

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/271530042>

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

CHAPTER · JANUARY 2011

DOI: 10.1007/978-94-007-0711-5_37

READS

50

4 AUTHORS:



[Alexander Golbraikh](#)

University of North Carolina at Chapel Hill

57 PUBLICATIONS 3,840 CITATIONS

SEE PROFILE



[Xiang Simon Wang](#)

Howard University

62 PUBLICATIONS 416 CITATIONS

SEE PROFILE



[Hao Zhu](#)

Rutgers, The State University of New Jersey

44 PUBLICATIONS 1,080 CITATIONS

SEE PROFILE



[Alexander Tropsha](#)

University of North Carolina at Chapel Hill

240 PUBLICATIONS 10,070 CITATIONS

SEE PROFILE

Metadata of the chapter that will be visualized online

Book Title		Handbook of Computational Chemistry
Chapter Number		38
Book Copyright Year		2011
Copyright Holder		Springer-Verlag GmbH Berlin Heidelberg
Title		Predictive QSAR Modeling: Methods and Applications in Drug Discovery and Chemical Risk Assessment
Author	Degree	Dr.
	Given Name	Alexander
	Particle	
	Family Name	Golbraikh
	Suffix	
	Phone	
	Fax	
	Email	golbraik@email.unc.edu
Affiliation	Division	Laboratory for Molecular Modeling and Carolina Center for Exploratory Cheminformatics Research, Division of Medicinal Chemistry and Natural Products
	Organization	UNC Eshelman School of Pharmacy, University of North Carolina
	Street	CB # 7568 Beard Hall
	Postcode	NC 27599
	City	Chapel Hill
	State	North Carolina
	Country	USA
Author	Degree	Dr.
	Given Name	Xiang Simon
	Particle	
	Family Name	Wang
	Suffix	
	Phone	
	Fax	
	Email	xswang@email.unc.edu
Affiliation	Division	Laboratory for Molecular Modeling and Carolina Center for Exploratory Cheminformatics Research, Division of Medicinal Chemistry and Natural Products
	Organization	UNC Eshelman School of Pharmacy, University of North Carolina
	Street	CB # 7568 Beard Hall
	Postcode	NC 27599
	City	Chapel Hill
	State	North Carolina
	Country	USA

Author	Degree	Dr.
	Given Name	Hao
	Particle	
	Family Name	Zhu
	Suffix	
	Phone	
	Fax	
	Email	haozhu@email.unc.edu
	Division	Laboratory for Molecular Modeling and Carolina Center for Exploratory Chemin- formatics Research, Division of Medicinal Chemistry and Natural Products
	Organization	UNC Eshelman School of Pharmacy, University of North Carolina
Affiliation	Street	CB # 7568 Beard Hall
	Postcode	NC 27599
	City	Chapel Hill
	State	North Carolina
	Country	USA
Author	Degree	Dr.
	Given Name	Alexander
	Particle	
	Family Name	Tropsha
	Suffix	
	Phone	919-966-2955
	Fax	919-966-0204
	Email	alex_tropsha@unc.edu
	Division	Division of Medicinal Chemistry and Natural Products
	Organization	UNC Eshelman School of Pharmacy, University of North Carolina
Affiliation	Street	CB # 7568 Beard Hall
	Postcode	NC 27599
	City	Chapel Hill
	State	North Carolina
	Country	USA

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

AU1

Alexander Golbraikh¹ · Xiang Simon Wang¹ · Hao Zhu¹ · Alexander Tropsha²

¹Laboratory 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, North Carolina, USA

²Division of Medicinal Chemistry and Natural Products, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, North Carolina, USA

1	QSAR Methodology: Summary of Approaches for Model	
2	Building and Validation	3
3	Data Preparation	3
4	The Problem of Outliers	7
5	QSAR Model Development	7
6	QSAR Methods	8
7	Target Functions	10
8	Continuous QSAR Models	11
9	Target Functions and Validation Criteria for Classification QSAR Models	11
10	Target Functions and Validation Criteria for Category QSAR Models	12
11	Applicability Domains	13
12	Y-randomization	14
13	External Validation	15
14	"Good Practices" in QSAR Modeling: Examples of Models and Their Application to Virtual	
15	Screening and Lead Identification	15
16	QSAR-Aided Discovery of Novel Anticonvulsant Compounds	16
17	QSAR-Enabled Discovery of Novel Anticancer Agents	18
18	QSAR Enabled Discovery of Novel Geranylgeranyltransferase I Inhibitors (GGTIs)	18
19	"Good Practices" in QSAR Model Development: Applications to Toxicity Modeling	20
20	Quantitative Structure In Vitro–In Vivo Relationship Modeling	21
21	Using "Hybrid" Descriptors for QSIIR Modeling of Rodent Carcinogenicity	22
22	Using "Hybrid" Descriptors for the QSIIR Modeling of Rodent Acute Toxicity	23

38

2

Predictive QSAR Modeling

23	Collaborative and Consensus Modeling of Aquatic Toxicity	24
24	Universal Statistical Figures of Merit for All Models	25
25	Consensus QSAR Models of Aquatic Toxicity; comparison Between	
26	Methods and Models	26
27	<i>Conclusions: Emerging Chemical/Biological Data and QSAR Research Strategies</i>	27
28	<i>References</i>	28

Uncorrected Proof

AU2

AU3

Quantitative structure–activity relationship (QSAR) modeling is the major cheminformatics 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.

44 QSAR Methodology: Summary of Approaches for Model 45 Building and Validation 46

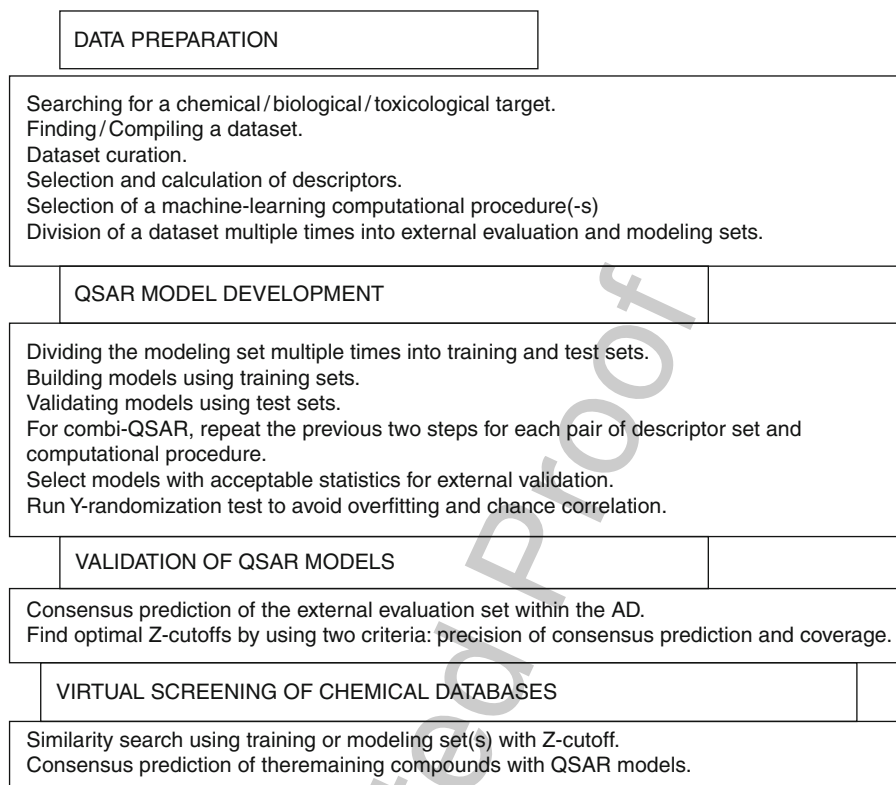
In order to find new leads in the process of drug design and discovery, there is a need for efficient and robust computational procedures that can be used to screen chemical databases and virtual libraries against molecules with known activities or properties. For this purpose, quantitative structure–activity 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:

$$47 \quad \text{Predicted Biological Activity} = \text{Function (Chemical Structure)} \quad (38.1)$$

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. The process of QSAR model development can be divided into three parts: data preparation, data analysis, and model validation (● Fig. 38-1). 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.

67 Data Preparation 68

The first part of QSAR analysis includes selection of a molecular dataset for QSAR studies, acquiring or calculation of molecular descriptors (quantities characterizing molecular



■ Fig. 38-1
Major steps of QSAR modeling

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.

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

93 Prior to QSAR modeling, a dataset should be curated, that is, all structures should be veri-
94 fied with respect to their correct representation in the dataset; structures containing atoms, for
95 which there are no parameters for descriptor calculation should be removed; structures con-
96 sisting of several disconnected parts should be removed; salts should be removed; a problem of
97 isomerism should be addressed; and duplicate structures should be removed. There are differ-
98 ent tools available for dataset curation. For example, Molecular Operating Environment (MOE)
99 (2008) includes DatabaseWash tool. It allows changing molecules' names, adding or removing
100 hydrogen atoms, removing salts and heavy atoms, even if they are covalently connected to the
101 rest of the molecule, and changing or generating the tautomers and protomers (cf. the MOE
102 manual for more details). Various database curation tools are included in ChemAxon (2008)
103 as well. If commercial software tools such as MOE are unavailable (notably, ChemAxon soft-
104 ware is free to academic investigators), one can use standard UNIX/LINUX tools to perform
105 some of the dataset cleaning tasks (Tropsha and Golbraikh 2010). It is important to have some
106 freely available molecular format converters such as OpenBabel (2010) or MolConverter from
107 ChemAxon (2008). Major procedures for database curation are discussed in our recent paper
108 (Fourches et al. 2010).

109 After the dataset is selected and curated, the next task is the acquisition or calculation of
110 descriptors. According to an excellent monograph titled *Handbook of Molecular Descriptors* by
111 Roberto Todeschini and Vivian Consonni (2000) molecular descriptors can be grouped into
112 zero-dimensional [0D] (sometimes referred to as constitutional descriptors), one-dimensional
113 [1D] (e.g., counts of different molecular groups, physicochemical properties of compounds, etc.),
114 two-dimensional [2D] (invariants of molecular graphs, e.g., connectivity indices, information
115 indices, counts of paths and walks, etc.), three-dimensional [3D], which are based on geo-
116 metrical spatial properties of molecules [e.g., Comparative Molecular Field Analysis (CoMFA)
117 descriptors (Tripos 2010) which are values of steric and electrostatic fields around aligned
118 molecules, and different CoMFA-like descriptors (Klebe 1998; Kubinyi et al. 1998; Robinson
119 et al. 1999)], and some other descriptors. Some descriptors can be experimental or calculated
120 physicochemical properties of molecules such as molecular weight, molar refraction, energies
121 of HOMO and LUMO, normal boiling point, octanol/water partition coefficient, molecular
122 surface, molecular volume, etc.

123 Herein, we will not discuss different types of descriptors in detail but mention briefly major
124 descriptor software. Most of descriptors included in the *Handbook of Molecular Descriptors*
125 (Todeschini and Consonni 2000) can be calculated by the Dragon software (Dragon 2007).
126 Molconn-Z (2007) is another widely used descriptor calculation software which calculates more
127 than 800 descriptors. A relatively small, but diverse set of molecular descriptors can be cal-
128 culated by the MOE (2008) software. Chirality molecular topological descriptors (CMTDs)
129 developed in our laboratory append 2D descriptors by conformation-independent chirality and
130 ZE-isomerism topological indices (Golbraikh and Tropsha 2003; Golbraikh et al. 2001, 2002).
131 Another group of descriptors frequently used in our laboratory is atom-pair (AP) descriptors
132 (Carhart et al. 1985). Each descriptor is defined as a count of pairs of atoms of certain types being
133 away from each other on a certain topological distance (2D AP descriptors) or a Euclidean dis-
134 tance within certain intervals (3D AP descriptors); chirality AP descriptors can be calculated as
135 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 38-1), in which each compound is a row and each descriptor as well as activity is a column.

■ Table 38-1

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

The Problem of Outliers

183
184

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 QSAR studies: Similar compounds are 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

206
207

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. 38-2. 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

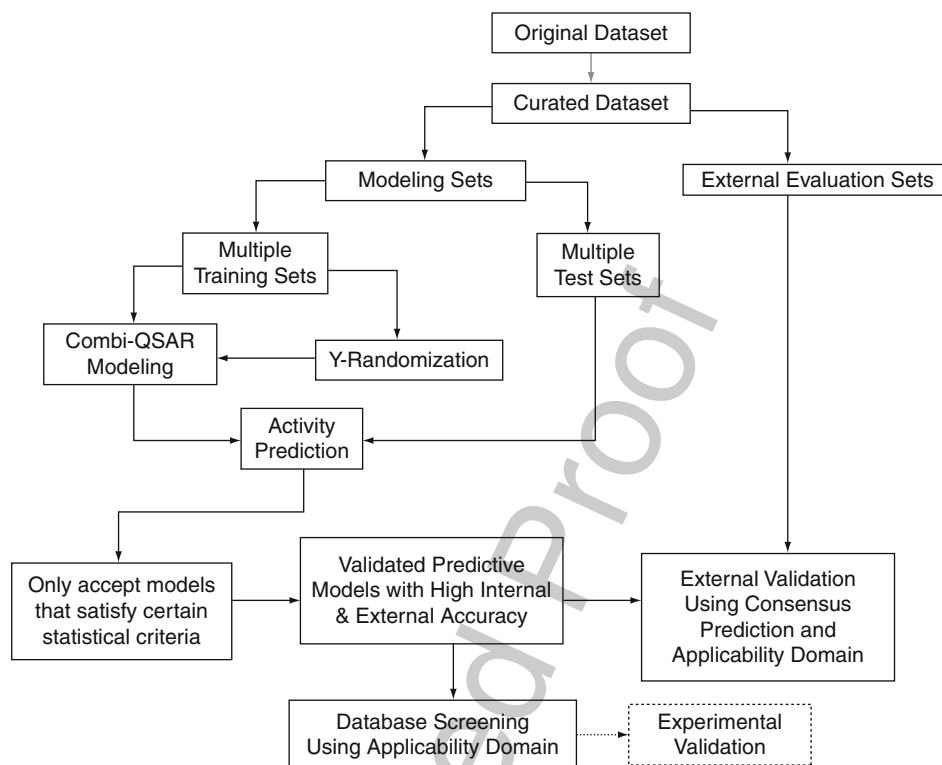
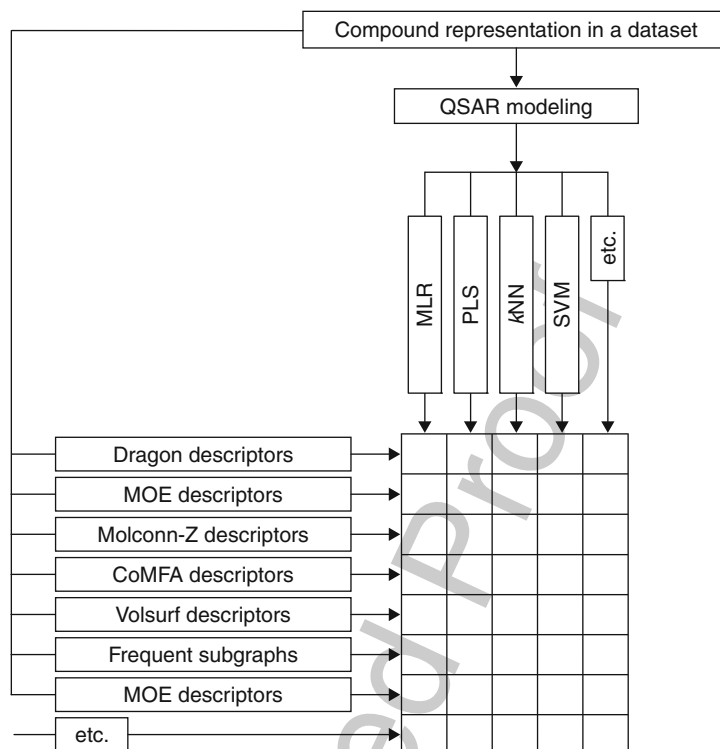


Fig. 38-2
Predictive QSAR modeling workflow

supervised data analysis techniques (combi-QSAR) (Fig. 38-3) 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 nonlinear approaches. Linear methods include simple and



■ Fig. 38-3
Combinatorial QSAR modeling

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. 38.1) 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.

297 Continuous QSAR Models

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

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

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

$$319 \quad R_{adj}^2 = 1 - (1 - R^2) \frac{n - 1}{n - c - 1}, \quad (38.2)$$

320 where n is the number of compounds in the dataset, and c is the number of variables (descrip-
 321 tors or principal components) included in the regression equation. It should be recognized that
 322 $R_{adj}^2 \leq R^2$. The higher the number of explanatory variables c is, the lower R_{adj}^2 is. R_{adj}^2 is partic-
 323 ularly important for linear QSAR models developed with variable selection. R_{adj}^2 is not a good
 324 criterion for variable selection k NN QSAR models, since contrary to regression methods, in
 325 the k NN algorithm descriptors are just selected or not selected, that is, their weights are either
 326 zero or one. As a result, much larger set of descriptors is selected by the k NN procedure than,
 327 for example, by stepwise regression.

328 Target Functions and Validation Criteria for Classification QSAR Models

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

$$331 \quad CCR_{\text{class}} = \frac{N_{\text{class}}^{\text{corr}}}{N_{\text{class}}^{\text{total}}} \quad (38.3)$$

332 and the p -value for each CCR_{class} value is not higher than a predefined threshold (in case of
 333 two classes, the CCR_{class} threshold should not be lower than 0.65–0.70, and generally, for any
 334 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

$$CCR = \frac{1}{K} \sum_{i=1}^K CCR_i = \frac{1}{K} \sum_{i=1}^K \frac{N_k^{\text{corr}}}{N_k^{\text{total}}} \quad (38.4)$$

AU4

along with the correct classification rate for each class (see Eq. 38.2). Criterion (38.4) 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 $N(\text{corr})/N(\text{total})$, where $N(\text{corr})$ and $N(\text{total})$ 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. 38.4, and at the same time it should be penalized by too high differences between CCR values for different classes.

344 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 ($i, j = 1, \dots, K$). Then the total error is calculated as follows:

$$E = \sum_{i=1}^K \sum_{j=1}^K n_{ij} f(|i - j|). \quad (38.5)$$

where $f(|i - j|)$ 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:

$$E = \sum_{i=1}^K \frac{1}{N_i} \sum_{j=1}^K n_{ij} f(|i - j|). \quad (38.6)$$

where N_i is the number of compounds of category i . QSAR procedure should minimize the total error of prediction calculated with Eqs. 38.5 or 38.6. In practice, the accuracy can be defined as $A = 1 - E/E_{\text{exp}}$, where E_{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.

373 Applicability Domains

374 Here we are approaching an extremely important problem of QSAR studies: model applicability
375 domain (AD). Formally, a QSAR model can predict the target property for any compound for
376 which chemical descriptors can be calculated. However, if a compound is highly dissimilar from
377 all compounds of the modeling set, reliable prediction of its activity is unlikely to be realized.
378 A concept of AD was developed and used to avoid such an unjustified extrapolation in activity
379 prediction. Applicability domains are one of the areas of intensive research. Different methods
380 of defining AD exist. Among others, the following definitions are considered by Jaworska and
381 colleagues (2005, 2008).

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

388 *Geometric Methods: Convex Hull AD.* AD is defined as a convex hull of points in the multidimensional descriptor space (Fechner et al. 2008).

390 The drawbacks of these definitions are as follows. Generally, the representative points are
391 distributed not in the entire hyperparallelepiped or convex hull, but only in a small part of it.
392 Another drawback is that structural outliers in the dataset can enormously increase the size of
393 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
395 compounds, the convex hull is not unique.

397 *Leverage-based AD.* 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
399 its leverage L is higher than $3K/N$, where K is the number of descriptors and N is the number
400 of compounds. The drawbacks of the leverage-based AD are as follows. (a) for each external
401 compound, it is necessary to recalculate leverage; (b) if there are cavities in the representative
402 point distribution area, a query compound the representative point of which is in this area will
403 be considered to be within the AD, while in fact it is far from all other compounds (Tropsha
404 and Golbraikh 2010).

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

$$408 \quad DT = \bar{y} + Z\sigma \quad (38.7)$$

409 Here, \bar{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
414 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

416 descriptor space. In this case, the same [Eq. 38.7](#) is used, $k = 1$, $Z = 0.5$, and Euclidean dis-
 417 tances are calculated using all descriptors. Thus, if the distance of the external compound from
 418 its nearest neighbor in the training set within either the entire descriptor space or the selected
 419 descriptor space exceeds these thresholds, the prediction is not made. We have also investigated
 420 changes of predictive power by changing the values of Z -cutoff. We have found that in general,
 421 starting from some Z -cutoff value, predictive power decreases while Z -cutoff value increases
 422 (Zhu et al. 2009), as expected. Instead of Euclidean distances, other distances and similarity
 423 measures can be used.

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

441 **Y-randomization**

442 To establish model robustness, Y-randomization (randomization of the response variable) test
 443 should be used. This test consists of repeating all the calculations with scrambled activities of
 444 the training set. Ideally, calculations should be repeated at least five (better, more) times. The
 445 goal of this procedure is to establish whether models built with real activities of the training
 446 set have good statistics not due to overfitting or chance correlation. If predictive power for the
 447 training or the test set of all models built with randomized activities of the training set is signif-
 448 icantly lower than that of models built with real activities of the training set, the latter ones are
 449 considered reliable. Using different parameters of the model development procedure, multiple
 450 QSAR models are built which have acceptable statistics. Suppose, the number of these models
 451 is m . Y-randomization test can also give n models with acceptable statistics. For acceptance of
 452 models developed with real activities of the training set, the condition $n \ll m$ should be satis-
 453 fied. In (Kovatcheva et al. 2005) and (de Cerqueira et al. 2006), we have introduced the measure
 454 of robustness $R = 1 - n/m$. If $R > 0.9$, the models are considered robust and their high pre-
 455 dictive accuracy cannot be explained by the chance correlation or overfitting. Y-randomization
 456 test is particularly important for small datasets. Unfortunately, in many publications on QSAR
 457 studies, Y-randomization test is not carried out but all QSAR practitioners must be strongly
 458 encouraged to use this simple procedure.

459 External Validation

460 Our previous experience suggests that the consensus prediction, which is the average of pre-
461 dicted activities over all predictive models, always provides the most stable results (Zhang et al.
462 2008; Zhu et al. 2008), and thus naturally avoids the need for (the best) model selection based
463 on the statistics for the training and test sets. The consensus prediction of biological activity for
464 an external compound on the basis of several QSAR models is more reliable and provides better
465 justification for the experimental exploration of hits.

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

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

489 We have used consensus prediction in many studies (de Cerqueira et al. 2006; Kovatcheva
490 et al. 2005; Shen et al. 2004; Votano et al. 2004; Zhang et al. 2007, 2008; Zhu et al. 2008) and
491 have shown that in most cases it gives better prediction and coverage than most of the individ-
492 ual predictive models. Thus, we recommend using consensus prediction for virtual screening
493 of chemical databases and combinatorial libraries for finding new lead compounds for drug
494 discovery.

495 "Good Practices" in QSAR Modeling: Examples of Models and Their 496 Application to Virtual Screening and Lead Identification

498 As discussed above, our experience in QSAR model development and validation has led us
499 to establishing a complex but straightforward workflow summarized in Fig. 38-2. The last
500 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. 38-2. We present several examples of these studies below to illustrate the use of QSAR models as virtual screening tools for lead identification.

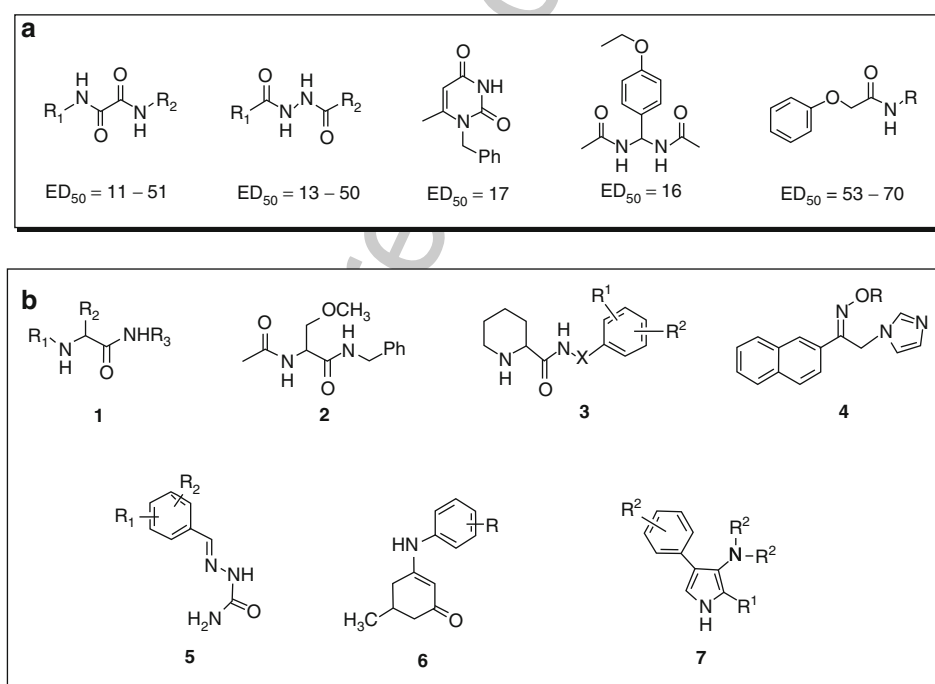
QSAR-Aided Discovery of Novel Anticonvulsant Compounds

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

545 additional compounds, which were close analogs of these two were either taken from the litera-
 546 ture or designed in our collaborator's laboratory. In total, seven compounds were resynthesized
 547 and sent to the NIH for the Maximum Electroshock (MES) test (a standard test for the anticon-
 548 vulsant activity, which was used for the training set compounds as well). The biological results
 549 indicated that upon initial and secondary screening, five out of seven compounds tested showed
 550 anticonvulsant activity with ED_{50} less than 100 mg/kg, which is considered promising. Interest-
 551 ingly, all seven compounds were also found to be very active in the same tests performed on rats
 552 (a complete set of experimental data on rats for the training set were not available, and therefore
 553 no QSAR models for rats were built).

554 Mining of the Maybridge database yielded two additional promising compounds that were
 555 synthesized and sent to the NIH for the MES anticonvulsant test. One of the compounds showed
 556 moderate anticonvulsant activity of ED_{50} between 30 and 100 mg/kg (in mice), while the other
 557 was found to be a *very* potent anticonvulsant agent with ED_{50} of 18 mg/kg in mice (ip). In sum-
 558 mary, both compounds were found to be very active in both mice and rats.  [Figure 38-4](#) shows
 559 chemical structures of experimentally confirmed hits that were identified by using validated
 560 QSAR models for virtual screening as applied to the anticonvulsant dataset. It is important
 561 to note that *none* of the compounds identified in external databases as potent anticonvulsants
 562 and validated experimentally belong to the same class of FAA molecules as the training set.
 563 This observation was very stimulating because it underscored the power of our methodology



■ Fig. 38-4

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

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.

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 phenanthrene-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 μ M implying an exceptionally high hit rate (80%). The same ten models were further applied to predict EC_{50} 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; Sebtì 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 (Sebtì 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. 38-3); the specific workflow as applied to the GGTI dataset is shown in Fig. 38-5. 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 *k*NN 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 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.

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. 38-6).

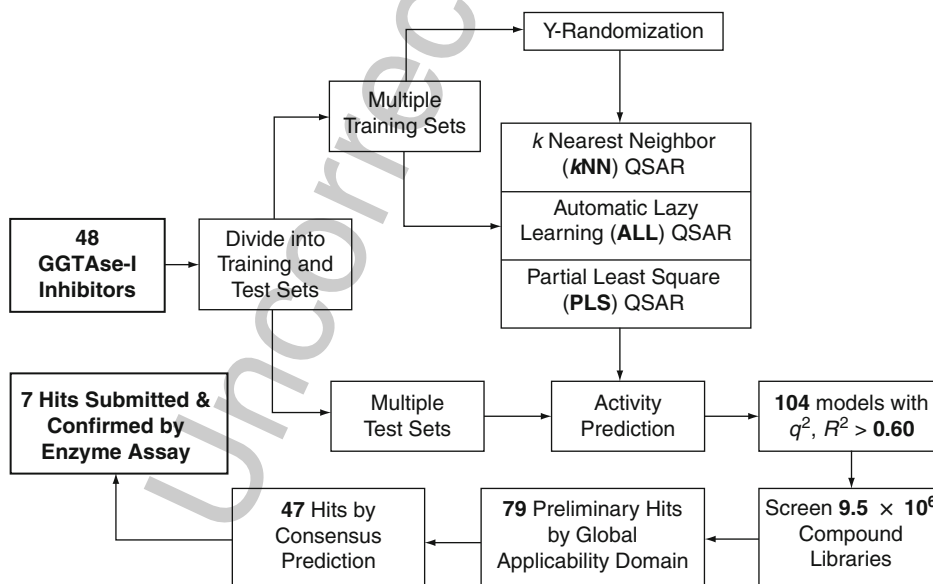
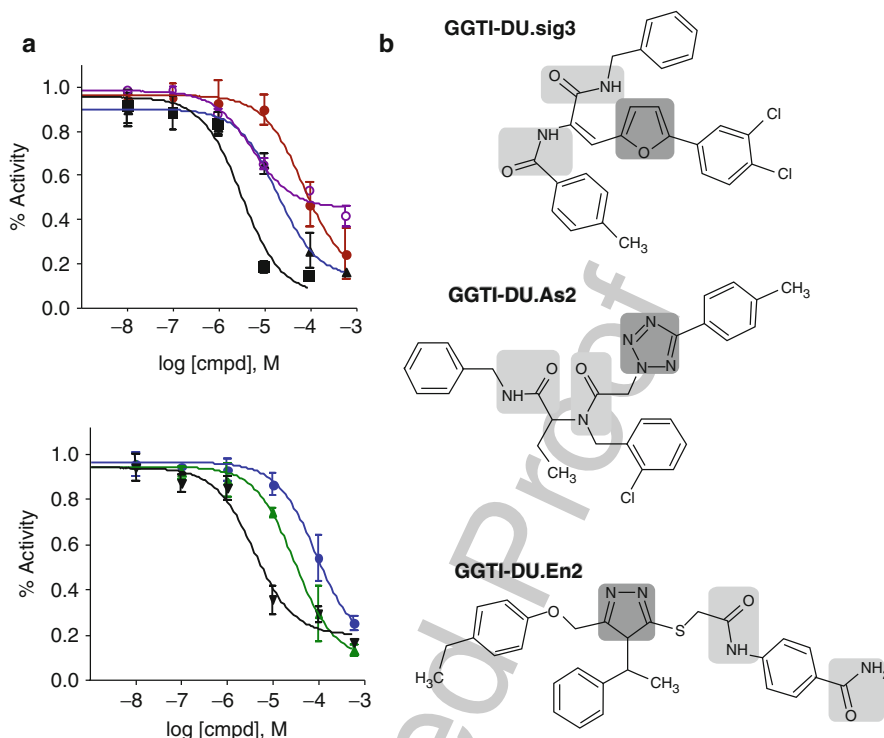


Fig. 38-5
The predictive QSAR modeling workflow illustrated for GGTIs



■ Fig. 38-6

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

630 The unexpected result was to identify several predicted actives that did not have a common
 631 ring feature in their structure. In fact, seven highly ranked hits had no apparent relationship with
 632 any of the training set molecules. They had furan, triazole, tetrazole, and pyridine cores in their
 633 scaffolds while all non-peptidomimetic compounds of the training set were based on a pyrazole
 634 core. Therefore, the seven hit compounds appeared to be the structurally novel hits. Figure
 635 38-6b shows chemical structures of the three representative confirmed hits with novel scaffolds
 636 highlighted. This study reconfirmed the observation that we already emphasized earlier with
 637 anticonvulsant compounds that contrary to the common belief, QSAR-based virtual screening
 638 is capable of identifying experimentally confirmed hit compounds with novel scaffolds.

639 "Good Practices" in QSAR Model Development: Applications to 640 Toxicity Modeling 641

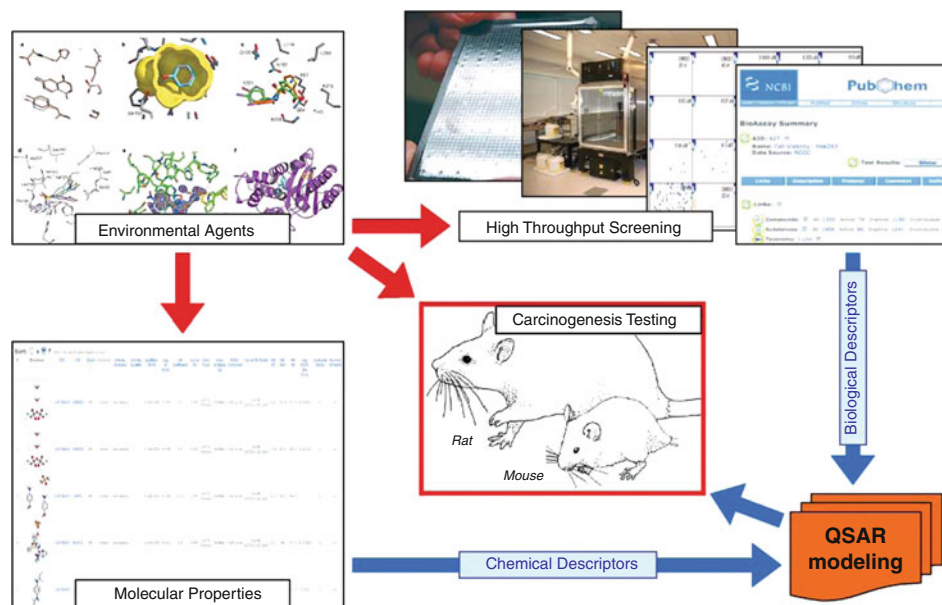
642 Many compounds entering clinical studies do not survive as a good pharmacological lead to
 643 become a marketed drug. Chemical toxicity and safety have been regarded as the major reason
 644 for attrition in the past decades (Kola and Landis 2004). However, evaluation of chemical

645 toxicity and safety in vivo at the early stage of drug discovery process is expensive and time
646 consuming. To replace the traditional animal toxicity testing and to understand the relevant
647 toxicological mechanisms, many in vitro toxicity screens and computational toxicity models
648 have been developed and implemented by academic institutes and pharmaceutical companies
649 (Cheeseman 2005; Dash et al. 2009; Dix et al. 2007; Inglese et al. 2006; Park et al. 2009; Riley and
650 Kenna 2004; Valerio 2009; Yang et al. 2009). In the past 15 years, innovative technologies that
651 enable rapid synthesis and high throughput screening of large libraries of compounds have been
652 adopted for toxicity studies. As a result, there has been a huge increase in the number of com-
653 pounds and the associated testing data in different in vitro screens. With this data, it becomes
654 feasible to reveal the relationship between the high throughput in vitro toxicity testing results
655 and the low throughput in vivo low dose toxicity evaluation for the same set of compounds.
656 Understanding these relationships could help us delineate the mechanisms underlying animal
657 toxicity of chemicals as well as potentially improve our ability to predict chemical toxicity using
658 short-term bioassays.

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

676 Quantitative Structure In Vitro–In Vivo Relationship Modeling

678 To stress a broad appeal of the conventional QSAR approach, it should be made clear that from
679 the statistical viewpoint QSAR modeling is a special case of general statistical data mining and
680 data modeling where the data is formatted to represent objects described by multiple descriptors
681 and the robust correlation between descriptors and a target property (e.g., chemical toxicity in
682 vivo) is sought. In previous computational toxicology studies, additional physicochemical prop-
683 erties, such as water partition coefficient (logP) (Klopman et al. 2003), water solubility (Stoner
684 et al. 2004), and melting point (Mayer and Reichenberg 2006) were used successfully to aug-
685 ment computed chemical descriptors and improve the predictive power of QSAR models. These
686 studies suggest that using experimental results as descriptors in QSAR modeling could prove
687 beneficial. The already available and rapidly growing HTS data for large and diverse chemical
688 libraries makes it possible to extend the scope of the conventional QSAR in toxicity studies
689 by using in vitro testing results as additional biological descriptors. Therefore, in some of the



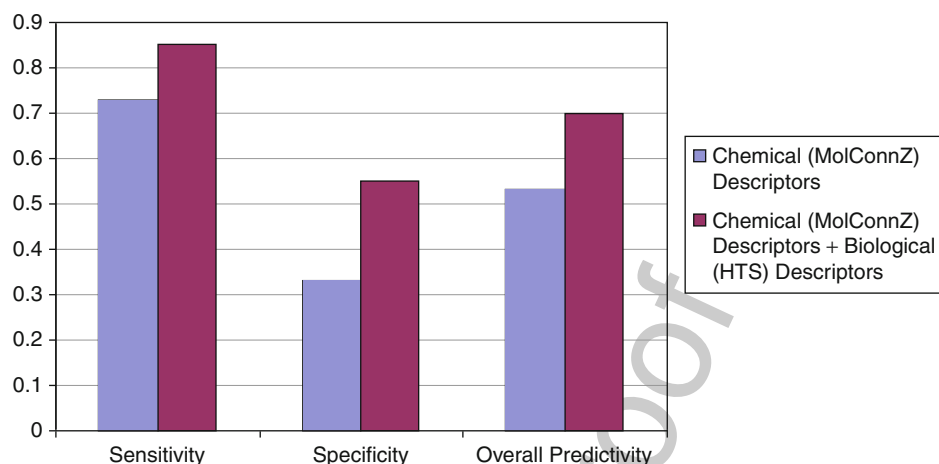
■ Fig. 38-7

Combining chemical and biological profiles as descriptors in QSIIR modeling of chemical carcinogenicity

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; Sjoström et al. 2008). Based on these reports, we proposed a new modeling workflow called Quantitative Structure In vitro–In vivo Relationship (QSIIR) 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 38-7 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



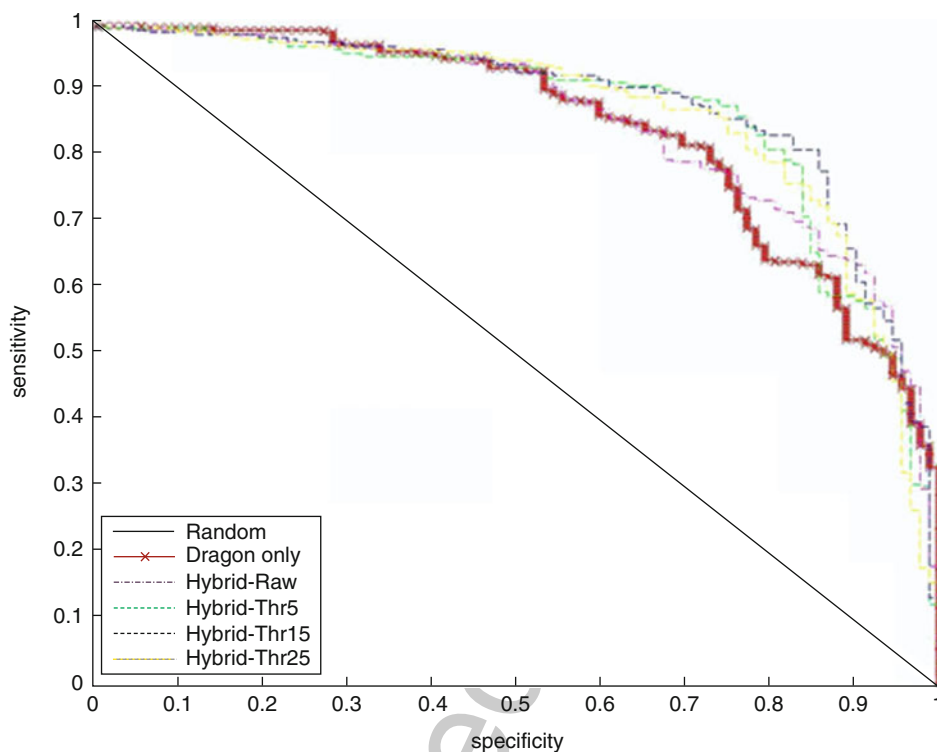
■ Fig. 38-8

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

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. 38-8). 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.

723 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). 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 contrast, the prediction accuracy was significantly improved to 85% when using hybrid descriptors.



■ Fig. 38-9

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

733 The receiver operating characteristic (ROC) curves of conventional QSAR models and different
 734 hybrid models are shown in Fig. 38-9. The sensitivities and specificities of hybrid models are
 735 clearly better than for conventional QSAR model for predicting the same external compounds.
 736 Furthermore, the prediction coverage increased from 76% when using chemical descriptors
 737 only to 93% when qHTS biological descriptors were also included. Our studies suggest that
 738 combining HTS profiles, especially the dose-response qHTS results, with conventional chemical
 739 descriptors could considerably improve the predictive power of computational approaches
 740 for rodent acute toxicity assessment.

741 Collaborative and Consensus Modeling of Aquatic Toxicity

742
 743 We discuss below the results of a recent important study of aquatic toxicity (Zhu et al. 2008). In
 744 our opinion, this particular study may serve as a useful example to illustrate the complexity and
 745 power of modern QSAR modeling approaches and highlight the importance of collaborative
 746 and consensual model development.

747 The combinational QSAR modeling approach has been applied to a diverse series of organic
 748 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, b). 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_{abs}^2 (squared leave-one-out cross-validation correlation coefficient) for the modeling set, R_{abs}^2 (frequently described as coefficient of determination) for the external validations sets, and MAE (mean absolute error) for the linear correlation between predicted (Y_{pred}) and experimental (Y_{exp}) data (here, $Y = pIGC_{50}$); these parameters are defined as follows:

$$Q_{abs}^2 = 1 - \frac{\sum_Y (Y_{exp} - Y_{LOO})^2}{\sum_Y (Y_{exp} - \langle Y \rangle_{exp})^2} \quad (38.8)$$

$$R_{abs}^2 = 1 - \frac{\sum_Y (Y_{exp} - Y_{pred})^2}{\sum_Y (Y_{exp} - \langle Y \rangle_{exp})^2} \quad (38.9)$$

$$MAE = \frac{\sum_Y |Y - Y_{pred}|}{n} \quad (38.10)$$

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_{abs}^2 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_{abs}^2 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

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

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

844 **Conclusions: Emerging Chemical/Biological Data and QSAR** 845 **Research Strategies** 846

847 In the past 15 years, innovative technologies that enable rapid synthesis and high throughput
848 screening of large libraries of compounds have been adopted in almost all major pharmaceutical
849 and biotech companies. As a result, there has been a huge increase in the number of com-
850 pounds available on a routine basis to quickly screen for novel drug candidates against new
851 targets or pathways. In contrast, such technologies have rarely become available to the aca-
852 demic research community, thus limiting its ability to conduct large-scale chemical genetics
853 or chemical genomics research. The NIH Molecular Libraries Roadmap Initiative has changed
854 this situation by forming the national Molecular Library Screening Centers Network (MLSCN)
855 (Austin et al. 2004) with the results of screening assays made publicly available via PubChem
856 (2010). These efforts have already led to the unprecedented growth of *available* databases of bio-
857 logically tested compounds [cf. our recent review where we list about 20 available databases of
858 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).

References

- (1997). *Addressing the curse of imbalanced training sets: One sided selection*. San Francisco: Morgan Kaufmann.
- Afantitis, A., Melagraki, G., Sarimveis, H., Koutentis, P. A., Markopoulos, J., & Iggleksi-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. *The 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. *The 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 trihydroxy-benzenes 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).

AUS

AU6

- 931 Carhart, R. E, Smith, D. H., & Venkataraghavan, 984
 932 R. (1985). Atom pairs as molecular features in 985
 933 structure-activity studies: Definition and appli- 986
 934 cations. *The Journal of Chemical Information and* 987
 935 *Computer Science*, 25, 64. 988
- 936 Casey, P. J., & Seabra, M. C. (1996). Protein prenyl- 989
 937 transferases. *The Journal of Biological Chemistry*, 990
 938 271, 5289. 991
- 939 Cheeseman, M. A. (2005). Thresholds as a unifying 992
 940 theme in regulatory toxicology. *Food Additives &* 993
 941 *Contaminants*, 22, 900. 994
- AU7 942 ChemAxon. (2008). <http://www.chemaxon.com>. 995
 943 ChEMBL Database. (2010). [http://www.ebi.ac.uk/](http://www.ebi.ac.uk/chembl/db/) 996
 944 [chembl/db/](http://www.ebi.ac.uk/chembl/db/). 997
- 945 Chen, C., Liaw, A., & Breiman, L. (2004). *Using* 998
 946 *random forest to learn imbalanced data*. 666. 999
 947 Berkeley: Department of Statistics, University of 1000
 948 California. 1001
- 949 Cherkasov, A. (2008). An updated steroid bench- 1002
 950 mark set and its application in the discovery 1003
 951 of novel nanomolar ligands of sex hormone- 1004
 952 binding globulin. *Journal of Medicinal Chem-* 1005
 953 *istry*, 51, 2047. 1006
- 954 Cho, S. J., Zheng, W., & Tropsha, A. (1998). Rational 1007
 955 combinatorial library design. 2. Rational design 1008
 956 of targeted combinatorial peptide libraries using 1009
 957 chemical similarity probe and the inverse QSAR 1010
 958 approaches. *The Journal of Chemical Information* 1011
 959 *and Computer Science*, 38, 259. 1012
- 960 Cox, A. D., & Der, C. J. (2002). Farnesyltransferase 1013
 961 inhibitors: Promises and realities. *Current Opin-* 1014
 962 *ion in Pharmacology*, 2, 388. 1015
- 963 Crivori, P., Cruciani, G., Carrupt, P. A., & Testa, B. 1016
 964 (2000). Predicting blood-brain barrier perme- 1017
 965 ation from three-dimensional molecular struc- 1018
 966 ture. *The Journal of Medicinal Chemistry*, 43, 1019
 967 2204. 1020
- 968 Cruciani, G., Pastor, M., & Guba, W. (2000). Vol- 1021
 969 Surf: A new tool for the pharmacokinetic opti- 1022
 970 mization of lead compounds. *The European* 1023
 971 *Journal of Pharmaceutical Sciences*, 11(Suppl 2), 1024
 972 S29–S39. 1025
- 973 Dash, A., Inman, W., Hoffmaster, K., Sevidal, S., 1026
 974 Kelly, J., Obach, R.S., et al. (2009). Liver tis- 1027
 975 sue engineering in the evaluation of drug safety. 1028
 976 *Expert Opinion in Drug Metabolism & Toxicol-* 1029
 977 *ogy*, 5, 1159. 1030
- 978 de Cerqueira, L. P., Golbraikh, A., Oloff, S., Xiao, Y., 1031
 979 & Tropsha, A. (2006). Combinatorial QSAR 1032
 980 modeling of P-Glycoprotein substrates. *The Jour-* 1033
 981 *nal of Chemical Information and Modeling*, 46, 1034
 982 1245. 1035
- AU8 983 Discovery Studio. (2010). 1036
 1037
- Dix, D. J., Houck, K. A., Martin, M. T., Richard, A. M., 984
 Setzer, R. W., & Kavlock, R. J. (2007). The Tox- 985
 Cast program for prioritizing toxicity testing of 986
 environmental chemicals. *Toxicological Sciences*, 987
 95, 5. 988
- Dragon. (2007). [http://www.taletemi.it/help/dragon](http://www.taletemi.it/help/dragon_help/index.html?IntroducingDRAGON) 989
[_help/index.html?IntroducingDRAGON](http://www.taletemi.it/help/dragon_help/index.html?IntroducingDRAGON). 990
- DSSTox. (2008). [http://www.epa.gov/nheerl/dsstox/](http://www.epa.gov/nheerl/dsstox/About.html) 991
[About.html](http://www.epa.gov/nheerl/dsstox/About.html). 992
- Durham, S. K., & Pearl, G. M. (2001). Computational 993
 methods to predict drug safety liabilities. *Cur-* 994
rent Opinion in Drug Discovery & Development, 995
 4, 110. 996
- Environmental Protection Agency. (1992). *Statistical* 997
training course for ground-water monitoring data 998
analysis EPA/530-R-93-003. Washington: Office 999
 of Solid Waste. 1000
- Fallon, A., Spada, C., & Gallagher, D. (1997). 1001
 Detection and Accommodation of Outliers in 1002
 Normally Distributed Data Sets. [http://ewr.cce.](http://ewr.cce.vt.edu/environmental/teach/smprimer/outlier/outlier.html) 1003
[vt.edu/environmental/teach/smprimer/outlier/](http://ewr.cce.vt.edu/environmental/teach/smprimer/outlier/outlier.html) 1004
[outlier.html](http://ewr.cce.vt.edu/environmental/teach/smprimer/outlier/outlier.html). Accessed 25 April 2005. 1005
- Fechner, N., Hinselmann, G., Schmiedl, C., & Zell, A. 1006
 (2008). Estimating the applicability domain 1007
 of kernel-based QSPR models using classical 1008
 descriptor vectors. *pdf. Chemistry Central Jour-* 1009
nal, 2(Suppl.1), P2. 1010
- Forsby, A., & Blaauboer, B. (2007). Integration of 1011
 in vitro neurotoxicity data with biokinetic mod- 1012
 elling for the estimation of in vivo neurotoxicity. 1013
Human & Experimental Toxicology, 26, 333. 1014
- Fourches, D., Muratov, E., & Tropsha, A. (2010). 1015
 Trust, but verify: On the importance of chemi- 1016
 cal structure curation in cheminformatics 1017
 and QSAR modeling research. *The Journal* 1018
of Chemical Information and Modeling, 50, 1019
 1189–1204. 1020
- Gasteiger, J. (2006). Of molecules and humans. *The* 1021
Journal of Medicinal Chemistry, 49, 6429. 1022
- Golbraikh, A., & Tropsha, A. (2002). Predictive 1023
 QSAR modeling based on diversity sampling of 1024
 experimental datasets for the training and test 1025
 set selection. *The Journal of Computer-Aided* 1026
Molecular Design, 16, 357. 1027
- Golbraikh, A., & Tropsha, A. (2003). QSAR modeling 1028
 using chirality descriptors derived from molec- 1029
 ular topology. *The Journal of Chemical Informa-* 1030
tion and Computer Science, 43, 144. 1031
- Golbraikh, A., Bonchev, D., & Tropsha, A. (2001). 1032
 Novel chirality descriptors derived from molec- 1033
 ular topology. *The Journal of Chemical Informa-* 1034
tion and Computer Science, 41, 147. 1035
- Golbraikh, A., Bonchev, D., & Tropsha, A. (2002). 1036
 Novel ZE-isomerism descriptors derived from 1037

- 1038 molecular topology and their application to
- 1039 QSAR analysis. *The Journal of Chemical Informa-*
- 1040 *tion and Computer Science*, 42, 769.
- 1041 Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y. D., Lee,
- 1042 K. H., & Tropsha, A. (2003). Rational selection
- 1043 of training and test sets for the development of
- 1044 validated QSAR models. *The Journal of*
- 1045 *Computer-Aided Molecular Design*, 17, 241.
- 1046 Gramatica, P. (2007). Principles of QSAR models
- 1047 validation: Internal and external. *Qsar & Com-*
- 1048 *binatorial Science*, 26, 694.
- 1049 Guha, R., & Van Drie, J. H. (2008a). Structure-
- 1050 activity landscape index: Identifying and quan-
- 1051 tifying activity cliffs. *The Journal of Chemical*
- 1052 *Information and Modeling*, 48, 646.
- 1053 Guha, R., & Van Drie, J. H. (2008b). Assessing how
- 1054 well a modeling protocol captures a structure-
- 1055 activity landscape. *The Journal of Chemical Infor-*
- 1056 *mation and Modeling*, 48, 1716.
- 1057 Hoffman, B., Cho, S. J., Zheng, W., Wyrick, S.,
- 1058 Nichols, D. E., Mailman, R. B., et al. (1999).
- 1059 Quantitative structure-activity relationship
- 1060 modeling of dopamine D(1) antagonists using
- 1061 comparative molecular field analysis, genetic
- 1062 algorithms-partial least-squares, and K nearest
- 1063 neighbor methods. *The Journal of Medicinal*
- 1064 *Chemistry*, 42, 3217.
- 1065 Hong, H., Xie, Q., Ge, W., Qian, F., Fang, H., Shi, L.,
- 1066 et al. (2008). Mold(2), molecular descriptors
- 1067 from 2D structures for chemoinformatics and
- 1068 toxicoinformatics. *The Journal of Chemical Infor-*
- 1069 *mation and Modeling*, 48, 1337.
- 1070 Horvath, D., Bonachera, F., Solov'ev, V., Gaudin, C.,
- 1071 & Varnek, A. (2007). Stochastic versus stepwise
- 1072 strategies for quantitative structure-activity
- 1073 relationship generation-how much effort
- 1074 may the mining for successful QSAR models
- 1075 take? *The Journal of Chemical Information and*
- 1076 *Modeling*, 47, 927.
- 1077 Hsieh, J. H., Wang, X. S., Teotico, D., Golbraikh, A.,
- 1078 & Tropsha, A. (2008). Differentiation of AmpC
- 1079 beta-lactamase binders vs. decoys using classi-
- 1080 fication kNN QSAR modeling and application
- 1081 of the QSAR classifier to virtual screening. *The*
- 1082 *Journal of Computer-Aided Molecular Design*, 22,
- 1083 593.
- 1084 Huan, J., Bandyopadhyay, D., Prins, J., Snoeyink, J.,
- 1085 Tropsha, A., & Wang, W. (2006). Distance-based
- 1086 identification of structure motifs in proteins
- 1087 using constrained frequent subgraph mining.
- 1088 *Computational Systems Bioinformatics Confer-*
- 1089 *ence*, 227.
- 1090 Inglese, J., Auld, D. S., Jadhav, A., Johnson, R.
- 1091 L., Simeonov, A., Yasgar, A., et al. (2006).
- 1092 Quantitative high-throughput screening: A
- titration-based approach that efficiently iden- 1093
- 1094 tifies biological activities in large chemical
- 1095 libraries. *Proceedings of the National Academy*
- 1096 *of Sciences of the United States of America*, 103,
- 11473.
- 1097
- 1098 Irwin, J. J., & Shoichet, B. K. (2005). ZINC-a free
- 1099 database of commercially available compounds
- 1100 for virtual screening. *The Journal of Chemical*
- 1101 *Information and Modeling*, 45, 177.
- 1102
- 1103 Jacobson-Kram, D., & Contrera, J. F. (2007). Genetic
- 1104 toxicity assessment: Employing the best science
- 1105 for human safety evaluation. Part I: Early screen-
- 1106 ing for potential human mutagens. *Toxicological*
- 1107 *Sciences*, 96, 16.
- 1108
- 1109 Jaworska, J., & Nikolova-Jeliazkova, N. (2008).
- 1110 Review of methods to assess a QSAR Appli-
- 1111 cability Domain. [http://ambit.acad.bg/nina/](http://ambit.acad.bg/nina/publications/2004/AppDomain_qsar04.ppt)
- 1112 [publications/2004/AppDomain_qsar04.ppt](http://ambit.acad.bg/nina/publications/2004/AppDomain_qsar04.ppt).
- 1113
- 1114 Jaworska, J., Nikolova-Jeliazkova, N., & Aldenberg, T.
- 1115 (2005). QSAR applicabilty domain estimation by
- 1116 projection of the training set descriptor space:
- 1117 A review. *Alternatives to Laboratory Animals*, 33,
- 1118 445.
- 1119
- 1120 Johnson, S. R. (2008). The trouble with QSAR (or
- 1121 how I learned to stop worrying and embrace fal-
- 1122 lacy). *The Journal of Chemical Information and*
- 1123 *Modeling*, 48, 25.
- 1124
- 1125 Klebe, G. (1998). Comparative molecular similarity
- 1126 indices: CoMSI. In H. Kubinyi, G. Folkers, &
- 1127 Y. Martin (Eds.), *3D QSAR in drug design* (pp.
- 1128 87-104). Great Britain: Kluwer.
- 1129
- 1130 Klopman, G., Zhu, H., Ecker, G., & Chiba, P. (2003).
- 1131 MCASE study of the multidrug resistance rever-
- 1132 sal activity of propafenone analogs. *The Journal*
- 1133 *of Computer-Aided Molecular Design*, 17, 291.
- 1134
- 1135 Kola, I., & Landis, J. (2004). Can the pharmaceutical
- 1136 industry reduce attrition rates? *Nature Reviews*
- 1137 *Drug Discovery*, 3, 711.
- 1138
- 1139 Kovatcheva, A., Golbraikh, A., Oloff, S., Feng, J.,
- 1140 Zheng, W., & Tropsha, A. (2005). QSAR model-
- 1141 ing of datasets with enantioselective compounds
- 1142 using chirality sensitive molecular descriptors.
- 1143 *SAR and QSAR in Environmental Research*,
- 1144 16, 93.
- 1145
- 1146 Kubinyi, H., Hamprecht, F. A., & Mietzner, T. (1998).
- 1147 Three-dimensional quantitative similarity-
- 1148 activity relationships (3D QSiAR) from SEAL
- 1149 similarity matrices. *The Journal of Medicinal*
- 1150 *Chemistry*, 41, 2553.
- 1151
- 1152 (2000). *Learning from imbalanced datasets: A com-*
- 1153 *parison of various strategies*. AAAI Workshop.
- 1154 Menlo Park: AAAI Press.
- 1155
- 1156 LigandScout. (2010).
- 1157
- 1158 Liu, T., Lin, Y., Wen, X., Jorissen, R. N., & Gilson,
- 1159 M. K. (2007). BindingDB: A web-accessible

AU10

AU11

AU12

- 1148 database of experimentally determined protein-
1149 ligand binding affinities. *Nucleic Acids Research*,
1150 35, D198–D201.
- 1151 Maggiora, G. M. (2006). On outliers and activity
1152 cliffs—why QSAR often disappoints. *The Journal*
1153 *of Medicinal Chemistry*, 46, 1535.
- 1154 Maybridge. (2005). [http://www.daylight.com/pro-](http://www.daylight.com/products/databases/Maybridge.html)
1155 [ducts/databases/Maybridge.html](http://www.daylight.com/products/databases/Maybridge.html).
- 1156 Mayer, P., & Reichenberg, F. (2006). Can highly
1157 hydrophobic organic substances cause aquatic
1158 baseline toxicity and can they contribute to
1159 mixture toxicity? *Environmental Toxicology &*
1160 *Chemistry*, 25, 2639.
- 1161 McGregor, M. J., & Pallai, P. V. (1997). Clustering of
1162 large databases of compounds: Using the MDL
1163 "Keys" as structural descriptors. *The Journal of*
1164 *Chemical Information and Computer Science*, 37,
1165 443.
- 1166 MDDR.SYMYX technologies. (2009). [http://www.](http://www.md.com/products/knowledge/drug_data_report/index.jsp)
1167 [md.com/products/knowledge/drug_data_re-](http://www.md.com/products/knowledge/drug_data_report/index.jsp)
1168 [port/index.jsp](http://www.md.com/products/knowledge/drug_data_report/index.jsp).
- 1169 Medina-Franco, J. L., Golbraikh, A., Oloff, S.,
1170 Castillo, R., & Tropsha, A. (2005). Quantitative
1171 structure-activity relationship analysis of pyridi-
1172 none HIV-1 reverse transcriptase inhibitors
1173 using the k nearest neighbor method and QSAR-
1174 based database mining. *The Journal of Computer-*
1175 *Aided Molecular Design*, 19, 229.
- 1176 Molconn-Z. (2007). <http://www.edusoft-lc.com/>.
- 1177 Molecular Operating Environment (MOE). (2008).
1178 <http://www.chemcomp.com/>.
- 1179 Muster, W., Breidenbach, A., Fischer, H., Kirchner, S.,
1180 Muller, L., & Pehler, A. (2008). Computational
1181 toxicology in drug development. *Drug Discovery*
1182 *Today*, 13, 303.
- 1183 nci. (2007). [http://dtp.nci.nih.gov/docs/3d_databa-](http://dtp.nci.nih.gov/docs/3d_database/structural_information/smiles_strings.html)
1184 [se/structural_information/smiles_strings.html](http://dtp.nci.nih.gov/docs/3d_database/structural_information/smiles_strings.html).
- 1185 Netzeva, T. I., Gallegos, S. A., & Worth, A. P. (2006).
1186 Comparison of the applicability domain of a
1187 quantitative structure-activity relationship for
1188 estrogenicity with a large chemical inventory.
1189 *Environmental Toxicology & Chemistry*, 25, 1223.
- 1190 Netzeva, T. I., & Schultz, T. W. (2005). QSARs for
1191 the aquatic toxicity of aromatic aldehydes from
1192 Tetrahymena data. *Chemosphere*, 61, 1632.
- 1193 (1996). *Neural networks in QSAR and drug design*. San
1194 Diego: Academic.
- 1195 Neural Networks. (2010). [http://www.](http://www.learnartificialneuralnetworks.com/)
1196 [learnartificialneuralnetworks.com/](http://www.learnartificialneuralnetworks.com/).
- 1197 Nikolova-Jeliazkova, N., & Jaworska, J. (2005). An
1198 approach to determining applicability domains
1199 for QSAR group contribution models: An analy-
1200 sis of SRC KOWWIN. *Alternatives to Laboratory*
1201 *Animals*, 33, 461.
- Olah, M., Rad, R., Ostopovici, L., Bora, A., Hadaruga,
1202 N., Hadaruga, D., et al. (2007). WOMBAT and
1203 WOMBAT-PK: Bioactivity databases for lead
1204 and drug discovery. In S. L. Schreiber, T. M.
1205 Kapoor, & G. Weiss (Eds.), *Chemical biology:*
1206 *From small molecules to systems biology and drug*
1207 *design* (pp. 760–786). Weinheim: Wiley-VCH.
1208
- Oloff, S., Mailman, R. B., & Tropsha, A. (2005).
1209 Application of validated QSAR models of D1
1210 dopaminergic antagonists for database min-
1211 ing. *The Journal of Medicinal Chemistry*, 48,
1212 7322.
- (2010). OpenBabel: The OpenSource Chemistry
1214 Toolbox. [Openbabel.org](http://openbabel.org). 2-1-2010.
1215
- Oprea, T., & Tropsha, A. (2006). Target, chemical and
1216 bioactivity databases – integration is key. *Drug*
1217 *Discovery Today*, 3, 357–365.
1218
- Organisation for Economic and Co-operation Devel-
1219 opment. (2008). OECD Quantitative Structure-
1220 Activity Relationships [(Q)SARs] Project. [http://](http://www.oecd.org/document/23/0,3343,en_2649_34365_33957015_1_1_1_1,00.html)
1221 [www.oecd.org/document/23/0,3343,en_2649_3](http://www.oecd.org/document/23/0,3343,en_2649_34365_33957015_1_1_1_1,00.html)
1222 [4365_33957015_1_1_1_1,00.html](http://www.oecd.org/document/23/0,3343,en_2649_34365_33957015_1_1_1_1,00.html).
1223
- Park, M. V., Lankveld, D. P., van, L. H., & de Jong,
1224 W. H. (2009). The status of in vitro toxicity
1225 studies in the risk assessment of nanomaterials.
1226 *Nanomedicine (Lond)*, 4, 669.
- Pastor, M., Cruciani, G., McLay, I., Pickett, S.,
1228 & Clementi, S. (2000). GRIND-INdependent
1229 descriptors (GRIND): A novel class of
1230 alignment-independent three-dimensional
1231 molecular descriptors. *The Journal of Medicinal*
1232 *Chemistry*, 43, 3233.
- PDSP. (2010). PDSP. <http://pdsp.med.unc.edu>.
1234
- Peterson, Y. K., Kelly, P., Weinbaum, C. A.,
1235 & Casey, P. J. (2006). A novel protein
1236 geranylgeranyltransferase-I inhibitor with
1237 high potency, selectivity, and cellular activ-
1238 ity. *The Journal of Biological Chemistry*, 281,
1239 12445.
- Peterson, Y. K., Wang, X. S., Casey, P. J., & Tropsha, A.
1241 (2009). Discovery of geranylgeranyltransferase-
1242 I inhibitors with novel scaffolds by the means
1243 of quantitative structure-activity relationship
1244 modeling, virtual screening, and experimental
1245 validation. *The Journal of Medicinal Chemistry*,
1246 52, 4210.
- Piersma, A. H., Janer, G., Wolterink, G., Bessems,
1248 J. G., Hakkert, B. C., & Slob, W. (2008). Quan-
1249 titative extrapolation of in vitro whole embryo
1250 culture embryotoxicity data to developmen-
1251 tal toxicity in vivo using the benchmark dose
1252 approach. *Toxicological Sciences*, 101, 91.
- PubChem. (2010). [http://pubchem.ncbi.nlm.nih.](http://pubchem.ncbi.nlm.nih.gov/)
1254 [gov/](http://pubchem.ncbi.nlm.nih.gov/).
1255

AU14

AU13

- 1256 Quinlan, J. R. (1993). *C4.5: Programs for machine*
1257 *learning*. San Mateo: Morgan Kaufmann.
- AU15 1258 Random Forests. (2001).
- 1259 Riley, R. J., & Kenna, J. G. (2004). Cellular mod-
1260 els for ADMET predictions and evaluation of
1261 drug-drug interactions. *Current Opinion in Drug*
1262 *Discovery & Development*, 7, 86.
- 1263 Robinson, D. D., Winn, P. J., Lyne, P. D., & Richards,
1264 W. G. (1999). Self-organizing molecular field
1265 analysis: A tool for structure-activity studies.
1266 *The Journal of Medicinal Chemistry*, 42, 573.
- 1267 Saliner, A. G., Netzeva, T. I., & Worth, A. P. (2006).
1268 Prediction of estrogenicity: Validation of a clas-
1269 sification model. *SAR and QSAR in Environmen-*
1270 *tal Research*, 17, 195.
- 1271 Salt, D. V., Yildiz, N., Livingstone, D. J., &
1272 Tinsley, C. J. (2006). The use of artificial neural
1273 networks in QSAR. *Pesticide Science*, 36, 161.
- 1274 Schirmer, K., Tanneberger, K., Kramer, N. I., Volker,
1275 D., Scholz, S., Hafner, C., et al. (2008). Devel-
1276 oping a list of reference chemicals for testing
1277 alternatives to whole fish toxicity tests. *Aquatic*
1278 *Toxicology*, 90, 128.
- AU16 1279 Schrodinger Software. (2010).
- 1280 Schultz, T. W. (1999). Structure-toxicity relationships
1281 for benzenes evaluated with *Tetrahymena pyri-*
1282 *formis*. *Chemical Research in Toxicology*, 12, 1262.
- 1283 Schultz, T. W., & Netzeva, T. I. (2004). Develop-
1284 ment and evaluation of QSARs for ecotoxic end-
1285 points: The benzene response-surface model for
1286 *Tetrahymena* toxicity. In M. T. D. Cronin & D. J.
1287 Livingstone (Eds.), *Modeling environmental fate*
1288 *and toxicity* (pp. 265–284). Boca Raton: CRC
1289 Press.
- 1290 Schultz, T. W., Cronin, M. T., Netzeva, T. I., & Aptula,
1291 A. O. (2002). Structure-toxicity relationships for
1292 aliphatic chemicals evaluated with *Tetrahymena*
1293 *pyriformis*. *Chemical Research in Toxicology*, 15,
1294 1602.
- 1295 Schultz, T. W., Hewitt, M., Netzeva, T. I., & Cronin,
1296 M. T. D. (2007). Assessing applicability domains
1297 of toxicological QSARs: Definition, confidence
1298 in predicted values, and the role of mechanisms
1299 of action. *QSAR & Combinatorial Science*, 26,
1300 238.
- 1301 Schultz, T. W., Netzeva, T. I., & Cronin, M. T.
1302 (2003). Selection of data sets for QSARs: Anal-
1303 yses of *Tetrahymena* toxicity from aromatic
1304 compounds. *SAR and QSAR in Environmental*
1305 *Research*, 14, 59.
- 1306 Schultz, T. W., Netzeva, T. I., Roberts, D. W.,
1307 & Cronin, M. T. (2005a). Structure-toxicity
1308 relationships for the effects to *Tetrahymena*
1309 *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., Sinks, G. D., & Miller, L. A.
(2001). Population growth impairment of sulfur-
containing compounds to *Tetrahymena pyri-*
formis. *Environmental Toxicology*, 16, 543.
- Sebti, S. M., & Hamilton, A. D. (2000). Farnesyltrans-
ferase and geranylgeranyltransferase I inhibitors
in cancer therapy: Important mechanistic and
bench to bedside issues. *Expert Opinion on Inves-*
tigational 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., Beguin, C., Golbraikh, A., Stables, J. P.,
Kohn, H., & Tropsha, A. (2004). Application
of predictive QSAR models to database min-
ing: Identification and experimental validation
of novel anticonvulsant compounds. *Journal of*
Medicinal Chemistry, 47, 2356.
- 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 meth-
ods. *The Journal of Medicinal Chemistry*, 45,
2811.
- Sisay, M. T., Peltason, L., Bajorath, J. (2009). Struc-
tural interpretation of activity cliffs revealed
by systematic analysis of structure-activity rela-
tionships in analog series. *The Journal of Chem-*
ical Information and Modeling, 49, 2179.
- Sjostrom, M., Kolman, A., Clemenson, C., & Cloth-
ier, 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 genotoxi-
city and carcinogenicity of marketed pharma-
ceuticals with reference to in silico predictivity.
Environmental and Molecular Mutagenesis, 50,
435.
- Stoner, C. L., Gifford, E., Stankovic, C., Lepsey, C. S.,
Broduehrer, J., Prasad, J. V. N. V., et al. (2004).
Implementation of an ADME enabling selection

- 1365 and visualization tool for drug discovery. *Journal*
1366 *of Pharmaceutical Sciences*, 93, 1131.
- 1367 Stouch, T. R., Kenyon, J. R., Johnson, S. R., Chen,
1368 X. Q., Doweyko, A., & Li, Y. (2003). In sil-
1369 ico ADME/Tox: why models fail. *The Journal of*
1370 *Computer-Aided Molecular Design*, 17, 83.
- 1371 Tetko, I. V., Bruneau, P., Mewes, H. W., Rohrer, D. C.,
1372 & Poda, G. I. (2006). Can we estimate the accu-
1373 racy of ADME-Tox predictions? *Drug Discovery*
1374 *Today*, 11, 700.
- AU19 1375 The Foundations of Cost-sensitive Learning. (2001).
- 1376 Thomas, C. J., Auld, D. S., Huang, R., Huang, W.,
1377 Jadhav, A., Johnson, R. L., et al. (2009). The
1378 pilot phase of the NIH chemical genomics center.
1379 *Current Topics in Medicinal Chemistry*, 9, 1181.
- 1380 Todeschini, R., & Consonni, V. (2000). *Handbook of*
1381 *molecular descriptors*. Weinheim: Wiley-VCH.
- AU20 1382 Tripos. (2010). Sybyl-X 1.0.
- 1383 Tropsha, A. (2005). Application of predictive QSAR
1384 models to database mining. In T. Oprea (Eds.),
1385 *Cheminformatics in drug discovery* (pp. 437–455).
1386 Wiley-VCH.
- AU21 1387 Tropsha, A., & Golbraikh, A. (2007). Predictive
1388 QSAR modeling workflow, model applicability
1389 domains, and virtual screening. *Current Phar-*
1390 *maceutical Design*, 13, 3494.
- 1391 Tropsha, A., & Golbraikh, A. (2010). Predictive quan-
1392 titative structure–activity relationships mod-
1393 eling: Development and validation of QSAR
1394 models. In J.-L. Faulon & A. Bender (Eds.),
1395 *Handbook of chemoinformatics algorithms*. The
1396 Netherlands: Leiden University, Chapman and
1397 Hall/CRC.
- 1398 Tropsha, A., & Golbraikh, A. (2010). Predictive quan-
1399 titative structure–activity relationships model-
1400 ing. *Data Preparation and the General Modeling*
1401 *Workflow*. In J.-L. Faulon & A. Bender (Eds.),
1402 *Handbook of chemoinformatics algorithms* (pp.
1403 175–214). The Netherlands: Leiden University,
1404 Chapman and Hall/CRC.
- 1405 Tropsha, A., & Zheng, W. (2001). Identification of the
1406 descriptor pharmacophores using variable selec-
1407 tion QSAR: Applications to database mining.
1408 *Current Pharmaceutical Design*, 7, 599.
- 1409 Valerio, L., Jr. (2008). Tools for evidence-based tox-
1410 icology: Computational-based strategies as a
1411 viable modality for decision support in chemical
1412 safety evaluation and risk assessment. *Human &*
1413 *Experimental Toxicology*, 27, 757.
- 1414 Valerio, L. G., Jr. (2009). In silico toxicology for the
1415 pharmaceutical sciences. *Toxicology and Applied*
1416 *Pharmacology*, 241, 356.
- 1417 Vapnik, V. (2000). *Nature of statistical learning the-*
1418 *ory*. 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 affini-
ties of structurally diverse compounds. *The*
Journal of Chemical Information and Computer
Science, 44, 758.
- Winter-Vann, A. M., & Casey, P. J. (2005). Post-
prenylation-processing enzymes as new tar-
gets 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 cytotoxic-
ity profiling using quantitative high-throughput
screening. *Environmental Health Perspect*, 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 prenyla-
tion: Molecular mechanisms and functional con-
sequences. *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. *Jour-*
nal 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 vari-
able selection quantitative structure–property
relationship approach based on the k-nearest-
neighbor principle. *The Journal of Chemical*
Information and Computer Science, 40, 185.
- Zhou, Z. H., & Liu, X.-Y. (2006). Training
cost-sensitive neural networks with

- 1473 methods addressing the class imbalance problem. *IEEE Transactions on*
 1474 *Knowledge and Data Engineering*,
 1475 18, 63.
 1476
- 1477 Zhu, H., Rusyn, I., Richard, A. M., & Tropsha, A.
 1478 (2008). Use of cell viability assay data improves
 1479 the prediction accuracy of conventional quantitative
 1480 structure-activity relationship models of
 1481 animal carcinogenicity. *Environmental Health*
 1482 *Perspect*, 116, 506.
- 1483 Zhu, H., Tropsha, A., Fourches, D., Varnek, A.,
 1484 Papa, E., Gramatica, P., et al. (2008). Combinatorial
 1485 QSAR modeling of chemical toxicants
 1486 tested against *tetrahymena pyriformis*. *Journal of*
 1487 *Chemical Information and Modeling*, 48, 766.
- AU22** 1488 Zhu, H., Tropsha, A., Fourches, D., Varnek, A., Papa,
 1489 E., Gramatica, P., et al. (2008). Combinatorial
 1490 QSAR modeling of chemical toxicants tested
 1491 against *tetrahymena pyriformis*. *The Journal of*
 1492 *Chemical Information and Modeling*.
- Zhu, H., Ye, L., Golbraikh, A., & Tropsha, A. (2009). 1493
 QSAR studies of chemical aquatic acute toxicity 1494
 using k Nearest Neighbor (kNN) Methodology. 1495
- Zhu, H., Ye, L., Richard, A., Golbraikh, A., Wright, F. 1496
 A., Rusyn, I., et al. (2009). A novel two-step hierarchical 1497
 quantitative structure-activity relationship 1498
 modeling work flow for predicting acute 1499
 toxicity of chemicals in rodents. *Environmental* 1500
Health Perspect, 117, 1257. 1501
- Zupan, J., & Gasteiger, J. (1999). *Neural networks* 1502
in chemistry and drug design. Weinheim: Wiley- 1503
 VCH. 1504
- Zvinavashe, E., Murk, A. J., & Rietjens, I. M. (2008). 1505 **AU24**
 Promises and pitfalls of quantitative structure- 1506
 activity relationship approaches for predicting 1507
 metabolism and toxicity. *Chemical Research in* 1508
Toxicology. 1509
- Zvinavashe, E., Murk, A. J., & Rietjens, I. M. (2009). 1510
 On the number of EINECS compounds that can 1511
 be covered by (Q)SAR models for acute toxicity. 1512
Toxicology Letters, 184, 67. 1513

Author Query Form

Handbook of Computational Chemistry

Chapter No. 00038

Query Refs.	Details Required	Author's response
AU1	Kindly check and confirm “author names & affiliations”.	
AU2	Please check if edit to the sentence starting “Quantitative structure...” is okay.	
AU3	Please confirm if this paragraph can be treated as an “Abstract”.	
AU4	“Formula” has been changed to “Eq.” in this cross citation and in all occurrences. Please check if appropriate.	
AU5	Kindly provide “author name & page range” for “Addressing the curse (1997)”.	
AU6	Kindly provide “url” for “C5.0 (2008)” if feasible.	
AU7	Kindly provide “Date of Access” for “ChemAxon (2008); ChEMBL Database (2010); Dragon (2007); DSSTox (2008); Fallon et al. (1997); Jaworska and Nikolova-Jeliazkova (2008); Maybridge (2005); MDDR.SYMYX (2009); Molconn (2007); Molecular Operating (2008); nci (2007); Neural Networks (2010); Organisation (2008); PDSP (2010); PubChem (2010)”.	
AU8	Kindly provide “url” for “Discovery Studio (2010)” if feasible.	
AU9	Kindly confirm “volume number & page range” inserted in “Fourches et al. (2010)”.	
AU10	Kindly provide “volume number” for “Huan et al. (2006)”.	
AU11	Kindly provide “author name & page range” for “Learning From Imbalanced (2000)”.	
AU12	Kindly provide “url” for “LigandScout (2010)” if feasible.	
AU13	Kindly provide “author name & page range” for “Neural Networks (1996)”.	
AU14	Kindly provide “author name, publisher name & publisher location” for “OpenBabel (2010)”.	

AU15	Kindly provide “ url” for “Random Forests (2001)” if feasible.	
AU16	Kindly provide “url” for “Schrodinger Software (2010)” if feasible.	
AU17	Kindly provide “year, volume number & page range” for “Sedykh et al. (in press)”.	
AU18	Kindly confirm “publisher location” inserted in “Smola and Schoelkopf (2004)”.	
AU19	Kindly provide “url” for “The Foundations (2001)” if feasible.	
AU20	Kindly provide “url” for “Tripos (2010)” if feasible.	
AU21	Kindly provide “publisher location” for “Tropsha (2005)”.	
AU22	Kindly provide “volume number & page range” for “Zhu et al. (2008)” if feasible.	
AU23	Kindly provide “Journal title, volume number & pagerange” for “Zhu et al. (2009)” if feasible.	
AU24	Kindly provide “volume number & page range” for “Zvinavashe et al. (2008)” if feasible.	