Science in China Series B: Chemistry

Springer

www.scichina.com
chem.scichina.com
www.springerlink.com

Progress and perspectives of quantitative structure-activity relationships used for ecological risk assessment of toxic organic compounds

CHEN JingWen†, LI XueHua, YU HaiYing, WANG YaNan & QIAO XianLiang

Key Laboratory of Industrial Ecology and Environmental Engineering (MOE), Department of Environmental Science and Technology, Dalian University of Technology, Dalian 116024, China

Structure-activity relationship (SAR) and quantitative structure-activity relationship (QSAR), collectively referred to as (Q)SARs, play an important role in ecological risk assessment (ERA) of organic chemicals. (Q)SARs can fill the data gap for physical-chemical, environmental behavioral and ecotoxicological parameters of organic compounds; they can decrease experimental expenses and reduce the extent of experimental testing (especially animal testing); they can also be used to assess the uncertainty of the experimental data. With the development for several decades, (Q)SARs in environmental sciences show three features: application orientation, multidisciplinary integration, and intelligence. Progress of (Q)SAR technology for ERA of toxic organic compounds, including endpoint selection and mathematic methods for establishing simple, transparent, easily interpretable and portable (Q)SAR models, is reviewed. The recent development on defining application domains and diagnosing outliers is summarized. Model characterization with respect to goodness-of-fit, stability and predictive power is specially presented. The purpose of the review is to promote the development of (Q)SARs orientated to ERA of organic chemicals.

(Q)SARs, ecological risk assessment, application domain, stability, predictive power

1 Ecological risk assessment of synthetic organic compounds

It is estimated that there are currently more than 80000 synthetic organic compounds in common use, and the number is increasing yearly by 500—1000. In fact, by the end of 2007, more than 33 million of chemicals had been registered in Chemical Abstracts Service (CAS: http://www.cas.org), and most of which are synthetic organic compounds. The pollution caused by synthetic organic chemicals has brought about severe lessons to human beings. Persistent toxic substances (PTS) in the environment have become an important issue affecting the survival and development of human in the 21st century. It is evident that ecological risk assessment (ERA) for the synthetic organic compounds can provide a precaution against the pollution[1]. ERA includes three primary phases as defined by U.S. Environmental Protection Agency (US EPA): problem formulation, analysis, and risk characterization[2]. It is obvious that data on physicochemical properties, environmental behavior and ecotoxicology of organic compounds, are indispensable for the ERA. However, the data have three aspects of problems:

(1) Lack of the data[3]. For example, for more than 80% of the common used synthetic organic chemicals, their environmental behavioral and ecotoxicological data are currently not available. Experimental determination

Received August 14, 2007; accepted December 3, 2007
doi: 10.1007/s11426-008-0076-6
†Corresponding author (email: jwchen@dlut.edu.cn)
Supported by the National Basic Research Program (973) of China (Grant No. 2006CB403302)

of the parameters is time-consuming, lags behind the regulatory needs, and cannot meet the "precautionary principle" for chemicals management.

(2) Large expense of testing[4]. According to the estimation by the "Registration, Evaluation and Authorization of Chemicals (REACH)" legislation that has been brought into effect since June 2007, the testing fee for a single chemical was about 85000 euros, while the full test for a new chemical cost about 570000 euros. In addition to the high expense, comprehensive testing of all the chemicals does not accord with the principle and trend of reducing testing (especially animal testing).

(3) Uncertainty in data[5]. For instance, the scientists in U.S. Geological Survey found variations of up to 4 orders of magnitude in the reported octanol-water partition coefficient (K_{OW}) values for DDT and its metabolite DDE. The large uncertainty in the physicochemical data will inevitably lead to high uncertainty of ERA results.

Molecular structures are internal factors governing the physicochemical properties, environmental behavior and ecotoxicology of organic compounds[6]. Compounds with similar molecular structures should have similar physicochemical properties, environmental fate and ecotoxicological effects, i.e., there are inherent relations between molecular structures and their physicochemical properties, environmental behavioral and ecotoxicological parameters[6]. The relations can be characterized as mathematical models, termed as structure-activity relationships (SARs) or quantitative structure-activity relationships (QSARs), collectively referred to as (Q)SARs. Thus, (Q)SARs can fill the data gap for physicochemical, environmental behavioral and ecotoxicological parameters of organic compounds; they also can decrease experimental expenses and reduce the extent of experimental testing (especially animal testing). Furthermore, (Q)SARs can be used as supporting tools to evaluate the adequacy of the available empirical data of organic compounds, which is also one of the four specific functions of (Q)SARs for ERA[7]. For example, interhomologue consistency for physicochemical properties of polychlorinated biphenyls (PCBs) was illustrated with simple (Q)SARs that use molar mass and the number of chlorine substitutions in ortho-positions as descriptors[8]. Thus (Q)SAR technology is of greate importance to ERA of organic compounds.

2 Principles, methods and development of (Q)SARs

2.1 Principles and methods of (Q)SARs

It has long been realized that there are inherent relationships between molecular structures of organic chemicals, and their physicochemical properties or biological activities. In 1930s, Hammett et al.[9,10] established the linear free energy relationships (LFERs) and brought forward Hammett substituent constants σ, which laid a thermodynamic foundation for (Q)SARs. In the 1950s, Taft[11] further developed LFERs by introducing steric substituent constants E_S. LFERs are extra-thermodynamics, which means that the objective relations between thermodynamic parameters and molecular activities cannot be deduced by thermodynamic theories[12]. The LFER theory has been successfully employed to model partition coefficients[13,14] and reaction rate constants[15,16] of organic pollutants in different environmental media.

As shown in Figure 1, establishment of (Q)SAR models consists of several steps. It is of fundamental importance to obtain and select appropriate molecular structural descriptors to characterize molecular structures. There are two general methods for selecting molecular structural descriptors. The first method relys on experience, the observable or perceivable characters of the studied molecules, and the possible underlying mechanism. For example, photohydrolysis is a main pathway for photolysis of halogenated aromatic compounds, thus various quantum chemical descriptors characterizing the C-X bonds were computed and employed for QSARs development on photolysis quantum yields of halogenated compounds[17,18]. The second method relies on the established models. There are four types of models that have been commonly used in QSARs development: the Hansch model, the linear solvation energy relationship (LSER) model, the Free-Wilson model[19], and the 3-dimensional (3D) analytical method, such as the comparative molecular field analysis (CoMFA) model[20,21].

Hansch extended the use of QSARs to biological activity based on the LFER theory, and suggested that electronic (σ), steric (E_s) and hydrophobic effects of given substituents govern biological activities[22−24]. These effects can be added with each other independently to form different linear or non-linear Hansch mod-

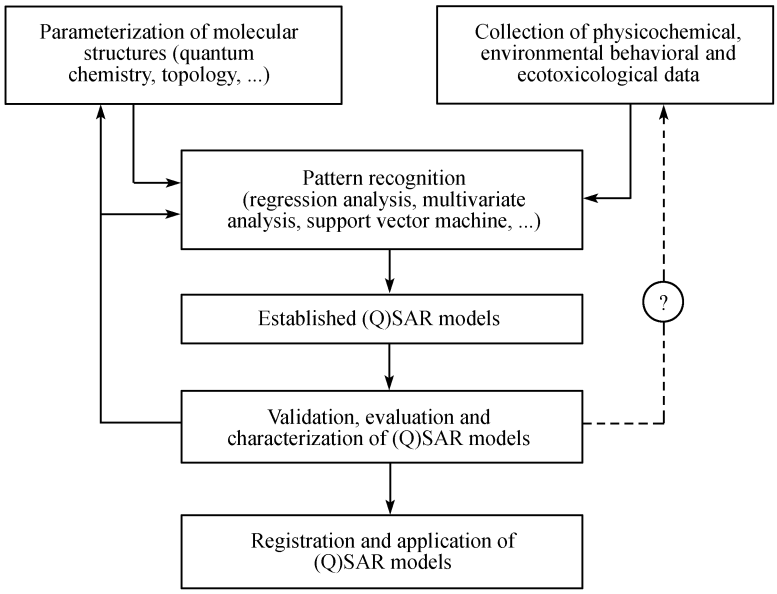

Figure 1 Flow diagram for establishing (Q)SAR models.

els[24,25], which were widely used to establish QSAR models[26,27].

The LSER model developed by Kamlet et al.[28−31] employed cavity term, polarizability, hydrogen-bonding acidity, and hydrogen-bonding basicity, represented by solvatochromic parameters, to describe molecular properties relevant to solvation/partition processes. Abraham developed new LSER parameter scales[32]. Using theoretical molecular structural descriptors to replace the empirical solvatochromic descriptors, Wilson and Famini[33,34] developed the theoretical linear solvation energy relationship (TLSER) models. The LSER or TLSER models were successfully used to estimate water solubility (S_W)[28,29], octanol-water partition coefficients (K_{OW})[30,31], high performance liquid chromatography retention factors[35], and nonreactive toxicity of organic compounds[36,37].

The Free-Wilson model[19] was advanced that the activities of series of chemical analogs change with the amount, the specific side chain arrangements and the performance characteristics. This approach is easy to be performed but only suitable for multi-substituted chemicals.

The CoMFA is a typical three-dimensional (3D) QSAR technique that ultimately allows one to design and predict activities of molecules, for which the foundation is that interaction force fields of series bioactive molecules with the same receptor are similar[20,21]. Even 3D structures of receptors are unclear, one can deduce the properties of receptors, design new chemicals and quantitatively estimate activities of chemicals by studying circumambient interaction force fields of bioactive molecules and quantifying bioactivities. CoMFA has been widely used in quantitative drugs and pesticides design[38,39]. In ecotoxicology, CoMFA has been mainly used to predict toxicity of organic pollutants, such as estrogenic activities of endocrine disrupting chemicals[40,41].

2.2 Development of (Q)SARs

In the early time, (Q)SAR was mainly studied in the field of drug design. Since the 1970s, (Q)SAR technology has been introduced into environmental science to meet the need of ERA for numerous and ever-increasing number of organic chemicals. The steady development of (Q)SAR technique in the last three decades presents the following three features.

(1) Objective-oriented characteristics and applicability. In the field of environmental science or engineering, (Q)SAR study mainly concentrated on exposure assessment (transfer and transformation of organic pollutions in multimedia environment) and effect assessment (mainly ecotoxicological effects). Some early QSAR models mainly focused on properties related with environmental partitioning (such as S_W[28,29], K_{OW}[30,31,42], bioconcentration factors (BCF)[43], octanol-air partition coefficients (K_{OA})[44,45] and soil or sediment adsorption coefficients (K_{OC})[46,47]) and acute

toxicity to aquatic organisms (median lethal concentration (LC_{50})[48] or median effect concentration (EC_{50})[49]). Later, QSARs studies were extended to environmental endocrine disruptor activity[50−52], and reaction kinetics of organic pollutants, such as biodegradability[53], photolysis rate constants[54] or quantum yields[17,18], reduction rate constants by zero-valent iron[15], oxidation rate constants by hydroxyl radicals[16], etc.

In 1993, the journal *SAR and QSAR in Environmental Research* started publication in France. Since 1984, international workshop on QSARs in environmental sciences has been held biyearly[55]. In 2003, 23 reviews on QSARs were published by the journal *Environmental Toxicology and Chemistry* (vol. 22, no. 8), which covered physicochemical properties, environmental behavior, biological activity and ecotoxicological effects of organic chemicals. Progress and applications of QSARs in environmental field were reviewed concentratively and detailedly by the monograph, indicating that (Q)SARs in environmental sciences are in the ascendent.

(2) Interdisciplinary integration. (Q)SAR is an interdisciplinary research field, converging many disciplines such as chemical informatics (chemometrics, computational chemistry), physical chemistry, biochemistry, toxicology, computer science and mathematics, etc.

From the point of view of molecular structure characterization, the aforetime empirical descriptors, such as hydrophobic (π)[22,24,56], electronic (σ)[10,57,58] and steric (E_S)[11,59−61] constants and solvatochromic parameters[28−30,62,63], have evolved into theoretical derived molecular descriptors such as topological indexes[64,65] and quantum-chemical descriptors[66] that are widely used nowadays. As an example, the software Dragon can be used to calculate various molecular structural descriptors describing compound structural diversity, representing 0−3 dimension molecular structures and containing information on atom and bond types, connectivity, partial charges and atomic spatial coordinates. More than 1000 descriptors, including constitutional (atom and group counts) descriptors, functional groups, atom-centered fragments, topological descriptors, molecular walk counts, BCUT descriptors, Galvez topological charge indices, 2D autocorrelations, charge descriptors, aromaticity indices, Randic molecular profiles, geometrical descriptors, radial distribution function (RDF) descriptors, 3D-MoRSE descriptors, WHIM descriptors, and GETAWAY descriptors, can be computed by this software[67]. It is thus the interdisciplinary intergration that drives the evolvement of molecular structural characterization, which lays a beneficial foundation for development of (Q)SAR models.

As to the methods for establishing QSAR models, early QSAR studies mainly employed various linear regression technologies[68−70], and subsequent multivariate analysis methods including factor analysis, principal component analysis (PCA)[71], discriminant analysis[71,72], cluster analysis[71,73], and partial least squares (PLS)[74] regression analysis. In recent decades, various nonlinear analysis methods[75], such as artificial neural network (ANN)[71,76] and support vector machine (SVM)[77,78] were also adopted in establishing QSAR models. Genetic algorithms (GA)[79,80] were used for vairable selection to obtain optimal QSAR models. Some combined algorithms, such as GA-PLS[81,82], GA-SVM[83], GA-BP[84], SVM-PLS[85], etc., were brought forward for the ease of establishing QSAR models. All the machine learning methods gradually improve the model establishment technology of QSARs. Meanwhile, the progress in biochemistry, toxicology and other relevant subjects has deepened the understanding of the modes of toxicity action and promoted the development of (Q)SAR technology.

(3) Intelligence. In recent years, many intelligent, distinctive, user-friendly and user-oriented softwares for development and application of (Q)SARs have been exploited by different goverments, companies and research institutes due to the development of computer technology. The Organization for Economic Cooperation and Development (OECD) surveyed the (Q)SAR softwares for regulatory purposes of chemicals, and found there were more than 40 software copyrights for America, 3 for England, 6 for France, 8 for Canada, 1 for Bulgaria. Considering the (Q)SAR softwares for miscellaneous purposes, the number is conservatively estimated to be larger than 200.

One key direction of (Q)SAR development in the future is related to decision support systems[86]. The systems should incorporate validated (Q)SAR models that meet the consensus of (Q)SAR criteria, have transparent databases with flexible search engines, be web based, and have user friendly interfaces. Such

decision support shystems can facilitate non-(Q)SAR developers for the selection and application of the models in management and decision-making processes.

3 Current development and application of (Q)SARs all over the world

As (Q)SARs can contribute to implementing the "precautionary principle" of organic chemicals management, reducing animal tests, and decreasing testing fee, they have been widely studied and applied to ERA and regulation of toxic organic chemicals in many countries[7,87]. By 2002, countries like America, Canada, Australia, Germany, Denmark, Japan, and Holland applied (Q)SARs in various degrees to predict physicochemical properties, environmental fate and toxicity to aquatic organisms[88], and the endpoints include K_{OW}, K_{OC}, S_W, the boiling point (B_p), the melting point (M_p), the vapor pressure (P), the Henry's law constant (K_H), the oxidation rate constant by hydroxyl radical, the biodegradation rate constant, the *BCF*, the hydrolysis rate constant, etc.[88].

REACH legislation prescribes 3 principles for chemicals management[4]: a) "No data, no market". Chemicals companies must provide information about their safety before the new chemicals are manufactured and put on market. b) Reducing testing, in particular, refining, reducing or replacing animal testing. On one hand, this can reduce testing fee; on the other hand, it can meet the proposal of protecting animal welfare. c) Application of (Q)SAR technology.

According to the REACH, results of (Q)SARs may be used instead of testing when the following conditions are met[4,89]: a) results should be derived from a (Q)SAR model whose scientific validity has been established; b) the substance should fall within the applicability domain of the (Q)SAR model; c) results should be adequate for the purpose of classification, labelling and risk assessment; d) adequate and reliable documentation of the applied method should be provided. Four specific applications of (Q)SARs, including data evaluation, decision for further testing, estimating specific parameters, and identifying data needs on effects of potential concern, were identified in the comprehensive technical guidance document (TGD) of EU in 1996[7].

The European Chemicals Bureau (ECB) (http://ecb.jrc.it/) is the focal point for data and the assessment procedure on dangerous chemicals, which plays an important role in the development of the new legislation on chemicals, i.e. in the establishment of REACH. In recent years, ECB has implemented a lot of research central to the development and application of (Q)SARs, including: a) reporting format, validation and assessment of (Q)SAR models, b) classification technologies of chemicals, c) analogue or read-across technologies of physicochemical properties, environmental behavioral and toxicology parameters, which are involved in the regulatory uses of (Q)SARs for different purposes.

OECD also carried out research on safety of chemicals and application of (Q)SARs. In November 2004, OECD put forward a set of five principles for (Q)SAR validation[90]. In February 2007, OECD issued a guidance document on (Q)SAR validation[91]. In addition, OECD organized case studies on regulatory uses of (Q)SARs in the assessment of existing and new chemicals in the member countries including Australia, Canada, Czech Republic, Denmark, Italy, Germany, Japan, Holland, America, England and EU Commission. The report of the case studies was published in August 2006[92].

In the U. S., many government organizations develop and apply (Q)SAR technologies[88]. The organizations include US EPA, the U. S. Air Force, the Agency for Toxic Substances and Disease Registry (ATSDR), the Toxic Substance Control Act Interagency Testing Committee, the National Oceanic Atmospheric Administration (NOAA), Consumer Product Safety Commission (CPSC), Food and Drug Administration (FDA), National Cancer Institue (NCI), and National Toxicology Program, etc.

In Fiscal Year 2002, the U. S. Congress directed the EPA to provide funds for the research and development of alternatives to traditional toxicological testing procedures on chemicals screening and priority pollutants identification, mainly referring to (Q)SARs technology. The national center for computational toxicology (NCCT) was founded as a part of EPA's Office of Research and Development (ORD) to implement EPA's research in the field of computational toxicology. Over the years, the budget for computational toxicology has been increased gradually in the U.S.

US EPA and Syracuse Research Corporation developed the software EPI Suite ™ (http://www.epa.gov/oppt/exposure), which includes subprograms for predicting K_{OW}, K_{OC}, H, S_W, B_p, M_p, P, *BCF*, biode-

gradability, hydrolysis rate constants, removal efficiency in wastewater treatment plants, etc. US EPA also used QSARs to predict biological effects of high production volume (HPV) chemicals and premanufacture notification (PMN) chemicals, including absorption, distribution, metabolism, excretion, acute toxicity effect, irritation, sensitization, chronic/sub-chronic toxicity effects, reproduction effect, developmental toxicity, carcinogenicity, mutagenicity, etc. Moreover, US EPA developed QSAR models to predict estrogenic effects of chemicals. Information on the research, development and application of QSARs in other countries can be found in the references[86,88].

There are more papers than patents for presenting research results of (Q)SARs. A search performed in the end of 2006 using "QSAR" as keywords in titles and abstracts resulted in 22 patents in the Worldwide Database of European Patent Office (EPO), 11 patents in the database of World Intellectual Property Organization (WIPO), and 8 patents in database of United States Patent and Trademark Office (USPTO).

To summarize, the developed countries have paid much more attention to the research, development and application of (Q)SAR technology for ERA and regulatory purposes. In China, the national natural science foundation has ratified a few research projects relevant with the development of QSARs in the field of environmental chemistry. There are several research groups that mainly concentrate on (Q)SARs for ERA and regulatory purposes, including those from the Nanjing University[93,94], Dalian University of Technology[95,96], Hunan University[97], Lanzhou University[77,98], Changchun Institute of Applied Chemistry (Chinese Academy of Sciences)[99] and Northeast Normal University[100]. However, the (Q)SAR studies implemented in China were not as systematic and thorough as those in the developed countries. In China, more studies are specially necessary in the practical application aspects of (Q)SARs.

4 New progress and perspectives on (Q)SARs

The practical application of (Q)SAR technologies in ERA involves many factors. The workshop held in Setubal in March 2002 was designed to develop more definitive guidance on the use and development of (Q)SARs. The so-called "Setubal principles" proposed that (Q)SARs for regulatory purposes should be associated with the following information[7,86,89]: a) a defined endpoint, b) an unambiguous algorithm, c) a defined domain of applicability, d) appropriate measures of goodness-of-fit, robustness and predictivity, and e) a mechanistic interpretation, if possible. These were officially defined as guidance principles on the development and application of (Q)SARs by OECD in 2004[89]. The (Q)SAR models that meet the principles can thus be used in a much broader scope, including ERA, chemical screening, and priority setting[86,87]. Centering on the above principles, the related works are reviewed as follows.

4.1 Environmental endpoints of (Q)SAR models

Environmental endpoints of (Q)SARs are defined as any physicochemical, environmental behavioral and ecological parameters that can be measured or predicted. These endpoints can be determined by normative experiments in standard conditions. The (Q)SAR models with unambiguous endpoints may help to judge if the predicted values are fit for specific ERA.

It is well known that high-quality experimental data are essential for the development of high quality (Q)SAR models[101]. In terms of environmental endpoints, ideally such data should be measured by a single standardized protocol, even in the same laboratory and by the same workers[102]. The variability in the data, as a result of interlaboratory testing, can introduce errors and unknown bias into (Q)SAR models[103]. Meanwhile, efforts should be made to ensure structural diversity of chemicals, both between and within chemical classes. Such structural diversity allows for development of more robust (Q)SAR models[104]. However, due to the limitation of experimental data, endpoint values from different literature sources were usually collected for developing (Q)SAR models, which can expand data range, increase structural diversity, but also lead to inaccurate predictions. Consequently, the experimental error must be considered in (Q)SAR modelling so as to assure that the goodness-of-fit is within the variation of determined values to prevent over-fitting.

4.2 Algorithm of QSAR models

The best algorithms of QSARs modelling are those that are simple, transparent, easily interpretable, and easily portable. A transparent model can be defined as one that is based on fundamental physicochemical properties

with a clear and unambiguous statement of how the model has been formulated[101]. Such a model is capable of mechanistic interpretation and portability from one user to another, preferably without the requirement for specialist software[102]. Meanwhile, it allows the user to examine and understand how the environmental endpoint is modeled. The transparent characteritics are usuallly achieved by proper mathematical algorithms[102].

Different statistical techniques vary in their transparency that relates to the amount of processed informaton obtainable from the statistical methodology[102]. Generally, the transparency of different statistical methodology follows the order: multiple regression analysis (MLR) > principal component analysis and partial least square regression (PCA & PLS) > artificial neural network (ANN) > genetic algorithm (GA)[105−107]. Nevertheless, the performance of a QSAR model is also related to its robustness. The term robustness means the range of the method's applicability and hence the relative freedom from conditions and its order is just contrary to that of transparency[102]. Therefore, the selection of statistical techniques for (Q)SARs development should consider the environmental endpoint, the purpose of application, the requirement of transparence and robustness.

4.3 Mechanism of (Q)SAR models

The establishment of (Q)SARs should be based on the proper analysis and understanding of mechanisms, and vice versa, the well-established (Q)SAR models should facilitate mechanism interpretations. Mechanisms can make clear the molecular structural factors determining the endpoints needed for ERA. The mechanism of (Q)SAR models can be related with the following two aspects:

(1) Molecular structural descriptors used in (Q)SARs should be capable of mechanistic evaluation. In other words, a (Q)SAR model should be interpretable in terms of the parameters employed[101,102]. For example, compared with molecular connectivity indicies and other topological indices, the fundamental physicochemical properties (such as molecular weight) and quantum chemical descriptors are easier to be interpreted[108].

(2) Interdisciplinary integration with subjects like biochemistry and toxicology can deepen the understanding of the modes of toxicity actions, and then improve mechanistic interpretabilities of (Q)SAR models.

4.4 Application domain of (Q)SAR models

(1) Characterizing application domain

Characterizing application domains (AD) of (Q)SAR models is one of the main difficulties in practical application of the models in ERA. Regardless of the diversity of the training data used, it is important to realize that (Q)SAR models are only valid in the domain they were trained and validated. Extrapolation is dangerous and can lead to grossly erroneous model predictions[109]. The concept of the AD 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"[110]. Thus the AD can be defined as "the group of chemicals for which the model is valid"[111]. Practically, it requires an operational and computerized quantitative procedure to define the AD of (Q)SAR models[112].

The AD can be initially defined on the basis of the descriptors used to establish (Q)SAR models, which can be termed as the descriptor domain[113]. Thus the method for selection of training set affects the descriptor domain.

Secondly, the structural domain can be obtained by taking into consideration of structural similarity of chemicals in the training set and the test set[114]. The chemicals with the highest molecular structural similarity to the training set are the best predicted[114]. Sometimes, structural similarity is determined by empirical knowledge or assumed action modes[115]. Therefore, it may result in different structural domains with different methods for definitions of molecular structural similarity.

That the molecular descriptors are within the descriptor space and the molecular structures are similar to those of the training set are prerequisites for differentiating whether a compound is within the AD[116]. Furthermore, for reliability and validity of predictions, the mechanistic domain was also introduced[116]. The definition of the mechanistic domain is often achieved by the definition of substructures, with the assumption that substances that are chemically similar have a similar mechanism of toxic action[116]. It is the most critical criterion for accuracy and confidence of (Q)SAR models.

Finally, if metabolic activation of the chemicals is a part of the (Q)SAR model, the reliability of simulated

metabolism (metabolites, pathways, and maps) should be taken into account in assessing the reliability of the predictions, leading to the metabolism domain[110].

To sum up, the most conservative AD of (Q)SAR models is the intersection determined by the descriptor space, structural similarity, mechanistic domain, and metabolism domain. Practically, depending on the ultimate use of the model predictions and the consequences of a wrong decision, some of the components defining the model domain can be bypassed[110]. This will expand the model application domain but will reduce the confidence level of the predictions[110].

(2) Outlier detection

It is very important to detect outliers as they may affect QSAR model performance. An outlier of a model can be defined as a compound that is in some way different from the rest of the substances used to establish the model and for which the model is not valid. The reasons may be mechanism differences, or chemical diversities, even error data. From multivariate statistics point of view, outliers can be classified into three types: X outliers, Y outliers and X/Y outliers[109]. Briefly, a substance is an X outlier if the molecular descriptors for this substance do not conform to the ''pattern'' (covariance structure) of the training data. Y outliers are substances for which the reference value of the response is invalid. X/Y outliers are substances for which the relationship between the descriptors (X variables) and the environmental property (Y variable) is not the same as in the (rest of the) training data, e.g. due to different toxicity mechanisms[109].

Outlier detection is of great importance to determining AD of QSAR models accurately. As Y outliers need to be judged by experience and X/Y outliers cannot be detected directly, more attention should be paid to the diagnosis of X outliers. Two methods for diagnosing X outliers are available: Hotelling's T^2 and DModX[117]. Hotelling's T^2 is a multivariate generalization of Student's t-test[118]. DModX represents the distances of a chemical to the model in X space. From the distances and the corresponding distances in the training set, it is possible to calculate an approximate probability that a (new) substance belongs to the model domain[109]. The two methods are usually used jointly, and the difference between them lies in the fact that Hotelling's T^2 is founded with the explained variances while DModX is derived from the unexplained X-variances (residuals)[117]. Through the diagnostics it is also possible to discriminate between strong (Hotelling's T^2) and moderate (DModX) outliers, depending on which tool is used for their detection[117]. In addition, standard residuals of regression analysis can be used as simple criteria for identifying outliers.

It should be noted that outliers of QSARs abound for endpoints, and have actually been extremely useful in their development[101]. The analysis of outliers proved to be the spur for the further analysis and identification of mechanisms of action[101]. There should be valid reasons for outlier removal to improve modelling performance. Further, to assess the effects of excluding outliers, QSAR models should be examined before and after the removal. If outliers are merely statistical artifacts, then the models will not change significantly following their removal[101].

4.5 Characterization of QSAR models

To characterize established QSAR models, the quality of goodness-of-fit should be statistically assessed, internal validation should be performed to assess the model stability or robustness, and external validation should be implemented to assess the predictive power[119].

(1) Measure of goodness-of-fit

The traditional parameters used for the measure of goodness-of-fit are as follows:

a) Determination coefficient (R^2)/adjusted determination coefficient (R_{adj}^2): R^2 is a measure of the quality of fit between model-predicted and experimental values. However, some high R^2 values resulting from low degrees of freedom due to excessary predictor variables involved in the model may lead to a poor predictor for its QSAR model[120]. Thus the adjusted R^2 by the freedom degree (R_{adj}^2) should be adopted. The higher the R_{adj}^2 value, the better is the goodness-of-fit.

b) Summary square error (SSE): this index reflects the deviation of predicted values from observed values, and depends on the number of data points[119].

c) Root mean square error (RMSE), mean absolute residual (MAR), and standard error (SE)/standard deviation (SD) are commonly used to indicate the precision of prediction. They are dependent on the endpoint data range and scatter, and can be affected by outliers[119].

d) F-test: It is a variance test method of the overall significance level and is only applicable to QSAR models derived from multivariate linear regression (MLR)[119].

The above-mentioned indices can simply measure goodness-of-fit, but cannot identify over-fitting or under-fitting. The termed under-fitting means that a model does not fully reveal the relationships between the independent variables and predictor variables, which will reduce the predictability. Over-fitting means that some of the relations that appear statistically significant are actually just noise. A model with over-fitting does not replicate well and does a lousy job of predicting environmental endpoints. It may occur in the way of too many variables included in the final model. Over-fitting is a common problem in development of QSAR models, especially for non-linear models[121]. Stability analysis can be used to determine and solve the over-fitting problem.

(2) Stability analysis and internal validation of QSAR models

Stability analysis of QSAR models is closely linked to the problem of model over- or under-fitting[119,122]. Usually, model instability is analyzed, which can be defined as the sensitivity of a model, with respect to individual and subsets of compounds in the training set. If predicted values go beyond the confidence interval of a model, the corresponding compounds can lead to model instability[119]. There were very few studies on quantitative measures of model instability, except that Kolossov and Stanforth[119] introduced Model Instability Coefficient (MIC) and Model Value Instability Coefficient (MVIC) by considering model descriptors and predicted values, respectively. If the two values are lower than 100%, the model is considered to be stable, otherwise, it is unstable.

Instability of QSAR models is often indirectly measured by internal validations. The following internal validation methods can estimate the degree of instability more or less.

a) Leave-many-out cross-validation (LMO-CV)[120]: n objects of original data set are divided in G cancellation groups of qual size m ($=n/G$). A large number of models are developed with each of the n - m objects in the training set and m objects in the validation set. For each corresponding model, m objects are predicted and the cross-validation coefficient Q^2 is computed. Q^2 is considered as an indicator of the robustness and predictive power of the model[120]. Generally, a $Q^2 > 0.5$ can be regarded as good and a $Q^2 > 0.9$ as excellent[117].

b) Leave-one-out cross-validation (LOO-CV): just a single object is removed ($G = n$), and the other steps are the same with LMO-CV. Statistical theory predicts that LMO-CV performs better in variable selection than LOO-CV[123,124]. Compared with LMO-CV with medium size of m, LOO-CV and LMO-CV with very low values of m tend to overfit the data and decrease the predictive ability due to inclusion of more variables or latent variables[124].

c) Bootstrapping[120,125]: the basic premise of this method is that the data set is representative of the population from which it was drawn. In a typical bootstrap validation, G groups of m objects are generated by a repeated random selection of n objects from the original dataset. The model obtained on the m objects randomly selected is used to predict the target properties of the exluded samples. As in LMO-CV validation, a high Q^2 in bootstrap validation is a demonstration of the model robustness.

d) Y-randomization test[120]: this is a widely used technique to estimate the robustness of a QSAR model. The dependent variable (Y-vector) is random shuffled and a new model is developed using the original predictor variable matrix. The process is repeated 50—100 times. It is expected that the resulting models should generally have low R_{adj}^2 and low cross-validation coefficient Q^2 values. If all models obtained in the Y-randomization test have relatively high R_{adj}^2 and Q^2, it implies that an acceptable QSAR model cannot be obtained for the given data set by the current modeling methods.

Usually $R_{adj}^2 > Q^2$, and the differences ($R_{adj}^2 - Q^2$) do not exceed 0.3[117]. A substantially larger difference indicates an over-fitting model, presence of irrelevant X variables, or outliers in the data[117].

(3) Assessing predictive power

Predictive power of QSARs relies on the goodness-of-fit and robustness, and highly depends on its AD. The most demanding way to assess the predictive power is external validation that can be performed in the following two ways.

a) Using independent external data set for validation is the most standard method to assess model predictive power. It is also recommended for the strongest evaluation of model applicability for prediction in new chemicals, which can guarantee the application of the model for ERA. However, generally it depends on whether the data set is large enough to permit an independent exter-

nal validation set[117].

b) In view of the difficulty of finding independent external validation dataset, an alternative method is to split the original data set into a training set, used for establishing QSAR models, and a test set, for external validation[120]. The underlying goal in this method is to ensure that both the training and testing sets separately span the whole descriptor space occupied by the entire data set and the chemical domains in the two sets are not too dissimilar[120]. Practicable approaches for creating training and test sets span from straightforward random selection[120,126] through systematic clustering techniques (neural network)[120,127], to experimental design (D-Optimal)[120,128].

Performance of external validation is indicated by cross-validation coefficient Q^2 and determination coefficient R^2 between the predicted and observed environmental endpoints for an external test set. However, it has been shown that there exists no correlation between Q^2 and R^2. It must be stressed that the high value of Q^2 appears to be the necessary but not the sufficient condition for the model to have a high predictive power[129,130].

In summary, goodness-of-fit, robustness, and predictive power are necessary to defining the quality of QSAR models. Based on the three aspects of assessment, Kolossov and Stanforth[119] proposed an overall statistical quality index for QSAR models. Only the QSAR models with high statistical quality can be used for screening, management and ERA of organic chemicals. Further efforts on QSAR characterization and validation are still necessary[119,131,132].

5 Registration and application criterion of (Q)SAR models

To promote the development and application of (Q)SARs, registration and regulation rules of current (Q)SAR models are required. According to Cronin et al.[101], all the environmental endpoint data used to build (Q)SAR models should be reported in the publications, which can establish the transparency of the (Q)SAR models and ensure that it cannot be abused through extrapolation, also allow others to use the data[101]. All significant physico-chemical descriptors used in the (Q)SAR modelling should also be listed[101]. The reported (Q)SARs should accompany with a full description of the chemical structures, either in terms of IUPAC name, SMILES, or CAS numbers[101]. More importantly, the goodness of fit, stability, and predictive power of the (Q)SAR model should be assessed. The registered (Q)SARs can only be used within the application domain of compounds. Application of the (Q)SARs should balance between accuracy and AD of the models, and consider the purpose of usages by defining acceptable uncertainty levels that depend on the variance of the existing experimental data.

6 Conclusion

For the ERA of toxic organic chemicals, (Q)SARs play an important role in filling the data gap of environmental endpoints, decreasing experimental expense, reducing and replacing testing (especially animal testing), and assessing the uncertainty of experimental data. The development of (Q)SAR technology shows the features of application-orientation, interdisciplinary integration, and intelligence. The developed countries have paid much more attention to the research, development and application of (Q)SAR technology for ERA and regulatory acceptance. REACH regulation prescribed detailed principles for its application. The further development of the technology should pay attention to its applicability, application guidelines for ERA, and particularly the methods for characterization, validation and registration of (Q)SARs.

1 Macleod M, Mckone T E, Foster K L, Maddalena R L, Parkerton T F, Mackay D. Applications of contaminant fate and bioaccumulation models in assessing ecological risks of chemicals: A case study for gasoline hydrocarbons. Environ Sci Technol, 2004, 38(23): 6225－6233

2 U. S. Environmental Protection Agency. Guidelines for ecological risk assessment. In: Risk Assessment Forum. Washington: U. S. Environmental Protection Agency, 1998, 63(93): 26846－26924

3 Verhaar H J M, Solbe J, Speksnijder J, Van Leeuwen C J, Hermens J L M. Classifying environmental pollutants: Part 3. External validation of the classification system. Chemosphere, 2000, 40(8): 875－883

4 Enterprise & Industry Directorate General and Environment Directorate General. European Commission, REACH in brief. 2002, September. Available online at: http://ecb.jrc.it/REACH/

5 Linkov I, Ames M R, Crouch E A C, Satterstrom F K. Uncertainty in octanol-water partition coefficient: Implications for risk assessment and remedial costs. Environ Sci Technol, 2005, 39(18): 6917－6922

6 Tunkel J, Mayo K, Austin C, Hickerson A, Howard P. Practical considerations on the use of predictive models for regulatory purposes. Environ Sci Technol, 2005, 39(7): 2188－2199

7 Cronin M T D, Walker J D, Jaworska J S, Comber M H I, Watts C D, Worth A P. Use of QSARs in international decision-making frameworks to predict ecologic effects and environmental fate of chemical substances. Environ Health Persp, 2003, 111(10): 1376－1390

8 Li N Q, Wania F, Lei Y D, Daly G L. A Comprehensive and critical compilation, evaluation, and selection of physical-chemical property data for selected polychlorinated biphenyls. J Phy Chem Ref Data, 2003, 32(4): 1545－1590

9 Hammett L P. Some relations between reaction rates and equilibrium constants. Chem Rev, 1935, 17 (1): 125－136

10 Hammett L P. The effect of structure upon the reactions of organic compounds. Benzene derivatives. J Am Chem Soc, 1937, 59(1): 96－103

11 Taft R M. Polar and steric substituent constants for aliphatic and *o*-benzoate groups from rates of esterification and hydrolysis of esters. J Am Chem Soc, 1952, 74(12): 3120－3128

12 Kaliszan R. Quantitative structure-retention relationships applied to reversed-phase high-performance liquid chromatography. J Chromatogr A, 1993, 656(1-2): 417－435

13 Goss K -U, Schwarzenbach R P. Linear free energy relationships used to evaluate equilibrium partitioning of organic compounds. Environ Sci Technol, 2001. 35(7): 1－9

14 Nguyen T H, Goss K -U, Ball W P. Polyparameter linear free energy relationships for estimating the equilibrium partition of organic compounds between water and the natural organic matter in soils and sediments. Environ Sci Technol, 2005, 39(4): 913－924

15 Chen J W, Pei J, Quan X, Zhao Y Z, Chen S. Linear free energy relationships on rate constants for dechlorination by zero-valent iron. SAR QSAR Environ Res, 2002, 13(6): 597－606

16 Yan C L, Chen J W, Huang L P, Ding G H, Huang X Y. Linear free energy relationships on rate constants for the gas-phase reactions of hydroxyl radicals with PAHs and PCDD/Fs. Chemosphere, 2005, 61(10): 1523－1528

17 Chen J W, Peijnenburg W J G M, Quan X, Chen S, Zhao Y Z, Yang F L. The use of PLS algorithms and quantum chemical parameters derived from PM3 Hamiltonian in QSPR studies on direct photolysis quantum yields of substituted aromatic halides. Chemosphere, 2000, 40(12): 1319－1326

18 Chen J W, Quan X, Schramm K -W, Kettrup A, Yang F L. Quantitative structure-property relationships (QSPRs) on direct photolysis of PCDDs. Chemosphere, 2001, 45(2): 151－159

19 Free S M, Wilson J M. A mathematical contribution to structure-activity studies. J Med Chem, 1964, 7(4): 395－399

20 Cramer R D, Patterson D E, Bunce J D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc, 1988, 110(18): 5959－5967

21 Marshall G R, Cramer III R D. Three-dimensional structure-activity relationships. Trends Pharmacol Sci, 1988, 9(8): 285－289

22 Hansch C, Maloney P P, Fujita T, Muir R M. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. Nature, 1962, 194(4824): 178－180

23 Hansch C, Muir R M, Fujita T, Maloney P P, Geiger F, Streich M. The correlation of biological activity of plant growth regulators and chloromycetin derivatives with Hammett constants and partition coefficients. J Am Chem Soc, 1963, 85(18): 2817－2824

24 Fujita T, Iwasa J, Hansch C. A new substituent constant, π, derived from partition coefficients. J Am Chem Soc, 1964, 86(23): 5175－5180

25 Hansch C, Clayton J M. Lipophilic character and biological activity of drugs II: the parabolic case. J Pharm Sci, 1973, 62(1): 1－21

26 Hermens J, Leeueanch P, Musch A. Quantitative structure-activity relationships and mixture toxicity studies of chloro- and alkylanilines at an acute lethal toxicity level to the guppy (*Poecilia reticulata*). Ecotox Environ Safe, 1984, 8(4): 388－394

27 Schultz W T, Bryant S E, Lin D T. Structure-toxicity relationships for tetrahymana: aliphatic aldehydes. B Environ Contam Tox, 1994, 52(2): 279－285

28 Kamlet M J, Abraham M H, Doherty R M, Taft R W. Solubility properties in polymers and biological media. 4. Correlations of octanol/water partition coefficients with solvatochromic parameters. J Am Chem Soc, 1984, 106(2): 464－466

29 Kamlet M J, Doherty R M, Abboud J -L M, Abraham M H, Taft R W. Solubility: a new look. Chemtech, 1986, 16(9): 566－576

30 Kamlet M J, Doherty R M, Carr P W, Mackay D, Abraham M H, Taft R W. Linear solvation energy relationships. 44. Parameter estimation rules that allow accurate prediction of octanol/water partition coefficients and other solubility and toxicity properties of polychlorinated biphenyls and polycyclic aromatic hydrocarbons. Environ Sci Technol, 1988, 22(5): 503－509

31 Leahy D E. Intrinsic molecular volume as a measure of the cavity term in linear solvation energy relationship: octanol-water partition coefficients and aqueous solubilities. J Pharm Sci, 1986, 75(7): 629－639

32 Abraham M H, Ibrahim A, Zissimos A M. Determination of sets of solute descriptors from chromatographic measurements. J Chromatogr A, 2004, 1037(1-2): 29－47

33 Wilson L Y, Famini G R. Using theoretical descriptors in quantitative structure-activity relationships: some toxicological indices. J Med Chem, 1991, 34(5): 1668－1674

34 Famini G R, Renski C A, Wilson L Y. Using theoretical descriptors in quantitative structure-activity relationships: Some physicochemical properties. J Phys Org Chem, 1992, 5(7): 395－408

35 Reta M, Carr P W, Sadek P C, Rutan S C. Comparative study of hydrocarbon, fluorocarbon, and aromatic bonded RP-HPLC stationary phases by linear solvation energy relationships. Anal Chem, 1999, 71(16): 3484－3496

36 Kamlet M J, Doherty R M, Veith G D, Taft R W, Abraham M H. Solubility properties in polymers and biological media. 7. An analysis of toxicant properties that influence inhibition of bioluminescence in *Photobacterium phosphoreum* (the Microtox test). Environ Sci Technol, 1986, 20(7): 690－695

37 Kamlet M J, Doherty R M, Abraham M H, Veith G D, Abraham D J, Taft R W. Solubility properties in polymers and biological media. 8. An analysis of the factors that influence toxicities of organic nonelectrolytes to the Golden Orfe Fish (*Leuciscus idus melanotus*). Environ Sci Technol, 1987, 21(2): 149－155

38 Balakrishnan A, Polli J E. Apical sodium dependent bile acid transporter: a potential prodrug target. Mol Pharmaceutics (Review), 2006, 3(3): 223－230

39 Webb S R, Durst G L, Pernich D, Hall J. C. Interaction of cyclohexanediones with acetyl coenzyme-a carboxylase and an artificial tar-

get-site antibody mimic: a Comparative molecular field analysis. J Agric Food Chem, 2000, 48(6): 2506－2511
40 Yu S J, Keenan S M, Tong W, Welsh W J. Influence of the structural diversity of data sets on the statistical quality of three-dimensional quantitative structure-activity relationship (3D-QSAR) models: Predicting the estrogenic activity of xenoestrogens. Chem Res Toxicol, 2002, 15(10): 1229－1234
41 Tong W, Lowis D R, Perkins R, Chen Y, Welsh W J, Goddette D W, Heritage T W, Sheehan D M. Evaluation of quantitative structure-activity relationship methods for large-scale prediction of chemicals binding to the estrogen receptor. J Chem Inf Comput Sci, 1998, 38(4): 669－677
42 Chen J W, Quan X, Zhao Y Z, Yan Y L, Yang F L. Quantitative structure-property relationship studies on *n*-octanol/water partitioning coefficients of PCDD/Fs. Chemosphere, 2001, 44(6): 1369－1374
43 Pavan M, Worth A P, Netzeva T I. Review of QSAR Models for Bioconcentration. JRC report EUR EN I-21020. 2006
44 Chen J W, Harner T, Ding G H, Quan X, Schramm K W, Kettrup A. Universal predictive models on octanol-air partition coefficients at different temperatures for persistent organic pollutants. Environ Toxicol Chem, 2004, 23(10): 2309－2317
45 Li X H, Chen J W, Zhang L, Qiao X L, Huang L P. The Fragment constant method for predicting octanol-air partition coefficients of persistent organic pollutants at different temperatures. J Phys Chem Ref Data, 2006, 35(3): 1365－1384
46 Meylan W M, Howard P H, Boethling R S. Molecular topology/fragment contribution method for predicting soil sorption coefficients. Environ Sci Technol, 1992, 26(8): 1560－1567
47 Schüürmann G, Ebert R -U, Kühne R. Prediction of the sorption of organic compounds into soil organic matter from molecular structure. Environ Sci Technol, 2006, 40(22): 7005－7011
48 Hermens J L M, Leeuwangh P, Musch A. Quantitative structure–activity relationships and mixture toxicity studies of chloro- and alkylanilines at an acute lethal toxicity level to the guppy (*Poecilia reticulata*). Ecotoxicol Environ Safe, 1984, 8: 388－394
49 Bradbury S P, Russom C L, Ankley G T, Schultz T W, Walker J D. Overview of data and conceptual approaches for derivation of quantitative structure-activity relationships for ecotoxicological effects of organic chemicals. Environ Toxicol Chem, 2003, 22(8): 1789－1798
50 Tong W, Fang H, Hong H, Xie Q, Perkins R, Anson1 J, Sheehan D M. Regulatory application of SAR/QSAR for priority setting of endocrine disruptors: A perspective. Pure Appl Chem, 2003, 75: 2375－2388
51 Asikainen A, Ruuskanen J, Tuppurainen K. Consensus kNN QSAR: A versatile method for predicting the estrogenic activity of organic compounds in silico. A comparative study with five estrogen receptors and a large, diverse set of ligands. Environ Sci Technol, 2004, 38(24): 6724－6729
52 Liu H X, Papa E, Gramatica P. QSAR prediction of estrogen activity for a large set of diverse chemicals under the guidance of OECD principles. Chem Res Toxicol, 2006, 19(11): 1540－1548
53 Raymond J W, Rogers T N, Shonnard D R, Kline A A. A review of structure-based biodegradation estimation methods. J Hazard Mater, 2001, 84(2-3): 189－215
54 Chen J W, Peijnenburg W J G M, Quan X, Chen S, Martens D, Schramm K W, Kettrup A. Is it possible to develop a QSPR model for direct photolysis half-lives of PAHs under irradiation of sunlight?. Environ Pollut, 2001, 114(1): 137－143
55 Walker J D. International workshops on QSARs in the environmental sciences──The first 20 years. QSAR Comb Sci, 2003, 22(4): 415－421
56 Nys G G, Rekker R F. Statistical analysis of a series of partition coefficients with special reference to the predictability of folding of drug molecules. The introduction of hydrophobic fragmental constants (f values). Eur J Med Chem, 1973, 8: 521－535
57 Taft R W, Lewis I C. The general applicability of a fixed scale of inductive effects. II. Inductive effects of dipolar substituents in the reactivities of *m*- and *p*-substituted derivatives of benzene. J Am Chem Soc, 1958, 80(10): 2436－2443
58 Hansch C, Leo A, Taft R W. A survey of Hammett substituent constants and resonance and field parameters. Chem Rev, 1991, 91(2): 165－195
59 Hancock C K, Meyers E A, Yager B J. Quantitative separation of hyperconjugation effects from steric substituent constants. J Am Chem Soc, 1961, 83(20): 4211－4213
60 Charton M. The nature of the ortho effect. II. Composition of the Taft steric parameters. J Am Chem Soc, 1969, 91(3): 615－618
61 Ghose A K, Crippen G M. Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. J Chem Inf Comput Sci, 1987, 27(1): 21－35
62 Kamlet M J, Taft R W. The solvatochromic comparison method. I. The beta-scale of solvent hydrogen-bond acceptor (*HBA*) basicities. J Am Chem Soc, 1976, 98(2): 377－383
63 Taft R W, Kamlet M J. The solvatochromic comparison method. 2. The alpha-scale of solvent hydrogen-bond donor (*HBD*) acidities. J Am Chem Soc, 1976, 98(10): 2886－2894
64 Balaban A T. Using real numbers as vertex invariants for third-generation topological indexes. J Chem Inf Comput Sci, 1992, 32(1): 23－28
65 Kier L B, Hall L H. The nature of structure-activity relationships and their relation to molecular connectivity. Eur J Med Chem, 1977, 12: 307－312
66 Karelson M, Lobanov V S, Katritzky A R. Quantum-chemical descriptors in QSAR/QSPR studies. Chem Rev, 1996, 96(3): 1027－1043
67 Todeschini R, Consonni V. Handbook of Molecular Descriptors. Wiley-VCH: Weinheim, Germany, 2000
68 Ren R E, Wang H W. Multivariate Statistical Analysis──Theory, Method, Case (in Chinese). Beijing: National Defence Industry Press, 1999
69 Livingstone D J, Salt D W. Judging the significance of multiple linear regression models. J Med Chem, 2005, 48(3): 661－663
70 Dudek A Z, Arodz T, Galvez J. Computational methods in developing quantitative structure-activity relationships (QSAR): A review. Comb Chem High T Scr, 2006, 9(3): 213－228
71 Xu L, Shao X G. Methods of Chemometrics (in Chinese). Beijing: Science Press, 2004
72 Guha R, Jurs P C. Determining the validity of a QSAR Model-a classification approach. J Chem Inf Model, 2005, 45(1): 65－73

73 Barnard J M, Downs G M. Clustering of chemical structures on the basis of two-dimensional similarity measures. J Chem Inf Comput Sci, 1992, 32(6): 644－649
74 Wang H W. Partial Least-Squares Regression-Method and Applications (in Chinese). Beijing: Defense Industry Press, 1999.
75 Vapnik V. An overview of statistical learning theory. IEEE T Neural Networ, 1999, 10(5): 988－999
76 Kövesdi I, Dominguez -Rodriguez M F, Ôrfi L, Naray-Szabo G, Varro A, Papp J G, Matyus P. Application of neural networks in structure-activity relationships. Med Res Rev, 1999, 19(3): 249－269
77 Luan F. Application of support vector machines (SVM) and Radial basis function neural networks (RBFNN) in Chemistry, Environmental Chemistry and Medicinal Chemistry. Doctoral Dissertation (in Chinese). Lanzhou: Lanzhou University, 2006
78 Yang S, Lu W, Chen N. Support vector regression based QSPR for the prediction of some physicochemical properties of alkyl benzenes. J Mol Struct, 2005, 719(1-3): 119－127
79 O'Hara-Mays P. Genetic Algorithms in Molecular Modeling. Edited by James Devillers. Principles of QSAR and Drug Design, Vol. 1. New York: Academic Press, Harcourt Brace & Company. 1996. 1－327
80 Leardi R. Genetic algorithms in chemometrics and chemistry: A review. J Chemometr, 2001, 15(7): 559－569
81 Liu H X, Zhang R S, Yao X J, Liu M C, Hu Z D, Fan B T. Prediction of the isoelectric point of an amino acid based on GA-PLS and SVMs. J Chem Inf Comput Sci, 2004, 44 (1): 161－167
82 Wanchana S, Yamashita F, Hashida M. QSAR analysis of the inhibition of recombinant CYP 3A4 activity by structurally diverse compounds using a genetic algorithm-combined partial least squares method. Pharm Res, 2003, 20(9): 1401－1408
83 Liu J J, Cutler G, Li W, Pan Z, Peng S, Hoey T, Chen L, Ling X B. Multiclass cancer classification and biomarker discovery using GA-based algorithms. Bioinformatics, 2005, 21(11): 2691－2697
84 McInerney M, Dhawan, A P. Use of genetic algorithms with back propagation in training of feed-forward neural networks. In: IEEE International Conference on Neural Networks, 1993. 203－208
85 Wang H, Yu J. Application study on nonlinear dynamic FIR modeling using hybrid SVM-PLS method. In: Proceedings of the World Congress on Intelligent Control and Automation (WCICA) 4, 2004. 3479－3482
86 Jaworska J S, Comber M, Auer C, Van Leeuwen C J. Summary of a workshop on regulatory acceptance of QSARs for human health and environmental endpoints. Environ Health Persp, 2003, 111(10): 1358－1360
87 Cronin M T D, Jaworska J S, Walker J D, Comber M H I, Watts C D, Worth A P. Use of QSARs in international decision-making frameworks to predict health effects of chemical substances. Environ Health Persp, 2003, 111(10): 1391－1401
88 Walker J W L, Carlsen E, Simon-Hettich B. Global government applications of analogues, SARs and QSARs to predict aquatic toxicity, chemical or physical properties, environmental fate parameters and health effects of organic chemicals. SAR QSAR Environ Res, 2002, 13(6): 607－616
89 Worth A P, Bassan A, De Bruijn J, Saliner A G, Netzeva T, Patlewicz G, Pavan M, Tsakovska I, Eisenreich S. The role of the European Chemicals Bureau in promoting the regulatory use of QSARs methods. SAR QSAR Environ Res, 2007, 18(1-2): 111－125
90 Organisation for Economic Co-Operation and Development (OECD). Report from the Expert Group on (Q)SARs on the Principles for the Validation of (Q)SARs, 2004. Available online at: http://appli1.oecd.org/olis/2004doc.nsf/linkto/env-jm-mono(2004)24
91 Organisation for Economic Co-Operation and Development (OECD). Guidance document on the validation of (Quantitative) Structure-Activity Relationships [(Q)SARs] models, 2007. Available online at: http://www.oecd.org/dataoecd/55/22/38131728.pdf
92 Organisation for Economic Co-Operation and Development (OECD). Testing and assessment Report on the regulatory uses and applications in OECD member countries of (Quantitative) Structure-Activity Relationship[(Q)SARs] models in the assessment of new and existing chemicals, 2006. Available online at: http://appli1.oecd.org/olis/2006doc.nsf/linkto/env-jm-mono(2006)25
93 Wang L S, Han S K. Quantitative Structure-Activity Relationships of Organic Compounds (in Chinese). Beijing: China Environmental Science Press, 1993
94 Wang L S. Chemistry of Organic Pollution (in Chinese). Beijing: Higher Education Press, 2004
95 Chen J W. Quantitative Structure-Property Relationships and Quantitative Structure-Activity Relationships of Organic Pollutants (in Chinese). Dalian: Dalian University of Technology Press, 1999
96 Ding G H. Application of PLS and GA on QSAR of Selected Organic Pollutants (in Chinese). Doctoral Dissertation. Dalian: Dalian University of Technology, 2006
97 Lv Q Z, Shen G L, Yu R Q. Genetic training of network using chaos concept: Application to QSAR studies of vibration modes of tetrahedral halides. J Comput Chem, 2002, 23(14): 1357－1365
98 Zhao C Y. Applications of QSAR in life analytical chemistry and environmental chemistry. Doctoral Dissertation (in Chinese). Lanzhou: Lanzhou University, 2003
99 Yao Y Y, Xu L, Yang Y Q, Yuan X S. Study on structure-activity relationships of organic compounds: Three new topological indices and their applications. J Chem Inf Comput Sci, 1993, 33(4): 590－594
100 Lu G H, Yuan X, Zhao Y H. QSAR study on the toxicity of substituted benzenes to the algae (*scenedesmus obliquus*). Chemsphere, 2001, 44(3): 437－440
101 Cronin M T D, Schultz T W. Pitfalls in QSAR. J Mol Struct, 2003, 622(1-2): 39－51
102 Schultz T W, Cronin M T D. Essential and desirable characteristics of ecotoxicity quantitative structure-activity relationships. Environ Toxicol Chem, 2003, 22(3): 599－607
103 Cronin M T D, Schultz T W. Validation of *Vibrio fischeri* acute toxicity data: Mechanism of action-based QSARs for nonpolar narcotics and polar narcotic phenols. Sci Total Environ, 1997, 204(1): 75－88
104 Walker J D, Jaworska J, Comber M H I, Schultz T W, Dearden J C. Guidelines for developing and using quantitative structure-activity relationships. Environ Toxicol Chem, 2003, 22(8): 1653－1665
105 Livingstone D J. Data Analysis for Chemists: Applications to QSAR and Chemical Product Design. Oxford: Oxford University Press, 1995

106 Cronin M T D, Schultz T W. Development of quantitative structure-activity relationships for the toxicity of aromatic compounds to *Tetrahymena pyriformis*: Comparative assessment of methodologies. Chem Res Toxicol, 2001, 14(9):1284－1295
107 Burden F R, Winkler D A. A quantitative structure-activity relationships model for the acute toxicity of substituted benzenes to *Tetrahymena pyriformis* using Bayesian-regularized neural networks. Chem Res Toxicol, 2000, 13(6): 436－440
108 Kholodovych V, Smith J R, Knight D, Abramson S, Kohn J, Welsh W J. Accurate predictions of cellular response using QSPR: A feasibility test of rational design of polymeric biomaterials. Polymer, 2004, 45(22): 7367－7379
109 Furusjö E, Svenson A, Rahmberg M, Andersson M. The importance of outlier detection and training set slection for reliable environmental QSAR prediction. Chemosphere, 2006, 63(1): 99－108
110 Dimitrov S, Dimitrova G, Pavlov T, Dimitrova N, Patlewicz G, Niemela J, Mekenyan O. A stepwise approach for defining the applicability domain of SAR and QSAR models. J Chem Inf Model, 2005, 45(4): 839－849
111 EC (European Commission). Technical Guidance Document on Risk Assessment in support of Commission Directive 93/67/EEC on Risk Assessment for new notified substances and Commission Regulation (EC) No 1488/94 on Risk Assessment for existing substances, and Directive 98/8/EC of the European Parliament and of the Council concerning the placing of bio-cidal products on the market, Parts 3. 2003
112 Jaworska J S, Nikolova -Jeliazkova N. Aldenberg T. QSAR applicability domain estimation by projection of the training set in descriptor space: A review. Atla-Altern Lab Anim, 2005, 33(5): 445－459
113 Netzeva T I, Saliner A G, Worth A P. Comparison of the applicability domain of a quantitative structure-activity relationship for estrogenicity with a large chemical inventory. Environ Toxicol Chem, 2006, 25(5): 1223－1230
114 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(6): 1912－1928
115 Dimitrov S, Koleva Y, Schiltz T W, Walker J D, Mekenyan O. Interspecies quantitative structure-activity relationships model for aldehydes: Aquatic toxicity. Environ Toxicol Chem, 2004, 23(2): 463－470
116 Schultz T W, Hewitt M, Netzeva T I, Cronin M T D. Assessing applicability domains of toxicological QSARs: Definition, confidence in predicted values, and the role of mechanisms of action. QSAR Comb Sci, 2007, 26(2): 238－254
117 Eriksson L, Jaworska 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 Persp, 2003, 111(10): 1361－1375
118 Jackson J E. A User's Guide to Principal Components. New York: John Wiley. 1991
119 Kolossov E, Stanforth R. The quality of QSAR models: Problems and solutions. SAR QSAR Environ Res, 2007, 18(1-2): 89－100
120 Tropsha A, Gramatica P, Gombar V K. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb Sci, 2003, 22(1): 69－77
121 Livingstone D J, Manallack D T, Tetko I V. Data modeling with neural networks: Advantages and limitations. J Comput Aid Mol Des, 1997, 11(2): 135－142
122 Hawkins D M. The problem of overfitting. J Chem Inf Comput Sci. 2004, 44(1): 1－12
123 Zhang P. Model selection via multifold cross validation. Ann Statist, 1993, 21: 299－313
124 Baumann K, Korff M, Albert H. A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. Part II. Practical applications. J Chemometr, 2002, 16(7): 351－360
125 Wehrens R, Putter H, Buydens L M C. The bootstrap: A tutorial. Chemom Intell Lab Systems, 2000(1), 54: 35－52
126 Yasri A, Hartsough D. Toward an optimal procedure for variable selection and QSAR model building. J Chem Inf Comput Sci, 2001, 41(5): 1218－1227
127 Burden F R, Ford M G, Whitley D C, Winkler D A. Use of automatic relevance determination in QSAR studies using bayesian neural networks. J Chem Inf Comput Sci, 2000, 40(6): 1423－1430
128 Mitchell T J. An algorithm for the construction of "D-optimal" experimental design. Technometrics, 2000, 42(1): 48－54
129 Kubinyi H, Hamprecht F A, Mietzner T. Three-dimensional quantitative similarity-activity relationships (3D QSAR) from SEAL similarity matrices. J Med Chem, 1998, 41(14): 2553－2564
130 Golbraikh A, Tropsha A. Beware of q^2! J Mol Graph Model, 2002, 20(4): 269－276
131 Schultz T W, Netzeva T I, Cronin M T D. Evaluation of QSARs for ecotoxicity: A method for assigning quality and confidence. SAR QSAR Environ Res, 2004, 15(5-6): 385－397
132 Deardon J C, Roberts D W. Larger molecules penetrate membranes more readily. J Pharm Pharmacol, 2006, 58: 60