Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure–Activity Relationships

The Report and Recommendations of ECVAM Workshop 52[1,2]

Tatiana I. Netzeva,[3] Andrew P. Worth,[3] Tom Aldenberg,[4] Romualdo Benigni,[5] Mark T.D. Cronin,[6] Paola Gramatica,[7] Joanna S. Jaworska,[8] Scott Kahn,[9] Gilles Klopman,[10] Carol A. Marchant,[11] Glenn Myatt,[12] Nina Nikolova-Jeliazkova,[13] Grace Y. Patlewicz,[14] Roger Perkins,[15] David W. Roberts,[16] Terry W. Schultz,[17] David T. Stanton,[18] Johannes J.M. van de Sandt,[19] Weida Tong,[15] Gilman Veith[20] and Chihae Yang[12]

[3]ECVAM, Institute for Health & Consumer Protection, European Commission Joint Research Centre, Ispra, Italy; [4]RIVM, Bilthoven, The Netherlands; [5]Experimental and Computational Carcinogenesis Unit, Environment and Health Department, Istituto Superiore di Sanità, Rome, Italy; [6]School of Pharmacy and Chemistry, John Moores University, Liverpool, UK; [7]QSAR and Environmental Chemistry Research Unit, Department of Structural and Functional Biology, University of Insubria, Varese, Italy; [8]Central Product Safety, Procter & Gamble, Strombeek–Bever, Belgium; [9]Accelrys Inc., San Diego, CA, USA; [10]MULTICASE Inc., Beachwood, OH, USA; [11]Lhasa Ltd, Department of Chemistry, University of Leeds, Leeds, UK; [12]Leadscope Inc., Columbus, OH, USA; [13]Institute of Parallel Processing, Bulgarian Academy of Sciences, Sofia, Bulgaria; [14]SEAC, Unilever, Colworth House, Sharnbrook, UK; [15]Center for Toxicoinformatics, Division of Biometry and Risk Assessment, National Center for Toxicological Research, Food and Drug Administration, Jefferson, AR, USA; [16]Bebington, Wirral, Merseyside, UK; [17]Biological Activity Testing & Modeling Laboratory, College of Veterinary Medicine, University of Tennessee, Knoxville, TN, USA; [18]Miami Valley Laboratory, Procter & Gamble, Cincinnati, OH, USA; [19]Food and Chemical Risk Analysis Department, TNO, Zeist, The Netherlands; [20]Environment, Health and Safety Division, OECD, Paris, France

Address for correspondence: *A. Worth, ECB, Institute for Health & Consumer Protection, European Commission Joint Research Centre, 21020 Ispra (VA), Italy.*
E-mail: *andrew.worth@jrc.it*

Preface

This is the 52nd report of a series of workshops organised by the European Centre for the Validation of Alternative Methods (ECVAM). The main objective of ECVAM, as defined in 1993 by its Scientific Advisory Committee, is to promote the scientific and regulatory acceptance of alternative methods which are of importance to the biosciences, and that *reduce*, *refine* or *replace* the use of laboratory animals.

The ECVAM workshop on the quantitative structure-activity relationship applicability domain was held at ECVAM on 29 September–1 October 2004, under the chairmanship of Andrew Worth. The workshop was attended by experts from academia, industry, international organisations and regulatory authorities. The aim of the workshop was to review the state of the art of methods for identifying the domain of applicability of structure-activity relationships (SARs) and quantitative structure-activity relationships (QSARs), collectively referred to as (Q)SARs. The report is intended to provide a source of input to the development of an OECD Guidance Document on (Q)SAR Validation. The report also makes recommendations for further research needed to understand and apply the concept of the (Q)SAR applicability domain (AD).

Introduction

(Q)SARs are theoretical models that can be used to predict the physicochemical, biological and environmental properties of chemicals.

Address for reprints: ECVAM, Institute for Health & Consumer Protection, European Commission Joint Research Centre, 21020 Ispra (VA), Italy.

[1]ECVAM — The European Centre for the Validation of Alternative Methods. [2]This document represents the agreed report of the participants as individual scientists.

A QSAR expresses in a mathematical form the quantitative relationship that may exist between the chemical structure of a series of chemicals and their measured effect or activity. Multiple linear regression analysis is often used as the method for developing such a relationship, although partial least squares (PLS) analysis, neural networks and other mathematical tools are often used as well. A SAR expresses the qualitative relationship between a two-dimensional molecular fragment (structural alert), or a three-dimensional arrangement of molecular features (pharmacophore), and the presence or absence of a particular effect or activity.

As a result of recent policy developments in the European Union (EU), it is expected that the use of (Q)SARs for regulatory purposes will increase. On 29 October 2003, the European Commission (EC) adopted a legislative proposal (1) for a new chemical management system called REACH (Registration, Evaluation and Authorisation of Chemicals), which is intended to harmonise the information requirements applied to New and Existing Chemicals. Annex IX of the legislative proposal for REACH provides for the use of valid (Q)SARs for predicting the environmental and toxicological properties of chemicals, in the interests of time-effectiveness, cost-effectiveness and animal welfare.

The development of valid (Q)SARs for human health endpoints will also contribute to meeting the needs of the Seventh Amendment to the Cosmetics Directive (2). This lays down deadlines for the replacement of animal tests via the gradual imposition of testing bans on cosmetics (i.e. products or ingredients), which are reinforced by the gradual imposition of marketing bans.

According to a recent assessment by the European Chemicals Bureau (ECB), which, like ECVAM, is part of the EC's Joint Research Centre (JRC), approximately 3.9 million additional vertebrate test animals could be used as a consequence of the implementation of REACH, if alternative methods are not accepted by regulatory authorities and adopted by industry (3). However, a considerable reduction in animal use could be obtained if alternatives were applied more extensively. According to the ECB report (3), a "standard scenario" based on the average acceptance of (Q)SARs and related techniques (for example, read-across) would lead to a saving of 1.3 million test animals, whereas the maximum acceptance of these techniques would enhance this saving potential to 1.9 million test animals.

The recent chemical policy developments are placing an enormous challenge on (Q)SAR developers, regulators and the EC, and have raised the need to develop internationally accepted guidance on good (Q)SAR modelling practices. The JRC, being responsible for the provision of independent scientific advice to policy makers in the EU, established an activity (called a "JRC Action") on (Q)SARs in January 2003, with the overall aim of promoting the availability of valid (Q)SARs for regulatory use. One activity of the JRC Action on (Q)SARs is the development of technical guidance on (Q)SAR validation. This guidance document is being developed within the framework of the Organisation for Economic Cooperation and Development (OECD) Group on (Q)SARs (4).

In November 2004, the OECD Member Countries adopted five principles for the validation of (Q)SAR models for regulatory purposes, now referred to as the OECD Principles for (Q)SAR Validation. According to these principles, and in order to facilitate its consideration for regulatory purposes, a (Q)SAR model should be associated with the following information:

1. a defined endpoint;
2. an unambiguous algorithm;
3. a defined domain of applicability;
4. appropriate measures of goodness-of-fit, robustness and predictivity; and
5. a mechanistic interpretation, if possible.

Principle 3 expresses the need to define an AD for (Q)SARs. This need is based on the fact that (Q)SARs are reductionist models, which are inevitably associated with limitations in terms of the types of chemical structures, physicochemical properties and mechanisms of action for which they can generate reliable predictions.

The ECVAM workshop on the (Q)SAR AD was organised to help to determine what types of information are needed to define (Q)SAR ADs, and to review the current status of methods for defining the ADs of (Q)SARs. This workshop report is intended to serve as a source of input to the OECD guidance document on (Q)SAR validation, the aim of which is to provide detailed guidance on how to apply the (Q)SAR validation principles to various types of models.

The Concept of the (Q)SAR Applicability Domain

In the (Q)SAR field, the AD is widely understood to express the scope and limitations of a model, i.e. the range of chemical structures for which the model is considered to be applicable. However, in the (Q)SAR literature, it is not always apparent whether (or to what extent) the AD concept has been applied. In some cases, the AD concept is implicit in the original publication; for example, the model has been developed from a training set of chemicals that belong to a single chemical class or

that are considered to share a common mechanism of action. In other cases, the AD concept has been explicitly defined. In such cases, the most commonly adopted approach has been to define the AD of the model with structural rules and/or a range of (continuous) descriptor variables (see, for example, 5). If continuous descriptor variables are used, it is possible to define the AD in terms of coverage of the training set in the model descriptor space (6). Such approximations are statistically based, since interpolated estimates are considered to be more reliable than extrapolated ones.

Other approaches have been based on: a) the application of multiple linear regression (MLR) analysis in combination with the distance approach (see, for example, 7); b) definition of a "tolerance volume" around a model by using PLS analysis (8); and c) decision tree analysis (9).

Various approaches for defining the AD have been based on similarity analysis. A comprehensive review of such approaches has been produced by Nikolova & Jaworska (10). All of these approaches are based on the premise that a QSAR prediction is reliable if the chemical for which a prediction is being made is "similar" to the compounds in the training set. The assessment of chemical similarity is not trivial, since the concept of "similarity" is sometimes used in a subjective manner, and in cases where the concept is used in a quantitative manner, different measures of chemical similarity have been proposed. Furthermore, in addition to structural and/or physicochemical similarity, it is also possible to consider similarity in terms of the response and/or mode of action.

The AD concept is applied in several commercially available (Q)SAR prediction systems, including MCASE (developed by Multicase Inc., Beachwood, OH, USA), TOPKAT (developed by Accelrys Inc., San Diego, CA, USA), and the Chem-Tox platform (developed by Leadscope Inc., Columbus, OH, USA).

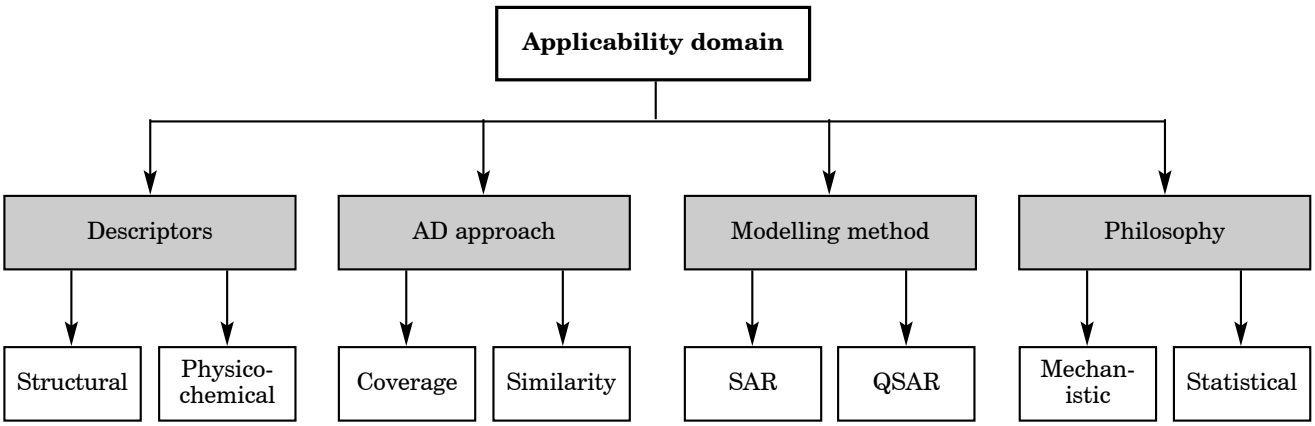
It can therefore be seen that, where the AD concept has been applied, it has been applied in different ways, depending on the type of (Q)SAR model and modelling approach. Given the regulatory need for transparency in the reporting of (Q)SAR models, including their ADs, and given the diversity of approaches for defining an AD, it can be concluded that there is a need to develop a single, but flexible, conceptual framework capable of expressing the ADs of various types of models, developed by different approaches and possibly for different purposes. Figure 1 summarises the multiple aspects of the AD concept. For any given (Q)SAR, one or more of these aspects could be relevant. For example, a structural fragment could be used to derive a qualitative model (SAR), but could also be used in a quantitative manner to develop a QSAR. Figure 1 should not be over-interpreted to imply that certain aspects are mutually exclusive: for example, it is possible to develop a QSAR based on structural descriptors (count variables for the presence/absence of specified structural features) and/or continuous physicochemical descriptors. Furthermore, while (Q)SARs are sometimes referred to as "mechanistically based" or "statistically based", this should be taken to reflect the philosophy and approach adopted in the development of the model, but does not imply that the two types of models are mutually exclusive. In fact, many (Q)SARs are both statistically and mechanistically based.

A general definition of the (Q)SAR AD

Taking into account the multiple aspects of the AD concept, the authors of this report proposed the following general definition for the concept of the AD:

"The applicability domain of a (Q)SAR model is the response and chemical structure space in which the model makes predictions with a given reliability".

Figure 1: Aspects of the (quantitative) structure-activity relationship ([Q]SAR) applicability domain



In this definition, chemical structure can be expressed by physicochemical and/or fragmental information, and response can be any physicochemical, biological or environmental effect that is being predicted.

The importance of the AD in the (Q)SAR life cycle

The AD is an important consideration in all three phases of the (Q)SAR life-cycle (development, validation and application), as illustrated by Figure 2. The concept should be applied during model development, to ensure that a domain is defined as broadly as possible for a desired level of predictivity. It should be noted that, for a model with a given number of descriptors, there is generally a trade-off between the breadth of the domain and the level of predictivity. Thus, in general, one would either aim to develop a model with broad applicability, sacrificing to some extent the level of predictivity, or one would aim to develop a model with narrow applicability (for example, a specific class of chemicals), but with greater predictivity. Both types of model, sometimes (rather confusingly) called "global" and "local" models, respectively, can be useful, depending on the desired application.

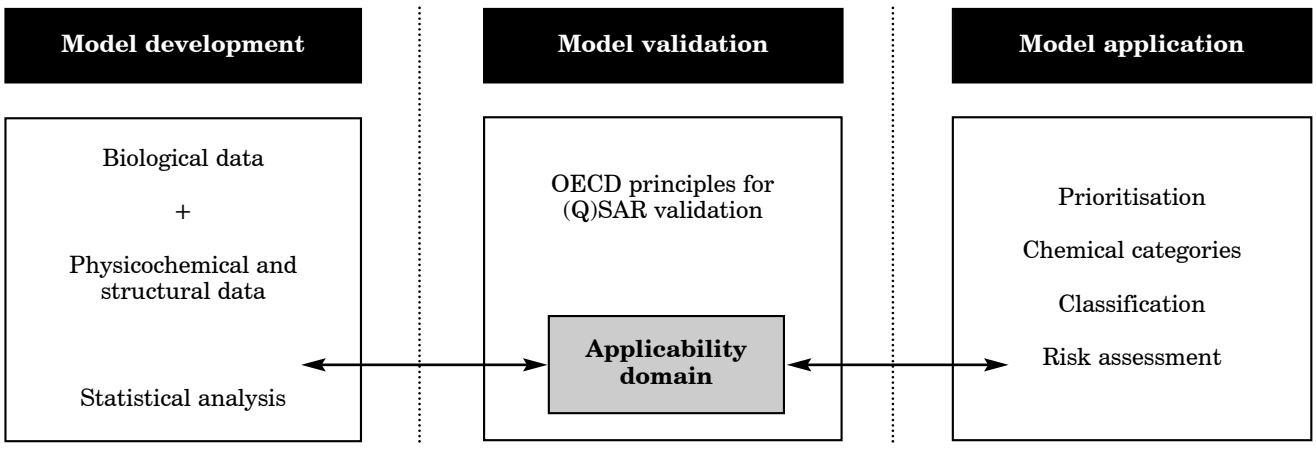
The AD is important during (Q)SAR validation, in the sense that a predefined AD can be verified and possibly refined. In particular, if an external validation is being performed, predictions must be made for compounds that do not form part of the training set. However, to ensure that the external "validation" set is appropriate for model validation, the test chemical structures should fall within the AD of the model, as deduced by analysis of the training set. An open question is whether external validation should also include chemical structures that are considered to fall outside the defined AD, to check whether the boundary is correctly defined, and to investigate the effect of extending the AD on the predictivity of the model.

The ultimate reason for having a well-defined AD is to assist the regulatory application of (Q)SARs to particular chemicals. The decision to use a (Q)SAR for regulatory purposes will generally require an assessment of whether the chemical of interest (for example, a chemical registered under REACH) fits within the AD of the model. This is an essential piece of information during the regulatory assessment of chemicals, because it informs the model user as to whether the endpoint of interest can be reliably predicted for the chemical of interest. Furthermore, in the case where multiple (Q)SARs are available for a given chemical, the model user may also wish to compare the reliability of the predictions made by different models.

"Mechanistic QSARs" based on Physicochemical Descriptors

The term "mechanism of toxic action" can be defined as the action of a toxicant at the molecular level, whereas "mode of [toxic] action" refers to a more general effect or physiological response at a higher level of biological organisation. For example, there is discussion of the mode of action of narcotics, which is displayed at the organism level as a general decrease in activity and, within this mode of action, several distinct mechanisms are sometimes considered (which may result from different interactions at the molecular level; 11). In this report, the terms "mechanism" and "mode of action" are used interchangeably, even though "mechanism" is sometimes used to provide a more detailed or lower-level description of events in the cause-to-effect chain than is "mode of action".

Figure 2: The central place of the applicability domain in various stages of the (quantitative) structure-activity relationship ([Q]SAR) life cycle



The terms "mechanistic" and "mechanistically based" have been used in relation to toxicological QSARs for about two decades, with different meanings (12). Initially, the expression "different mechanism of toxic action" was used to explain outliers to simple regression-derived acute toxicity QSARs developed with data from a particular chemical class. Subsequently, rules based on two-dimensional structures were used to identify chemicals considered to elicit toxicity by the same mechanism or mode of action. As the number of potential chemical descriptors increased, the term "mechanistically based" became increasingly used to refer to QSARs developed by using descriptors that were interpretable in terms of the physicochemical properties they encoded and the causal link between the physicochemical properties and the endpoint modelled. Examples of "interpretable" descriptors include the octanol–water partition coefficient ($\log K_{ow}$) for hydrophobicity, and the energy of the lowest unoccupied molecular orbital (E_{LUMO}) for soft electrophilicity. Models based exclusively on "interpretable" descriptors are often referred to as "mechanistic QSARs".

In principle, if one develops a mechanistically based QSAR, it should be possible to define an AD with data for fewer chemicals than would be the case if one developed a QSAR that is only statistically based. In the development of a mechanistically based QSAR for a specific endpoint, it is hypothesised that the endpoint is a function of certain physicochemical properties, because the endpoint results from a particular molecular mechanism of action. Data on selected chemicals can then be compiled to test the hypothesis for the key physicochemical properties. If the additional data support the hypothesis, there would be a mechanistic aspect to the AD, and the boundaries of the AD could probably be defined by using data for fewer chemicals.

In the development of a purely statistically based QSAR, no assumptions are made about the cause of the endpoint, or more than one cause is anticipated. Thus, it is necessary to test a larger number of chemicals to capture the variation in the descriptor space before using statistics in the definition of the AD.

There are relatively few regulatory endpoints for which "mechanistic QSARs" have been proposed, due to gaps in our understanding of underlying mechanisms of action and the scarcity of high-quality data sets suitable for hypothesis testing. Two examples are acute aquatic toxicity and skin sensitisation.

Mechanistic (Q)SARs for aquatic toxicity

In terms of their acute aquatic toxicity, the majority of industrial organic chemicals are considered to exhibit a narcosis mechanism of toxic action (13). Narcotic chemicals cause only non-covalent and reversible alterations at the theoretical site of action, which is considered to be the cell membrane. In the modelling of the narcosis mode of action, octanol is often regarded as an appropriate surrogate for the target lipid, and the $\log K_{ow}$ is used as a descriptor for the chemical interaction with the membrane. The AD of a QSAR for narcosis can be expressed either as a set of exclusion rules (i.e. all compounds that do not fall in certain classes are narcotics (14), or as a set of inclusion rules that identify chemicals (not necessary classes) capable of exhibiting the narcosis mode of action (5). The first approach can be applicable to more chemicals, but the AD risks being reduced by additional exclusion rules, added to account for chemicals exhibiting mechanisms that were not known or that were not taken into account when the rules were developed. The second approach can give a higher confidence in the domain, but it may restrict the number of chemicals that can be predicted by the model.

As toxicity data sets grew larger with the testing of more so-called "reactive" chemicals (i.e. chemicals with measured toxicity significantly greater than that predicted by narcosis models), the development of QSARs took different approaches. In one approach, QSARs were developed on the basis of generic electrophilic and hydrophobic terms (the response–surface approach), which sacrificed fit in order to expand the AD while maintaining interpretability of descriptors. In a second approach, QSARs were developed on the basis of larger numbers of descriptors which sacrificed the interpretability of descriptors in order to expand the AD, while maintaining fit. A third approach was to develop QSARs that represent well-studied molecular mechanisms of action, while maintaining both fit and interpretability of descriptors. The ADs of such QSARs were limited to narrowly-defined sets of chemicals (defined in terms of classical organic chemical reactions, such as Michael addition).

Mechanistic (Q)SARs for skin sensitisation

Another field where "mechanistic" QSARs have been developed is the modelling of skin sensitisation. For chemicals that act as skin sensitisers, electrophilic or pro-electrophilic behaviour is almost always the key step in the mechanism of action. Therefore, it is natural to group the chemicals into ADs based on the various (pro)electrophilic mechanisms, such as SN2 electrophiles, Michael-type acceptors, SNAr electrophiles, activated esters, and "poison ivy" type pro-electrophiles.

As yet, there is no "global" QSAR for any of these natural domains. In principle, it should be possible to develop, for example, a "global Michael-type acceptor QSAR" for skin sensitisation, covering, for example,

α,β-unsaturated aldehydes, ketones, nitriles, nitroaliphatics and sulphones. In practice, the development of such a QSAR is difficult, because it most likely requires extensive new testing. Therefore, the modelling of skin sensitisation is often based on a set of structure-based rules and QSARs for several structural domains (for example, the aldehydes domain and the sulphonate esters domain) which cut across the natural mechanistic domains (15–17). For example, the aldehydes domain cuts across Schiff base formers (a domain which also includes non-aldehydes such as diketones and pyruvate esters) and Michael acceptors (a domain including many non-aldehydes). Within a given structure domain, a QSAR is typically developed for a "tested domain", i.e. a subset of chemicals that occupy a smaller region of the descriptor space of the entire structure domain.

Statistically Based QSARs Based on Physicochemical Descriptors

This section describes a variety of interpolation methods that have been developed for statistically based QSARs (6). Interpolation is a mathematical term that describes the process of predicting the value of a function at a point from its known values at two or more surrounding points. Interpolation methods make estimations from the training set of data, which are represented as a set of points in n-dimensional descriptor space, where n is the number of descriptors in the model.

An interpolation region in one-dimensional descriptor space is simply the interval between the minimum and the maximum values of the training data set. Interpolation regions in multivariate descriptor space are more complex. Four major approaches have been recognised to estimate interpolation regions in multivariate space. These are based on ranges, geometry, distances and probability density distribution functions (6).

Range-based methods

The simplest method for describing the AD is to consider ranges of individual descriptors. This defines an n-dimensional hyper-rectangle with sides parallel to the coordinate axes. The data distribution is assumed to be uniform. Two limitations of this approach are that interior empty space is not detected and there is no correction for correlations (linear or non-linear) between descriptors.

Principal components analysis (PCA) is a mathematical method in which the original data set is transformed by rotation of the axes, to correct for the correlations between the descriptors. The PCA procedure involves centring the data around the standard mean, and manipulating the covariance matrix of the transformed data to form a new coordinate system, in which the new axes, called principal components (PCs), are orthogonal to one another. The PCs are aligned with the directions of the greatest variations in the data set. An n-dimensional hyper-rectangle can then be defined with sides parallel to the PCs and with the data points between the minimum and maximum value of each PC. The hyper-rectangle also includes empty space, but this is less empty than the hyper-rectangle based on the original descriptor ranges.

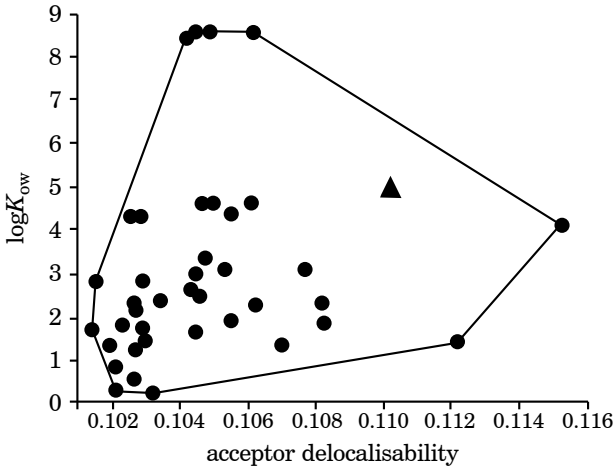
A variation of the PCA domain is implemented as the optimum prediction space (OPS) in the TOPKAT software (18). In the case of the OPS, the data are also centred around the average of each parameter, and the PCA procedure is applied to generate the new orthogonal coordinate system (called the OPS coordinate system). The minimum and maximum values of the data points on each axis of the OPS coordinate system define the OPS boundary. In addition, to deal with data sets comprising non-uniformly distributed data, the confidence of the prediction is estimated in terms of the property sensitive object similarity (PSS) between the training set and a queried point. The PSS is the TOPKAT implementation of a heuristic solution to reflect dense and sparse regions of the data set. A "similarity search" enables the user to check the performance of TOPKAT in predicting the effects of a chemical which is structurally similar to the test structure. The user is also given literature references to the original sources of information.

Geometric methods

The most straightforward empirical method for defining the coverage of an n-dimensional set is the convex hull, which is the smallest convex area that contains the original set. This method is illustrated in Figure 3, in which the two-dimensional interpolation space is defined by two descriptors ($\log K_{ow}$ and acceptor delocalisability), which were found to be important predictors of acute toxicity to fish (19). Even within the convex hull defined by these two descriptors, there are regions with a high density of data points, and regions where the data are sparse. To address these limitations, more-sophisticated methods for defining the domain have been developed.

Calculation of the convex hull is a computational geometry problem (20). Efficient algorithms for convex hull calculation are available for two and three dimensions, but the order of complexity (O) of the algorithms also increases with increasing numbers of data points and dimensions. For n points and d dimensions, the complexity is of the order of $O^{(n[d/2]+1)}$. A disadvantage of this approach is that potential empty spaces within the convex hull cannot be identified.

Figure 3: Representation of the applicability domain by a two-dimensional plot



The training data are represented by circles, and a new chemical to be predicted by the model is represented by the triangle. The training set was used to derive a two-dimensional linear regression model for acute fish toxicity on the basis of two descriptors: logK_{ow} and acceptor delocalisability of the ether oxygen of the ester group (19).

Distance-based methods

Distance-based approaches calculate the distance from a query data point to a data set. The decision as to whether a data point is close to the data set depends on whether there is a criterion for the distance to be below a defined threshold.

Regions at a constant distance are called iso-distance contours. The shape of the iso-distance contours depends on the particular distance measure used and the particular approach for measuring the distance between a query point and a data set. Examples of different approaches include: a) distance to the mean; b) average distance between the query point and all data set points; and c) maximum distance between the query point and all data set points.

Distance-based approaches can be used to separate regions of varying density by imposing cut-off values. However, these regions do not reflect the actual information density of the data set, and the cut-off values do not correspond with the density of the data.

Three distance-based approaches have been found to be most useful in QSAR research, namely, the Euclidean, Mahalanobis and city-block distance measures (6). Related approaches are based on the Hotellings test and the leverage, which is calculated from the Hat Matrix.

Euclidean, Mahalanobis and city-block distances

The Euclidean distance is the square root of the squared differences between corresponding elements of the rows (or columns) in the distance matrix. This is probably the most commonly used distance metric. The Mahalanobis distance is a weighted Euclidean distance, where the weighting is determined by the sample variance–covariance matrix. The block distance is the sum of the absolute differences between corresponding elements of the rows (or columns) in the distance matrix. The block distance is also known as the city-block or Manhattan distance.

Methods based on the Euclidean and Mahalanobis distance measures identify the interpolation regions by assuming that the data are normally distributed. In contrast to the Euclidean distance, the Mahalanobis distance takes into account the correlation between descriptor axes. The city-block distance assumes a uniform distribution of data points.

The Euclidean distance, calculated according to Equation 1, places an equal weight on each dimension (descriptor) in the model data space.

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2}$$

(Equation 1)

where $d(i,j)$ is the distance between two points i and j; x_{ip} is the value of point i along axis p; and x_{jp} is the value of point j along axis p.

To increase the accuracy and relevance of the similarity measure, a correction is made to account for the fact that not all descriptors to the overall model are equally important. A simple correction is to use the weighted Euclidean distance (21), as indicated in Equation 2:

$$d(i,j) = \sqrt{w_1(x_{i1} - x_{j1})^2 + w_2(x_{i2} - x_{j2})^2 + \ldots + w_p(x_{ip} - x_{jp})^2}$$

(Equation 2)

where w_n is the weight assigned based on the importance of the nth descriptor in the model.

In Equation 2, the overall weights of the descriptors in the QSAR model are used. These are obtained by calculating the QSAR model coefficients with auto-scaled (mean-centred and variance-normalised) descriptors. The magnitude of each resulting QSAR model coefficient reflects the relative contribution of a single descriptor to the calculated value of the modelled property. The weights can be obtained by normalising all coefficients, so that the most important descriptor has a coefficient of 1.0.

To compare the usefulness of conventional and the weighted-Euclidean distances, a data set was

compiled for a series of similar QSAR models for the prediction of boiling points (21–23). Two main observations were made. Firstly, the conventional Euclidean distance did not always correctly determine when a chemical was in the domain of the model (i.e. when it should be accurately predicted). Chemicals with large distances could still be predicted accurately. In other words, a model would not be considered to be applicable to these new chemicals, despite the fact that it predicted them accurately. Secondly, the use of weighted-Euclidean distances improved this situation. In all cases where the external prediction set member had a relative distance greater than 1.0, it also had a less accurate prediction.

These observations lead to two general conclusions. Firstly, it is necessary to account for the influence of each descriptor in the model when quantifying similarity based on the model data space. The descriptors do not contribute equally to the model prediction, so they should not be expected to contribute equally to the assessment of molecular similarity. Such a modification appears to avoid the discounting of a model when it is actually applicable to a new chemical. Secondly, it is insufficient to determine molecular similarity only in the model data space. It is possible that a new chemical can be very similar to the training set chemicals in all respects that the model considers, but it could have an additional feature that also affected the property in question, and that was not properly accounted for by the model.

Hotelling's test and leverage

The Hotelling's T^2 statistic is the multivariate equivalent of Student's *t* statistic, and provides a check for observations adhering to multivariate normality (24). A similar statistic is the leverage value (25), which is proportional to the Hotelling T^2 and to the Mahalanobis distance. Both Hotelling T^2 and leverage correct for co-linearity in the descriptors through the use of the covariance matrix.

The model space can be represented by a two-dimensional matrix comprising *n* chemicals (rows) and *k* variables (columns), called the descriptor matrix (X). The leverage of a chemical provides a measure of the distance of the chemical from the centroid of X. Chemicals close to the centroid are less influential in model building than are extreme points. The leverages of all chemicals in the data set are generated by manipulating X according to Equation 3, to give the so-called Influence Matrix or Hat Matrix (H).

$$H = X(X^TX)^{-1}X^T \qquad \text{(Equation 3)}$$

where X is the descriptor matrix, X^T is the transpose of X, and $(A)^{-1}$ is the inverse of matrix A, where $A = (X^TX)$.

The leverages or hat values (h_i) of the chemicals (i) in the descriptor space are the diagonal elements of H, and can be computed by Equation 4 (25).

$$h_{ii} = x_i^T(X^TX)^{-1}x_i \qquad \text{(Equation 4)}$$

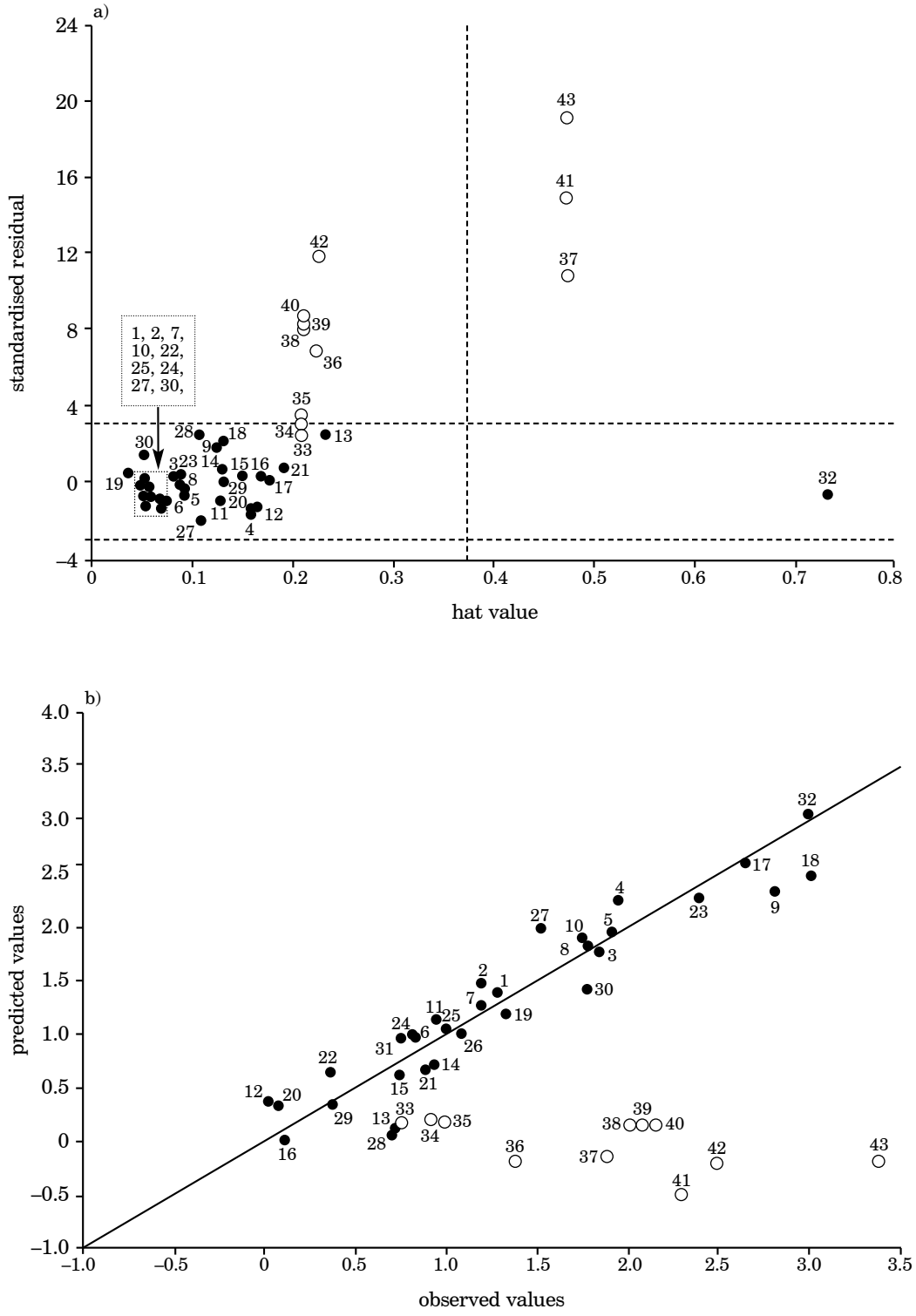
where x_i is the descriptor row-vector of the query chemical.

A "warning leverage" (h^*) is generally fixed at $3p/n$, where *n* is the number of training chemicals, and *p* the number of model variables plus one. A chemical with high leverage in the training set greatly influences the regression line: the fitted regression line is forced near to the observed value and its residual (observed-predicted value) is small, so the chemical does not appear to be an outlier, even though it may actually be outside the AD. In contrast, if a chemical in the test set has a hat value greater than the warning leverage h^*, this means that the prediction is the result of substantial extrapolation and therefore may not be reliable.

The Hotelling T^2 and leverage statistics can be used in the assessment of whether test chemicals fall outside the QSAR AD, as illustrated by Gramatica *et al.* (26), Tropsha *et al.* (27), and Eriksson *et al.* (8). The observation that a chemical has a hat value greater than the warning leverage indicates that the chemical falls outside the AD. However, the observation that a chemical has a hat value less than the warning leverage does not necessarily indicate that the chemical falls within the AD. Chemicals may also fall outside the AD, if they are outliers as defined by their large standardised residuals. To identify chemicals that are outside the AD on the basis of both leverages and standardised residuals, the Williams plot is sometimes used, as illustrated in Figure 4.

The model used for this illustration, taken from Kulkarni *et al.* (28), uses three descriptors for 32 chemicals (six chemicals were excluded as outliers) to predict acute toxicity to the fish, *Pimephales promelas*, where $p = 4$, $n = 32$, and $h^* = 3 \times 4/32 = 0.375$. When the model was redeveloped (29) by using a training and a test set, four points with extreme leverages were identified on the Williams graph (Figure 4a). Of these, one chemical (compound 32) with a large leverage from the training set, was predicted correctly, but would be expected to have a disproportionate influence on the regression line, whereas three chemicals (compounds 37, 41, and 43) with a large leverage from the test set, were not well predicted. At the same time, there were several additional outliers which were not identified on the basis of leverage alone (Figure 4b). In the assessment of a new chemical for which a prediction can be made, but for which there is no experimental value, it is not possible to determine the standardised residual, so the conclusion can only be based on the leverage. Thus, the leverage can be useful in identifying some of the chemicals

Figure 4: Examination of outliers in a regression-based quantitative structure-activity relationship



● = *training*, ○ = *test*.

a) Williams plot, i.e. plot of standardised residuals versus hat values, with a warning leverage of 0.375.

b) Plot of predicted versus observed toxicity.

Both plots were derived by re-analysis of data from Kulkarni et al. *(28).*

that fall outside the AD, but it does not necessarily identify all of them.

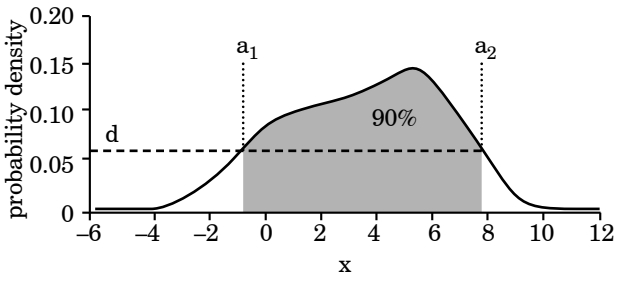
Probability density distribution-based methods

The probability density function of a data set can be estimated by parametric or non-parametric methods (6). Parametric methods assume that the density function has the shape of a standard distribution (for example, a Gaussian or Poisson distribution). Alternatively, a number of non-parametric techniques are available which do not make any assumptions about the data distribution. Non-parametric techniques allow the probability density to be estimated solely from data by kernel density estimation or mixture density methods. For the assessment of QSAR ADs, emphasis has been on the investigation of non-parametric techniques.

Probability density methods are the only methods capable of identifying internal empty regions within the convex hull of a QSAR AD. Furthermore, if empty regions are located close to the convex hull border, probability density methods can generate concave regions to reflect the actual data distribution. The first step is to estimate the probability density of the data set. Density estimation is an area of extensive research, but most methods focus on low dimensional (1D, 2D, 3D) densities, unless some further assumptions are made (30). An algorithm for multivariate kernel density estimation has been developed by Gray & Moore (31, 32).

The next step after the probability density estimation is to find the smallest region that comprises some predefined fraction of the total probability mass. The smallest interval (in 1D) or multidimensional region (>1D), comprising (1-α)*100 percent of the probability mass, where ($0 < \alpha < 1$), is known as the (1-α)-highest density region (HDR). A 90% HDR is illustrated in Figure 5.

Figure 5: Probability density and the 90% highest density region (HDR)



d = the probability density corresponding to the upper (a_2) and lower (a_1) limits of x, which define the boundaries of the HDR. x defines the 1D property space.

It is not a trivial task to calculate the HDR, because it becomes increasingly computationally intensive for higher dimensions, unless one assumes a Gaussian model or another parametric distribution (33). An example of probability density estimation applied to the model of Dimitrov *et al.* (19) is illustrated in Figure 6.

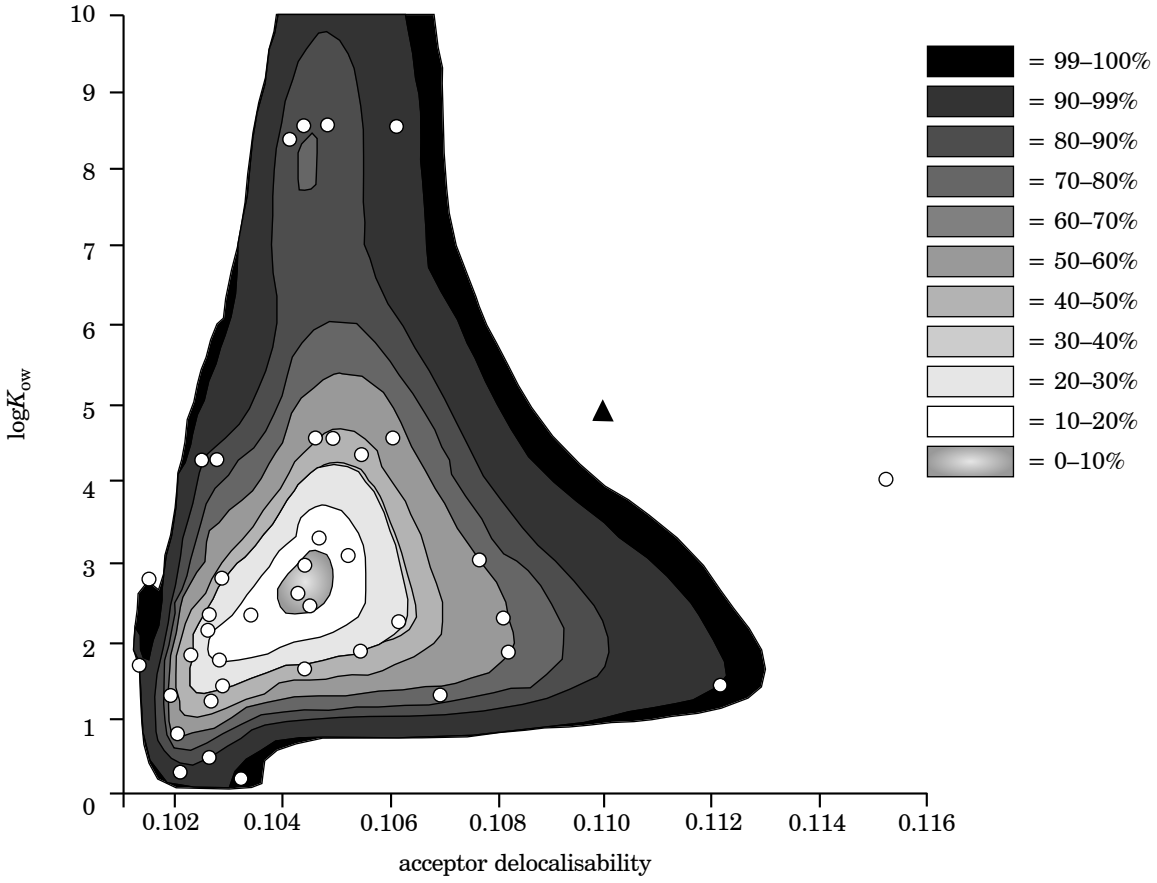
Another probabilistic approach for assessing the AD of regression-based QSARs, based on the "joint applicability domain", has been proposed by T. Aldenberg (unpublished results). In this approach, the probability contours for the joint distribution of *X* (predictor) and *Y* (response) is calculated on the basis of the bivariate or multivariate distribution. This is then used to identify data points as inside or outside the domain.

Probability cuts-offs can be provided by the well-known graphical device in exploratory data analysis called the box-and-whisker (or simply box) plot (34). For univariate empirical data, "near" outside values are characterised by being beyond 1.5 times the interquartile range above the third quartile (or below the first quartile). Extreme ("far") outside values are those located beyond 3.0 times the interquartile range above the third quartile, or below the first quartile. When a data set displays several near or far outside values, it may contain erroneous data, and/or the data may come from another distribution (for example, a more skewed distribution), in which case a transformation may be needed.

The idea of "outside values" can be transferred to the multivariate normal distribution. Points within the (elliptical) contour that capture 99% of the probability distribution are inside points. They are in the "code green" zone. The points between this boundary and the outer ellipse covering 99.99% of the cases are "outside". A model should be used for those values with caution. This is the "code orange" zone. Data points outside the outer ellipse are in the "code red" zone. The model should not be used for these predictions.

Figure 7 illustrates the joint applicability domain of a single-descriptor QSAR, in which the predictor is $\log K_{ow}$ and the response is mutagenicity (log TA98). The original model by Debnath *et al.* (35) was based on a data set of 88 chemicals. The labelled points are 18 chemicals taken from Glende *et al.* (36). It can be seen that eight of the labelled compounds fall in the "green zone", eight fall in the "orange zone", and two in the "red zone". At present, the joint AD approach has been found to be useful when applied to single-descriptor QSARs. It is considered possible to extend the joint AD approach to more-complex models (involving the *x*-dependent conditional distribution of *Y)*, thus providing a more-sophisticated estimate of the predictor (descriptor) domain, but this needs further research.

Figure 6: Interpolation region estimated with a kernel density approach



The training set was used to derive a two-dimensional linear regression model for acute fish toxicity on the basis of two descriptors: logK_{ow} and acceptor delocalisability of the ether oxygen of the ester group (19). The triangle represents a new chemical whose toxicity is to be predicted by the qualitative structure-activity relationship.

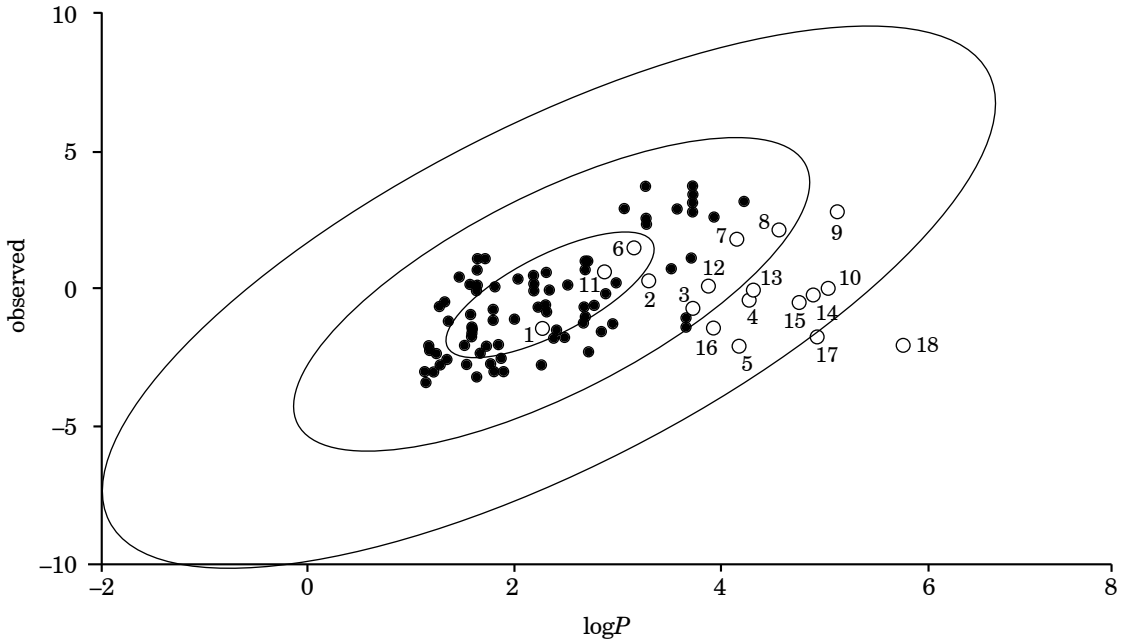
Statistical QSARs based on Structural Descriptors

It is important to discriminate between QSARs that predict physical properties and those that predict (bio)chemical activities, because there are conceptual differences between the two that affect the assessment of their ADs. Physical properties (for example, molecular weight, solubility, partitioning properties) are global properties, in that every atom of the molecule contributes to the observed property, and while the relative locations of the atoms is relevant, the effect is mostly limited to immediate neighbouring atoms. The ability to define the ADs of such QSARs is limited by lack of prior knowledge of the contribution of the selected descriptors to the properties of tested molecules. The applicability of a QSAR is compromised when the values of one or more descriptors fall outside the range of values used in the derivation of the model.

Chemical and biochemical activities (for example, chemical reactivity, metabolism, biodegradation, some toxic and pharmacological properties, and active membrane transport) are properties which are determined primarily by a specific part of the molecule. These properties result from a specific "chemical functionality" that must be present in the molecule, and which enables the molecule to bind or react in a defined way. Thus, in contrast to models for physical properties, not all molecules will exhibit the biological property, since they need the proper structural feature(s) to be active. If the "chemical functionality" is unknown, or several functionalities need to be present for the activity, the ability to define the AD is complicated by the difficulty of ensuring that the appropriate structural features are represented in both the training and test sets.

The possible occurrence of unknown fragments in a test set is inevitable when applying QSARs based on structural fragments. It is widely accepted that the accuracy of prediction for molecules with unknown fragments is lower than the accuracy for those that contain known fragments. Therefore, it is considered important to inform the user of the

Figure 7: Joint applicability domain of a single-descriptor quantitative structure-activity relationship



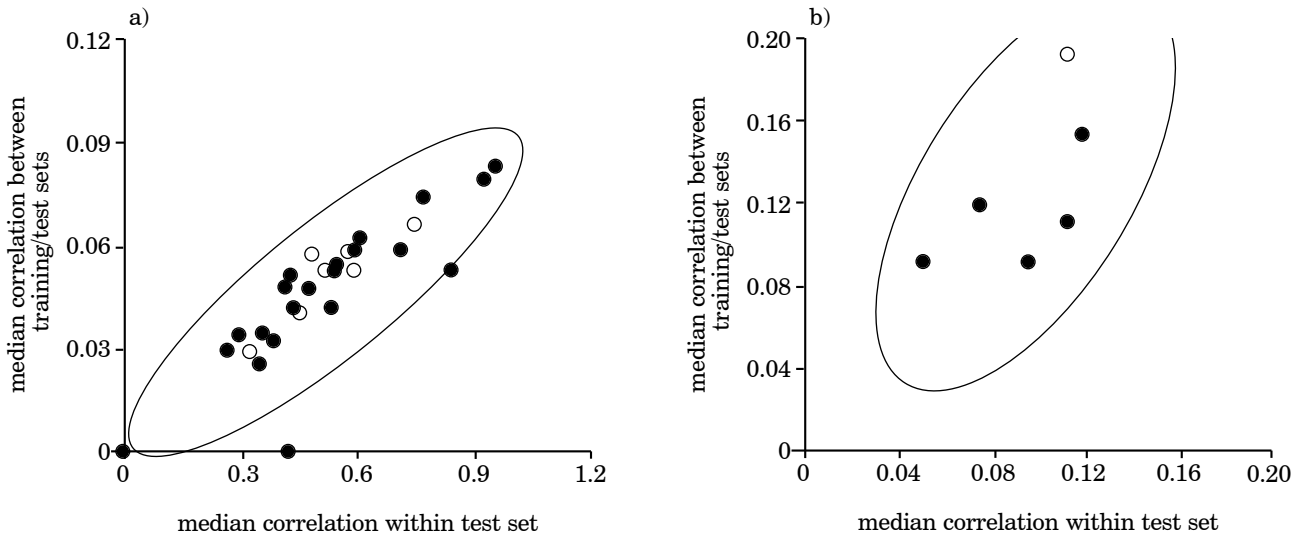
The elliptical contours contain a given fraction of the bivariate normal probability mass: 50%, 99% (used as the AD) and 99.99%. The model may be used for predictions lying inside the inner two ellipses (green zone). However, the model should not be applied outside the outer ellipse (red zone), and should only be used with caution for predictions lying between the outer two ellipses (orange zone).

QSAR model when an unknown fragment appears. The generation of such a warning is given with QSAR models in the MULTICASE platform. MCASE (and MC4PC) evaluate the structural features of a set of non-congeneric molecules and identify the substructural fragments, called biophores, that are considered responsible for the observed activity (37). Chemicals containing the same biophore are grouped into subsets for which independent QSAR models are developed. The descriptors of these models are called modulators, and consist of fragments found within the individual sets, as well as calculated transport, partitioning and quantum mechanical properties that may be relevant to the chemical activities of the individual biophores. The result of this operation is a set of QSAR models for the congeneric sets of molecules containing the same biophore, which is identified as a "chemical functionality" responsible for the observed property. The validity of the model is expressed as the probability that the corresponding biophore is indeed related to activity. Model predictions are also accompanied by an assessment of whether every group of three bonded non-hydrogen atoms in the test structure has been seen and therefore evaluated by the model builder, or not seen and therefore of unknown effect on the prediction results (38, 39).

Another way of expressing the AD of a QSAR with structural fragments is by calculating similarity measures. An example based on the experience of Leadscope Inc. includes the generation of a warning for unknown substituents, as well as determination of whether a compound is sufficiently similar to the training set. For the latter, the Tanimoto score is calculated. A test compound or the entire test set can be compared with the entire training set. In Figure 8, a test set is compared with the training set for two models. For each compound in the test set, a pairwise similarity score is computed. The median pairwise similarity within the test set is then obtained for each test compound. Next, for each compound in the test set, a pairwise similarity score is calculated for all the compounds in the training set. The median pairwise similarity between the test and training sets is calculated for each test compound. Figure 8 shows the correlation of median similarities "within" the test set against the median similarities "between" test and training sets. Those compounds within the density ellipse are structurally similar to those in the training set, and can be considered to lie within the AD of the model.

The AD of QSARs with structural fragments can also be defined in terms of the coverage of fragmen-

Figure 8: Applicability domain based on a measure of similarity defined by Leadscope Inc. (Columbus, OH, USA)



a) Global; b) aromatic amines.

● *= points with correct predictions;* ○ *= points with incorrect predictions.*

tal space. In this case, the above-mentioned interpolation methods (for example, ranges, distances, leverages and probability density approaches) are applicable. However, some data pre-processing is generally needed, due to the nature of the data. For example, a scaling of descriptors (for example, in the range 0–1) is useful when the descriptors display different numerical ranges, to ensure that all the variables have the same chance of influencing a regression model. Another important pre-treatment is to analyse the correlations between the descriptors, and, if the descriptors are highly correlated, to apply PCA to develop new orthogonal axes (new descriptors).

A comparative study of different interpolation methods (40) was applied to the Syracuse Research Corporation (SRC) KOWWIN model for logK_{ow} prediction, which uses the group-contribution method (41). In this study, it was shown that the probability density approach is more restrictive than the range, distance and leverage methods. As a criterion of success, the root mean square error (RMSE) of the compounds from an external validation set (the same for all methods) that fall in the AD of the model was considered (at a cut-off threshold equal to the lowest probability density value of a training set data point). The result was not surprising, since probability density approaches do not require descriptors to be normally distributed, and are therefore well suited for descriptors based on structural fragments. However, the result was also attributed to a dramatic reduction in the number of structures that were classified "in the AD" (nearly half of the total number of chemical structures that were considered to be in the AD by the other methods).

The SAR Applicability Domain

The term SAR describes a qualitative relationship, which means that no mathematical model needs to be applied in order for a prediction for a new chemical to be made. The simplest example is the structural alert. Although a SAR may be qualitative, it is not necessarily the case that the derivation of the SAR itself is achieved by non-statistical means. A structural alert, for example, can be identified by the automated statistical analysis of a training set of chemicals or by expert judgement. In the former case, useful structural alerts can be generated, even in the absence of mechanistic understanding. In the latter case, additional information relating to the known or putative mechanism of action of a chemical can be considered, and this additional information may compensate for gaps in the available data set.

An example of the structural alert approach has been published by Ashby and co-workers for the identification of chemicals with carcinogenic potential based on their DNA reactivity, either directly or following metabolic activation (42, 43). Such struc-

tural alerts can be applied to new chemicals by expert judgement, or can be incorporated into computer systems (44). The automated use of structural alerts facilitates their rapid and reproducible use in the absence of human error.

Several computer systems for toxicity prediction make use of structural alerts (in addition to rules based on physicochemical properties). DEREK for Windows (45) is an example of such a system, and is used here for illustration purposes. Other systems include HazardExpert (46), the OncoLogic carcinogenicity prediction program (47), and the decision support system for irritation and corrosivity developed by the German Bundesinstitut für Risikobewertung (BfR), formerly called the Bundesinstitut für gesundheitlichen Verbraucherschutz und Veterinärmedizin (BgVV; 48).

DEREK for Windows is an expert system that makes use of a knowledge base composed of structural alerts, examples and rules, each of which may contribute to the toxicity predictions made by the system. Each alert in the knowledge base describes the relationship between a structural feature, or toxicophore, and the toxicological endpoint with which it is associated. When a chemical is processed, the system reports any matches of alerts present in the knowledge base with the query structure. For example, decanoyl chloride is predicted by DEREK for Windows to be a possible skin sensitiser in humans, as a result of the presence of alert describing the relationship between a carboxylic acid halide group and the occurrence of skin sensitisation.

The AD for an alert of this type can be defined simply in terms of the scope of the alert. If a chemical contains the alert, then it lies within the domain; if it does not contain the alert, then it lies outside the domain, in which case no conclusion for or against toxicity can be drawn. This scenario is analogous to the situation with any QSAR model. For example, a query chemical may lie outside the AD of a QSAR describing the skin sensitisation of carboxylic acid halides, either because it is a carboxylic acid halide, but possesses some property which is not adequately represented in the model, or because it is a member of an entirely different chemical class. In either case, no conclusion can be drawn about the skin sensitisation potential of the chemical, because activity may still occur by some entirely different mechanism.

The skin sensitisation alert for carboxylic acid halides in DEREK for Windows is comparatively simple in scope and, as a result, many chemicals which contain this functional group will activate the alert. In practice, it is unlikely that all such compounds in the chemical universe will exhibit skin sensitisation. However, in the absence of toxicity data for chemicals of sufficient structural diversity, more-stringent constraints to the scope of the alert cannot currently be defined. This is equivalent to the generation of a QSAR model from a training set of chemicals which identifies, for example, an electronic descriptor as the primary determinant of the observed biological activity. Other physicochemical properties, such as steric parameters, may also be influential, but, unless sufficient variation in these parameters is represented within the training set, their importance may not be identified during development of the QSAR model. As a consequence, a query chemical with steric properties which differ significantly from the chemicals in the training set, would appear to lie within the AD, but the resulting prediction could be unreliable. More-refined alerts can be derived for chemical classes where more toxicity data and other supporting evidence are available. Refinements of the alert provide information on the boundaries of an alert, and can take at least two forms. One type of refinement refers to the presence of particular functional groups and their locations, which leads to some compounds within the general alert class being excluded as active. Another type of refinement refers to a range of physicochemical values, outside which the alert is not considered a reliable indicator of activity.

As an illustration of the association of physicochemical ranges with structural alerts, DEREK for Windows makes use of SAR rules which are dependent on physicochemical and toxicologically relevant biological properties. For example, query chemicals with a molecular weight above 1000 are considered unlikely to result in oestrogenic activity, according to reported screening filters (49). The rule concerned in this case can be applied universally, on the basis that a molecular weight can be unambiguously calculated for any chemical of defined composition, and on the mechanistic understanding that chemicals above a certain size will be physically too large to bind to the oestrogen receptor. On the other hand, the system considers that the likelihood of skin sensitisation in humans is reduced for chemicals with a percutaneous absorption below 10^{-5}cm/hour. Currently, percutaneous absorption is determined from the Potts & Guy equation (50), as applied to all chemicals for which molecular weight and logK_{ow} values can be calculated. The algorithm used to calculate the logK_{ow} value for use in the Potts & Guy equation, can itself be considered a QSAR model and will therefore be associated with its own AD.

For QSARs, one approach for avoiding the inappropriate application of a model in cases where a query chemical appears to fall within the AD, involves the use of a similarity measure to compare the query chemical with those present in the training set. This similarity measure should ideally reflect the mechanistic basis of the QSAR, although the use of structural analogy alone may be adequate in situations where the mechanism is unclear. The same approach could also be applied to the applica-

tion of structural alerts for a particular query chemical, provided that the training set for each alert is available. Currently, each DEREK for Windows alert includes links only to selected chemicals from the training set, chosen to reflect the scope of the alert, in the interests of conciseness and, in some instances, data confidentiality.

The Applicability Domain of Decision Trees and Decision Forests

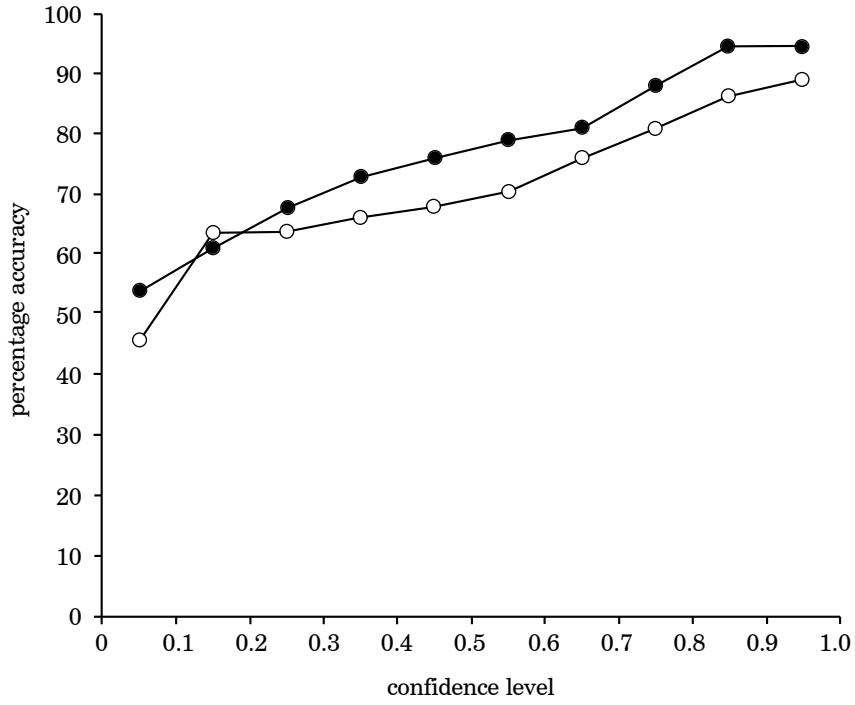
This section addresses the definition of ADs for models based on decision tree (DT) and decision forest (DF) approaches.

An approach developed by Tong and colleagues has been applied to a novel DF consensus modelling method (51, 52), which uses the consensus prediction of multiple, comparable and heterogeneous DTs. The critical assumption in consensus modelling is that multiple models will effectively identify and encode more aspects of the relationship than will a single model. The DF method attempts to minimise overfitting by combining DTs and by maximising the differences among individual DTs, thereby cancelling some random noise. The approach specifies the AD in terms of prediction confidence and domain extrapolation.

Prediction confidence is a measure of the certainty of prediction of a specific chemical. In the DF method, prediction confidence is probabilistically calculated for each unknown chemical by averaging the predictions over all the DTs that are combined to form the model. Figure 9 gives an example to illustrate how prediction accuracy and prediction confidence are related. Prediction accuracy is plotted versus prediction confidence for both a DT and a DF, for a problem where oestrogen receptor binding activity was modelled by applying 2000 runs of 10-fold cross-validation to a data set containing 232 chemicals (ER232). A strong trend of increasing accuracy with increasing confidence is apparent for both the DT and the DF, as is the substantially higher accuracy for the DF across the entire range of confidence levels.

Domain extrapolation is the prediction accuracy for a chemical that is outside the training domain, i.e. the model space defined by the training set chemicals. Domain extrapolation can be probabilistically calculated as the average Euclidian distance that an unknown chemical's descriptors in tree

Figure 9: Decision forest prediction accuracy versus confidence level for oestrogen receptor binding



The statistics were calculated by applying 2000 runs of 10-fold cross-validation to a data set containing 232 chemicals.

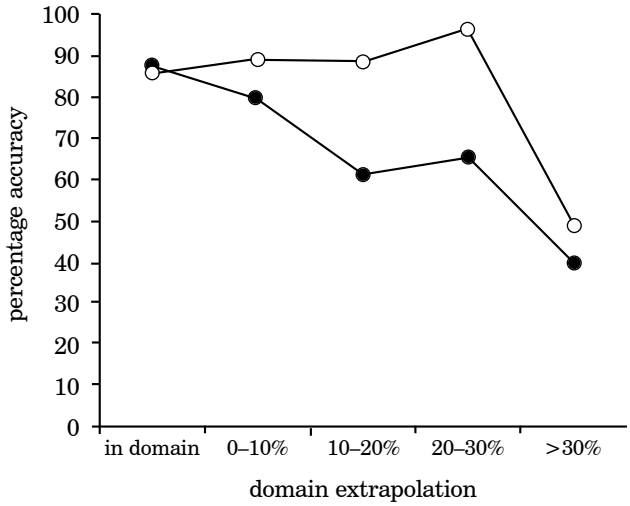
● = *forest;* ○ = *tree.*

paths are outside the range of those same descriptors, based on all chemicals in the training set that determines the AD. Figure 10 shows the results of evaluation of DF domain extrapolation for two oestrogen receptor binding data sets, ER232 containing 232 chemicals, and ER1092 containing 1092 chemicals. Specifically, Figure 10 compares the overall prediction accuracy for chemicals within the training domain with accuracy for chemicals falling several degrees of extrapolation outside the focused domain. In general, the further away the chemicals are from the training domain, the smaller the prediction accuracy, and the larger the data set, the greater the degree of extrapolation for a given prediction accuracy.

Conclusions

1. The validity of a (Q)SAR model depends on its goodness-of-fit, robustness and predictivity. The degree of predictivity needs to be considered in conjunction with the breadth of the AD, since there is generally a trade-off between the two.

Figure 10: Decision forest prediction accuracy versus domain extrapolation for binding to two oestrogen receptors (ER232 and ER1092)



The statistics were generated by performing 2000 runs of 10-fold cross-validation to two data sets (ER232 [○] and ER1092 [●]) containing 232 and 1092 chemicals, respectively (9). Domain extrapolation for a chemical is defined as a percentage away from the training domain, while the prediction accuracy for the domain is calculated by dividing correct predictions by the total number of chemicals in this domain.

2. When determining the properties of an individual chemical by means of a (Q)SAR, the AD of the (Q)SAR is an essential piece of information in judging the reliability of the prediction for that chemical.

3. A general definition of the (Q)SAR AD is proposed by the authors of this report, to cover different types of models and different types of (statistical) modelling approaches. However, it is recognised that the specific definitions for individual models will be model-dependent.

4. The starting point for the definition of any specific (Q)SAR AD is the training domain, i.e. the chemical space of the training set. Therefore, to develop an adequate definition of the AD of a given model, the full training set comprising both structures and descriptors is required.

5. Despite the model-specificity of (Q)SAR ADs, a single conceptual framework can be developed, based on the various elements that can be included in the definition of an AD. The elements identified so far, include the modelling method (SAR, QSAR, decision tree), the philosophy of the modelling approach (mechanistically based, statistically based), the types of descriptor used (structural, physicochemical), and the general AD approach (coverage/interpolation, chemical similarity). These elements are not intended to be mutually exclusive, so the definition of a given AD could be composed of multiple elements.

6. There are various methods for AD estimation that are dependent on a number of factors, including the model dimensionality, the descriptors used and the underlying data distribution. These various methods will agree to the extent that the underlying assumptions in model development are met.

7. There will always be an uncertainty associated with any method for assessing (Q)SAR ADs, just as there is always uncertainty associated with individual (Q)SAR predictions. One type of uncertainty is the “unexpected deviation from the model”, and this relates to the fact that a prediction may fall within the defined AD of a model, and yet still be unreliable, due to the fact that the chemical has some additional property/feature, not accounted for by the model. Another type of uncertainty relates to the fact that a chemical falling outside the defined AD of a given model may still exhibit the response being modelled, because it elicits this response by a mechanism not accounted for by the model in question.

8. Mechanistic information can be useful as a supplement to information provided by mathematical/statistical methods. For example, mechanistic information may be useful when assessing the reliability of predictions in "empty spaces", or when rationalising unexpected deviations from the model.

Recommendations

1. There is a need to develop a global similarity test to determine whether the structural features in a new test compound are covered in the original training set of chemicals (for example, a quantitative measure of uniqueness relative to the training set).

2. Further work is needed to elaborate the conceptual framework proposed in this report.

3. Further work is needed to explore the possibility of associating confidence limits with the AD. Confidence limits could be a useful addition to the AD, since it could be useful to have "fuzzy" boundaries rather than simply "black and white" boundaries.

4. There is a need to develop automated tools, to help (Q)SAR users to appreciate the limitations of the (Q)SAR models they are applying. Mathematical and statistical approaches are particularly well-suited to automation via computer-based tools.

5. The definition of the AD should be the responsibility of the model builder rather than the model user. The reason is that the model developer generally has a better understanding of the training set, the method(s) used for model development, and the limitations of these methods.

6. There is a need for training to improve awareness of the AD concept and its implications for the assessment and application of (Q)SAR models, and to familiarise end-users with automated tools for the assessment of ADs.

7. This report should be used as an input to the development of the OECD Guidance Document on (Q)SAR Validation.

References

1. Anon. (2003). *Proposal Concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH)*. COM(2003)644 final. Brussels, Belgium: European Commission. Website http://europa.eu.int/eur-lex/en/com/pdf/2003/com2003_0644en.html (Accessed 20.12.04).
2. Anon. (2003). Directive 2003/15/EC of the European Parliament and of the Council of 27 February 2003 amending Council Directive 76/768/EEC on the approximation of the laws of the Member States relating to cosmetic products (Text with EEA relevance). *Official Journal of the European Union* **L66**, 26–35.
3. Van der Jagt, K., Munn, S., Tørsløv, J. & de Brujin, J. (2004). *Alternative Approaches can Reduce the Use of Test Animals under REACH. Addendum to the Report "Assessment of Additional Testing Needs under REACH. Effects of (Q)SARs, Risk Based Testing and Voluntary Industry Initiatives"*. JRC Report EUR 21405 EN, 25 pp. Ispra, Italy: European Commission Joint Research Centre. Website http://ecb.jrc.it (Accessed 16.3.05).
4. Worth, A.P., van Leeuwen, C.J. & Hartung, T. (2004). The prospects for using (Q)SARs in a changing political environment: high expectations and a key role for the Commission's Joint Research Centre. *SAR & QSAR in Environmental Research* **15**, 331–343.
5. Schultz, T.W., Cronin, M.T.D., Netzeva, T.I. & Aptula, A.O. (2002). Structure-toxicity relationships for aliphatic chemicals evaluated with *Tetrahymena pyriformis*. *Chemical Research in Toxicology* **15**, 1602–1609.
6. Jaworska, J., Aldenberg, T. & Nikolova, N. (2005). Review of methods for assessing the applicability domains of SARs and QSARs. Final report to the Joint Research Centre (Contract No. ECVA-CCR. 496575-Z). Part 1: Review of statistical methods for QSAR AD estimation by the training set. Website http://ecb.jrc.it/QSAR/Documents (Accessed 16.3.05).
7. Gramatica, P., Pilutti, P. & Papa, E. (2003). Predicting the NO_3 radical tropospheric degradability of organic pollutants by theoretical molecular descriptors. *Atmospheric Environment* **37**, 3115–3124.
8. Eriksson, L., Jaworska, J., Worth, A.P., Cronin, M.T.D., McDowell, R.M. & Gramatica, P. (2003). Methods for reliability, uncertainty assessment, and applicability evaluations of classification and regression based QSARs. *Environmental Health Perspectives* **111**, 1361–1375.
9. Tong, W., Xie, Q., Hong, H., Shi, L., Fang, H. & Perkins, R. (2004). Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity. *Environmental Health Perspectives* **112**, 1249–1254.
10. Nikolova, N. & Jaworska, J. (2003). Approaches to measure chemical similarity: a review. *QSAR & Combinatorial Science* **22**, 1006–1026.
11. Cronin, M.T.D. (2003). Quantitative structure-activity relationships for acute aquatic toxicity: the role of mechanism of toxic action in successful modeling. In *Quantitative Structure-Activity Relationship (QSAR) Models of Mutagens and Carcinogens* (ed. R Benigni), pp. 235–258. Boca Raton, FL, USA: CRC Press
12. Schultz, T.W., Cronin, M.T.D., Walker, J.D. & Aptula, A.O. (2003). Quantitative structure-activity relationships (QSARs) in toxicology: a historical perspective. *Journal of Molecular Structure: THEOCHEM* **622**, 1–22.
13. Bradbury, S.P. & Lipnick, R.L. (1990). Introduction: structural properties for determining mechanisms of toxic action. *Environmental Health Perspectives* **87**, 181–182.

14. Schultz, T.W., Sinks, G.D. & Cronin, M.T.D. (1997). Identification of mechanisms of toxic action of phenols to *Tetrahymena pyriformis* from molecular descriptors. In *Quantitative Structure-Activity Relationships in Environmental Sciences, Vol. VII, Proceedings of QSAR 96, Elsinore, DK, June 24–28, 1996* (ed. F. Chen & G. Schüürmann), pp. 329–342. Pensacola, FL, USA: SETAC Press.
15. Patlewicz, G., Basketter, D.A., Smith, C.K., Hotchkiss, S.A. & Roberts, D.W. (2001). Skin-sensitization structure-activity relationships for aldehydes. *Contact Dermatitis* **44**, 331–336.
16. Roberts, D.W. & Patlewicz, G. (2002). Mechanism based structure-activity relationships for skin sensitisation: the carbonyl group domain. *SAR & QSAR in Environmental Research* **13**, 145–152.
17. Patlewicz, G.Y., Wright, Z.M., Basketter, D.A., Pease, C.K., Lepoittevin, J.P. & Arnau, E.G. (2002). Structure-activity relationships for selected fragrance allergens. *Contact Dermatitis* **47**, 219–226.
18. Anon. (2000). *US patent no. 6 036 349: Method and Apparatus for Validation of Model-based Predictions*. Issued March 14, 2000. Washington, DC: USA.
19. Dimitrov, S.D., Mekenyan, O.G., Sinks, G.D. & Schultz, T.W. (2003). Global modeling of narcotic chemicals: ciliate and fish toxicity. *Journal of Molecular Structure: THEOCHEM* **622**, 63–70.
20. Preparata, F.P. & Shamos, M.I. (1991). *Computational Geometry: An Introduction*, 390pp. New York, NY, USA: Springer Verlag.
21. Stanton, D.T. & Jurs, P.C. (1991). Computer-assisted prediction of normal boiling points of furans, tetrahydrofurans, and thiophenes. *Journal of Chemical Information and Computer Sciences* **31**, 301–310.
22. Stanton, D.T., Egolf, L.M. & Jurs, P.C. (1992). Computer-assisted prediction of normal boiling points of pyrans and pyrroles. *Journal of Chemical Information and Computer Sciences* **32**, 306–316.
23. Stanton, D.T. (2000). Development of a quantitative structure-property relationship model for estimating normal boiling points of small multifunctional organic molecules. *Journal of Chemical Information and Computer Sciences* **40**, 81–90.
24. Seber, G.A.F. (2004). *Multivariate Observations*, 686pp. New York, NY, USA: John Wiley & Sons.
25. Atkinson, A.C. (1991). *Plots, Transformation, Regression*, 282pp. Oxford, UK: Clarendon Press.
26. Gramatica, P., Pilutti, P. & Papa, E. (2004). Validated QSAR prediction of OH tropospheric degradation of VOCs: splitting into training-test sets and consensus modeling. *Journal of Chemical Information and Computer Sciences* **44**, 1794–1802.
27. Tropsha, A., Gramatica, P. & Gombar, V. (2003). The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR & Combinatorial Science* **2**, 69–77.
28. Kulkarni, S.A., Raje, D.V. & Chakrabarti, T. (2001). Quantitative structure-activity relationships based on functional and structural characteristics of organic compounds. *SAR and QSAR in Environmental Research* **12**, 565–591.
29. Gramatica, P. (2004). *Evaluation of Different Statistical Approaches to the Validation of Quantitative Structure-activity Relationships. Final report to the Joint Research Centre.* Contract No. ECVA-CCR. 496576-Z. 177pp. Website http://ecb.jrc.it/QSAR/Documents (Accessed 16.3.05).
30. Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, 176pp. London, UK: Chapman & Hall.
31. Gray, A. & Moore, A. (2003). Nonparametric Density Estimation: Toward Computational Tractability. In *Proceedings of SIAM International Conference on Data Mining, San Francisco, USA, 2003*, 9p. Website www.siam.org/meetings.sdm03/ (Accessed 16.3.05).
32. Gray, A. & Moore, A. (2003). Very fast multivariate kernel density estimation using via computational geometry. *Proceedings of Joint Statistics Meeting 2003*. Alexandria, VA, USA: The American Statistical Association (Website http://www.amstat.org/meetings/jsm/2003 (Accessed 16.3.05).
33. Chen, M-H. & Shao, Q-M. (1999). Monte Carlo estimation of bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics* **8**, 69–92.
34. Tukey, J.W. (1977). *Exploratory Data Analysis*, 688pp. Reading, UK: Addison-Wesley.
35. Debnath, A.K., Debnath, G., Shusterman, A.J. & Hansch, C. (1992). A QSAR investigation of the role of hydrophobicity in regulating mutagenicity in the Ames test. I. Mutagenicity of aromatic and heteroaromatic amines in Salmonella typhimurium TA98 and TA100. *Environmental and Molecular Mutagenesis* **19**, 37–52.
36. Glende, C, H. Schmitt, L. Erdinger, G. Engelhardt, & G. Boche (2001). Transformation of mutagenic aromatic amines into non-mutagenic species by alkyl substituents. Part I. Alkylation *ortho* to the amino function. *Mutation Research* **498**, 19–37.
37. Klopman, G. (1992). MULTICASE: a hierarchical computer automated structure evaluation program. *Quantitative Structure-Activity Relationships* **11**, 176–184.
38. Klopman, G. & Chakravarti, S.K. (2003). Structure-activity relationship study of a diverse set of estrogen receptor ligands (I) using MultiCASE expert system. *Chemosphere* **51**, 445–459
39. Klopman, G. & Chakravarti, S.K. (2003). Screening of high production volume chemicals for estrogen receptor binding affinity (II) by the MultiCASE expert system. *Chemosphere* **51**, 461–468.
40. Jaworska, J., Aldenberg, T. & Nikolova, N. (2005). Review of methods for assessing the applicability domains of SARs and QSARs. Final report to the Joint Research Centre (Contract No. ECVA-CCR.496575-Z). Part 2: An approach to determining applicability domain for QSAR group contribution models: an analysis of SRC KOWWIN. Website http://ecb.jrc.it/QSAR/Documents (Accessed 16.3.05).
41. Meylan, W.M. & Howard, P.H. (1995). Atom fragment contribution method for estimating octanol-water partition-coefficients. *Journal of Pharmaceutical Sciences* **84**, 83–92.
42. Ashby, J., Tennant R.W., Zeiger, E. & Stasiewicz, S. (1989). Classification according to chemical structure, mutagenicity to Salmonella and level of carcinogenicity of a further 42 chemicals tested for carcinogenicity by the U.S. National Toxicology Program. *Mutation Research* **223**, 73–103.
43. Tennant, R.W. & Ashby, J. (1991). Classification according to chemical structure, mutagenicity to Salmonella and level of carcinogenicity of a further 39 chemicals tested for carcinogenicity by the U.S. National Toxicology Program. *Mutation Research*

257, 209–227.

44. 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., Watson, W.P. & Yih, T.D. (1996). Computer prediction of possible toxic action from chemical structure: an update on the DEREK system. *Toxicology* **106**, 267–279.
45. Judson, P.N., Marchant, C.A. & Vessey, J.D. (2003). Using argumentation for absolute reasoning about the potential toxicity of chemicals. *Journal of Chemical Information and Computer Sciences* **43**, 1364–1370.
46. Smithing, M.P. & Darvas, F. (1992). HazardExpert: an expert system for predicting chemical toxicity. In *Food Safety Assessment* (ed. J.W. Finley, S.F. Robinson & D.J. Armstrong), ACS Symposium Series, pp. 191–200. Washington, DC, USA: American Chemical Society.
47. Woo, Y., Lai, D.Y., Argus, M.F. & Arcos, J.C. (1995). Development of structure-activity relationship rules for predicting carcinogenic potential of chemicals. *Toxicology Letters* **79**, 219–228.
48. Gerner, I., Zinke, S., Graetschel, G. & Schlede, E. (2000). Development of a decision support system for the introduction of alternative methods into local irritancy/corrosivity testing strategies. Creation of fundamental rules for a decision support system. *ATLA* **28**, 665–698.
49. Hong, H., Tong, W., Fang, H., Shi, L., Xie, Q., Wu, J., Perkins, R., Walker, J.D., Branham, W. & Sheehan, D.M. (2002). Prediction of estrogen receptor binding for 58,000 chemicals using an integrated system of a tree-based model with structural alerts. *Environmental Health Perspectives* **110**, 29–36.
50. Potts, R.O. & Guy, R.H. (1992). Predicting skin permeability. *Pharmaceutical Research* **9**, 663–669.
51. Tong, W., Hong, H., Fang, H., Xie, Q. & Perkins, R. (2003). Decision forest: combining the predictions of multiple independent decision tree models. *Journal of Chemical Information and Computer Sciences* **43**, 525–531.
52. Tong, W., Hong, H., Xie, Q., Xie, L., Fang, H. & Perkins, R. (2004) Assessing QSAR limitations: a regulatory perspective. *Current Computer Aided Drug Design* **1**, 65–72.