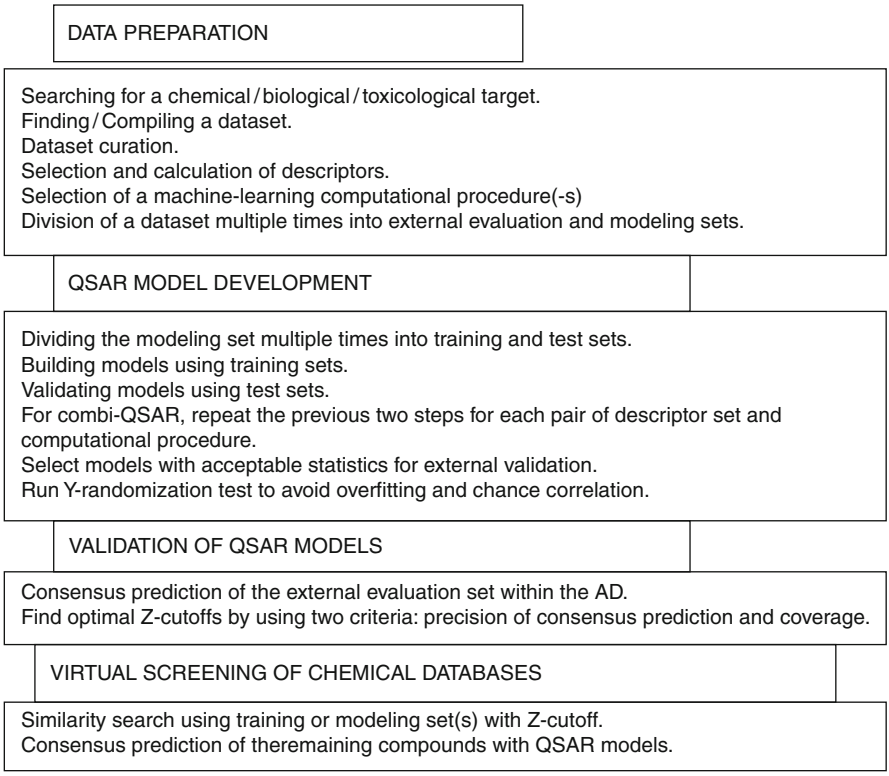


**Fig. 1** Major steps of QSAR modeling

The process of QSAR model development can be divided into three parts: data preparation, data analysis, and model validation (Fig. 1). Model validation should include establishment of model applicability domain (AD). Recently, the European Organization for Economic Co-operation and Development (OECD) developed a set of principles for the development and validation of QSAR models, which, in particular, requires "appropriate measures of goodness-of-fit, robustness, and predictivity" (Organisation 2008). The OECD guidance document especially emphasizes that QSAR models should be rigorously validated using external sets of compounds that were not used in the model development.

## Data Preparation

The first part of QSAR analysis includes selection of a molecular dataset for QSAR studies, acquiring or calculation of molecular descriptors (quantities characterizing molecular structures), and selection of a QSAR (statistical analysis and correlation) method. Datasets for QSAR studies can be found in research papers or electronic

databases available either publicly (PubChem 2010; BindingDB (Liu et al. 2007); ChEMBL 2010; DSSTox 2008; NIMH Psychoactive Drug Screening (PDSP) 2010) or commercially (e.g., Wombat (Olah et al. 2007) or MDDR 2009); more examples are given in a recent review (Oprea and Tropsha 2006). The dataset should include biological activity values for all compounds (e.g., binding energies to a receptor, or inhibition constants $IC_{50}$, or in case of toxicity modeling, lethal concentration in water $LC_{50}$, or lethal dose $LD_{50}$, etc.) preferably measured in the same lab using the same experimental method. If these experimental data are not available from one lab or one source, and the correlation between measurements made in different labs or by different methods cannot be established, they may not be used directly in QSAR studies. Instead, compounds in the dataset should be given a rank or assigned to categories of activities: for example, a compound can be very active, moderately active, or inactive. In the majority of such cases, binary classification is used, in which a compound is classified as either active or inactive. Another situation may arise, when compounds in the dataset naturally belong to different classes, for example, they are ligands to different receptors. In this case, the types of ligand specificity for a target can be considered as classes of compound activities, and the goal of QSAR analysis becomes to achieve accurate prediction of the target specificity for a new compound.

According to the nature of the activity data, QSAR studies can be divided into continuous (activities, i.e., response variable, takes many different values from within some interval), category (activities are represented by ranks or ordinal numbers), and classification (activities are different types of biological properties which cannot be rank ordered) approaches.

Prior to QSAR modeling, a dataset should be curated, that is, all structures should be verified with respect to their correct representation in the dataset; structures containing atoms, for which there are no parameters for descriptor calculation should be removed; structures consisting of several disconnected parts should be removed; salts should be removed; a problem of isomerism should be addressed; and duplicate structures should be removed. There are different tools available for dataset curation. For example, Molecular Operating Environment (MOE) (2008) includes DatabaseWash tool. It allows changing molecules' names, adding or removing hydrogen atoms, removing salts and heavy atoms, even if they are covalently connected to the rest of the molecule, and changing or generating the tautomers and protomers (cf. the MOE manual for more details). Various database curation tools are included in ChemAxon (2008) as well. If commercial software tools such as MOE are unavailable (notably, ChemAxon software is free to academic investigators), one can use standard UNIX/LINUX tools to perform some of the dataset cleaning tasks (Tropsha and Golbraikh 2010). It is important to have some freely available molecular format converters such as OpenBabel (2010) or MolConverter from ChemAxon (2008). Major procedures for database curation are discussed in our recent paper (Fourches et al. 2010).

After the dataset is selected and curated, the next task is the acquisition or calculation of descriptors. According to an excellent monograph titled Handbook of Molecular Descriptors by Roberto Todeschini and Vivian Consonni (2000)

molecular descriptors can be grouped into zero-dimensional [0D] (sometimes referred to as constitutional descriptors), one-dimensional [1D] (e.g., counts of different molecular groups, physicochemical properties of compounds, etc.), two-dimensional [2D] (invariants of molecular graphs, e.g., connectivity indices, information indices, counts of paths and walks, etc.), three-dimensional [3D], which are based on geometrical spatial properties of molecules [e.g., Comparative Molecular Field Analysis (CoMFA) descriptors (Tripos 2010) which are values of steric and electrostatic fields around aligned molecules, and different CoMFA-like descriptors (Klebe 1998; Kubinyi et al. 1998; Robinson et al. 1999)], and some other descriptors. Some descriptors can be experimental or calculated physicochemical properties of molecules such as molecular weight, molar refraction, energies of HOMO and LUMO, normal boiling point, octanol/water partition coefficient, molecular surface, molecular volume, etc.

Herein, we will not discuss different types of descriptors in detail but mention briefly major descriptor software. Most of descriptors included in the Handbook of Molecular Descriptors (Todeschini and Consonni 2000) can be calculated by the Dragon software (Dragon 2007). Molconn-Z (2007) is another widely used descriptor calculation software which calculates more than 800 descriptors. A relatively small, but diverse set of molecular descriptors can be calculated by the MOE (2008) software. Chirality molecular topological descriptors (CMTDs) developed in our laboratory append 2D descriptors by conformation-independent chirality and ZE-isomerism topological indices (Golbraikh and Tropsha 2003; Golbraikh et al. 2001, 2002). Another group of descriptors frequently used in our laboratory is atom-pair (AP) descriptors (Carhart et al. 1985). Each descriptor is defined as a count of pairs of atoms of certain types being away from each other on a certain topological distance (2D AP descriptors) or a Euclidean distance within certain intervals (3D AP descriptors); chirality AP descriptors can be calculated as well (Kovatcheva et al. 2005).

Many descriptors calculated from the knowledge of 3D structure of molecules (3D descriptors) have been developed and published as well. Although these are inherently more rigorous, one should keep in mind that their calculation is much more time and resource consuming. In many QSAR applications, the calculation of 3D descriptors should be preceded by conformational search and 3D structure alignment. However, even for rigid compounds, it is not generally known whether the alignment corresponds to real positions of molecules in the receptor binding site (Cherkasov 2008). There are different conformational analysis and pharmacophore modeling tools included in molecular modeling packages such as MOE (2008), Sybyl (there are LINUX and MS Windows versions) (Tripos 2010), Discovery Studio (2010), LigandScout (2010), etc. It has been demonstrated that in many cases QSAR models based on 2D descriptors have comparable (or even superior) predictivity than models based on 3D descriptors (Bures and Martin 1998; Golbraikh et al. 2001; Hoffman et al. 1999; Zheng and Tropsha 2000). Thus when 3D QSAR studies are necessary, if possible, 3D alignment of molecules should be preferably obtained by docking studies. VolSurf (Crivori et al. 2000; Cruciani et al. 2000) and GRIND (Pastor et al. 2000) descriptors are examples of alignment-free

3D descriptors. But their calculation still requires extensive conformational analysis of molecules. Both VolSurf and GRIND descriptors are available in Sybyl (VolSurf and Almond modules) (Tripos 2010). Various types of descriptors can be calculated by different modules of Schrodinger software (2010). Virtually, any molecular modeling software package contains sets of its own descriptors and there are many other descriptors not mentioned here that can be found in the specialized literature.

There are sets of descriptors that take values of zero or one depending on the presence or absence of certain predefined molecular features (or fragments) such as oxygen atoms, aromatic rings, rings, double bonds, triple bonds, halogens, and so on. These sets of descriptors are called molecular fingerprints or structural keys. Such descriptors can be represented by bit strings and many are found in popular software packages. For instance, several different sets of such descriptors are included in MOE (2008), Sybyl (Tripos 2010), and others, and examples of their use can be found in the published literature (McGregor and Pallai 1997; Waller 2004). Molecular holograms are similar to fingerprints; however, they use counts of features rather than their presence or absence. For example, holograms are included in the Sybyl HQSAR module (Tripos 2010). There are also more recent approaches when molecular features are not predefined a priori (as fingerprints discussed above) but are identified for each specific dataset. For example, frequent subgraph mining approaches developed independently at the University of North Carolina (Huan et al. 2006) and at the Louis Pasteur University in Strasbourg (Horvath et al. 2007) can find all frequent closed subgraphs (i.e., subgraph descriptors) for given datasets of compounds described as chemical graphs. A large and diverse set of 2D descriptors can be generated by MOLD2 software (Hong et al. 2008) available from FDA. A wide variety of descriptors are included in ADRIANA software (Gasteiger 2006).

Prior to QSAR studies, processing of descriptors is required. It includes: exclusion of descriptors having the same value for all compounds in the dataset as well as duplicate descriptors. To avoid higher influence on QSAR models of descriptors with higher variance, all descriptors are usually normalized (in most cases, range scaling or autoscaling is used). Molecular holograms or AP descriptors do not need to be normalized. Molecular field values around molecules are also not normalized. Preferably, descriptors with low variance and one of the highly correlated pair of descriptors should be excluded as well.

Finally, data for QSAR model development can be represented in a form of a table (see Table 1), in which each compound is a row and each descriptor as well as activity is a column.

## The Problem of Outliers

Success of QSAR modeling depends on the appropriate selection of a dataset for QSAR studies. In a recent editorial of the *Journal of Chemical Information and Modeling*, Maggiora (2006) noticed that one of the main deficiencies of many chemical datasets is that they do not fully satisfy the main hypothesis underlying all

**Table 1** QSAR table

| Compound | Descriptor 1 | Descriptor 2 | . . . | Descriptor N | Activity |
|---|---|---|---|---|---|
| 1 | $X_{11}$ | $X_{12}$ | . . . | $X_{1N}$ | $Y_1$ |
| 2 | $X_{21}$ | $X_{22}$ | . . . | $X_{2N}$ | $Y_2$ |
| . . . | . . . | . . . | . . . | . . . | . . . |
| M | $X_{M1}$ | $X_{M2}$ | . . . | $X_{MN}$ | $Y_M$ |

QSAR studies: Similar compoundsare expected to have similar biological activities or properties. Maggiora defines the "cliffs" in the descriptor space where the properties change so rapidly, that, in fact adding or deleting one small chemical group can lead to a dramatic change in the compound's property. In other words, small changes of descriptor values can lead to large changes in molecular properties. Generally, in this case there could be not just one outlier, but a subset of compounds properties of which are different from those on the other "side" of the cliff. In other words, cliffs are areas where the main QSAR hypothesis does not hold. So cliff detection remains a major QSAR problem that has not been adequately addressed in most of the reported studies.

There are two types of outliers we must be aware of: leverage (or structural) outliers and activity outliers. In case of activity outliers the problem of "cliffs" should be addressed as well. Recently, different approaches to find activity outliers have been published (Bajorath et al. 2009; Guha and Van Drie 2008a, b; Sisay et al. 2009). We have suggested that Grubb's (Environmental Protection Agency 1992) and Dixon's (Fallon et al. 1997) statistical tests can be used to find activity outliers (Tropsha and Golbraikh 2010). Structural outliers can be defined as compounds that are largely dissimilar to all other compounds in the descriptor space. The methods of finding them are similar to finding compounds out of QSAR model applicability domains (Tropsha and Golbraikh 2010) that is discussed below.

## QSAR Model Development

The ultimate goal of QSAR analysis is the development of validated models for accurate and precise prediction of biological activities of compounds which could be potential leads in the process of drug discovery. Eventually, predictions should be confirmed by experimental validation. The general QSAR modeling workflow is represented in Fig. 2. Following the data curation step, we start by randomly selecting a fraction of compounds (typically, 10–20 %) as an external evaluation set. The Sphere Exclusion protocol implemented in our laboratory (Golbraikh and Tropsha 2002; Golbraikh et al. 2003) is then used to rationally divide the remaining subset of compounds (the modeling set) multiple times into pairs of training and test sets that are used for model development and validation, respectively. We employ multiple QSAR techniques based on the combinatorial exploration of all possible pairs of descriptor sets and various supervised data analysis techniques

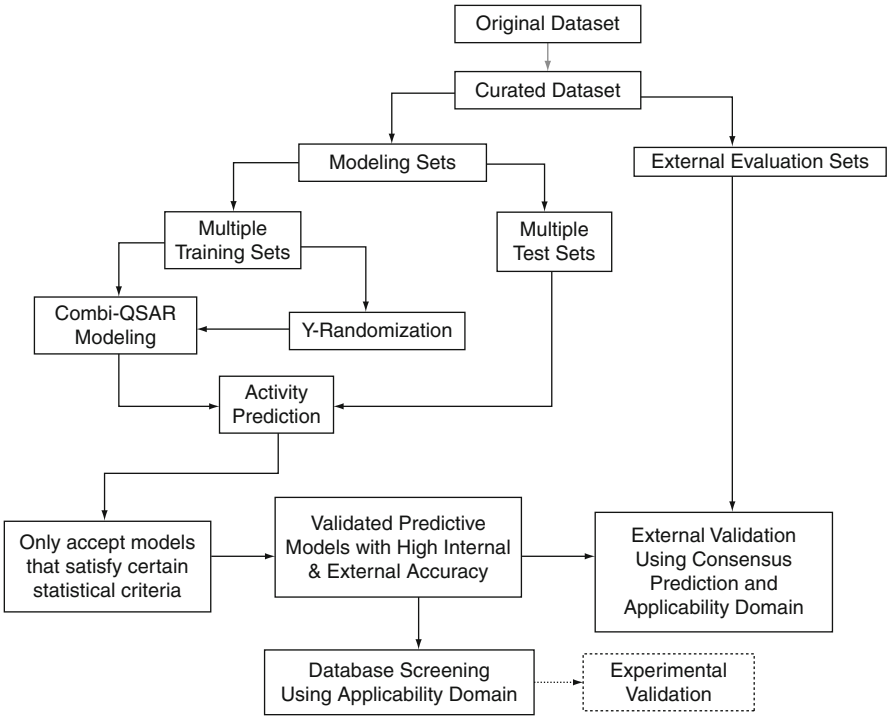


**Fig. 2** Predictive QSAR modeling workflow

(combi-QSAR) (Fig. 3) and select models characterized by high accuracy in predicting both training and test sets data. Validated models are finally tested using the external evaluation set. The critical step of the external validation is the use of applicability domains (ADs). If external validation demonstrates the significant predictive power of the models, we employ them for virtual screening of available chemical databases (e.g., ZINC (Irwin and Shoichet 2005)) to identify putative active compounds and work with collaborators who could validate such hits experimentally. The entire approach is described in detail in several recent papers and reviews (Tropsha 2005; Tropsha and Golbraikh 2007).

## QSAR Methods

QSAR modeling techniques employ various methods of multidimensional data analysis as well as supervised machine learning used in different areas of research in natural and social sciences such as biological sciences, geography, psychology, medicine, economics, signal processing, speech recognition, forensic studies, etc. Herein, it is impossible to discuss all the methods used in QSAR analysis. Instead, we will name only some of them. All these methods can be classified into linear and

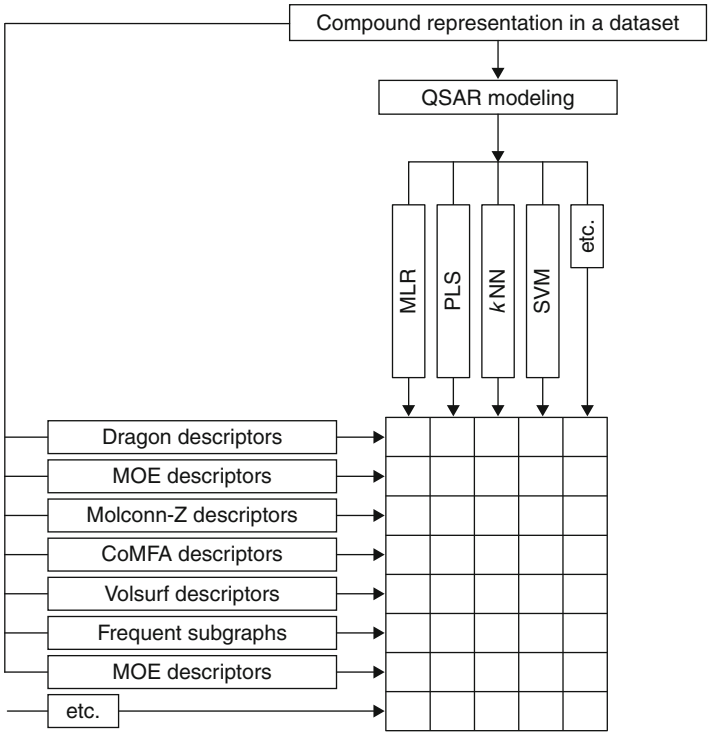


**Fig. 3** Combinatorial QSAR modeling

nonlinear approaches. Linear methods include simple and multiple linear regression (MLR), principal component regression (PCR), partial least squares (PLS), etc. The main distinctive characteristic of these methods is the linearity of the function approximating the biological activity (see Eq. 1) of their arguments (which are molecular descriptors). In linear discriminant analysis (LDA), linear combinations of descriptors are built, which define hyperplanes that separate representative points of different classes of compounds in the multidimensional descriptor space.

Nonlinear methods can be based derived from linear or based on more complex approaches that predict compound activities from their descriptors by the means of nonlinear relationships. For example, if nonlinear terms (like squares, products, or logarithms of some descriptors) are added to a linear regression, it becomes nonlinear regression. Many nonlinear methods are derived from linear methods via transforming them by a so-called kernel trick. Calculations are executed in a so-called feature space where linear methods are applied. The advantage of these methods is that there is no need to directly calculate the transformation functions. Examples of such methods include non-linear support vector machines (SVMs) and support vector regression (SVR) methods (Berk 2008; Vapnik 2000), nonlinear discriminant analysis, kernel-PCA, kernel-PLS, etc. In the multidimensional feature

space, SVM builds a soft margin hyperplane, which separates points belonging to two different classes, or more hyperplanes to separate points of larger number of classes. In contrast, SVR builds a hyperplane such that as many points as possible are within the margin. Good SVM tutorial was written by Burges (1998), and SVR tutorial by Smola and Schoelkopf (2004). Other non-linear methods include k-nearest neighbors QSAR, in which the activity of a compound is predicted as a (weighted) average of activities of its nearest neighbors. k-nearest neighbor methods can include stochastic (Zheng and Tropsha 2000) or stepwise variable (descriptor) selection (Ajmani et al. 2006).

Another large group of generally nonlinear methods are artificial neural networks (ANNs) (Neural Networks 1996; Salt et al. 2006; Zupan and Gasteiger 1999). Ensembles of ANNs can make use of bagging and boosting approaches (Agrafiotis et al. 2002). ANNs consist of groups of artificial neurons. In feed-forward back-propagation neural networks (Neural Networks 2010), neurons are organized in input, hidden, and output layers. Input layer neurons receive descriptor values of compounds, which are passed with different weights to the hidden layer neurons. A neuron activation function is then applied at each neuron to the sum of weighted inputs, and the results are passed to the output layer neurons, which calculate predicted activities of compounds. During training process, parameters of neuron functions and weights are adjusted so that the total error of predictions is minimized. There are network architectures with multiple hidden layers.

Recursive partitioning (RP) methods build decision trees in order to precisely assign compounds to their classes. The tree consists of one root node containing all objects (compounds), intermediate (or decision), and leaf (terminal) nodes. A measure of node purity is introduced; for example, it could be the ratio of counts of compounds belonging to majority and minority class in a node. At each node, the procedure tries to partition the data to increase the purity measure, that is, to make the difference between sum of child node purities and parent node purity as higher as possible. Analysis is based on descriptor value distributions between classes at the node. If such a partition at the node is impossible, it becomes a leaf node. Additional criteria may be imposed on the minimum number of compounds in a leaf node, etc. Compounds in each node satisfy certain descriptor criteria. After growing, some leaves are consecutively removed based on the improvement of classification at them (so-called pruning of a tree). Without pruning, the tree could be overfitted. Prediction process consists of moving a query compound up the tree (based on its descriptor values) until it reaches a leaf node. Predicted class of a compound is defined as that of the majority class in this node. There are also RP regression methods which are used, if response variable is continuous. There are several RP algorithms widely used such as Classification and Regression Trees (CART (Berk 2008)), C4.5 (Quinlan 1993), C5.0 (2008), etc.

Random Forest methods (Breiman 2001; Random Forests 2001) construct ensembles of trees based on multiple random selections of subsets of descriptors and bootstrapping of compounds. The compounds not selected in a particular bootstrapping are considered as a so-called out of bag set, and used as the test

set. The trees are not pruned. Best trees in the forest are chosen for consensus prediction of external compounds. The method can include bagging (Berk 2008; Breiman 1996) and boosting (Berk 2008; Breiman 1998) approaches.

## Target Functions

Based on the nature of the response variable, QSAR approaches can be grouped into classification, category, or continuous QSAR ( vide infra). Classes are different from categories in a sense that the former cannot be ordered in any scientifically meaningful way, while the latter can be rank ordered.

## Continuous QSAR Models

We suggested that the following validation criteria should be used for continuous QSAR models (Tropsha and Golbraikh 2010): (1) leave-one-out (LOO) cross-validated $q^2$ (which is also used as the target function, that is, it is optimized by the QSAR modeling procedure) (2) square of the correlation coefficient $R(R^2)$ between the predicted and observed activities of the test set; (3) coefficients of determination (predicted versus observed activities ($R_0^2$), and observed versus predicted activities ($R'^2_0$) for the test set) for regressions through the origin; (4) slopes k and k′ of regression lines through the origin (predicted versus observed activities, and observed versus predicted activities for the test set). In our studies, we consider models acceptable, if they have (1) $q^2 > 0.5$; (2) $R^2 > 0.6$; (3) $\left(R^2 - R_0^2\right)/R^2 < 0.1$ and $0.85 \leq k \leq 1.15$ or $\left(R^2 - R'^2_0\right)/R^2 < 01$ and $0.85 \leq k' \leq 1.15$; (4) $\left(\left|R_0^2 - R'^2_0\right| < 0.3\right.$. Sometimes, stricter criteria are used (Tropsha and Golbraikh 2010).

In some papers, other criteria are used. For example, sometimes standard error of prediction is used instead of (or together with) $R^2$. Standard error of prediction itself makes no sense until we compare it with the standard deviation for activities of the test set, which brings us back to the correlation coefficients. If used, mean absolute error (MAE) should be compared with the mean absolute deviation from the mean. Sometimes, F-ratio is calculated, which is the variance explained by the model divided by the unexplained variance. It is believed that the higher is the F-ratio, the better is the model. We suppose that when F-ratio is used, it must be always accompanied by the corresponding p-value.

Frequently, especially for linear models such as developed with multiple linear regression (MLR) or partial least squares (PLS) the adjusted $R^2$ is used:

$$R^2_{adj} = 1 - \left(1 - R^2\right)\frac{n-1}{n-c-1}, \tag{2}$$

where $n$ is the number of compounds in the dataset, and c is the number of variables (descriptors or principal components) included in the regression equation. It should

be recognized that $R^2_{adj} \leq R^2$. The higher the number of explanatory variables c is, the lower $R^2_{adj}$ is. $R^2_{adj}$ is particularly important for linear QSAR models developed with variable selection. $R^2_{adj}$ is not a good criterion for variable selection *k*NN QSAR models, since contrary to regression methods, in the *k*NN algorithm descriptors are just selected or not selected, that is, their weights are either zero or one. As a result, much larger set of descriptors is selected by the *k*NN procedure than, for example, by stepwise regression.

## Target Functions and Validation Criteria for Classification QSAR Models

We consider a classification QSAR model predictive, if the prediction accuracy characterized by the correct classification rate (CCR) for each class is sufficiently large:

$$\mathrm{CCR}_{\mathrm{class}} = \frac{N^{\mathrm{corr}}_{\mathrm{class}}}{N^{\mathrm{total}}_{\mathrm{class}}} \tag{3}$$

and the p-value for each $\mathrm{CCR}_{\mathrm{class}}$ value is not higher than a predefined threshold (in case of two classes, the $\mathrm{CCR}_{\mathrm{class}}$ threshold should not be lower than 0.65–0.70, and generally, for any number of classes, p-value should not be higher than 0.05 for each class).

For the classification QSAR with K classes, we shall use the following criterion

$$\mathrm{CCR} = \frac{1}{K}\sum_{i=1}^{K}\mathrm{CCR_i} = \frac{1}{K}\sum_{i=1}^{K}\frac{N^{\mathrm{corr}}_{k}}{N^{\mathrm{total}}_{k}} \tag{4}$$

along with the correct classification rate for each class (see Eq. 2). Criterion 4 is correct for both balanced and imbalanced (biased) datasets (i.e., when the number of compounds of each class is different). For imbalanced datasets, formula N(corr)/N(total), where N(corr) and N(total) are the number of compounds predicted correctly and the total number of compounds in the dataset) is incorrect. QSAR procedure should maximize the CCR value calculated according to Eq. 4, and at the same time it should be penalized by too high differences between CCR values for different classes.

## Target Functions and Validation Criteria for Category QSAR Models

Category QSAR with more than two classes should use target functions and validation criteria other than those used in classification QSAR. These target functions and validation criteria should consider errors as differences between

predicted and observed categories, or increasing functions of these differences. The total error of prediction over all compounds is the sum of all errors of predictions for individual compounds. Let $n_{ij}$ be the number of compounds of category i assigned by a model to category j (i, j = 1, . . . , K). Then the total error is calculated as follows:

$$E = \sum_{i=1}^{K} \sum_{j=1}^{K} n_{ij} f\left(|i-j|\right). \tag{5}$$

where $f\left(|i-j|\right)$ is the increasing function of errors. In case of biased datasets, it would be important to normalize the errors for compounds of category i on the number of compounds in this category:

$$E = \sum_{i=1}^{K} \frac{1}{N_i} \sum_{j=1}^{K} n_{ij} f\left(|i-j|\right). \tag{6}$$

where $N_i$ is the number of compounds of category i. QSAR procedure should minimize the total error of prediction calculated with 5 or 6. In practice, the accuracy can be defined as $A = 1 - E/E_{\text{exp}}$, where $E_{\text{exp}}$ is the expected total error. Thus, QSAR procedure should maximize the target function A penalized by too high differences between CCR values for different classes.

More detailed consideration of target functions and validation criteria as well as different aspects of cost-sensitive learning, weighting, penalties, as well as threshold moving in QSAR studies are discussed in our recent review (Tropsha and Golbraikh 2010). General aspects of cost-sensitive learning are discussed by Elkan (The Foundations 2001) and Chen et al. (2004). Oversampling of the minority class, that is, inclusion of compounds of the minority class in the dataset more than once, is considered by Yen and Lee (2006), and Kubat and Matwin (1997). The opposite approach, called undersampling, that is, removing part of the majority class from the dataset, is considered by Japkowicz (2000). Using moving threshold for dividing compounds into active and inactive classes when continuous property values are available but one desires to use classification modeling approaches is considered by Zhou and Liu (2006). In QSAR studies, threshold is usually moved toward the larger class, which is easier to predict correctly.

## Applicability Domains

Here we are approaching an extremely important problem of QSAR studies: model applicability domain (AD). Formally, a QSAR model can predict the target property for any compound for which chemical descriptors can be calculated. However, if a compound is highly dissimilar from all compounds of the modeling set, reliable prediction of its activity is unlikely to be realized. A concept of AD was

developed and used to avoid such an unjustified extrapolation in activity prediction. Applicability domains are one of the areas of intensive research. Different methods of defining AD exist. Among others, the following definitions are considered by Jaworska and colleagues (2005, 2008).

AD is defined as a hyperparallelepiped in the descriptor space in which representative points are distributed (Netzeva et al. 2006; Nikolova-Jeliazkova and Jaworska 2005; Saliner et al. 2006). Dimensionality of the hyperparallelepiped is equal to the number of descriptors, and the size of each dimension is defined by the minimum and maximum values of the corresponding descriptor or it stretches beyond these limits to some extent up to predefined thresholds.

AD is defined as a convex hull of points in the multidimensional descriptor space (Fechner et al. 2008).

The drawbacks of these definitions are as follows. Generally, the representative points are distributed not in the entire hyperparallelepiped or convex hull, but only in a small part of it. Another drawback is that structural outliers in the dataset can enormously increase the size of the hyperparallelepiped, and the area around the outlier will contain no other points. Consequently, for many compounds within the hyperparallelepiped or convex hull, prediction will be unreliable. Besides, if the number of linearly independent descriptors exceeds the number of compounds, the convex hull is not unique.

Leverage for a compound is defined as the corresponding diagonal element of the hat matrix (Afantitis et al. 2006). A compound is defined as outside of the AD, if its leverage L is higher than 3 K/N, where K is the number of descriptors and N is the number of compounds. The drawbacks of the leverage-based AD are as follows. (a) for each external compound, it is necessary to recalculate leverage; (b) if there are cavities in the representative point distribution area, a query compound the representative point of which is in this area will be considered to be within the AD, while in fact it is far from all other compounds (Tropsha and Golbraikh 2010).

In our studies, the AD is defined as the Euclidean distance threshold DT between a query compound and its closest k-nearest neighbors of the training set. It is calculated as follows:

$$DT = \overline{y} + Z\sigma \tag{7}$$

Here, $\overline{y}$ is the average Euclidean distance between each compound and its k-nearest neighbors in the training set k is optimized in the course of QSAR modeling, and the distances are calculated using descriptors selected by the optimized (model only), σ is the standard deviation of these Euclidean distances, and Z is an arbitrary cutoff parameter defined by a user (de Cerqueira et al. 2006; Hsieh et al. 2008; Kovatcheva et al. 2005; Zhang et al. 2008). We set the default value of this parameter Z at 0.5, which formally places the allowed distance threshold at the mean plus one-half of the standard deviation. We also define the AD in the entire descriptor space. In this case, the same Eq. 7 is used, $k = 1$, $Z = 0.5$, and Euclidean distances are calculated using all descriptors. Thus, if the distance of the external compound from its nearest neighbor in the training set within either the entire descriptor space or

the selected descriptor space exceeds these thresholds, the prediction is not made. We have also investigated changes of predictive power by changing the values of Z-cutoff. We have found that in general, starting from some Z-cutoff value, predictive power decreases while Z-cutoff value increases (Zhu et al. 2009), as expected. Instead of Euclidean distances, other distances and similarity measures can be used.

The predicted activity of a query compound by an ensemble of QSAR models is calculated as the average over all predicted values. In binary QSAR modeling, each model will predict the compound category as either 0 (inactive) or 1 (active); however, different models used in an ensemble may yield inconsistent predictions. Consequently, the averaged predicted activity value for an external compound resulting from the use of an ensemble of models may fall anywhere within the [0;1] range. For classification and category QSAR, the average predicted value is rounded to the closest integer (which is a class or category number); in the case of imbalanced datasets, rounding can be done using the moving threshold (vide supra). Predicted average classes or categories (before rounding) that are closer to the nearest integers are considered more reliable since such value indicates higher concordance between different models. For example, before rounding, one compound has the predicted value of 0.2, but the other has 0.4. Hence, both compounds are predicted to belong to class 0 but the prediction for the first compound is considered more reliable. Using these prediction values, additional constraint on the AD can be defined by a threshold of the absolute difference between the predicted and the rounded predicted activity. There are several other definitions of AD (Jaworska and Nikolova-Jeliazkova 2008; Tetko et al. 2006) based on probability density distributions, distances to models, etc.

## Y-randomization

To establish model robustness, Y-randomization (randomization of the response variable) test should be used. This test consists of repeating all the calculations with scrambled activities of the training set. Ideally, calculations should be repeated at least five (better, more) times. The goal of this procedure is to establish whether models built with real activities of the training set have good statistics not due to overfitting or chance correlation. If predictive power for the training or the test set of all models built with randomized activities of the training set is significantly lower than that of models built with real activities of the training set, the latter ones are considered reliable. Using different parameters of the model development procedure, multiple QSAR models are built which have acceptable statistics. Suppose, the number of these models is m. Y-randomization test can also give n models with acceptable statistics. For acceptance of models developed with real activities of the training set, the condition $n \ll m$ should be satisfied. In (Kovatcheva et al. 2005) and (de Cerqueira et al. 2006), we have introduced the measure of robustness $R = 1 - n/m$. If $R > 0.9$, the models are considered robust and their high predictive accuracy cannot be explained by the chance correlation or overfitting. Y-randomization test is particularly important for small datasets.

Unfortunately, in many publications on QSAR studies, Y-randomization test is not carried out but all QSAR practitioners must be strongly encouraged to use this simple procedure.

## External Validation

Our previous experience suggests that the consensus prediction, which is the average of predicted activities over all predictive models, always provides the most stable results (Zhang et al. 2008; Zhu et al. 2008), and thus naturally avoids the need for (the best) model selection based on the statistics for the training and test sets. The consensus prediction of biological activity for an external compound on the basis of several QSAR models is more reliable and provides better justification for the experimental exploration of hits.

External evaluation set compounds are predicted by models that have passed all validation criteria described above. Each compound is predicted by models for which the compound is within the AD. Actually, each external compound should be within the AD of the training set within the entire descriptor space as well (vide supra). A useful parameter for consensus prediction is the minimum number (or percentage) of models for which a compound is within the AD; it is defined by the user. If the compound is found within the AD of a lower number of models, it is considered to be outside of the AD. Prediction value is the average of predictions by all models. If a compound is predicted by more than one model, standard deviation of all predictions by these models is also calculated. For classification and category QSAR, the average prediction value is rounded to the closest integer (which is a class or category number); in case of imbalanced datasets, rounding can be done using the moving threshold.

Predicted average classes or categories (before rounding), which are closer to the nearest integers are considered more reliable (Zhang et al. 2008). Using these prediction values, AD can be defined by a threshold of the absolute difference between predicted and rounded predicted activity. For classification and category QSAR, the same prediction accuracy criteria are used as for the training and test sets. The situation is more complex for the continuous QSAR. In this case, if the range of activities of the external evaluation set is comparable to that for the modeling set, criteria (1)–(4) are used (see section "Target Functions"). Sometimes, however, the external evaluation set may have a much smaller range of activities than the modeling set, so it could be impossible to obtain sufficiently large $R^2$ value (and other acceptable statistical characteristics) for it. In this case, we recommend using the mean absolute error (MAE) or the standard error of prediction (SEP) as discussed in one of our previous publications (Tropsha and Golbraikh 2010).

We have used consensus prediction in many studies (de Cerqueira et al. 2006; Kovatcheva et al. 2005; Shen et al. 2004; Votano et al. 2004; Zhang et al. 2007, 2008; Zhu et al. 2008) and have shown that in most cases it gives better prediction and coverage than most of the individual predictive models. Thus, we recommend using

consensus prediction for virtual screening of chemical databases and combinatorial libraries for finding new lead compounds for drug discovery.

## "Good Practices" in QSAR Modeling: Examples of Models and Their Application to Virtual Screening and Lead Identification

As discussed above, our experience in QSAR model development and validation has led us to establishing a complex but straightforward workflow summarized in Fig. 2. The last critical component of this workflow is the use of models to identify tentative active hits that should be validated in experimental laboratories, and we strongly encourage every computational scientist to use this ultimate model validation strategy. We note that this approach shifts the emphasis from ensuring good (best) statistics for the model that fits known experimental data toward generating testable hypotheses about purported bioactive compounds. Thus, the output of the modeling has exactly same format as the input, that is, chemical structures and (predicted) activities making model interpretation and utilization completely seamless for medicinal chemists. In our recent studies, we have been fortunate to recruit experimental collaborators who have validated computational hits identified through our modeling of anticonvulsants (Shen et al. 2004), HIV-1 reverse transcriptase inhibitors (Medina-Franco et al. 2005), D1 antagonists (Oloff et al. 2005), antitumor compounds (Zhang et al. 2007), beta-lactamase inhibitors (Hsieh et al. 2008), geranylgeranyltransferase inhibitors (Peterson et al. 2009), and others. The discovery of novel bioactive chemical entities is the primary goal of computational drug discovery, and the development of validated and predictive QSAR models is critical to achieve this goal. We note that such studies could only be done if there is sufficient data available for a series of tested compounds such that robust validated models could be developed using the workflow described in Fig. 2. We present several examples of these studies below to illustrate the use of QSAR models as virtual screening tools for lead identification.

## QSAR-Aided Discovery of Novel Anticonvulsant Compounds

We have applied kNN (Zheng and Tropsha 2000) and simulated annealing – partial least squares (SA-PLS) (Cho et al. 1998) QSAR approaches to a dataset of 48 chemically diverse functionalized amino acids (FAAs) with anticonvulsant activity that were synthesized previously, and successful QSAR models of FAA anticonvulsants have been developed (Shen et al. 2002). Both methods utilized multiple descriptors such as molecular connectivity indices or atom-pair descriptors, which are derived from two-dimensional molecular topology. QSAR models with high internal accuracy were generated, with leave-one-out cross-validated $R^2(q^2)$ values ranging between 0.6 and 0.8. The $q^2$ values for the actual dataset were significantly higher than those obtained for the same dataset with randomly shuffled

activity values, indicating that models were statistically significant. The original dataset was further divided into several training and test sets, and highly predictive models providing $q^2$ values for the training sets greater than 0.5 and $R^2$ values for the test sets greater than 0.6.

In the second phase of modeling, we have applied the validated QSAR models to mining available chemical databases for new lead FAA anticonvulsant agents. Two databases have been explored: the National Cancer Institute (nci 2007) and Maybridge (2005) databases, including (at the time of that study) 237,771 and 55,273 chemical structures, respectively. Database mining was performed independently using ten individual QSAR models that have been extensively validated using several criteria of robustness and accuracy. Each individual model selected some number of hits as a result of independent database mining, and the consensus hits (i.e., those selected by all models) were further explored experimentally for their anticonvulsant activity. As a result of computational screening of the NCI database, 27 compounds were selected as potential anticonvulsant agents and submitted to our experimental collaborators. Of these 27 compounds, our collaborators chose two for synthesis and evaluation; their choice was based on the ease of synthesis and the fact that these two compounds had structural features that would not be expected to be found in active compounds based on prior experience. Several additional compounds, which were close analogs of these two were either taken from the literature or designed in our collaborator's laboratory. In total, seven compounds were resynthesized and sent to the NIH for the Maximum Electroshock (MES) test (a standard test for the anticonvulsant activity, which was used for the training set compounds as well). The biological results indicated that upon initial and secondary screening, five out of seven compounds tested showed anticonvulsant activity with $ED_{50}$ less than 100 mg/kg, which is considered promising. Interestingly, all seven compounds were also found to be very active in the same tests performed on rats (a complete set of experimental data on rats for the training set were not available, and therefore no QSAR models for rats were built).

Mining of the Maybridge database yielded two additional promisingcompounds that were synthesized and sent to the NIH for the MES anticonvulsant test. One of the compounds showed moderate anticonvulsant activity of $ED_{50}$ between 30 and 100 mg/kg (in mice), while the other was found to be a very potent anticonvulsant agent with $ED_{50}$ of 18 mg/kg in mice (ip). In summary, both compounds were found to be very active in both mice and rats. Figure 4 shows chemical structures of experimentally confirmed hits that were identified by using validated QSAR models for virtual screening as applied to the anticonvulsant dataset. It is important to note that none of the compounds identified in external databases as potent anticonvulsants and validated experimentally belong to the same class of FAA molecules as the training set. This observation was very stimulating because it underscored the power of our methodology to identify potent anticonvulsants of novel chemical classes as compared to the training set compounds, which is one of the most important goals of virtual screening.

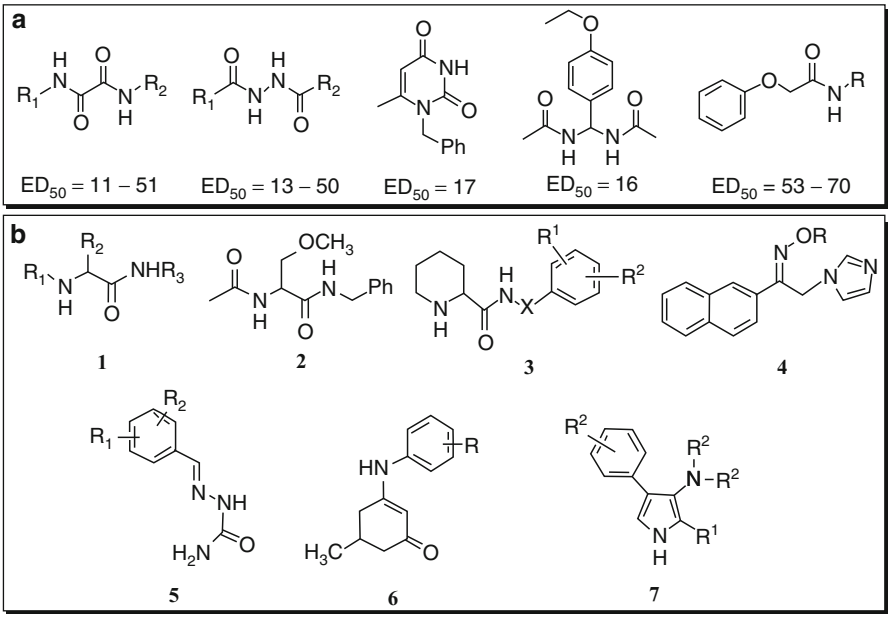


**Fig. 4** Uniqueness of scaffolds for QSAR-based experimentally confirmed virtual screening hits (a) as compared to training set compounds; (b) for the anticonvulsant dataset

## QSAR-Enabled Discovery of Novel Anticancer Agents

A combined approach of validated QSAR modeling and virtual screening was successfully applied to the discovery of novel tylophorine derivatives as anticancer agents (Zhang et al. 2007). QSAR models have been initially developed for 52 chemically diverse phenathrene-based tylophorine derivatives (PBTs) with known experimental $EC_{50}$ using chemical topological descriptors (calculated with the Molconn-Z program) and variable selection k-nearest neighbor (kNN) method. Several validation protocols have been applied to achieve robust QSAR models. The original dataset was divided into multiple training and test sets, and the models were considered acceptable only if the leave-one-out cross-validated $R^2(q^2)$ values were greater than 0.5 for the training sets and the correlation coefficient $R^2$ values were greater than 0.6 for the test sets. Furthermore, the $q^2$ values for the actual dataset were shown to be significantly higher than those obtained for the same dataset with randomized target properties (Y-randomization test), indicating that models were statistically significant. Ten best models were then employed to mine a commercially available ChemDiv Database (ca. 500 K compounds) resulting in 34 consensus hits with moderate to high predicted activities. Ten structurally diverse hits were experimentally tested and eight were confirmed active with the highest

experimental $EC_{50}$ of 1.8 μM implying an exceptionally high hit rate (80 %). The same ten models were further applied to predict EC50 for four new PBTs, and the correlation coefficient ($R^2$) between the experimental and predicted $EC_{50}$ for these compounds plus eight active consensus hits was shown to be as high as 0.57.

## QSAR Enabled Discovery of Novel Geranylgeranyltransferase I Inhibitors (GGTIs)

The proper functioning of proteins often relies on posttranslational modification of the polypeptide leading to changes in chemical characteristics. Found at the extreme carboxyl terminus of the protein, one posttranslational "program" utilized for over 140 proteins is the so-called CaaX box, where "C" is a cysteine, "aa" is any aliphatic dipeptide, and "X" is the terminal residue that directs which of two prenyl groups is added (Cox and Der 2002; Zhang and Casey 1996). Protein geranylgeranyltransferase type I (GGTase-I) transfers the 20-carbon geranylgeranyl group to proteins including critical signaling molecules from many classes, for example, the Ras superfamily (including K-Ras, Rho, Rap, Cdc42, and Rac), several G-protein gamma subunits, protein kinases (rhodopsin kinase, phosphorylase kinase, and GRK7), and protein phosphatases (Casey and Seabra 1996; Sebti and Hamilton 2000). Several GGTIs have been developed that inhibit C20 lipid modification of GGTase-I substrates. GGTIs have been primarily developed for use as cancer therapeutics, particularly in cancers that have high levels, or activating mutations of geranylgeranylated proteins (Sebti and Hamilton 2000; Winter-Vann and Casey 2005).

The pharmacological data for 48 GGTIs reported in (Peterson et al. 2009) were generated as part of an iterative drug discovery program that led to GGTI-DU40 (Peterson et al. 2006). The structure of GGTI-DU40 can be discussed in the context of the CaaL peptide framework. There is a free amide group, a spacer domain relating to the dialiphatic motif, and critical sulfur as found in the requisite cysteine residue of GGTase-I's substrates. Four additional GGTIs included in the data set were peptidomimetics as well. Importantly, the modeling set included compounds with different (chemical scaffolds), which in theory (and as we have established in our study, in practice) should have enabled the identification of chemically diverse hits from virtual screening.

Three different modeling techniques have been used to model GGTIs following our general combi-QSAR strategy (Fig. 3); the specific workflow as applied to the GGTI dataset is shown in Fig. 5. As the first step of our QSAR-based virtual screening, the preliminary filtering of the 9.5 million compounds in our screening library yielded 79 initial hits. This was done by using the global applicability domain of all 48 GGTIs in the modeling set. After consensus predictions by 104 validated kNN models, their predicted activities ($pIC_{50}$) were found ranging from 4.51 to 5.96. Only 47 hits, including two pairs of stereoisomers, showed high predicted activity ($pIC_{50} > 5.50$) as well as high model coverage and were designated as the final hits. Concurrently, two additional QSAR models were employed to reevaluate

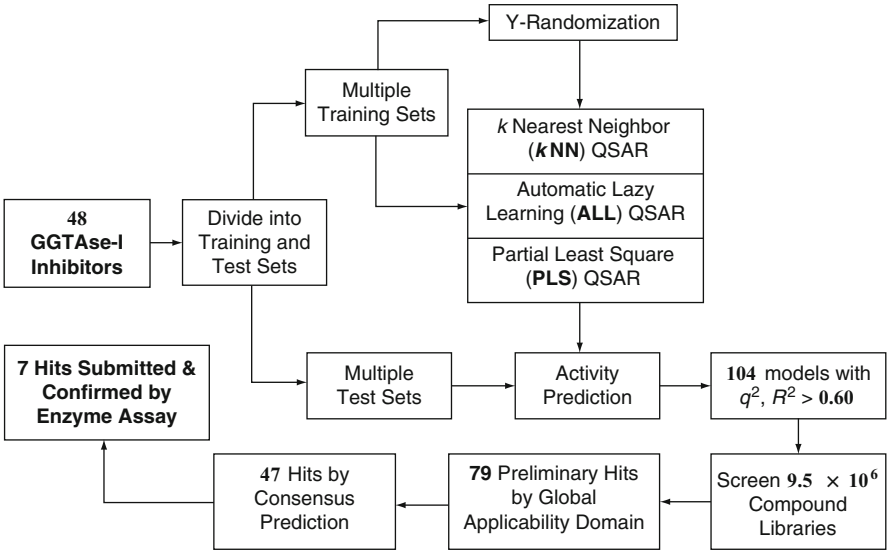


**Fig. 5** The predictive QSAR modeling workflow illustrated for GGTIs Using purified recombinant GGTase-I as an enzyme source and GGpp and Ras-CVLL as substrates, seven hit compounds were tested in vitro as a matter of the experimental validation. The selection was based on high predicted activity, availability, and structural uniqueness. All tested compounds showed inhibition of GGTase-I with the $pIC_{50}$ ranging from 3.63 to 5.44 (cf. Fig. 6)

those 79 hits in order to identify the consensus hits among all three methods. In the end, seven compounds were prioritized for experimental validation based on high predicted activity, uniqueness of structure, and availability.

The unexpected result was to identify several predicted actives that did not have a common ring feature in their structure. In fact, seven highly ranked hits had no apparent relationship with any of the training set molecules. They had furan, triazole, tetrazole, and pyridine cores in their scaffolds while all non-peptidomimetic compounds of the training set were based on a pyrazole core. Therefore, the seven hit compounds appeared to be the structurally novel hits. Figure 6b shows chemical structures of the three representative confirmed hits with novel scaffolds highlighted. This study reconfirmed the observation that we already emphasized earlier with anticonvulsant compounds that contrary to the common belief, QSAR-based virtual screening is capable of identifying experimentally confirmed hit compounds with novel scaffolds.

## "Good Practices" in QSAR Model Development: Applications to Toxicity Modeling

Many compounds entering clinical studies do not survive as a good pharmacological lead to become a marketed drug. Chemical toxicity and safety have been regarded

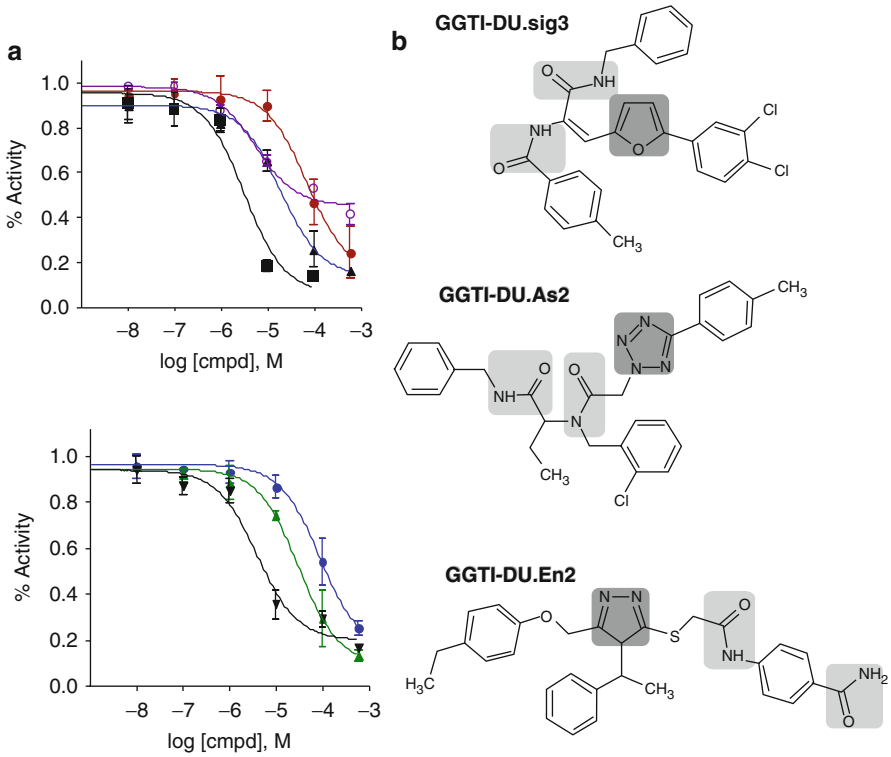


**Fig. 6** Experimental validations of computational GGTI hits using GGTase-I in vitro activity assay. (**a**) Inhibition curves; (**b**) Chemical structures of three representative confirmed hits; the novel scaffolds in the structures are highlighted

as the major reason for attrition in the past decades (Kola and Landis 2004). However, evaluation of chemical toxicity and safety in vivo at the early stage of drug discovery process is expensive and time consuming. To replace the traditional animal toxicity testing and to understand the relevant toxicological mechanisms, many in vitro toxicity screens and computational toxicity models have been developed and implemented by academic institutes and pharmaceutical companies (Cheeseman 2005; Dash et al. 2009; Dix et al. 2007; Inglese et al. 2006; Park et al. 2009; Riley and Kenna 2004; Valerio 2009; Yang et al. 2009). In the past 15 years, innovative technologies that enable rapid synthesis and high throughput screening of large libraries of compounds have been adopted for toxicity studies. As a result, there has been a huge increase in the number of compounds and the associated testing data in different in vitro screens. With this data, it becomes feasible to reveal the relationship between the high throughput in vitro toxicity testing results and the low throughput in vivo low dose toxicity evaluation for the same set of compounds. Understanding these relationships could help us delineate

the mechanisms underlying animal toxicity of chemicals as well as potentially improve our ability to predict chemical toxicity using short-term bioassays.

The unique advantage of using a computational toxicity model in risk analysis is that a chemical could be evaluated for its toxicity potential even before it is synthesized. The computational toxicity tools based on QSAR models have been used to assist in predictive toxicological profiling of pharmaceutical substances for understanding drug safety liabilities (Durham and Pearl 2001; Jacobson-Kram and Contrera 2007; Muster et al. 2008; Valerio 2009), supporting regulatory decision making on chemical safety and risk of toxicity (Bailey et al. 2005), and are effectively enhancing an already rigorous US regulatory safety review of pharmaceutical substances (Valerio 2008). Predictive QSAR models of chemical toxicity are beginning to be used to evaluate compounds' safety in the pharmaceutical industry and environmental agencies (Durham and Pearl 2001; Snyder 2009). However, it has been reported that most QSAR models do not work well for evaluating in vivo toxicity, especially for external compounds (Zvinavashe et al. 2008, 2009). Several reviews were published recently that challenge the feasibility and reliability of QSAR models of chemical toxicity (Johnson 2008; Stouch et al. 2003). At the same time, experimental data resulting from short-term high throughput screening assays are emerging prompting the development of novel modeling approaches that can combine short-term assay data and conventional chemical descriptors of molecules to develop enhanced QSAR models of animal toxicity. We briefly review these emerging approaches and applications below.

## Quantitative Structure In Vitro–In Vivo Relationship Modeling

To stress a broad appeal of the conventional QSAR approach, it should be made clear that from the statistical viewpoint QSAR modeling is a special case of general statistical data mining and data modeling where the data is formatted to represent objects described by multiple descriptors and the robust correlation between descriptors and a target property (e.g., chemical toxicity in vivo) is sought. In previous computational toxicology studies, additional physicochemical properties, such as water partition coefficient (logP) (Klopman et al. 2003), water solubility (Stoner et al. 2004), and melting point (Mayer and Reichenberg 2006) were used successfully to augment computed chemical descriptors and improve the predictive power of QSAR models. These studies suggest that using experimental results as descriptors in QSAR modeling could prove beneficial. The already available and rapidly growing HTS data for large and diverse chemical libraries makes it possible to extend the scope of the conventional QSAR in toxicity studies by using in vitro testing results as additional biological descriptors. Therefore, in some of the most recent toxicology studies, the relationships between various in vitro and in vivo toxicity testing results were generated (Forsby and Blaauboer 2007; Piersma et al. 2008; Schirmer et al. 2008; Sjostrom et al. 2008). Based on these reports, we proposed a new modeling workflow called Quantitative Structure In vitro–In vivo Relationship (QSIIRQuantitative structure in vitro–in vivo relationship (QSIIR)

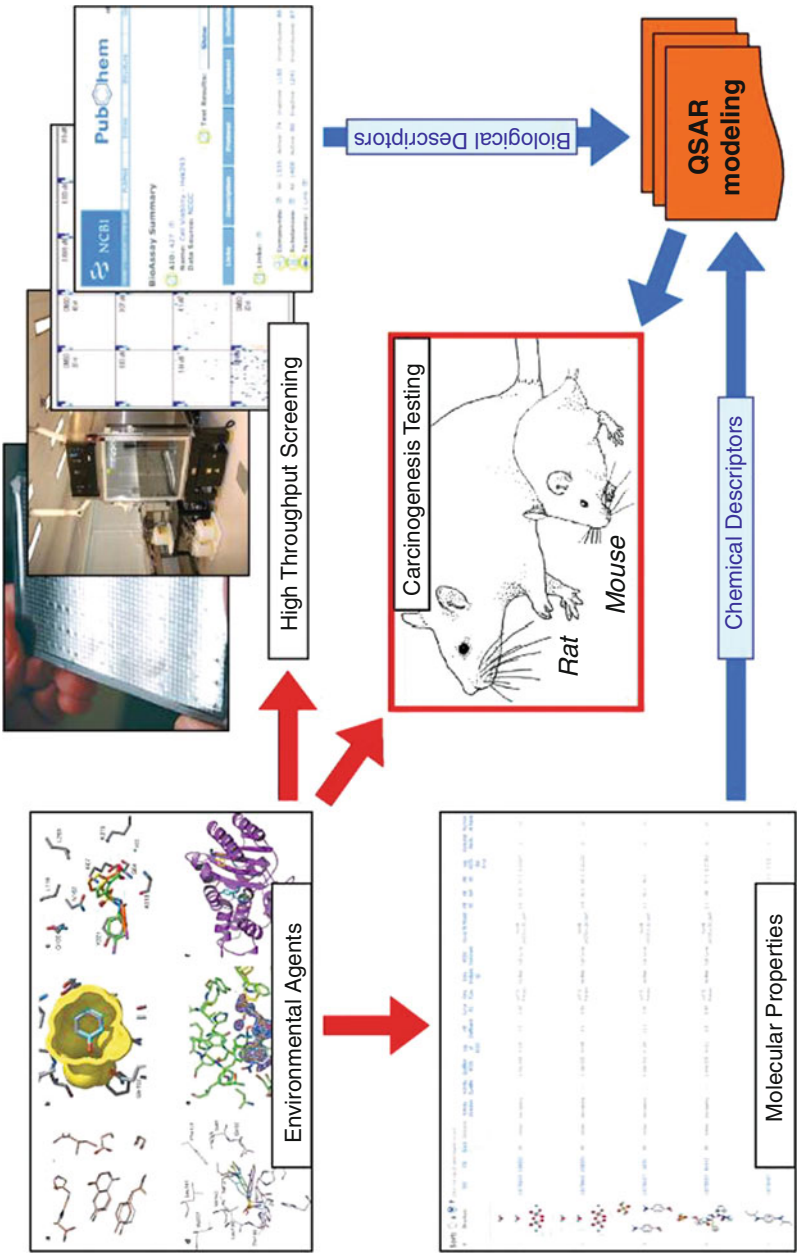


**Fig. 7** Combining chemical and biological profiles as descriptors in QSIIR modeling of chemical carcinogenicity

modeling) and used it in animal toxicity modeling studies (Zhu et al. 2008, 2009). The target properties of QSIIR modeling were still biological activities, such as different toxicity end points, but the content and interpretation of "descriptors" and the resulting models is different. This focus on the prediction of the same target property from different (chemical, biological, and genomic) characteristics of environmental agents affords an opportunity to most fully explore the source-to-outcome continuum of the modern experimental toxicology using cheminformatics approaches. Figure 7– provides visual illustration of the integrated QSIIR approach to in vivo toxicity modeling.

## Using "Hybrid" Descriptors for QSIIR Modeling of Rodent Carcinogenicity

To explore efficient approaches for rapid evaluation of chemical toxicity and human health risk of environmental compounds, the National Toxicology Program (NTP), in collaboration with the National Center for Chemical Genomics (NCGC) has initiated an HTS Project (Inglese et al. 2006; Thomas et al. 2009). The first batch of HTS results for a set of 1,408 compounds tested in six human cell lines was released via PubChem. We have explored this data in terms of their utility for predicting adverse health effects of the environmental agents (Zhu et al. 2008). Initially, the classification k-nearest neighbor (kNN) QSAR modeling method was applied to the HTS data only for the curated dataset of 384 compounds. The resulting models had prediction accuracies for training, test (containing 275 compounds together), and external validation (109 compounds) sets as high as 89 %, 71 %, and 74 %, respectively. We then asked if HTS results could be of value in predicting rodent carcinogenicities. We identified 383 compounds for which data were available from both the Berkeley Carcinogenic Potency Database and NTP-HTS studies. We found that compounds classified by HTS as "actives" in at least one cell line were likely to be rodent carcinogens (sensitivity 77 %); however, HTS "inactives" were far less informative (specificity 46 %). Using chemical descriptors only, kNN QSAR modeling resulted in the overall external prediction accuracy of 62 % for rodent carcinogenicity. Importantly, the prediction accuracy of the model was significantly improved (to 73 %) when chemical descriptors were augmented by the HTS data, which were regarded as biological descriptors (Fig. 8). Thus, our studies suggested, for the first time, that combining HTS profiles with conventional chemical descriptors could considerably improve the predictive power of computational approaches in chemical toxicology.

## Using "Hybrid" Descriptors for the QSIIR Modeling of Rodent Acute Toxicity

We used the cell viability qHTS data from NCGC as mentioned in the above section for the same 1,408 compounds but in 13 cell lines (Xia et al. 2008).

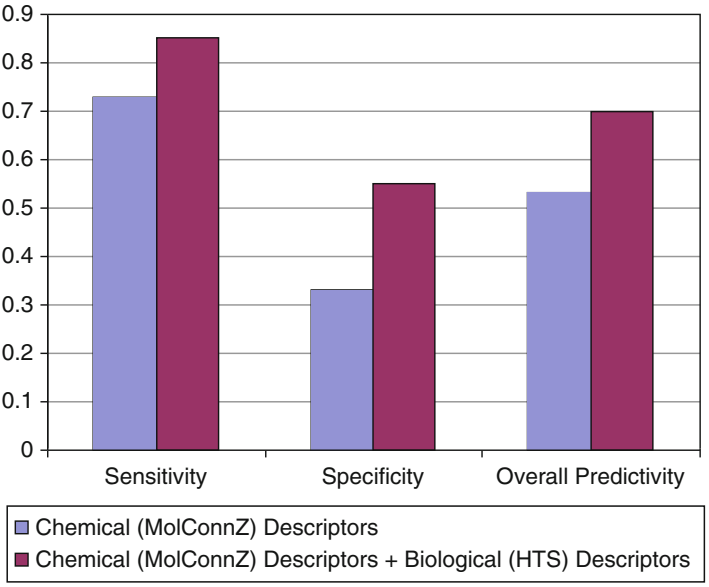


**Fig. 8** Comparison of the prediction power of QSTR models of chemical carcinogenicity for the independent validation set using conventional versus hybrid descriptors

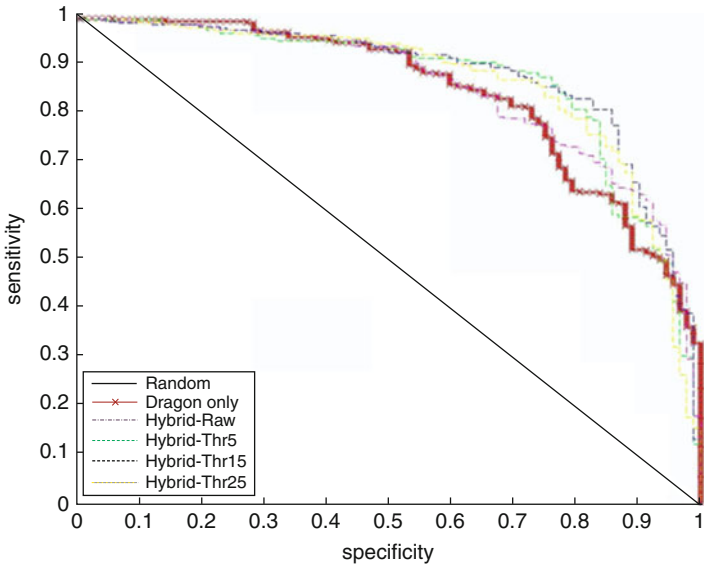


**Fig. 9** Acute toxicity modeling. The ROC curves for conventional QSAR model (bold line) and different hybrid models for the same external compounds.

Besides the carcinogenicity, we asked if HTS results could be of value in predicting rodent acute toxicity (Sedykh et al. in press). For this purpose, we have identified 690 of these compounds, for which rodent acute toxicity data (i.e., toxic or nontoxic) was also available. The classification kNN QSAR modeling method was applied to these compounds using either chemical descriptors alone or as a combination of chemical and qHTS biological (hybrid) descriptors as compound features. The external prediction accuracy of models built with chemical descriptors only was 76 %. In contract, the prediction accuracy was significantly improved to 85 % when using hybrid descriptors. The receiver operating characteristic (ROC) curves of conventional QSAR models and different hybrid models are shown in Fig. 9. The sensitivities and specificities of hybrid models are clearly better than for conventional QSAR model for predicting the same external compounds. Furthermore, the prediction coverage increased from 76 % when using chemical descriptors only to 93 % when qHTS biological descriptors were also included. Our studies suggest that combining HTS profiles, especially the dose–response qHTS results, with conventional chemical descriptors could considerably improve the predictive power of computational approaches for rodent acute toxicity assessment.

## Collaborative and Consensus Modeling of Aquatic Toxicity

We discuss below the results of a recent important study of aquatic toxicity (Zhu et al. 2008). In our opinion, this particular study may serve as a useful example to illustrate the complexity and power of modern QSAR modeling approaches and highlight the importance of collaborative and consensual model development.

The combinational QSAR modeling approach has been applied to a diverse series of organic compounds tested for aquatic toxicity in Tetrahymena pyriformis in the same laboratory over nearly a decade (Aptula et al. 2005; Netzeva and Schultz 2005; Schultz 1999; Schultz and Netzeva 2004; Schultz et al. 2001, 2002, 2003, 2005a, 2005b). The unique aspect of this research was that it was conducted in collaboration between six academic groups specializing in cheminformatics and computational toxicology. The common goals for our virtual collaboratory were to explore the relative strengths of various QSAR approaches in their ability to develop robust and externally predictive models of this particular toxicity end point. We have endeavored to develop the most statistically robust, validated, and externally predictive QSAR models of aquatic toxicity. The members of our collaboratory included scientists from the University of North Carolina at Chapel Hill in the United States (UNC); University of Louis Pasteur (ULP) in France; University of Insubria (UI) in Italy; University of Kalmar (UK) in Sweden; Virtual Computational Chemistry Laboratory (VCCLAB) in Germany; and the University of British Columbia (UBC) in Canada. Each group relied on its own QSAR modeling approaches to develop toxicity models using the same modeling set, and we agreed to evaluate the realistic model performance using the same external validation set(s).

The T. pyriformis toxicity dataset used in this study was compiled from several publications of the Schultz group as well as from data available at the Tetratox database Web site of (http://www.vet.utk.edu/TETRATOX/). After deleting duplicates as well as several compounds with conflicting test results and correcting several chemical structures in the original data sources, our final dataset included 983 unique compounds. The dataset was randomly divided into two parts: (1) the modeling set of 644 compounds; (2) the validation set including 339 compounds. The former set was used for model development by each participating group and the latter set was used to estimate the external prediction power of each model as a universal metric of model performance. In addition, when this project was already well underway, a new dataset had become available from the most recent publication by the Schultz group (Schultz et al. 2007). It provided us with an additional external set to evaluate the predictive power and reliability of all QSAR models. Among compounds reported in (Schultz et al. 2007) 110 were unique, that is, not present among the original set of 983 compounds; thus, these 110 compounds formed the second independent validation set for our study.

## Universal Statistical Figures of Merit for All Models

Different groups have employed different techniques and (sometimes) different statistical parameters to evaluate the performance of models developed independently for the modeling set (described below). To harmonize the results of this study, the same standard parameters were chosen to describe each model's performance as applied to the modeling and external test set predictions. Thus, we have employed $Q^2_{abs}$ (squared leave-one-out cross-validation correlation coefficient) for the modeling set, $R^2_{abs}$ (frequently described as coefficient of determination) for the external validations sets, and MAE (mean absolute error) for the linear correlation between predicted ($Y_{\text{pred}}$) and experimental ($Y_{\text{exp}}$) data (here, $Y = pIGC_{50}$); these parameters are defined as follows:

$$Q^2_{abs} = 1 - \sum_Y \left(Y_{\text{exp}} - Y_{LOO}\right)^2 / \sum_Y \left(Y_{\text{exp}} - < Y >_{\text{exp}}\right)^2 \tag{8}$$

$$R^2_{abs} = 1 - \sum_Y \left(Y_{\text{exp}} - Y_{\text{pred}}\right)^2 / \sum_Y \left(Y_{\text{exp}} - < Y >_{\text{exp}}\right)^2 \tag{9}$$

$$\text{MAE} = \sum_Y \left|Y - Y_{pred}\right| / n \tag{10}$$

Many other statistical characteristics can be used to evaluate model performance; however, we restricted ourselves to these three parameters that provide minimal but sufficient information concerning any model's ability to reproduce both the trends

in experimental data for the test sets as well as mean accuracy of predicting all experimental values. The models were considered acceptable if $R^2_{abs}$ exceeded 0.5.

## Consensus QSAR Models of Aquatic Toxicity; Comparison Between Methods and Models

The objective of this study from methodological prospective was to explore the suitability of different QSAR modeling tools for the analysis of a dataset with an important toxicological end point. Typically, such datasets are analyzed with one (or several) modeling techniques, with a great emphasis on the (high value of) statistical parameters of the training set models. In this study, we went well beyond the modeling studies reported in the original publications in several respects. First, we have compiled all reported data on chemical toxicity against T. pyriformis in a single large dataset and attempted to develop global QSAR models for the entire set. Second, we have employed multiple QSAR modeling techniques thanks to the engagement of six collaborating groups. Third, we have focused on defining model performance criteria not only using training set data but most importantly using external validation sets that were not used in model development in any way (unlike any common cross-validation procedure) (Gramatica 2007). This focus afforded us the opportunity to evaluate and compare all models using simple and objective universal criteria of external predictive accuracy, which in our opinion is the most important single figure of merit for a QSAR model that is of practical significance for experimental toxicologists. Fourth, we have explored the significance of applicability domains and the power of consensus modeling in maximizing the accuracy of external predictivity of our models.

We believe that results of our analysis lend a strong support for our strategy. Indeed, all models performed quite well for the training set with even the lowest $Q^2_{abs}$ among them as high as 0.72. However, there was much greater variation between these models when looking at their (universal and objective) performance criteria as applied to the validation sets.

Of 15 QSAR approaches used in this study, nine implemented method-specific applicability domains. Models that did not define the AD showed a reduced predictive accuracy for the validation set II even though they yielded reasonable results for the validation set I. On average, the use of applicability domains improved the performance of individual models although the improvement came at the expense of the lower chemistry space coverage.

For the most part all models succeeded in achieving reasonable accuracy of external prediction especially when using the AD. It then appeared natural to bring all models together to explore the power of consensus prediction. Thus, the consensus model was constructed by averaging all available predicted values taking into account the applicability domain of each individual model. In this case, we could use only 9 of 15 models that had the AD defined. Since each model had its unique way of defining the AD, each external compound could be found within

the AD of anywhere between one and nine models so for averaging we only used models covering the compound. The advantage of this data treatment is that the overall coverage of the prediction is still high because it was rare to have an external compound outside of the ADs of all available models. The results showed that the prediction accuracy for both the modeling set and the validation sets was the best compared to any individual model. The same observation could be made for the correlation coefficient $R^2_{abs}$. The coverage of this consensus model II was 100 % for all three data sets. This observation suggests that consensus models afford both high space coverage and high accuracy of prediction

In summary, this study presents an example of a fruitful international collaboration between researchers that use different techniques and approaches but share general principles of QSAR model development and validation. Significantly, we did not make any assumptions about the purported mechanisms of aquatic toxicity yet were able to develop statistically significant models for all experimentally tested compounds. In this regard it is relevant to cite an opinion expressed in an earlier publication by T. Schultz that "models that accurately predict acute toxicity without first identifying toxic mechanisms are highly desirable" (Schultz 1999). However, the most significant single result of our studies is the demonstrated superior performance of the consensus modeling approach when all models are used concurrently and predictions from individual models are averaged. We have shown that both the predictive accuracy and coverage of the final consensus QSAR models were superior as compared to these parameters for individual models. The consensus models appeared robust in terms of being insensitive to both incorporating individual models with low prediction accuracy and the inclusion or exclusion of the AD. Another important result of this study is the power of addressing complex problems in QSAR modeling by forming a virtual collaboratory of independent research groups leading to the formulation and empirical testing of best modeling practices. This latter endeavor is especially critical in light of the growing interest of regulatory agencies to developing most reliable and predictive models for environmental risk assessment (Yang et al. 2006) and placing such models in the public domain.

## Conclusions: Emerging Chemical/Biological Data and QSAR Research Strategies

In the past 15 years, innovative technologies that enable rapid synthesis and high throughput screening of large libraries of compounds have been adopted in almost all major pharmaceutical and biotech companies. As a result, there has been a huge increase in the number of compounds available on a routine basis to quickly screen for novel drug candidates against new targets or pathways. In contrast, such technologies have rarely become available to the academic research community, thus limiting its ability to conduct large-scale chemical genetics or chemical genomics research. The NIH Molecular Libraries Roadmap Initiative has changed this situation by forming the national Molecular Library Screening

Centers Network (MLSCN) (Austin et al. 2004) with the results of screening assays made publicly available via PubChem (2010). These efforts have already led to the unprecedented growth of available databases of biologically tested compounds [cf. our recent review where we list about 20 available databases of compounds with known bioactivity (Oprea and Tropsha 2006)].

This growth creates new challenges for QSAR modeling such as developing novel approaches for the analysis and visualization of large databases of screening data, novel biologically relevant chemical diversity or similarity measures, and novel tools for virtual screening of compound libraries to ensure high expected hit rates. Application studies discussed in this chapter have established that QSAR models could be used successfully as virtual screening tools to discover compounds with the desired biological activity in chemical databases or virtual libraries (Hsieh et al. 2008; Oloff et al. 2005; Shen et al. 2004; Tropsha 2005; Tropsha and Zheng 2001; Zhang et al. 2007). The discovery of novel bioactive chemical entities is the primary goal of computational drug discovery, and the development of validated and predictive QSAR models is critical to achieve this goal. Due to the significant recent increase in publicly available datasets of biologically active compounds and the critical need to improve the hit rate of experimental compound screening there is a strong need in developing widely accessible and reliable computational QSAR modeling techniques and specific end-point predictors.

**Acknowledgments** The studies described in this chapter were supported in parts by the NIH research grants R01GM066940 and R21GM076059 and EPA grants EPA (RD832720 and RD833825).

## Bibliography

Afantitis, A., Melagraki, G., Sarimveis, H., Koutentis, P. A., Markopoulos, J., & Igglessi-Markopoulou, O. (2006). A novel QSAR model for predicting induction of apoptosis by 4-aryl-4H-chromenes. *Bioorganic & Medicinal Chemistry, 14*, 6686.

Agrafiotis, D. K., Cedeno, W., & Lobanov, V. S. (2002). On the use of neural network ensembles in QSAR and QSPR. *Journal of Chemical Information and Computer Science, 42*, 903.

Ajmani, S., Jadhav, K., & Kulkarni, S. A. (2006). Three-dimensional QSAR using the k-nearest neighbor method and its interpretation. *Journal of Chemical Information and Modeling, 46*, 24.

Aptula, A. O., Roberts, D. W., Cronin, M. T. D., & Schultz, T. W. (2005). Chemistry-toxicity relationships for the effects of Di-and trihydroxybenzenes to Tetrahymena pyriformis. *Chemical Research in Toxicology, 18*, 844.

Austin, C. P., Brady, L. S., Insel, T. R., & Collins, F. S. (2004). NIH molecular libraries initiative. *Science, 306*, 1138.

Bailey, A. B., Chanderbhan, R., Collazo-Braier, N., Cheeseman, M. A., & Twaroski, M. L. (2005). The use of structure-activity relationship analysis in the food contact notification program. *Regulatory Toxicology and Pharmacology, 42*, 225.

Bajorath, J., Peltason, L., Wawer, M., Guha, R., Lajiness, M. S., & Van Drie, J. H. (2009). Navigating structure-activity landscapes. *Drug Discovery Today, 14*, 698.

Berk, R. A. (2008). *Classification and Regression Trees (CART). Statistical learning from a regression perspective*. New York: Springer.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*, 123.

Breiman, L. (1998). Arcing classifiers. *The Annals of Statistics, 26*, 801.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5.
Bures, M. G., & Martin, Y. C. (1998). Computational methods in molecular diversity and combinatorial chemistry. *Current Opinion in Chemical Biology, 2*, 376.
Burges, J. C. (1998). Tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery, 2*, 121.
C5.0.(2008).
Carhart, R. E., Smith, D. H., & Venkataraghavan, R. (1985). Atom pairs as molecular features in structure-activity studies: Definition and applications. *Journal of Chemical Information and Computer Science, 25*, 64.
Casey, P. J., & Seabra, M. C. (1996). Protein prenyltransferases. *The Journal of Biological Chemistry, 271*, 5289.
Cheeseman, M. A. (2005). Thresholds as a unifying theme in regulatory toxicology. *Food Additives & Contaminants, 22*, 900.
ChemAxon. (2008). http://www.chemaxon.com.
ChEMBL Database. (2010). http://www.ebi.ac.uk/chembldb/.
Chen, C., Liaw, A., & Breiman, L. (2004). *Using random forest to learn imbalanced data* (Vol. 666). Berkeley: Department of Statistics, University of California.
Cherkasov, A. (2008). An updated steroid benchmark set and its application in the discovery of novel nanomolar ligands of sex hormone-binding globulin. *Journal of Medicinal Chemistry, 51*, 2047.
Cho, S. J., Zheng, W., & Tropsha, A. (1998). Rational combinatorial library design. 2. Rational design of targeted combinatorial peptide libraries using chemical similarity probe and the inverse QSAR approaches. *Journal of Chemical Information and Computer Science, 38*, 259.
Cox, A. D., & Der, C. J. (2002). Farnesyltransferase inhibitors: Promises and realities. *Current Opinion in Pharmacology, 2*, 388.
Crivori, P., Cruciani, G., Carrupt, P. A., & Testa, B. (2000). Predicting blood–brain barrier permeation from three-dimensional molecular structure. *Journal of Medicinal Chemistry, 43*, 2204.
Cruciani, G., Pastor, M., & Guba, W. (2000). VolSurf: A new tool for the pharmacokinetic optimization of lead compounds1. *European Journal of Pharmaceutical Sciences, 11*(Suppl 2), S29–S39.
Dash, A., Inman, W., Hoffmaster, K., Sevidal, S., Kelly, J., Obach, R. S., et al. (2009). Liver tissue engineering in the evaluation of drug safety. *Expert Opinion on Drug Metabolism & Toxicology, 5*, 1159.
de Cerqueira, L. P., Golbraikh, A., Oloff, S., Xiao, Y., & Tropsha, A. (2006). Combinatorial QSAR modeling of P-Glycoprotein substrates. *Journal of Chemical Information and Modeling, 46*, 1245.
Discovery Studio. (2010).
Dix, D. J., Houck, K. A., Martin, M. T., Richard, A. M., Setzer, R. W., & Kavlock, R. J. (2007). The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicological Sciences, 95*, 5.
Dragon. (2007). http://www.talete.mi.it/help/dragon_help/index.html?IntroducingDRAGON
DSSTox. (2008). http://www.epa.gov/nheerl/dsstox/About.html.
Durham, S. K., & Pearl, G. M. (2001). Computational methods to predict drug safety liabilities. *Current Opinion in Drug Discovery & Development, 4*, 110.
Environmental Protection Agency. (1992). *Statistical training course for ground-water monitoring data analysis EPA/530-R-93-003*. Washington: Office of Solid Waste.
Fallon, A., Spada, C., & Gallagher, D. (1997). Detection and Accommodation of Outliers in Normally Distributed Data Sets. http://ewr.cee.vt.edu/environmental/teach/smprimer/outlier/outlier.html. Accessed 25 April 2005.
Fechner, N., Hinselmann, G., Schmiedl, C., & Zell, A. (2008). Estimating the applicability domain of kernel-based QSPR models using classical descriptor vectors. *Chemistry Central Journal, 2*(Suppl.1), P2.

Forsby, A., & Blaauboer, B. (2007). Integration of in vitro neurotoxicity data with biokinetic modelling for the estimation of in vivo neurotoxicity. *Human & Experimental Toxicology, 26*, 333.

Fourches, D., Muratov, E., & Tropsha, A. (2010). Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. *Journal of Chemical Information and Modeling, 50*, 1189–1204.

Gasteiger, J. (2006). Of molecules and humans. *Journal of Medicinal Chemistry, 49*, 6429.

Golbraikh, A., & Tropsha, A. (2002). Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Journal of Computer-Aided Molecular Design, 16*, 357.

Golbraikh, A., & Tropsha, A. (2003). QSAR modeling using chirality descriptors derived from molecular topology. *Journal of Chemical Information and Computer Science, 43*, 144.

Golbraikh, A., Bonchev, D., & Tropsha, A. (2001). Novel chirality descriptors derived from molecular topology. *Journal of Chemical Information and Computer Science, 41*, 147.

Golbraikh, A., Bonchev, D., & Tropsha, A. (2002). Novel ZE-isomerism descriptors derived from molecular topology and their application to QSAR analysis. *Journal of Chemical Information and Computer Science, 42*, 769.

Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y. D., Lee, K. H., & Tropsha, A. (2003). Rational selection of training and test sets for the development of validated QSAR models. *Journal of Computer-Aided Molecular Design, 17*, 241.

Gramatica, P. (2007). Principles of QSAR models validation: Internal and external. *QSAR & Combinatorial Science, 26*, 694.

Guha, R., & Van Drie, J. H. (2008a). Structure–activity landscape index: Identifying and quantifying activity cliffs. *Journal of Chemical Information and Modeling, 48*, 646.

Guha, R., & Van Drie, J. H. (2008b). Assessing how well a modeling protocol captures a structure-activity landscape. *Journal of Chemical Information and Modeling, 48*, 1716.

Hoffman, B., Cho, S. J., Zheng, W., Wyrick, S., Nichols, D. E., Mailman, R. B., et al. (1999). Quantitative structure-activity relationship modeling of dopamine D(1) antagonists using comparative molecular field analysis, genetic algorithms-partial least-squares, and K nearest neighbor methods. *Journal of Medicinal Chemistry, 42*, 3217.

Hong, H., Xie, Q., Ge, W., Qian, F., Fang, H., Shi, L., et al. (2008). Mold(2), molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *Journal of Chemical Information and Modeling, 48*, 1337.

Horvath, D., Bonachera, F., Solov'ev, V., Gaudin, C., & Varnek, A. (2007). Stochastic versus stepwise strategies for quantitative structure-activity relationship generation–how much effort may the mining for successful QSAR models take? *Journal of Chemical Information and Modeling, 47*, 927.

Hsieh, J. H., Wang, X. S., Teotico, D., Golbraikh, A., & Tropsha, A. (2008). Differentiation of AmpC beta-lactamase binders vs. decoys using classification kNN QSAR modeling and application of the QSAR classifier to virtual screening. *Journal of Computer-Aided Molecular Design, 22*, 593.

Huan, J., Bandyopadhyay, D., Prins, J., Snoeyink, J., Tropsha, A., & Wang, W. (2006). Distance-based identification of structure motifs in proteins using constrained frequent subgraph mining. Computational Systems Bioinformatics Conference, 227.

Inglese, J., Auld, D. S., Jadhav, A., Johnson, R. L., Simeonov, A., Yasgar, A., et al. (2006). Quantitative high-throughput screening: A titration-based approach that efficiently identifies biological activities in large chemical libraries. *Proceedings of the National Academy of Sciences of the United States of America, 103*, 11473.

Irwin, J. J., & Shoichet, B. K. (2005). ZINC–a free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling, 45*, 177.

Jacobson-Kram, D., & Contrera, J. F. (2007). Genetic toxicity assessment: Employing the best science for human safety evaluation. Part I: Early screening for potential human mutagens. *Toxicological Sciences, 96*, 16.

Japkowicz, N. (2000). *Learning from imbalanced datasets: A comparison of various strategies. AAAI Workshop*. Menlo Park: AAAI Press.

Jaworska, J., & Nikolova-Jeliazkova, N. (2008). *Review of methods to assess a QSAR Applicability Domain*. http://ambit.acad.bg/nina/publications/2004/AppDomain_sar04.ppt

Jaworska, J., Nikolova-Jeliazkova, N., & Aldenberg, T. (2005). QSAR applicabilty domain estimation by projection of the training set descriptor space: A review. *Alternatives to Laboratory Animals, 33*, 445.

Johnson, S. R. (2008). The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). *Journal of Chemical Information and Modeling, 48*, 25.

Klebe, G. (1998). Comparative molecular similarity indices: CoMSI. In H. Kubinyi, G. Folkers, & Y. Martin (Eds.), *3D QSAR in drug design* (pp. 87–104). Great Britain: Kluver.

Klopman, G., Zhu, H., Ecker, G., & Chiba, P. (2003). MCASE study of the multidrug resistance reversal activity of propafenone analogs. *Journal of Computer-Aided Molecular Design, 17*, 291.

Kola, I., & Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery, 3*, 711.

Kovatcheva, A., Golbraikh, A., Oloff, S., Feng, J., Zheng, W., & Tropsha, A. (2005). QSAR modeling of datasets with enantioselective compounds using chirality sensitive molecular descriptors. *SAR and QSAR in Environmental Research, 16*, 93.

Kubat, M., & Matwin, S. (1997). *Addressing the curse of imbalanced training sets: One sided selection*. San Francisco: Morgan Kaufmann.

Kubinyi, H., Hamprecht, F. A., & Mietzner, T. (1998). Three-dimensional quantitative similarity-activity relationships (3D QSiAR) from SEAL similarity matrices. *Journal of Medicinal Chemistry, 41*, 2553.

LigandScout. (2010).

Liu, T., Lin, Y., Wen, X., Jorissen, R. N., & Gilson, M. K. (2007). BindingDB: A web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Research, 35*, D198–D201.

Maggiora, G. M. (2006). On outliers and activity cliffs–why QSAR often disappoints. *Journal of Medicinal Chemistry, 46*, 1535.

Maybridge. (2005). http://www.daylightcom/products/databases/Maybridge.html

Mayer, P., & Reichenberg, F. (2006). Can highly hydrophobic organic substances cause aquatic baseline toxicity and can they contribute to mixture toxicity? *Environmental Toxicology & Chemistry, 25*, 2639.

McGregor, M. J., & Pallai, P. V. (1997). Clustering of large databases of compounds: Using the MDL "Keys" as structural descriptors. *Journal of Chemical Information and Computer Science, 37*, 443.

MDDR. SYMYX technologies. (2009). http://www.mdl.com/products/knowledge/drug_data_report/index.jsp

Medina-Franco, J. L., Golbraikh, A., Oloff, S., Castillo, R., & Tropsha, A. (2005). Quantitative structure-activity relationship analysis of pyridinone HIV-1 reverse transcriptase inhibitors using the k nearest neighbor method and QSAR-based database mining. *Journal of Computer-Aided Molecular Design, 19*, 229.

Molconn-Z. (2007). http://www.edusoft-lc.com/

Molecular Operating Environment (MOE). (2008). http://www.chemcomp.com/

Muster, W., Breidenbach, A., Fischer, H., Kirchner, S., Muller, L., & Pahler, A. (2008). Computational toxicology in drug development. *Drug Discovery Today, 13*, 303.

nci. (2007). http://dtp.nci.nih.gov/docs/3d_database/structural_information/smiles_strings.html.

Netzeva, T. I., & Schultz, T. W. (2005). QSARs for the aquatic toxicity of aromatic aldehydes from Tetrahymena data. *Chemosphere, 61*, 1632.

Netzeva, T. I., Gallegos, S. A., & Worth, A. P. (2006). Comparison of the applicability domain of a quantitative structure-activity relationship for estrogenicity with a large chemical inventory. *Environmental Toxicology & Chemistry, 25*, 1223.

Neural Networks. (1996). *Neural networks in QSAR and drug design*. San Diego: Academic.

Neural Networks. (2010). http://www.learnartificialneuralnetworks.com/.
Nikolova-Jeliazkova, N., & Jaworska, J. (2005). An approach to determining applicability domains for QSAR group contribution models: An analysis of SRC KOWWIN. *Alternatives to Laboratory Animals, 33*, 461.
Olah, M., Rad, R., Ostopovici, L., Bora, A., Hadaruga, N., Hadaruga, D., et al. (2007). WOMBAT and WOMBAT-PK: Bioactivity databases for lead and drug discovery. In S. L. Schreiber, T. M. Kapoor, & G. Weiss (Eds.), *Chemical biology: From small molecules to systems biology and drug design* (pp. 760–786). Weinheim: Wiley-VCH.
Oloff, S., Mailman, R. B., & Tropsha, A. (2005). Application of validated QSAR models of D1 dopaminergic antagonists for database mining. *Journal of Medicinal Chemistry, 48*, 7322.
(2010). OpenBabel: The OpenSource Chemistry Toolbox. Openbabel.org. 2-1-2010.
Oprea, T., & Tropsha, A. (2006). Target, chemical and bioactivity databases – integration is key. *Drug Discovery Today, 3*, 357–365.
Organisation for Economic and Co-operation Development. (2008). OECD Quantitative Structure-Activity Relationships [(Q)SARs] Project. http://www.oecd.org/document/23/0,3343,en_2649_34365_33957015_1_1_1_1,00.html.
Park, M. V., Lankveld, D. P., Van, L. H., & de Jong, W. H. (2009). The status of in vitro toxicity studies in the risk assessment of nanomaterials. *Nanomedicine (London, England), 4*, 669.
Pastor, M., Cruciani, G., McLay, I., Pickett, S., & Clementi, S. (2000). GRid-INdependent descriptors (GRIND): A novel class of alignment-independent three-dimensional molecular descriptors. *Journal of Medicinal Chemistry, 43*, 3233.
PDSP. (2010). PDSP. http://pdsp.med.unc.edu.
Peterson, Y. K., Kelly, P., Weinbaum, C. A., & Casey, P. J. (2006). A novel protein geranylgeranyltransferase-I inhibitor with high potency, selectivity, and cellular activity. *The Journal of Biological Chemistry, 281*, 12445.
Peterson, Y. K., Wang, X. S., Casey, P. J., & Tropsha, A. (2009). Discovery of geranylgeranyltransferase-I inhibitors with novel scaffolds by the means of quantitative structure-activity relationship modeling, virtual screening, and experimental validation. *Journal of Medicinal Chemistry, 52*, 4210.
Piersma, A. H., Janer, G., Wolterink, G., Bessems, J. G., Hakkert, B. C., & Slob, W. (2008). Quantitative extrapolation of in vitro whole embryo culture embryotoxicity data to developmental toxicity in vivo using the benchmark dose approach. *Toxicological Sciences, 101*, 91.
PubChem. (2010). http://pubchem.ncbi.nlm.nih.gov/.
Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Mateo: Morgan Kaufmann.
Random Forests. (2001).
Riley, R. J., & Kenna, J. G. (2004). Cellular models for ADMET predictions and evaluation of drug-drug interactions. *Current Opinion in Drug Discovery & Development, 7*, 86.
Robinson, D. D., Winn, P. J., Lyne, P. D., & Richards, W. G. (1999). Self-organizing molecular field analysis: A tool for structure-activity studies. *Journal of Medicinal Chemistry, 42*, 573.
Saliner, A. G., Netzeva, T. I., & Worth, A. P. (2006). Prediction of estrogenicity: Validation of a classification model. *SAR and QSAR in Environmental Research, 17*, 195.
Salt, D. V., Yildiz, N., Livingstone, D. J., & Tinsley, C. J. (2006). The use of artificial neural networks in QSAR. *Pesticide Science, 36*, 161.
Schirmer, K., Tanneberger, K., Kramer, N. I., Volker, D., Scholz, S., Hafner, C., et al. (2008). Developing a list of reference chemicals for testing alternatives to whole fish toxicity tests. *Aquatic Toxicology, 90*, 128.
Schrodinger Software. (2010).
Schultz, T. W. (1999). Structure-toxicity relationships for benzenes evaluated with Tetrahymena pyriformis. *Chemical Research in Toxicology, 12*, 1262.
Schultz, T. W., & Netzeva, T. I. (2004). Development and evaluation of QSARs for ecotoxic endpoints: The benzene response-surface model for Tetrahymena toxicity. In M. T. D. Cronin & D. J. Livingstone (Eds.), *Modeling environmental fate and toxicity* (pp. 265–284). Boca Raton: CRC Press.

Schultz, T. W., Sinks, G. D., & Miller, L. A. (2001). Population growth impairment of sulfur-containing compounds to Tetrahymena pyriformis. *Environmental Toxicology, 16*, 543.

Schultz, T. W., Cronin, M. T., Netzeva, T. I., & Aptula, A. O. (2002). Structure-toxicity relationships for aliphatic chemicals evaluated with Tetrahymena pyriformis. *Chemical Research in Toxicology, 15*, 1602.

Schultz, T. W., Netzeva, T. I., & Cronin, M. T. (2003). Selection of data sets for QSARs: Analyses of Tetrahymena toxicity from aromatic compounds. *SAR and QSAR in Environmental Research, 14*, 59.

Schultz, T. W., Netzeva, T. I., Roberts, D. W., & Cronin, M. T. (2005a). Structure-toxicity relationships for the effects to Tetrahymena pyriformis of aliphatic, carbonyl-containing, alpha, beta-unsaturated chemicals. *Chemical Research in Toxicology, 18*, 330.

Schultz, T. W., Yarbrough, J. W., & Woldemeskel, M. (2005b). Toxicity to Tetrahymena and abiotic thiol reactivity of aromatic isothiocyanates. *Cell Biology and Toxicology, 21*, 181.

Schultz, T. W., Hewitt, M., Netzeva, T. I., & Cronin, M. T. D. (2007). Assessing applicability domains of toxicological QSARs: Definition, confidence in predicted values, and the role of mechanisms of action. *QSAR & Combinatorial Science, 26*, 238.

Sebti, S. M., & Hamilton, A. D. (2000). Farnesyltransferase and geranylgeranyltransferase I inhibitors in cancer therapy: Important mechanistic and bench to bedside issues. *Expert Opinion on Investigational Drugs, 9*, 2767.

Sedykh, A., Zhu, H., Tang, H., Zhang, L., Rusyn, I., Richard, A., et al. The use of dose–response qHTS data as biological descriptors improves the prediction accuracy of QSAR models of acute rat toxicity. Environmental Health Perspect, In press.

Shen, M., LeTiran, A., Xiao, Y., Golbraikh, A., Kohn, H., & Tropsha, A. (2002). Quantitative structure-activity relationship analysis of functionalized amino acid anticonvulsant agents using k nearest neighbor and simulated annealing PLS methods. *Journal of Medicinal Chemistry, 45*, 2811.

Shen, M., Beguin, C., Golbraikh, A., Stables, J. P., Kohn, H., & Tropsha, A. (2004). Application of predictive QSAR models to database mining: Identification and experimental validation of novel anticonvulsant compounds. *Journal of Medicinal Chemistry, 47*, 2356.

Sisay, M. T., Peltason, L., & Bajorath, J. (2009). Structural interpretation of activity cliffs revealed by systematic analysis of structure-activity relationships in analog series. *Journal of Chemical Information and Modeling, 49*, 2179.

Sjostrom, M., Kolman, A., Clemedson, C., & Clothier, R. (2008). Estimation of human blood LC50 values for use in modeling of in vitro-in vivo data of the ACuteTox project. *Toxicology In Vitro, 22*, 1405.

Smola, A. J., & Schoelkopf, B. A. (2004). Tutorial on support vector regression. Tuebingen: Max Planck Society - eDocument Server (Germany).

Snyder, R. D. (2009). An update on the genotoxicity and carcinogenicity of marketed pharmaceuticals with reference to in silico predictivity. *Environmental and Molecular Mutagenesis, 50*, 435.

Stoner, C. L., Gifford, E., Stankovic, C., Lepsy, C. S., Brodfuehrer, J., Prasad, J. V. N. V., et al. (2004). Implementation of an ADME enabling selection and visualization tool for drug discovery. *Journal of Pharmaceutical Sciences, 93*, 1131.

Stouch, T. R., Kenyon, J. R., Johnson, S. R., Chen, X. Q., Doweyko, A., & Li, Y. (2003). In silico ADME/Tox: why models fail. *Journal of Computer-Aided Molecular Design, 17*, 83.

Tetko, I. V., Bruneau, P., Mewes, H. W., Rohrer, D. C., & Poda, G. I. (2006). Can we estimate the accuracy of ADME-Tox predictions? *Drug Discovery Today, 11*, 700.

The Foundations of Cost-sensitive Learning. (2001).

Thomas, C. J., Auld, D. S., Huang, R., Huang, W., Jadhav, A., Johnson, R. L., et al. (2009). The pilot phase of the NIH chemical genomics center. *Current Topics in Medicinal Chemistry, 9*, 1181.

Todeschini, R., & Consonni, V. (2000). *Handbook of molecular descriptors*. Weinheim: Wiley-VCH.

Tripos. (2010). Sybyl-X 1.0

Tropsha, A. (2005). Application of predictive QSAR models to database mining. In T. Oprea (Ed.), *Cheminformatics in drug discovery* (pp. 437–455). Wiley-VCH.

Tropsha, A., & Golbraikh, A. (2007). Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Current Pharmaceutical Design, 13*, 3494.

Tropsha, A., & Golbraikh, A. (2010a). Predictive quantitative structure–activity relationships modeling: Development and validation of QSAR models. In J.-L. Faulon & A. Bender (Eds.), *Handbook of chemoinformatics algorithms*. The Netherlands: Leiden University, Chapman and Hall/CRC.

Tropsha, A., & Golbraikh, A. (2010b). Predictive quantitative structure–activity relationships modeling. Data Preparation and the General Modeling Workflow. In J.-L. Faulon & A. Bender (Eds.), *Handbook of chemoinformatics algorithms* (pp. 175–214). The Netherland: Leiden University, Chapman and Hall/CRC.

Tropsha, A., & Zheng, W. (2001). Identification of the descriptor pharmacophores using variable selection QSAR: Applications to database mining. *Current Pharmaceutical Design, 7*, 599.

Valerio, L., Jr. (2008). Tools for evidence-based toxicology: Computational-based strategies as a viable modality for decision support in chemical safety evaluation and risk assessment. *Human & Experimental Toxicology, 27*, 757.

Valerio, L. G., Jr. (2009). In silico toxicology for the pharmaceutical sciences. *Toxicology and Applied Pharmacology, 241*, 356.

Vapnik, V. (2000). *Nature of statistical learning theory*. New York: Springer.

Votano, J. R., Parham, M., Hall, L. H., Kier, L. B., Oloff, S., Tropsha, A., et al. (2004). Three new consensus QSAR models for the prediction of Ames genotoxicity. *Mutagenesis, 19*, 365.

Waller, C. L. (2004). A comparative QSAR study using CoMFA, HQSAR, and FRED/SKEYS paradigms for estrogen receptor binding affinities of structurally diverse compounds. *Journal of Chemical Information and Computer Science, 44*, 758.

Winter-Vann, A. M., & Casey, P. J. (2005). Post-prenylation-processing enzymes as new targets in oncogenesis. *Nature Reviews Cancer, 5*, 405.

Xia, M., Huang, R., Witt, K. L., Southall, N., Fostel, J., Cho, M. H., et al. (2008). Compound cytotoxicity profiling using quantitative high-throughput screening. *Environmental Health Perspectives, 116*, 284.

Yang, C., Richard, A. M., & Cross, K. P. (2006). The art of data mining the minefields of toxicity databases to link chemistry to biology. *Current Computer-Aided Drug Design, 2*, 135.

Yang, C., Valerio, L. G., Jr., & Arvidson, K. B. (2009). Computational toxicology approaches at the US food and drug administration. *Alternatives to Laboratory Animals, 37*, 523.

Yen, S.-J., & Lee, Y.-S. (2006). Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. *Lecture Notes in Control and Information Sciences, 344*, 731.

Zhang, F. L., & Casey, P. J. (1996). Protein prenylation: Molecular mechanisms and functional consequences. *Annual Review of Biochemistry, 65*, 241.

Zhang, S., Wei, L., Bastow, K., Zheng, W., Brossi, A., Lee, K. H., et al. (2007). Antitumor agents 252. Application of validated QSAR models to database mining: Discovery of novel tylophorine derivatives as potential anticancer agents. *Journal of Computer-Aided Molecular Design, 21*, 97.

Zhang, L., Zhu, H., Oprea, T. I., Golbraikh, A., & Tropsha, A. (2008). QSAR modeling of the blood–brain barrier permeability for diverse organic compounds. *Pharmaceutical Research, 25*, 1902.

Zheng, W., & Tropsha, A. (2000). Novel variable selection quantitative structure–property relationship approach based on the k-nearest-neighbor principle. *Journal of Chemical Information and Computer Science, 40*, 185.

Zhou, Z. H., & Liu, X.-Y. (2006). Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering, 18*, 63.

Zhu, H., Rusyn, I., Richard, A. M., & Tropsha, A. (2008a). Use of cell viability assay data improves the prediction accuracy of conventional quantitative structure-activity relationship models of animal carcinogenicity. *Environmental Health Perspectives, 116*, 506.

Zhu, H., Tropsha, A., Fourches, D., Varnek, A., Papa, E., Gramatica, P., et al. (2008b). Combinatorial QSAR modeling of chemical toxicants tested against tetrahymena pyriformis. *Journal of Chemical Information and Modeling, 48*, 766.

Zhu, H., Ye, L., Golbraikh, A., & Tropsha, A. (2009). QSAR studies of chemical aquatic acute toxicity using k Nearest Neighbor (kNN) Methodology

Zhu, H., Ye, L., Richard, A., Golbraikh, A., Wright, F. A., Rusyn, I., et al. (2009a). A novel two-step hierarchical quantitative structure-activity relationship modeling work flow for predicting acute toxicity of chemicals in rodents. *Environmental Health Perspectives, 117*, 1257.

Zupan, J., & Gasteiger, J. (1999). *Neural networks in chemistry and drug design*. Weinheim: Wiley-VCH.

Zvinavashe, E., Murk, A. J., & Rietjens, I. M. (2008). Promises and pitfalls of quantitative structure-activity relationship approaches for predicting metabolism and toxicity. *Chemical Research in Toxicology, 18*, 844.

Zvinavashe, E., Murk, A. J., & Rietjens, I. M. (2009). On the number of EINECS compounds that can be covered by (Q)SAR models for acute toxicity. *Toxicology Letters, 184*, 67.