

A Stepwise Approach for Defining the Applicability Domain of SAR and QSAR Models

Sabcho Dimitrov,[†] Gergana Dimitrova,[†] Todor Pavlov,[†] Nadezhda Dimitrova,[†] Grace Patlewicz,[‡]
Jay Niemela,[§] and Ovanes Mekenyan^{*,†}

Laboratory of Mathematical Chemistry, University “Prof. As. Zlatarov”, 8010 Bourgas, Bulgaria, Safety and
Environmental Assurance Centre (SEAC), Unilever Colworth, Sharnbrook,
Bedford MK44 1LQUK, United Kingdom, and Danish Institute for Food and Veterinary Research,
19 Mørkhøj Bygade, DK-2860 Søborg, Denmark

Received February 1, 2005

A stepwise approach for determining the model applicability domain is proposed. Four stages are applied to account for the diversity and complexity of the current SAR/QSAR models, reflecting their mechanistic rationality (including metabolic activation of chemicals) and transparency. General parametric requirements are imposed in the first stage, specifying in the domain only those chemicals that fall in the range of variation of the physicochemical properties of the chemicals in the training set. The second stage defines the structural similarity between chemicals that are correctly predicted by the model. The structural neighborhood of atom-centered fragments is used to determine this similarity. The third stage in defining the domain is based on a mechanistic understanding of the modeled phenomenon. Here, the model domain combines the reliability of specific reactive groups hypothesized to cause the effect and the domain of explanatory variables determining the parametric requirements in order for functional groups to elicit their reactivity. Finally, the reliability of simulated metabolism (metabolites, pathways, and maps) is taken into account in assessing the reliability of predictions, if metabolic activation of chemicals is a part of the (Q)SAR model. Some of the stages of the proposed approach for defining the model domain can be eliminated depending on the availability and quality of the experimental data used to derive the model, the specificity of (Q)SARs, and the goals of their ultimate application. The performance of the proposed definition of the model domain is tested using several examples of (Q)SARs that have been externally validated, including models for predicting acute toxicity, skin sensitization, and biodegradation. The results clearly showed that credibility in predictions of QSAR models for chemicals belonging to their domain is much higher than for chemicals outside this domain.

INTRODUCTION

One of the general criteria for selecting (Q)SARs for use in the risk assessment process is the determination of its applicability domain.¹ The concept for the applicability domain of a model is closely related to the term *model validation*. The latter is defined as the “substantiation that a model within its domain of applicability possesses a satisfactory range of accuracy within the intended application of the model”.^{2,3} Two decades later, the domain of a (Q)SAR was defined as “the group of chemicals for which the model is valid”.¹ This loop of definitions based on the alternate use of model validity and applicability domains reminds one of the chicken and the egg dilemma and emphasizes the interconnectivity of these two characteristics.

(Q)SAR models are confined by these definitions since they are derived from structurally limited training sets. Therefore, it is unlikely that a model will be applicable to every chemical. The key is to understand the scope of a model and whether it is appropriate to make a prediction for a given query chemical. Recently, international activities

in the area of (Q)SAR validation have resulted in a number of suggestions for the formal validation of predictive models.^{4–6} Although the requirements for the use of standardized assays and unambiguous endpoints are generally accepted by modelers, the demand for testing substances that span the descriptor’s space of interest well is not a trivial task because it is not possible to guess in advance which combination of explanatory variables will be relevant to the (Q)SAR modeling.

An additional problem concerning the use of (Q)SARs in risk assessment is the formal definition of structural similarity between chemicals in the training and prediction sets. The simplest definition of structural similarity, and consequently, the definition of structural domain, consists of a list of atoms for which a certain (Q)SAR can be used to predict a modeled endpoint. Such “inclusion” rules may be accompanied by “exclusion” rules that determine the compound classes for which the (Q)SAR is not applicable. In some cases, the structural domains of applicability of (Q)SARs are determined on the basis of empirical knowledge or hypothesized modes of action.^{7–9} For example, the QSAR model derived for aldehyde acute aquatic toxicity⁹ is defined for chemicals with aldehyde moiety CH=O , but not for acrolein; crotonaldehyde; α,β -unsaturated alkynals, $\text{R-C}\equiv\text{CCH=O}$; or aldehydes halogenated at the β C site, R-C(Hal)HCH=O .

* Corresponding author e-mail: omekenyan@btu.bg, tel.: ++359 56 880230, fax: ++359 56 880249.

[†] University “Prof. As. Zlatarov”.

[‡] SEAC.

[§] Danish Institute for Food and Veterinary Research.

Such structural rules define the group of substances for which the model is valid. In addition, the domain may be represented by the ranges of the molecular parameters for chemicals in the training set, that is, the descriptor space of the training set. The model may not be applicable to certain regions of the rectangular domain as defined by the range of variation of the variables in the training set, and determination of the optimal prediction space of the model¹⁰ is required. The similarity of the molecule to the nearest or the number of nearest molecule(s) in the training set can be used as a good discriminator of prediction accuracy.¹¹ It was found that molecules with the highest similarity or with more neighbors in the training set are better predicted. The descriptors used to determine the similarity do not have to be the same as the descriptors used in the (Q)SAR. Another dichotomy measure of confidence for the predictions from the regression model was proposed.¹² The residuals from the training set, classified as *bad* and *good*, are used to build a classifier, which is then used to determine the class of the residual for a new compound. The proposed approach restricts itself to using the same descriptors that were used in the original quantitative model. These attempts to increase creditability in model predictions traced a new perspective in the (Q)SAR methodology where the main (Q)SAR model used to predict activity is accompanied by a satellite model determining the confidence of the obtained predictions.

Finally, although the metabolism is usually not accounted for explicitly in traditional (Q)SAR models, it is difficult to identify a toxicological endpoint for which metabolism can be ignored. If the simulation of metabolism is part of the model, such as is the case for CATABOL,¹³ TIMES,¹⁴ and METEOR,^{15,16} the domain of the metabolic simulator can be defined in terms of the reliability of the computer-simulated reactions used to mimic the molecular transformations. The intersection of the structural, mechanistic, and metabolism domains is perhaps the most conservative definition of the applicability domain for a (Q)SAR model. Depending on the use of the model predictions, some of these components of the model domain can be potentially ignored.

The aim of this study is to formulate a stepwise approach to determine the applicability domain of a (Q)SAR model. The first stage is based on general requirements for variation of the physicochemical properties of chemicals. The second stage accounts for structural similarity between chemicals that are correctly predicted. The structural neighborhood of atom-centered fragments is used to define this similarity. The third stage is based on a mechanistic understanding of the modeled phenomenon. Here, the model domain may account for specific reaction groups and the interpolation domain of explanatory variables used by the (Q)SAR model. Finally, if metabolism simulation is a part of the (Q)SAR model, the reliability of generated metabolites and metabolic maps should be taken into account in assessing the reliability of predictions.

The value of this proposed definition of the model domain is highlighted using several examples of (Q)SARs that have been externally validated. Examples of (Q)SARs include those for predicting acute toxicity, skin sensitization, and biodegradation.

METHODS

The four stages of the domain procedure are presented in this section.

Stage 1. General Parametric Requirements. The variations of molecular parameters that may affect the quality of the measured endpoint significantly are included here, such as molecular weight, absorption, water solubility, volatility, and ionic dissociation. Such properties are not usually the driving forces for the studied phenomenon, but they may implicitly affect the measured endpoint, for example, by reducing the bioavailability of chemicals. The range of variation of these properties can be extracted from the training set or from expert knowledge if the training set is structurally limited. Formally, the set of such requirements, D_{GR} , was described as

$$D_{GR} = \{Pr_i^{\text{Min}}, Pr_i^{\text{Max}}, i = 1, 2, \dots, I\} \quad (1)$$

where Pr_i^{Min} and Pr_i^{Max} determine the accepted range of variation of i th molecular parameters. The application of the D_{GR} rules on a certain set of chemicals S will identify a subset of chemicals S_{GR} that fall in the range of physical and chemical properties where subsequent application of the model may result in predictions of higher reliability:

$$S_{GR} = D_{GR}(S) \quad (2)$$

Stage 2. Structural Domain. In 1861, Butlerov put forward a theory that laid the foundation of present-day principles of molecular structure:

- a. The atoms in a molecule are combined in a definite order.
- b. Atoms combine in accordance with their valences.
- c. The properties of a substance depend not only on the nature of the atoms and their number but also on their arrangement.

The basic principles of the theory of chemical structure allow the introduction of the concept of mutual influence among atoms. It is common knowledge that directly bound atoms affect each other in addition to there being a mutual influence among all atoms. For example, a chlorine atom bound to a carbon atom differs substantially depending on whether the carbon atom is part of an alkane, alkene, or aromatic-ring system.

On the basis of the concept of mutual influence between atoms in a molecule, the structural domain of an atom participating in a number of molecules, ions, or radicals may be defined by the set of atom-centered fragments containing this atom and its first, second, and so forth neighbors.¹⁷ This approach partitions the molecules into atom-centered fragments, and different binary vector similarity measures can be used to quantify the structural similarity between pairs of molecules.¹⁸ The drawback of such similarity measures is that their application for similarity between a vector and a set of vectors (or a matrix) requires further adaptation, such as averaging of partial similarities, a search for maximum (minimum) similarity, and so forth.

The effect of neighbors decreases with the increase of their distance to a specified atom; it also depends on the atomic type of the neighbors. For example, conjugated systems of bonds and aromatic-ring systems may transmit the electron withdrawing or donating effect of an atom much further than

Table 1. Defining the Structural Domain Accounting for the Second Neighbors of the N Atom for a Set of Chemicals

Set of chemicals used for defining the structural domain of N-atom	Extracted substructures defining the structural domain of N-atom, $D = \{F_{N-atom}^2\}$

to its first neighbors. The selection of an optimal fragmentation of a molecule is not a trivial task. Sometimes, third neighbors are considered, and as a result, all seven-atomic molecules are considered as a single fragment. To overcome the problem for the optimal fragmentation of a molecule, the following rules are proposed here to reflect the effect of different neighbors on the specified atom:

1. Hydrogen atoms are treated as an inherent part of the atom to which they are bound (i.e., effectively ignored and excluded from the list of first neighbors).

2. The first, second, ..., and n th neighbors are selected to determine the atom-centered fragment.

3. If one of the first, second, ..., and n th neighbors is an aromatic carbon atom ($C\{ar\}$), then the whole aromatic ring containing this atom is considered as a single neighbor.

4. If the n th neighbor is a $C\{sp^3\}$ or $C\{ar\}$ atom, then the $(n + 1)$ th, $(n + 2)$ th, and further neighbors are assumed to have an insignificant effect on the properties of the centered atom.

5. If the n th neighbor is not a $C\{sp^3\}$ or $C\{ar\}$ atom, then the atom-centered fragment is propagated until a $C\{sp^3\}$ or $C\{ar\}$ atom is reached.

In these rules, aromaticity is restricted to six-membered rings including heterocycles. Tautomerism is not accounted for.

The application of these rules allows the extraction of a set of atom-centered fragments that could be used to characterize the structural domain of the atoms presented in a certain set of chemicals S :

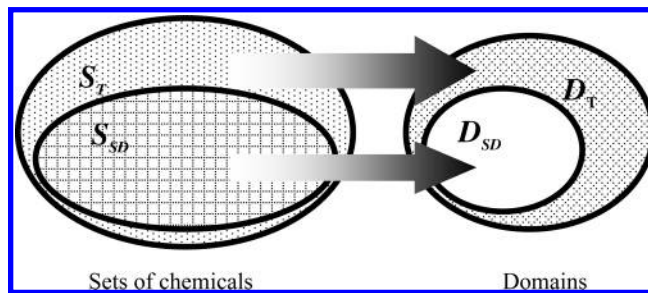
$$D = \{F_i^n, i = 1, 2, \dots, I\} \quad (3)$$

where F_i^n is a set of atom-centered fragments for i th atom accounting for its n th neighbors. If the atom-centered fragments for each atom constituting an external chemical are determined to be elements of D , then this chemical is considered to belong to the structural domain as defined by

chemicals from S . An example for defining the domain of a nitrogen atom for a set of chemicals is illustrated in Table 1.

Chemicals from an external set that are in and out of the structural domain of the N atom (defined in Table 1) are listed in Table 2.

The set of chemicals S_T used to build a certain (Q)SAR model is known as the model training set. For a subset of these chemicals $S_{SD} \subset S_T$, the model correctly predicts the modeled endpoint. The two sets S_T and S_{SD} can be used to determine the sets of atom-centered fragments D_T and D_{SD} for defining the structural domains of the training set and (Q)SAR model, respectively. An extraction of the domains is sketched in Figure 1. Because $S_{SD} \subset S_T$, it follows that

**Figure 1.** Extraction of the domain of training chemicals (D_T) and the structural domain of the model (D_{SD}).

$$D_{SD} \subset D_T.$$

The application of the structural rules of D_T on chemicals from S_T will classify all chemicals as elements of S_T :

$$S_T = D_T(S_T). \quad (4)$$

The same holds when D_{SD} rules are applied on the chemicals of S_{SD} :

$$S_{SD} = D_{SD}(S_{SD}) \quad (5)$$

Table 2. Evaluating the Inclusion of Chemicals from External Sets

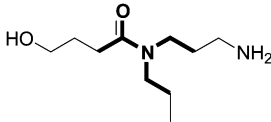
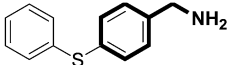

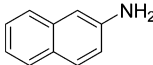
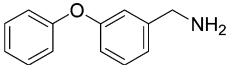
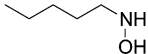
Chemicals that belong to the structural domain of N-atom (centered fragments are highlighted)	Chemicals that do not belong to the structural domain of N-atom
  	  

Table 3. Classification of Chemicals by Using the Rules of the Structural Domain D_{SD} of the Model

	observed, S_T	
	in domain, S_{SD}	out of domain, S_T/S_{SD}
predicted	In Domain, $D_{SD}(S_{SD}) = S_{SD}$	$D_{SD}(S_T/S_{SD})$ (false positives)
	Out of Domain $S_{SD}/D_{SD}(S_{SD}) = \emptyset$ (false negatives)	$(S_T/S_{SD})/D_{SD}(S_T/S_{SD})$

Classification of chemicals from the training set S_T by the rules of D_{SD} will result in the following set of chemicals:

$$S_{SD}^{Class} = D_{SD}(S_T) = D_{SD}(S_{SD} \cup S_T/S_{SD}) = D_{SD}(S_{SD}) \cup D_{SD}(S_T/S_{SD}) \quad (6)$$

where S_T/S_{SD} is the complement of S_{SD} in S_T and $D_{SD}(S_T/S_{SD})$ is the set of chemicals from the training set that are incorrectly identified to belong to the model domain. The misclassification $D_{SD}(S_T/S_{SD})$ could be conditioned by a number of factors such as the experimental error, model inadequacy, and empirical approach used to determine the structural domain. The classification of chemicals from the training set S_T by using the rules D_{SD} of the structural domain of the model is summarized in contingency Table 3.

The goodness of the determined model domain can be quantified by the following statistics:

Adjusted Pearson's contingency coefficient C^ :*

$$C^* = \sqrt{\frac{\varphi^2}{1 + \varphi^2}} / C_{MAX} \quad (7)$$

where n is the number of rows or columns, which ever is

$$C_{MAX} = \sqrt{\frac{n-1}{n}} \quad (8)$$

$$\varphi = \sqrt{\frac{\chi^2}{N}} \quad (9)$$

smaller, and N is the number of data in the table.

Sensitivity: The probability that the chemical belongs to the structural domain of the model, D_{SD} , given that the chemical is really an element of S_{SD} .

$$\text{Sensitivity} = \frac{\text{Card}[D_{SD}(S_{SD})]}{\text{Card}(S_{SD})} \quad (10)$$

where $\text{Card}(X)$ denotes the number of elements of set X . By default, as can be seen from eq 5 and Table 3, the sensitivity is equal to 1.

Specificity: The probability that the chemical does not belong to the structural domain of the model, S_{SD} , given that the chemical is not really an element of S_T .

$$\text{Specificity} = \frac{\text{Card}[(S_T/S_{SD})/D_{SD}(S_T/S_{SD})]}{\text{Card}(S_T/S_{SD})} \quad (11)$$

False negatives: The probability that a chemical will be classified out of the structural domain of the model, given that the chemical is really in this domain.

$$\text{False negatives} = \frac{\text{Card}[S_{SD}/D_{SD}(S_{SD})]}{\text{Card}(S_{SD})} \quad (12)$$

By default, as can be seen from eq 5 and Table 3, false negatives are zero.

False positives: The probability that a chemical will be classified in the structural domain of the model, given that the chemical is really out of this domain.

$$\text{False positives} = \frac{\text{Card}[D_{SD}(S_T/S_{SD})]}{\text{Card}(S_T/S_{SD})} \quad (13)$$

Stage 3. Mechanistic Domain. This is the most multifarious and insusceptible-to-standardization part of the model domain because it is responsible for reflecting the individual characteristics of the model: structure and mathematical formalism of the model, computational method used for its derivation, accepted hypotheses, and so forth. The suggested approach here is only an attempt to reduce the diversity in this matter.

Two subdomains could be distinguished in the mechanistic domain D_{MH} , the domain of functional (reactive) groups D_{FG} and the domain of explanatory variables D_{EV} , as defined by eq 14:

$$D_{MH} = D_{FG} \cup D_{EV} \quad (14)$$

Domain of Functional (Reactive) Groups. The approach for determining the empirical structural domain is focused on individual atoms and their closest neighbors. The focus here will be on functional groups whose reactivity modulates the studied endpoint or structural fragments used in group-contribution models. An examination of mechanistic knowledge about the individual reactive groups and their impact on the model performance can be used to determine the domain of functional groups D_{FG} . An empirical measure of the reliability of functional groups on the model adequacy can be evaluated by relating the number of their successful applications to total number of their applications within the training set. The obtained reliability of a functional group is defined as a probabilistic estimate for its correct application in predicting the model endpoint:

$$R_i^{FG} = \frac{N_{i,succ}^{FG}}{N_{i,succ}^{FG} + N_{i,fail}^{FG}} \quad (15)$$

where $N_{i,succ}^{FG}$ and $N_{i,fail}^{FG}$ are the numbers of successful and unsuccessful applications of i th functional group, respectively. The set of these reliabilities coupled with an appropriate threshold T_{FG} for their significance, $R_i^{FG} \geq T_{FG}$, can be used to determine the domain of the functional groups:

$$D_{FG} = \{R_i^{FG}, T_{FG}, i = 1, 2, \dots, I\} \quad (16)$$

where I is the number of functional groups. The main drawback of this approach is that the values of the reliabilities R_i^{FG} for groups that are not well-presented in the training set (i.e., with a low number of occurrence) will be questionable. For such groups, it is better to use expert knowledge to assign their reliability.

A natural evolution of the domain of the functional groups is the domain of interaction mechanisms. Here, all reactive groups causing an effect by the same interaction mechanism will be combined. Eventually, reliabilities will be defined for interactions mechanisms rather than for different reactive groups.

Domain of Explanatory Variables. Parameter Interpolation Space. In many (Q)SARs, the functional groups are considered as necessary conditions for eliciting activity; the presence of such groups triggers the application of a model that assesses the potency by making use of explanatory variables, such as quantum-chemical descriptors, topological indexes, molecular parameters, and so forth. Thus, such a model evaluates the extent to which the reaction (functional) group is fired. For example, such models are frequently used in regression QSARs, neural networks, or more complicated pattern recognition models in the form of decision trees or forests with sets of logic boxes. In this case, the domain of the functional groups could be determined by the interpolation domain of the model descriptors. In a recently developed probabilistic approach,¹⁹ the population density of chemicals within the training set is assessed to determine a probability

density (or density function) $f(X)$:

$$\int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} f(X) dx_1 \dots dx_N = 1 \quad (17)$$

where X is a vector of N explanatory variables used in the (Q)SAR. The density function (eq 17) provides a measure for the probability that the descriptors of chemicals from the training set get values in a certain range of the descriptor space. This function can also be used to determine the periphery of the descriptor space, that is, the threshold for acceptable interpolation space ($T_{X \notin PS\%}$), where the probability of obtaining the explanatory variables of chemicals from the training set is very small, for example 0.1. This will determine the most unpopulated border range of the interpolation domain of the descriptor space that contains only $T_{X \notin PS\%}$ (e.g., 10%) of training chemicals:

$$P(X \in PS_{T_{X \notin PS\%}}) = P(X \in PS_{10\%}) = \int_{DS_{10\%}} \dots \int f(X) dx_1 \dots dx_N = 0.1 \quad (18)$$

where $PS_{10\%}$ is the 10% periphery space of the interpolation domain of the model. In the same way, it is possible to determine peripheries containing 20, 30, and so forth percent of the training chemicals. This clustering can be used to define the optimal descriptor space of the (Q)SARs by excluding sparsely populated peripheries of the parametric interpolation domain of the model (Figure 2).

Local Performance of the Model. The application of the model for chemicals falling in the interpolation domain of the model descriptors does not guarantee reliable predictions for these chemicals because of model insufficiency or experimental error. A more sophisticated measure of the model performance within optimal descriptor space can be obtained by analyzing the so-called local performance of the model. The latter is determined for training chemicals located within a suitably chosen neighborhood of the chemical under investigation. One of the components of the local performance is the probability $P_{Sph(Y,r)}^{Corr}$ for the model prediction to be correct within the hypersphere, with center Y defined by the descriptors of the screened chemical and radius r in the interpolation domain of the model:

$$P_{Sph(Y,r)}^{Corr} = \frac{N_{Corr}^{Sph(Y,r)}}{N_{Corr}^{Sph(Y,r)} + N_{Noncorr}^{Sph(Y,r)}} \quad (19)$$

where $N_{Corr}^{Sph(Y,r)}$ and $N_{Noncorr}^{Sph(Y,r)}$ are the numbers of correctly and noncorrectly predicted training chemicals in the defined hypersphere, respectively (Figure 3). Depending on the goal and type of model predictions, other statistics, such as sensitivity, specificity, concordance, and so forth, can also be introduced here.

The second component of the model performance should take into account the number of chemicals used for calculating the correctness of the local performance, $P_{Sph(Y,r)}^{Corr}$. A suitable measure of the density of training chemicals used to calculate the local correctness of predictions could be obtained by the following equation:

$$P[X \in Sph(Y,r)] = \int_{Sph(Y,r)} \dots \int f(X) dx_1 \dots dx_N \quad (20)$$

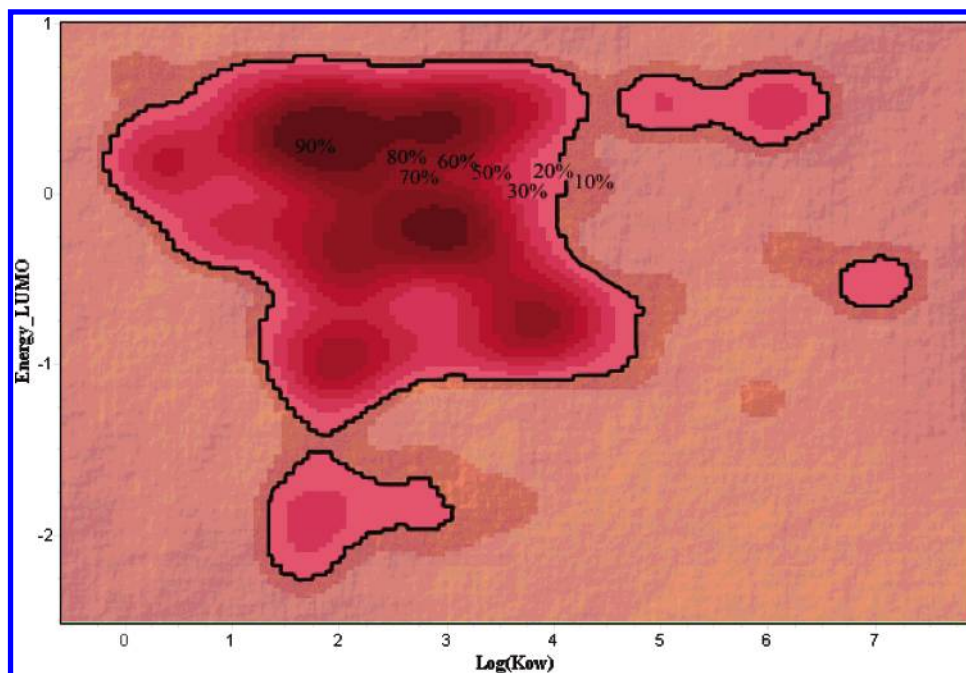


Figure 2. Optimal descriptor space confining 80% of interpolation area as defined by the Duluth fish toxicity database²⁷ used as a training set for deriving acute toxicity models. The descriptor space coordinates E_{LUMO} and $\log K_{\text{OW}}$ are defined by the response-surface QSAR model predicting the toxicity of chemicals causing effect by nonspecific interaction mechanisms (narcotics).²⁴

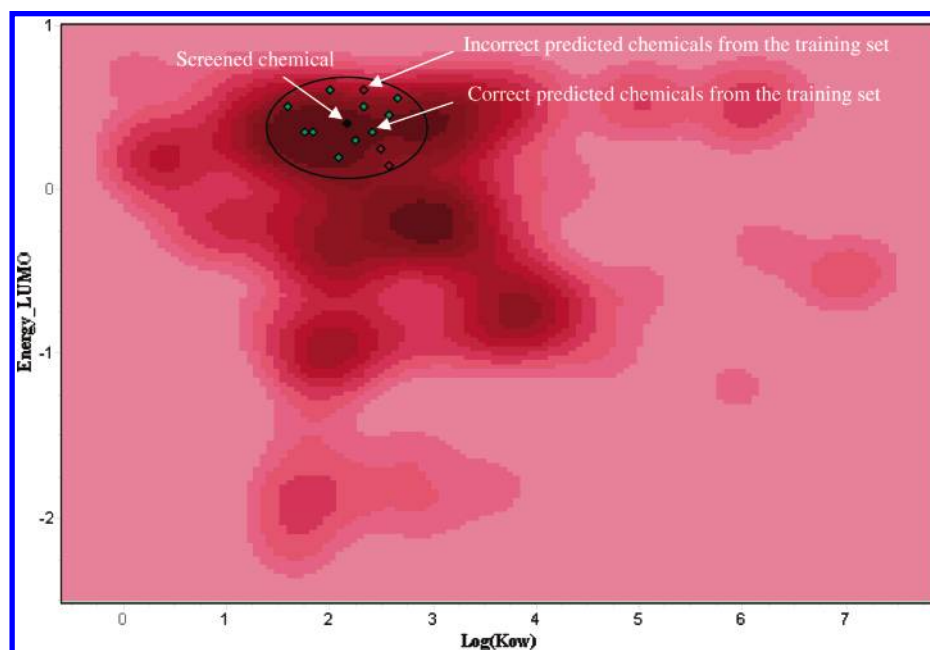


Figure 3. Local performance of the model in the interpolation domain of the model descriptor space is assessed for a hypersphere, the center (Y) of which is defined by the screened chemical and radius r .

where $P[X \in \text{Sph}(Y, r)]$ is the probability that descriptors of chemicals from the training set get values within $\text{Sph}(Y, r)$. The local correctness of predictions and local population density depend on the sphere radius, which should be determined on the basis of the general rules for forming frequency distributions.²⁰

As a measure of the local performance of the model, one can use the product between the probability that the prediction is correct, $P_{\text{Sph}(Y, r)}^{\text{Corr}}$ (i.e., local correctness of predictions as defined by eq 19), and the probability that descriptors of training chemicals have values close to those of the query

chemical $P[X \in \text{Sph}(Y, r)]$ (i.e., local population density as defined by eq 20):

$$P_{\text{LP}} = P_{\text{Sph}(Y, r)}^{\text{Corr}} P[X \in \text{Sph}(Y, r)] \quad (21)$$

In the case where functional groups or some structural symptoms are set as a premise for eliciting activity and potency is predicted by making use of (Q)SARs based on continuous explanatory variables, the mechanistic domain is determined by combining the domains of functional groups

and explanatory variables (in turn, including parameter interpolation space and local performance of the model):

$$D_{MH} = \{R^{FG}, P(X \notin PS_{T_{X \notin PS\%}}), P_{LP}\} \quad (22)$$

The mechanistic domain providing acceptable predictions needs to be confined by the thresholds for fragment reliability (T_{FG}), interpolation space for explanatory variables ($T_{X \notin PS\%}$), and local model performance ($T_{X \in Sph(Y,r)}$, $T_{Sph(Y,r)}^{Corr}$).

Confidence Level and Goodness of Predictions. A variety of QSAR models have been developed by making use of linear and nonlinear regression. The regression analysis assumes that predictor variables are independent and measured without experimental error; the experimental error of dependent variables and residuals follow normal distribution, homoscedasticity, and so forth. In this case, the regression analysis provides other useful measures for assessing goodness of predictions made by regression models. One of these measures is the confidence interval of prediction $\hat{y}(X)$:

$$\hat{y}(X) \pm t_{\nu,\alpha} s_{\hat{y}(X)} \quad (23)$$

where $t_{\nu,\alpha}$ is a Student's t distribution and $s_{\hat{y}(X)}$ is the standard deviation of the prediction $\hat{y}(X)$.²¹ The range of confidence intervals $2t_{\nu,\alpha} s_{\hat{y}(X)}$ of predictions can also be included in the mechanistic domain of the model:

$$D_{MH} = \{R^{FG}, P(X \notin PS_{T_{X \notin PS\%}}), P[X \in Sph(Y,r)], P_{Sph(Y,r)}^{Corr}, 2t_{\nu,\alpha} s_{\hat{y}(X)}\} \quad (24)$$

The set of these probabilities and confidence intervals coupled with an appropriate set of thresholds can be used to determine the domain of the functional groups:

$$D_{MH} = \{R^{FG}, P(X \notin PS_{T_{X \notin PS\%}}), P[X \in Sph(Y,r)], P_{Sph(Y,r)}^{Corr}, 2t_{\nu,\alpha} s_{\hat{y}(X)}, T_{FG}, T_{X \notin PS\%}, T_{X \in Sph(Y,r)}, T_{Sph(Y,r)}^{Corr}\} \quad (25)$$

where T_{FG} , $T_{X \notin PS\%}$, $T_{X \in Sph(Y,r)}$, and $T_{Sph(Y,r)}^{Corr}$ are thresholds for the fragment reliability, interpolation space, local correctness of predictions, and local population density, respectively.

The rules of the mechanistic domain defined by eq 22, 24, or 25 can be used to improve the determination of the model applicability domain already evaluated on the basis of general requirements and chemical structures:

$$S_{MH \cap SD \cap GR} = D_{MH}\{D_{SD}[D_{GR}(S)]\} \quad (26)$$

where $S_{MH \cap SD \cap GR}$ is a subset of a set of chemicals of S classified to belong to the applicability domain of the model.

Stage 4. Domain of Metabolic Simulators. Although the metabolism is usually not accounted for explicitly in traditional (Q)SAR models, it is difficult to identify a toxicological endpoint for which metabolism can be completely ignored. If the metabolism simulation is a part of the model, then a measure of the reliability of the predicted metabolic fate of chemicals should be extracted from the data used to build the simulator. Depending on the specificity of the simulator and available information, this measure can be quantitative (e.g., probabilistic) or discrete (e.g., expressed by levels of confidence).

In CATABOL²² and TIMES,¹⁴ models of the metabolic pathways are generated on the basis of a set of hierarchically

ordered principal transformations including abiotic transformations, enzyme-catalyzed phase I and phase II reactions, and reactions with protein nucleophiles. The reliability of the i th transformation R_i^{TR} is evaluated on the basis of the number of times that the transformation succeeds in producing an experimentally observed metabolite within the training set of metabolic pathways S_{MP} and the number of times it fails:¹⁴

$$R_i^{TR} = \frac{N_{i,succ}^{TR}}{N_{i,succ}^{TR} + N_{i,fail}^{TR}} \quad (27)$$

where $N_{i,succ}^{TR}$ and $N_{i,fail}^{TR}$ are the numbers of successful and unsuccessful applications of the transformation, respectively.

If the experimental data for the modeled metabolism are scarce and insufficient to obtain an objective measure of transformation reliability, as was the case with the development of the simulator of skin sensitization metabolism, then R_i^{TR} could be estimated on the basis of expert knowledge. Transformation reliabilities can be used to determine the reliability of a predicted metabolite R_l^M by multiplying transformation reliabilities across the pathways between parent chemicals and metabolites:

$$R_l^M = \prod_{j=1}^J R_j^{TR} \quad (28)$$

where J is the level of metabolic transformations across the tree from parent ($j = 1$) to the J th metabolite. The reliability of the generated metabolic map for the k th parent chemical is estimated by normalizing reliabilities associated with all of the metabolites generated in the map:

$$R_k^{Map} = \frac{\sum_{j=1}^K R_k^M}{K} \quad (29)$$

where the summation is over all K metabolites in the map.

The set of reliabilities of transformations R_i^{TR} , generated metabolites R_k^M , or metabolic maps R_l^{Map} can be used to determine the domain of the metabolic simulator D_{MS} :

$$D_{MS} = \{R_i^{TR} (i = 1, 2, \dots, I), R_l^M (l = 1, 2, \dots, L), R_m^{Map} (k = 1, 2, \dots, K), T_{TR}, T_m, T_{Map}\} \quad (30)$$

where T_{TR} , T_m , and T_{Map} are thresholds for the trustworthiness of the results produced by the metabolism simulator. The application of rules of the metabolism domain can improve the identification of the overall model applicability domain further:

$$S_{MS \cap MH \cap SD \cap GR} = D_{MS}(D_{MH}\{D_{SD}[D_{GR}(S)]\}) \quad (31)$$

This is the most conservative definition of the applicability domain of the model. Figure 4 illustrates the proposed stepwise concept for determining the applicability domain of the (Q)SAR model. Depending on the ultimate use of the model predictions and the consequences of a wrong decision (based on model predictions), some of the components defining the model domain can be bypassed. This will expand

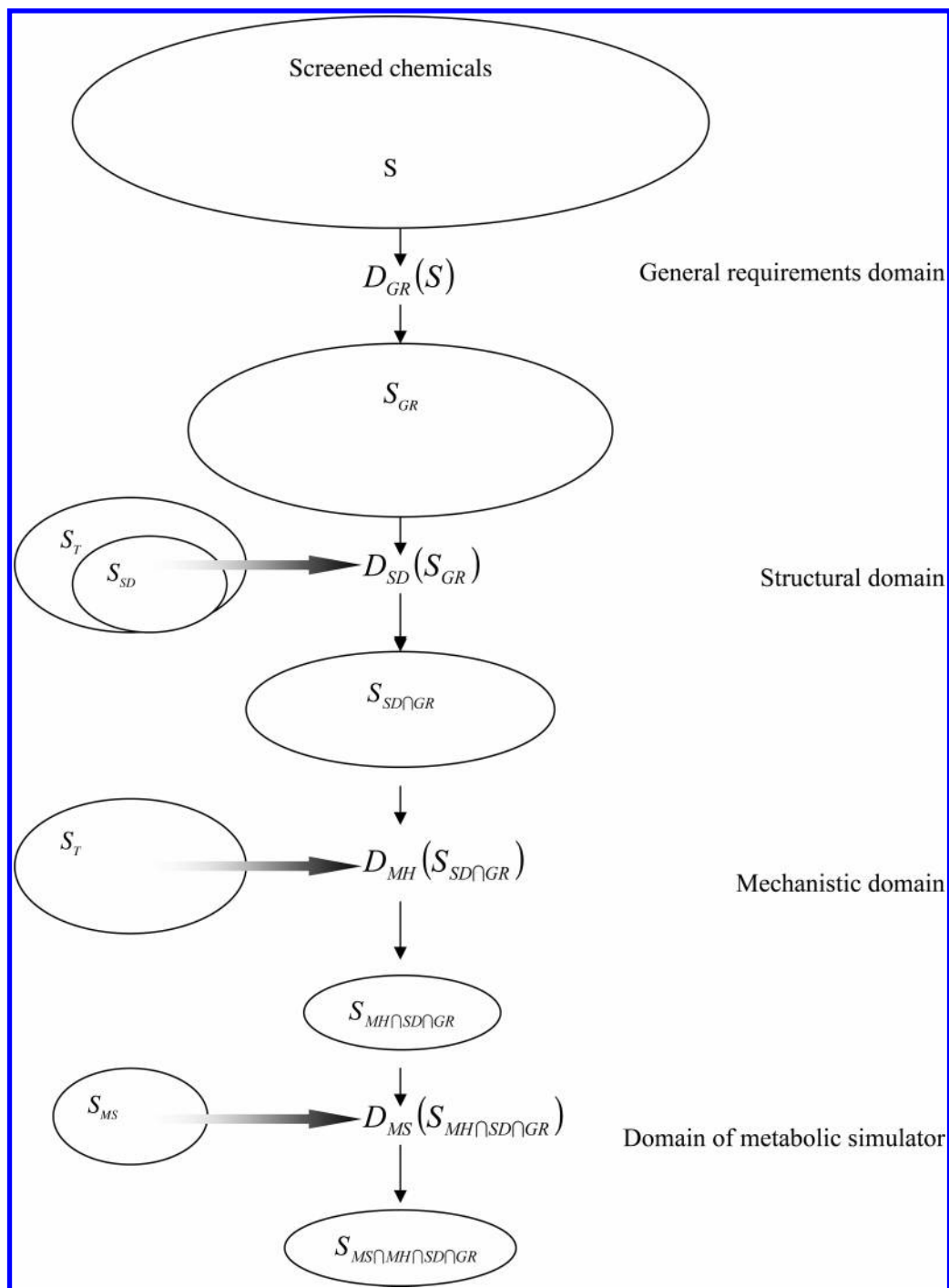


Figure 4. Stepwise concept for determination of the applicability domain of the (Q)SAR model.

the model domain but will reduce the confidence level of the predictions.

RESULTS AND DISCUSSIONS

The presented approach was applied to quantify the applicability domain of the models representing different (Q)-SAR approaches: regression models (commonly used for modeling acute aquatic toxicity), a combination of nonlinear regression and in silico simulation of bacterial catabolism, and an expert system for the prediction of skin sensitization potential. External validation sets were used to analyze the effect of the model domain on the reliability of prediction of these models. The examples were selected to represent different types of (Q)SARs and endpoints and, as a result,

the different steps of the proposed approach. The general parametric requirements (including molecular weight, water solubility, octanol–water partitioning, etc.) were not illustrated because this stage of the model domain did not appear to be restrictive for the domain considered here. It should be noted that in determining the model domain, the degree of accuracy of the model also needs to be specified. In turn, the definition of the accuracy depends on the type of modeled endpoint (categorical or continuous), the intended application of the predictions, and the magnitude of the experimental error associated with the training set. Any attempt to force the approach for defining a domain where the model predicts better than the experimental error will fail.

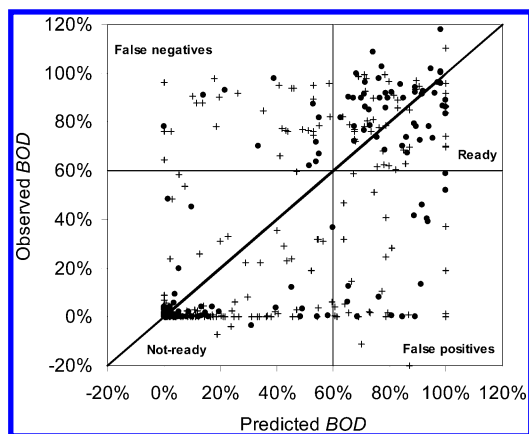


Figure 5. Comparison of predicted versus observed BOD for chemicals falling in and out of the model domain: (●) chemicals in model domain; (+) chemicals out of the model domain.

CATABOL Biodegradation Model. The CATABOL model^{13,22} was trained to predict the bacterial degradation of organic chemicals, the most probable metabolic pathways, and quantities of stable metabolites within 28 days on the basis of 745 chemicals from the Ministry of International Trade and Industry (Japan) (MITI) database.²³ Biodegradation is expressed as oxygen uptake relative to the theoretical uptake (%BOD). The model was able to classify 86% of ready biodegradable chemicals correctly (predicted and observed %BOD $\geq 60\%$) from the training set, whereas the correctness of classification of not-ready chemicals (predicted and observed %BOD $< 60\%$) was 91%. The overall correctness of the classification of training set chemicals or concordance was 89%.

Chemicals from the training set that were correctly classified as ready or not-ready biodegradable were used to extract the structural domain of the model accounting for the first and second neighbors of atom-centered fragments. The attempt to account for the third and next neighbors will result in an extremely conservative domain and will significantly reduce the practical value of the model.

An external set of 347 chemicals from the MITI database was used to demonstrate the effect of a model domain on the accuracy of the predictions of chemicals as ready or not-ready biodegradable. A total of 140 chemicals were identified to be within the domain of the CATABOL model accounting for the first neighbors of atom-centered fragments. The correctness of classification for ready and not-ready chemicals in the model domain was 82%, whereas the correct classification was obtained for 74% of the chemicals that were recognized to be out of the model domain. Figure 5 illustrates these results. As can be expected, accounting for the second neighbors resulted in a more conservative model domain. In this case, only 36 chemicals were identified to be in the model domain, but the correctness of classification rose to 96%.

Acute Aquatic Toxicity. An examination of the ciliate and fish acute toxicity databases resulted in the development of a system of QSAR models, each of them relating toxicity to molecular descriptors for a class of chemicals.^{9,23,24} On the basis of the mode of toxic action, QSARs for neutral narcotics, amine narcotics, esters, polar narcotics, and aldehydes were derived. Both global and local descriptors, such as hydrophobicity ($\log K_{OW}$), bioconcentration factor

($\log BCF_{tox}$),²⁵ and lowest unoccupied molecular orbital (E_{LUMO}), were used to model narcotic toxicity. Because inclusion and exclusion rules based on molecular structure were used to determine the mode of action and the appropriate QSAR, the role of the structural domain will not be analyzed here. The analysis of the reproducibility of fish toxicity revealed that about 95% of repeated observations belong to the interval < 0.5 log units. As a consequence, in the example presented below, a correct prediction was assumed for those chemicals for which the residuals between predicted and observed values were less than 0.5 log units.

A set of 149 narcotics with observed 50% lethal concentrations for *Oryzias latipes*²³ was used to demonstrate the effect of a mechanistic model domain on the accuracy of predictions. A randomly selected 33% portion of the training data was kept out of the model training and used as an external validation set. The rest of the training data (67% of the total data) were fitted by the following model:

$$\log(1/LC_{50}) = b_0 + b_1 \log BCF_{tox} + b_2 E_{LUMO} \quad (32)$$

The procedure was performed three times until each observation was left out once. The mechanistic domain was determined by making use of the interpolation space of the descriptors and the local performance of the model. The optimal interpolation space was defined by excluding the sparsely populated periphery of the descriptor space containing no more than 20% of the training chemicals, $P(X \in PS_{20\%})$. The local reliability of a prediction was estimated by the product between the probability of the prediction to be correct, $P_{Sph(Y,r)}^{Corr}$ (see also eq 19), and the probability that descriptors of the training chemicals have values close to those of the query chemical $P[X \in Sph(Y,r)]$ (see eq 20). The predictions were accepted to be in the model domain if they were within the interpolation space and the local model performance {product of $P_{Sph(Y,r)}^{Corr}$ and $P[X \in Sph(Y,r)]$ } was greater than 0.2. These thresholds were selected after an analysis of the effect of different cutoff values (0.1, 0.2, 0.3, etc.) on the correctness of the predictions for chemicals from the training sets. A significant improvement of the predictions for cutoff 0.2 was found as compared with those for 0.1; the correctness of the predictions for threshold 0.3 was practically the same as for that of 0.2. The same approach was used to obtain the threshold for the sparsely populated periphery.

The application of the mechanistic domain increased the correctness of the predictions from 74%, when the models (eq 32) were applied to all external chemicals, up to 86% for chemicals within the model domain. Figure 6 illustrates the effect of different components of the domain on the quality of the prediction for the external validation sets. As can be seen from the figure, most of the outliers were correctly identified by both components of the mechanistic domain.

TIMES Skin Sensitization Model. A QSAR model for estimating skin sensitization potency has been developed that considers the potential of parent chemicals and their activated metabolites to react with skin proteins.²⁶ The biological activity of the 633 training set chemicals was categorized into one of three classes: significant, weak, or nonsensitizing chemicals. The metabolic simulator was built to mimic the enzyme activation of chemicals in skin, which predicts the

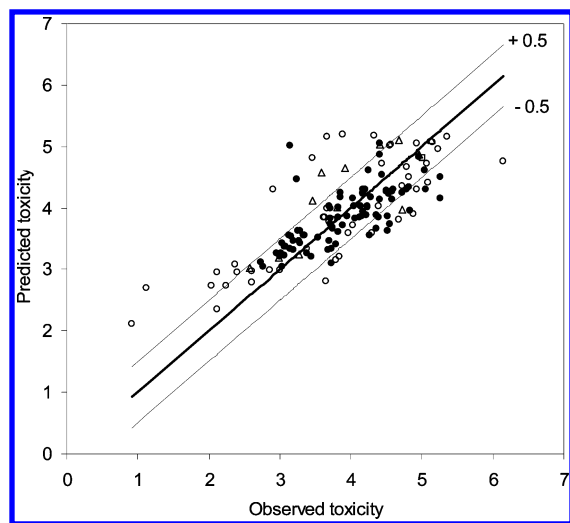


Figure 6. Effect of different components of the mechanistic model domain on the quality of the prediction for acute aquatic toxicity: (●) chemicals in model domain; (□) classified out of model domain by optimal prediction space; (△) chemicals classified out of model domain by local correctness of predictions; (○) chemicals classified out of model domain by optimal prediction space and local performance.

possible molecular transformations and generates the plausible metabolites. Because of the limited quantitative skin metabolism data reported for chemicals, transformation reliabilities were assigned on the basis of available information or expert knowledge, in one of the following confidence levels:

- $R_i^{\text{TR}} = 1.00$ was assigned when a reference source supported enzymatic transformation or when this was a commonly known abiotic transformation.
- $R_i^{\text{TR}} = 0.75$ was assigned to well-known enzymatic reactions that, however, were not supported by a reference source.
- $R_i^{\text{TR}} = 0.50$ was assigned to transformations that are expected to occur but where the enzymatic system causing them is unknown.
- $R_i^{\text{TR}} = 0.25$ was assigned to transformations for which there is some expert expectation that is feasible.
- $R_i^{\text{TR}} = 0.00$ was assigned to transformations for which there is no knowledge for feasibility, although these were included in the simulator to better describe activity variation.

The skin sensitization model was able to correctly classify about 80% of the chemicals with significant sensitizing effect, 34% of the weak sensitizers, and 72% of the nonsensitizing chemicals. Because of the low predictability of the model for weak sensitizers, the overall correctness of prediction was reduced to 65%.

Chemicals from the training set that were correctly classified as strong sensitizers, weak sensitizers, and nonsensitizers were used to extract the structural domain of the model. The first or second neighbors of the atom-centered fragments accounted for the structural similarity. Subsequently, the domain of the skin metabolism simulator was also incorporated in the description of the model domain:

$$S_{\text{MS} \cap \text{SD}} = D_{\text{MS}}[D_{\text{SD}}(S)] \quad (33)$$

where the domain of the metabolism simulator accounted

Table 4. Goodness of the Model Domain Determination When First or Second Neighbors of Centered Atoms Are Considered

statistics	significant neighbors of centered atoms	
	first	second
C^*	0.80	0.95
sensitivity	100%	100%
specificity	59%	87%
false negatives	0%	0%
false positives	41%	13%

for the reliability of the generated metabolites causing sensitization:

$$D_{\text{MS}} = \{R_l^{\text{M}} (l = 1, 2, \dots, L)\} \quad (34)$$

Table 4 summarizes the statistics of the determination of the model domain applied for training set chemicals. Two alternatives have been explored for determining the model domain by accounting for the first and second neighbors of atom-centered fragments. The categorization of the chemicals by the model identified all correct predictions as belonging to the model domain. An additional number of chemicals were incorrectly classified as a part of the model domain (false positives). This misclassification may perhaps be explained by experimental error, model inadequacy, or the empirical approach used to determine the structural domain. Regardless of the empirical nature of the outlined approach, the goodness of the determined model domain was very high if the atom-centered fragments included second neighbors. As expected, determining the model domain on the basis of the first neighbors of atom-centered fragments only resulted in a significant increase in the number of chemicals incorrectly classified as elements of the model domain.

The usefulness of the definition of the model domain and the predictability of the model were examined on the basis of data for sensitization potency of 96 randomly selected chemicals not used in the model building. The structural domain extracted by accounting for the first neighbors of centered atoms identified only 21 chemicals as belonging to the model domain. The overall correctness of the predictions for these chemicals was 71% as compared to 52% if the structural domain of the model was ignored.

The analysis of the training chemicals revealed that an appropriate threshold of 0.25 for the probability of generating a sensitizing metabolite R_l^{M} could be used to improve the predictability within the model domain further. The subsequent application of the domain of the metabolism simulator reduced the chemicals in the model domain to 19, four of which were incorrectly classified. Thus, the ultimate overall correctness of predictions for the external chemicals in the model domain was 79%.

CONCLUSIONS

A stepwise approach for determining the applicability domain of a (Q)SAR model is proposed, distinguishing chemicals for which the models provide highly reliable predictions. The approach accounts for the complexity of the current QSAR models, reflecting their mechanistic rationality (including metabolic activation of chemicals) and transparency. The general requirements specified in the first stage are based on the variation of the physicochemical properties of chemicals in the training set. The second stage

defines the structural similarity between chemicals that are correctly predicted by the model. The structural neighborhood of atom-centered fragments is used to determine this similarity. The third stage in defining the domain is based on a mechanistic understanding of the modeled phenomenon. Here, the model domain combines the reliability of specific reactive groups hypothesized to cause the effect and the domain of explanatory variables. The latter is defined by the parameter interpolation space and local performance of the model in this space. Finally, if metabolic activation of the chemicals is a part of the (Q)SAR model, the reliability of simulated metabolism (metabolites, pathways, and maps) is taken into account in assessing the reliability of the predictions. Some of the stages of the proposed scheme for defining the model domain can be skipped depending on the availability and quality of the experimental data used to derive the model, the specificity of the (Q)SARs, and their ultimate application.

The discrimination of chemicals belonging to the model domain may be at the cost of misclassification of some correctly predicted chemicals as being out of the model domain. The rate of this misclassification could be reduced by an appropriate statistical analysis of the application of the model domain on the training chemicals. In this respect, some components of the approach proposed here are inapplicable for size-restricted data sets that do not allow for statistically significant analysis.

The value of this proposed definition of the model domain is highlighted using several examples of (Q)SARs that have been externally validated, including models for predicting acute toxicity, skin sensitization, and biodegradation. The results clearly showed that predictability of QSAR models for chemicals belonging to a defined domain of applicability is much higher than for chemicals being out of this domain. Although the ideas implemented in the approach for the determination of the model domain were oriented to restrict the application of the model to chemicals for which there is insufficient information for correct prediction, a similar strategy can be used to determine the domains for expanding the applicability of the model, model improvement, model discrimination, and so forth.

REFERENCES AND NOTES

- (1) ECB. *Technical Guidance Document on Risk Assessment, Part III*; European Commission, Joint Research Center: Ispra, Italy, 2003.
- (2) Schlesinger, S.; Crosbie, R. R.; Gagne, R. E.; Innis, G. S.; Lalwani, C. S.; Loch, J.; Sylvester, R. J.; Wright, R. D.; Kheir, N.; Bartos, D. *Terminol. Model Creditability Simulation* **1979**, 32, 103–104.
- (3) Sargent, R. G. Verifying and validating simulation models. In *Proceedings of the 1996 Winter Simulation Conference*, Dec 8–11, 1996; Charnes, J. M., Morrice, D. M., Brunner, D. T., Swain, J. J., Eds.; IEEE: New York, 1997; pp 55–64.
- (4) Balci, O. Verification, Validation, and Accreditation. In *Proceedings of the 1998 Winter Simulation Conference*, Dec 13–16, 1998; Mdeiros, D. J., Carson, J. S., Manivannan, M. S., Eds.; IEEE: New York, 1999; pp 51–48.
- (5) Eriksson, L.; Jaworska, J. J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs. *Environ. Health Perspect.* **2003**, 111, 1361–1375.
- (6) Worth, A. P.; Cronin, M. T. D.; van Leewen, C. J. A framework for promoting the acceptance and regulatory use of (quantitative) structure–activity relationships. *Predicting Chemical Toxicity and Fate*; Cronin, M. T. D., Livingston, D. J., Eds.; Taylor and Francis: London, 2004, in press.
- (7) Schultz, T. W.; Sinks, G. D.; Cronin, M. T. D. Identification of Mechanisms of Toxic Action of Phenols to Tetrahymena pyriformis from Molecular Descriptors. In *Proceedings of QSAR 96*, Elsinore, Denmark, 1996; Chen, F., Schüürmann, G., Eds.; SETAC Press: Pensacola, FL, 1997; pp 329–342.
- (8) Dimitrov, S. D.; Mekenyan, O. G.; Schultz, T. W. Interspecies modeling of narcotics toxicity to aquatic animals. *Bull. Environ. Contam. Toxicol.* **2000**, 65, 399–406.
- (9) Dimitrov, S.; Koleva, Y.; Schultz, T. W.; Walker, J. D.; Mekenyan, O. Interspecies Quantitative Structure–Activity Relationships Model for Aldehydes: Aquatic Toxicity. *Environ. Toxicol. Chem.* **2004**, 23, 463–470.
- (10) Gombar, V. K.; Enslein, K.; Blanke, B. W.; Reid, D. A. In *Proceedings of QSAR 96*, Elsinore, Denmark, 1996; Chen, F., Schüürmann, G., Eds.; SETAC Press: Pensacola, FL, 1997; pp 399–411.
- (11) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, 44, 1912–1928.
- (12) Guha, R.; Jurs, P. C. Determining the Validity of a QSAR Model—A Classification Approach. *J. Chem. Inf. Model.* **2005**, 45, 65–73.
- (13) Jaworska, J.; Dimitrov, S.; Nikolova, N.; Mekenyan, O. Probabilistic Assessment of Biodegradability Based on Metabolic Pathways: CATABOL System. *SAR QSAR Environ. Res.* **2002**, 13 (2), 307–323.
- (14) Mekenyan, O. G.; Dimitrov, S.; Pavlov, T. S.; Veith, G. D. A Systematic Approach to Simulating Metabolism in Computational Toxicology. I. The TIMES Heuristic Modelling Framework. *Curr. Pharm. Des.* **2004**, 10 (11), 1273–1293.
- (15) Sanderson, D. M.; Earnshaw, C. G. Computer prediction of possible toxic action from chemical structure; The DEREK system. *Hum. Exp. Toxicol.* **1991**, 10, 261–273.
- (16) Ridings, J. E.; Barratt, M. D.; Cary, R.; Earnshaw, C. G.; Eggington, C. E.; Ellis, M. K.; Judson, P. N.; Langowski, J. J.; Marchant, C. A.; Payne, M. P.; Yih, T. D. Computer prediction of possible toxic action from chemical structure: an update on the DEREK system. *Toxicology* **1996**, 106, 267–279.
- (17) Kühne, R.; Kleint, F.; Ebert, R.-E.; Schüürmann, G. Calculation of Compound Properties Using Experimental Data From Sufficiently Similar Chemicals. In *Software Developments in Chemistry*, Proceedings of the 10th Workshop “Computer in Chemistry”, Hochfilzen, Tirol, Austria, November 19–21, 1995; pp 125–134.
- (18) Tubbs, J. D. A note on binary template matching. *Pattern Recognit.* **1989**, 22 (4), 359–365.
- (19) Mekenyan, O.; Dimitrov, S.; Serafimova, R.; Thompson, E.; Kotov, S.; Dimitrova, N.; Walker, J. D. Identification of the structural requirements for mutagenicity, by incorporating molecular flexibility and metabolic activation of chemicals I. TA100 Model. *Chem. Res. Toxicol.* **2004**, 17, 753–766.
- (20) Johnson, N. L.; Leone, F. C. *Statistics and Experimental Design in Engineering and Physical Sciences*, Second Edition; John Wiley & Sons: New York, 1997; Vol. I.
- (21) Seber, G. A. F. *Linear Regression Analysis*; John Wiley & Sons: New York, 1997.
- (22) Dimitrov, S.; Kamenska, V.; Walker, J. D.; Windle, W.; Purdy, R.; Lewis, M.; Mekenyan, O. Predicting the biodegradation products of perfluorinated chemicals using CATABOL. *SAR QSAR Environ. Res.* **2004**, 15 (1), 69–82.
- (23) Chemicals Investigation and Testing Institute. Biodegradation and Bioaccumulation Data of Existing Chemicals Based on the CSDL Japan, ISBN 4-98074-101-1. Japan Chemical Industry Ecology–Toxicology & Information Center, 1992.
- (24) Dimitrov, S. D.; Mekenyan, O. G.; Sinks, G. D.; Schultz, T. W. Global modeling of narcotic chemicals: ciliate and fish toxicity. *THEOCHEM* **2003**, 622, 63–70.
- (25) Dimitrov, S. D.; Mekenyan, O. G.; Walker, J. D. Nonlinear modeling of bioconcentration factor using partition coefficient for narcotic chemicals. *SAR QSAR Environ. Res.* **2002**, 13 (1), 177–184.
- (26) Dimitrov, S.; Dimitrova, G.; Mekenyan, O.; Comber, M.; Low, L.; Phillips, R.; Patlewicz, G.; Kern, P.; Niemela, J. Skin sensitization: modeling based on skin metabolism simulation. *Int. J. Toxicol.* in press.
- (27) Geiger, D. L.; Brooke, L. T.; Call, D. J. *Center for Lake Superior Environmental Studies. Acute toxicities of organic chemicals to Fathead minnows (Pimephales promelas)*. University of Wisconsin—Superior: Superior, WI, 1990; Vol. 5; also Vols. 1–4.