For Official Use **ENV/JM/MONO(2007)2**

Organisation de Coopération et de Développement Economiques
Organisation for Economic Co-operation and Development

15-Feb-2007

English, French

ENVIRONMENT DIRECTORATE
JOINT MEETING OF THE CHEMICALS COMMITTEE AND
THE WORKING PARTY ON CHEMICALS, PESTICIDES AND BIOTECHNOLOGY

GUIDANCE DOCUMENT ON THE VALIDATION OF (QUANTITATIVE)STRUCTURE-ACTIVITY RELATIONSHIPS [(Q)SAR] MODELS

JT03221944

ENV/JM/MONO(2007)2
For Official Use

English, French

OECD Environment Health and Safety Publications

Series on Testing and Assessment

No. 69

GUIDANCE DOCUMENT ON THE VALIDATION OF (QUANTITATIVE) STRUCTURE-ACTIVITY RELATIONSHIP [(Q)SAR] MODELS

INTER-ORGANIZATION PROGRAMME FOR THE SOUND MANAGEMENT OF CHEMICALS

A cooperative agreement among UNEP, ILO, FAO, WHO, UNIDO, UNITAR and OECD

Environment Directorate

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

Paris, 2007

Also published in the Series on Testing and Assessment:

No. 1, *Guidance Document for the Development of OECD Guidelines for Testing of Chemicals (1993; reformatted 1995, revised 2006)*

No. 2, *Detailed Review Paper on Biodegradability Testing (1995)*

No. 3, *Guidance Document for Aquatic Effects Assessment (1995)*

No. 4, *Report of the OECD Workshop on Environmental Hazard/Risk Assessment (1995)*

No. 5, *Report of the SETAC/OECD Workshop on Avian Toxicity Testing (1996)*

No. 6, *Report of the Final Ring-test of the Daphnia magna Reproduction Test (1997)*

No. 7, *Guidance Document on Direct Phototransformation of Chemicals in Water (1997)*

No. 8, *Report of the OECD Workshop on Sharing Information about New Industrial Chemicals Assessment (1997)*

No. 9, *Guidance Document for the Conduct of Studies of Occupational Exposure to Pesticides during Agricultural Application (1997)*

No. 10, *Report of the OECD Workshop on Statistical Analysis of Aquatic Toxicity Data (1998)*

No. 11, *Detailed Review Paper on Aquatic Testing Methods for Pesticides and industrial Chemicals (1998)*

No. 12, *Detailed Review Document on Classification Systems for Germ Cell Mutagenicity in OECD Member Countries (1998)*

No. 13, *Detailed Review Document on Classification Systems for Sensitising Substances in OECD Member Countries 1998)*

No. 14, *Detailed Review Document on Classification Systems for Eye Irritation/Corrosion in OECD Member Countries (1998)*

No. 15, *Detailed Review Document on Classification Systems for Reproductive Toxicity in OECD Member Countries (1998)*

No. 16, *Detailed Review Document on Classification Systems for Skin Irritation/Corrosion in OECD Member Countries (1998)*

No. 17, *Environmental Exposure Assessment Strategies for Existing Industrial Chemicals in OECD Member Countries (1999)*

No. 18, *Report of the OECD Workshop on Improving the Use of Monitoring Data in the Exposure Assessment of Industrial Chemicals (2000)*

No. 19, *Guidance Document on the Recognition, Assessment and Use of Clinical Signs as Humane Endpoints for Experimental Animals used in Safety Evaluation (1999)*

No. 20, *Revised Draft Guidance Document for Neurotoxicity Testing (2004)*

No. 21, *Detailed Review Paper: Appraisal of Test Methods for Sex Hormone Disrupting Chemicals (2000)*

No. 22, *Guidance Document for the Performance of Out-door Monolith Lysimeter Studies (2000)*

No. 23, *Guidance Document on Aquatic Toxicity Testing of Difficult Substances and Mixtures (2000)*

No. 24, *Guidance Document on Acute Oral Toxicity Testing (2001)*

No. 25, *Detailed Review Document on Hazard Classification Systems for Specifics Target Organ Systemic Toxicity Repeated Exposure in OECD Member Countries (2001)*

No. 26, *Revised Analysis of Responses Received from Member Countries to the Questionnaire on Regulatory Acute Toxicity Data Needs (2001)*

No 27, *Guidance Document on the Use of the Harmonised System for the Classification of Chemicals Which are Hazardous for the Aquatic Environment (2001)*

No 28, *Guidance Document for the Conduct of Skin Absorption Studies (2004)*

No 29, *Guidance Document on Transformation/Dissolution of Metals and Metal Compounds in Aqueous Media (2001)*

No 30, *Detailed Review Document on Hazard Classification Systems for Mixtures (2001)*

No 31, *Detailed Review Paper on Non-Genotoxic Carcinogens Detection: The Performance of In-Vitro Cell Transformation Assays (draft)*

No. 32, *Guidance Notes for Analysis and Evaluation of Repeat-Dose Toxicity Studies (2000)*

No. 33, *Harmonised Integrated Classification System for Human Health and Environmental Hazards of Chemical Substances and Mixtures (2001)*

No. 34, *Guidance Document on the Development, Validation and Regulatory Acceptance of New and Updated Internationally Acceptable Test Methods in Hazard Assessment (2005)*

No. 35, *Guidance notes for analysis and evaluation of chronic toxicity and carcinogenicity studies (2002)*

No. 36, *Report of the OECD/UNEP Workshop on the use of Multimedia Models for estimating overall Environmental Persistence and long range Transport in the context of PBTS/POPS Assessment (2002)*

No. 37, *Detailed Review Document on Classification Systems for Substances Which Pose an Aspiration Hazard (2002)*

No. 38, *Detailed Background Review of the Uterotrophic Assay Summary of the Available Literature in Support of the Project of the OECD Task Force on Endocrine Disrupters Testing and Assessment (EDTA) to Standardise and Validate the Uterotrophic Assay (2003)*

No. 39, *Guidance Document on Acute Inhalation Toxicity Testing (in preparation)*

No. 40, *Detailed Review Document on Classification in OECD Member Countries of Substances and Mixtures Which Cause Respiratory Tract Irritation and Corrosion (2003)*

No. 41, *Detailed Review Document on Classification in OECD Member Countries of Substances and Mixtures which in Contact with Water Release Toxic Gases (2003)*

No. 42, *Guidance Document on Reporting Summary Information on Environmental, Occupational and Consumer Exposure (2003)*

No. 43, *Draft Guidance Document on Reproductive Toxicity Testing and Assessment (in preparation)*

No. 44, *Description of Selected Key Generic Terms Used in Chemical Hazard/Risk Assessment (2003)*

No. 45, *Guidance Document on the Use of Multimedia Models for Estimating Overall Environmental Persistence and Long-range Transport (2004)*

No. 46, *Detailed Review Paper on Amphibian Metamorphosis Assay for the Detection of Thyroid Active Substances (2004)*

No. 47, *Detailed Review Paper on Fish Screening Assays for the Detection of Endocrine Active Substances (2004)*

No. 48, *New Chemical Assessment Comparisons and Implications for Work Sharing (2004)*

No. 49, *Report from the Expert Group on (Quantitative) Structure-Activity Relationships [(Q)SARs] on the Principles for the Validation of (Q)SARs (2004)*

No. 50, *Report of the OECD/IPCS Workshop on Toxicogenomics (2005)*

No. 51, *Approaches to Exposure Assessment in OECD Member Countries: Report from the Policy Dialogue on Exposure Assessment in June 2005 (2006)*

No. 52, *Comparison of emission estimation methods used in Pollutant Release and Transfer Registers (PRTRs) and Emission Scenario Documents (ESDs): Case study of pulp and paper and textile sectors (2006)*

No. 53, *Guidance Document on Simulated Freshwater Lentic Field Tests (Outdoor Microcosms and Mesocosms) (2006)*

No. 54, *Current Approaches in the Statistical Analysis of Ecotoxicity Data: A Guidance to Application (2006)*

No. 55, *Detailed Review Paper on Aquatic Arthropods in Life Cycle Toxicity Tests with an Emphasis on Developmental, Reproductive and Endocrine Disruptive Effects (2006)*

No. 56, *Guidance Document on the Breakdown of Organic Matter in Litter Bags (2006)*

No. 57, *Detailed Review Paper on Thyroid Hormone Disruption Assays (2006)*

No. 58, *Report on the Regulatory Uses and Applications in OECD Member Countries of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models in the Assessment of New and Existing Chemicals (2006)*

No. 59, *Report of the Validation of the Updated Test Guideline 407: Repeat Dose 28-Day Oral Toxicity Study in Laboratory Rats (2006)*

No. 60, *Report of the Initial Work Towards the Validation of the 21-Day Fish Screening Assay for the Detection of Endocrine Active Substances (Phase 1A) (2006)*

No. 61, *Report of the Validation of the 21-Day Fish Screening Assay fort he Detection of Endocrine Active Substances (Phase 1B) (2006)*

No. 62, *Final OECD Report of the Initial Work Towards the Validation of the Rat Hershberger Assay : Phase-1, Androgenic Response to Testosterone Propionate, and Anti-Androgenic Effects of Flutamide (2006)*

No. 63, *Guidance Document on the Definition of Residue (2006)*

No. 64, *Guidance Document on Overview of Residue Chemistry Studies (2006)*

No. 65, *OECD Report of the Initial Work Towards the Validation of the Rodent Utertrophic Assay - Phase 1 (2006)*

No. 66, *OECD Report of the Validation of the Rodent Uterotrophic Bioassay: Phase 2. Testing of Potent and Weak Oestrogen Agonists by Multiple Laboratories (2006)*

No. 67, *Additional data supporting the Test Guideline on the Uterotrophic Bioassay in rodents (draft)*

No. 68, *Summary Report of the Uterotrophic Bioassay Peer Review Panel, including Agreement of the Working Group of the National Coordinators of the Test Guidelines Programme on the follow up of this report (2006)*

About the OECD

The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organisation in which representatives of 30 industrialised countries in North America, Europe and the Asia and Pacific region, as well as the European Commission, meet to co-ordinate and harmonise policies, discuss issues of mutual concern, and work together to respond to international problems. Most of the OECD's work is carried out by more than 200 specialised committees and working groups composed of member country delegates. Observers from several countries with special status at the OECD, and from interested international organisations, attend many of the OECD's workshops and other meetings. Committees and working groups are served by the OECD Secretariat, located in Paris, France, which is organised into directorates and divisions.

The Environment, Health and Safety Division publishes free-of-charge documents in ten different series: **Testing and Assessment**; **Good Laboratory Practice and Compliance Monitoring**; **Pesticides and Biocides**; **Risk Management**; **Harmonisation of Regulatory Oversight in Biotechnology**; **Safety of Novel Foods and Feeds; Chemical Accidents; Pollutant Release and Transfer Registers; Emission Scenario Documents; and the Safety of Manufactured Nanomaterials.** More information about the Environment, Health and Safety Programme and EHS publications is available on the OECD's World Wide Web site (http://www.oecd.org/ehs/).

This publication was produced within the framework of the Inter-Organisation Programme for the Sound Management of Chemicals (IOMC).

The Inter-Organisation Programme for the Sound Management of Chemicals (IOMC) was established in 1995 following recommendations made by the 1992 UN Conference on Environment and Development to strengthen co-operation and increase international co-ordination in the field of chemical safety. The participating organisations are FAO, ILO, OECD, UNEP, UNIDO, UNITAR and WHO. The World Bank and UNDP are observers. The purpose of the IOMC is to promote co-ordination of the policies and activities pursued by the Participating Organisations, jointly or separately, to achieve the sound management of chemicals in relation to human health and the environment.

FOREWORD

The introduction of a new technology into formal decision-making processes involving chemicals requires a solid scientific foundation and technical guidance on useful approaches for implementation. (Quantitative) Structure-Activity Relationship [(Q)SAR] technology is not really a new technology and it has enjoyed more than 20 years of use in some regulatory applications. However, advances in computers and the Internet together with the growing gap between the need for empirical data and the availability of testing resources seem to be ushering in a new international emphasis on (Q)SAR-based technologies for initial risk assessments.

The solid scientific foundation for (Q)SAR technology is the underlying premise in chemistry that similar chemical structures are expected to exhibit similar chemical behaviour. This simple premise becomes especially important whenever there are not enough empirical data for hazard identification and risk assessment purposes. With tens of thousands of chemical structures to assess and many different empirical tests needed to understand chemical behaviour, the concept of grouping similar chemicals together and extending existing data through models for similar chemical structures which have not been tested seems to be a prudent approach.

The scientific underpinning of (Q)SAR technology is made complex, however, because of the complexity of methods to measure similarity and the number of forms chemical behaviour can take in toxicology. To keep (Q)SAR applications on a solid scientific foundation, an international effort to articulate principles for (Q)SAR technology and to develop a guidance document for use of (Q)SAR in regulatory applications. This document presents those principles and helpful guides for validating (Q)SAR technology for a variety of applications. The reader will find that transparency in the validation process and objective determination of the reliability of (Q)SAR models are crucial to extending the regulatory acceptance of (Q)SAR models.

The first draft of this document was produced by the Joint Research Centre (JRC) of the European Commission. The draft was developed by the OECD Steering Group for (Q)SARs and overseen by the OECD Ad Hoc Group on (Q)SARs. The second draft was circulated to the members of the Ad Hoc Group for their input in March 2006. Comments were received from Germany, Italy, Japan and the United States and discussed at the meeting of the Ad Hoc Group in June 2006, and the revised draft was be circulated to the Ad Hoc Group on (Q)SARs for final review in August 2006 and endorsed in November 2006.

This document is published on the responsibility of the Joint Meeting of the Chemicals Committee and the Working Party on Chemicals, Pesticides and Biotechnology of the OECD.

This document has been produced with the financial assistance of the European Union. The views expressed herein can in no way be taken to reflect the official opinion of the European Union.

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION

Purpose of this document

1. (Quantitative) Structure-Activity Relationship [(Q)SAR] represents a technology aimed at providing estimates of many laboratory test results before the tests are conducted. The computer-based (Q)SARs give a virtual glimpse of the information a particular test might yield, and this new capability offers all stakeholders in the regulation of chemicals new opportunities in setting priorities for limited testing resources. In some cases, as (Q)SAR performance evolves, the estimated values from (Q)SARs will increasingly be used to inform initial risk assessments. Anticipating the benefits of adding the *in silico* technology of (Q)SAR to the well established *in vitro* and *in vivo* test guidelines, experts have been meeting to discuss the barriers to acceptance of (Q)SARs by regulatory agencies. A critical element of regulatory acceptance is the creation of a flexible scientific validation process for (Q)SARs which allows individual regulatory agencies to establish the reliability of (Q)SAR estimates for specific authorities and regulatory constraints.

2. This document provides a discussion of the "OECD Principles for the Validation, for Regulatory Purposes, of (Q)SAR Models" (see Annex A) and provides guidance on how individual regulatory agencies can evaluate specific (Q)SAR models with respect to those principles. The **purpose** of this document is to provide detailed but non-prescriptive guidance that explains and illustrates the application of the validation principles to different types of (Q)SAR models. This document is **needed** to provide a harmonised framework for (Q)SAR validation studies to explain and illustrate with examples how the validation principles can be interpreted for different types of (Q)SAR models.

3. While this document provides non-prescriptive guidance on the processes of validation which address the performance of a wide variety of (Q)SAR models, the document in not intended to establish specific criteria for judging the scientific validity or for regulatory acceptance of individual (Q)SAR models. This document defers explicitly to the appropriate regulatory authorities in the member countries to establish criteria for validity and acceptance.

4. The **audiences** for which this document is intended include the regulatory decision makers who wish to understand the usefulness and the uncertainty of (Q)SAR estimates results as well as the (Q)SAR specialist who participates in regulatory decisions and who is likely to carry out (Q)SAR validation exercises for specific regulatory applications. Additionally, the document is intended to inform the registrants for chemicals who may need to prepare explanations for registration dossiers with respect to the specific (Q)SAR models used for individual chemicals. All stakeholders of the (Q)SAR validation process, regardless of their familiarity with (Q)SAR models will benefit from the transparency and documentation of the process by which a particular (Q)SAR model is judged adequate for a regulatory decision and of the (Q)SAR estimates produced for specific chemicals being assessed. To that end, this guidance document provides a historical overview and highlights key scientific issues involved in acceptance of a (Q)SAR model. The audiences are encouraged to make full use of appendices and references where information on specific applications can be found for a variety of models.

Regulatory Acceptance of (Q)SAR

5. The OECD workgroups on (Q)SAR and the Joint Meeting have concurred that the validation of (Q)SAR models for regulatory purposes are best carried out by the regulatory authorities of the member countries. In the foreseeable future, the acceptance of (Q)SARs as a nontesting alternative source of data in making decisions will be based on the reliability and transparency of a specific (Q)SAR model within a specific regulatory context. Consequently, the validation principles for (Q)SAR models are intended to guide regulatory agencies in the evaluation of performance of (Q)SAR for specific decision processes at a higher level than the criteria used to judge statistical validity. Nonetheless, transparent communication of the statistical performance of a (Q)SAR model is the cornerstone for reliable use in regulation and this document describes numerous useful approaches. As acceptance of (Q)SAR grows to fill the need for data, it is anticipated that statistical validity will remain crucial while mechanistic interpretation of the models and explanation of the (Q)SAR results will be required.

The OECD Principles of (Q)SAR Validation

6. In November 2004, the 37th OECD's Joint Meeting of the Chemicals Committee and the Working Party on Chemicals, Pesticides and Biotechnology (Joint Meeting) agreed on the OECD Principles for the Validation, for Regulatory Purposes, of (Q)SAR Models (see Annex A). Flexibility will be needed in the interpretation and application of each OECD principle because ultimately, the proper integration of (Q)SARs into any type of regulatory/decision-making framework depends upon the needs and constraints of the specific regulatory authority. For example, the need for such flexibility is given in a case study by the US EPA presented in a case studies report on the regulatory uses and applications of (Q)SAR models in OECD member countries (OECD, 2006).

7. The agreed OECD principles are as follows:

"To facilitate the consideration of a (Q)SAR model for regulatory purposes, it 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;*
5. *a mechanistic interpretation, if possible."*

It was also agreed that these principles should be read in conjunction with the associated explanatory notes for each of principles (see Annex A) and that the check list developed by the Expert Group provides useful guidance on the interpretation and application of principles (See Annex B). The principles for (Q)SAR validation and the associated check list are intended to identify the types of information that are considered useful for the regulatory review of (Q)SARs. Taken together, the principles and the check list constitute a conceptual framework to guide the validation of (Q)SARs, but they are not intended to provide criteria for the regulatory acceptance of (Q)SARs. The definition of acceptance criteria, where considered necessary, is the responsibility of individual authorities within the member countries.

8. According to Principle 1, a (Q)SAR should be associated with a "defined endpoint", where endpoint refers to any physicochemical, biological or environmental effect that can be measured and therefore modelled. The intent of this principle is to ensure transparency in the endpoint being predicted by a given model, since a given endpoint could be determined by different experimental protocols and under different experimental conditions. Ideally, (Q)SARs should be developed from homogeneous datasets in

which the experimental data have been generated by a single protocol. However, this is rarely feasible in practice, and data produced by different protocols are often combined.

9. According to Principle 2, a (Q)SAR should be expressed in the form of an unambiguous algorithm. The intent of this principle is to ensure transparency in the description of the model algorithm. In the case of commercially-developed models, this information is not always made publicly available.

10. According to Principle 3, a (Q)SAR should be associated with a "defined domain of applicability". The need to define an applicability domain expresses 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 the models can generate reliable predictions. This principle does not imply that a given model should only be associated with a single applicability domain. As discussed in Chapter 4, the boundaries of the domain can vary according to the method used to define it and the desired trade-off between the breadth of model applicability and the overall reliability of predictions.

11. According to Principle 4, a (Q)SAR should be associated with "appropriate measures of goodness-of–fit, robustness and predictivity." This principle expresses the need to provide two types of information: a) the internal performance of a model (as represented by goodness-of-fit and robustness), determined by using a training set; and b) the predictivity of a model, determined by using an appropriate test set. As discussed in Chapter 5, there is no absolute measure of predictivity that is suitable for all purposes, since predictivity can vary according to the statistical methods and parameters used in the assessment.

12. According to Principle 5, a (Q)SAR should be associated with a "mechanistic interpretation", wherever such an interpretation can be made. Clearly, it is not always possible to provide a mechanistic interpretation of a given (Q)SAR. The intent of this principle is therefore to ensure that there is an assessment of the mechanistic associations between the descriptors used in a model and the endpoint being predicted, and that any association is documented. Where a mechanistic interpretation is possible, it can also form part of the defined applicability domain (Principle 3).

Historical Background

13. A set of principles for assessing the validity of (Q)SARs (Setubal principles) were proposed at an international workshop on the "Regulatory Acceptance of QSARs for Human Health and Environment Endpoints", organised by the International Council of Chemical Associations (ICCA) and the European Chemical Industry Council (CEFIC), and held in Setubal, Portugal, on 4-6 March, 2002 (Jaworska *et al*., 2003; Eriksson *et al*., 2003; Cronin *et al*., 2003a, 2003b).

14. The regulatory use of structure-activity relationships (SARs) and quantitative structure-activity relationships (QSARs), collectively referred to as (Q)SARs, varies considerably among OECD member countries, and even between different agencies within the same member country. This is partly due to different regulatory frameworks, which impose different requirements and work under different constraints, but also because an internationally harmonised conceptual framework for assessing (Q)SARs has been lacking. The lack of such a framework led to the widespread recognition of the need for an internationally-agreed set of principles for (Q)SAR validation. The development of a set of agreed principles was considered important, not only to provide regulatory bodies with a scientific basis for making decisions on the acceptability (or otherwise) of data generated by (Q)SARs, but also to promote the mutual acceptance of (Q)SAR models by improving the transparency and consistency of (Q)SAR reporting.

15. In November 2002, the 34th Joint Meeting agreed to start a new OECD activity aimed at increasing the regulatory acceptance of (Q)SARs, and to establish an Expert Group for this work.

16. The first Meeting of the Expert Group was hosted by the European Commission's Joint Research Centre (JRC), in Ispra, Italy, on 31 March – 2 April, 2003. Following the request of the 34th JM, the participants of the first Expert Group Meeting proposed a (two-year) work plan for the OECD work on (Q)SARs. The work plan included three Work Items:

- Work Item 1: Apply the specific development/validation principles agreed at the ICCA Workshop on Regulatory Acceptance of (Q)SARs, and the general validation principles for new and updated test methods, to selected (Q)SARs in use;
- Work Item 2: Develop guidance documents for development, validation and regulatory application of (Q)SARs; and,
- Work Item 3: Identify practical approaches to enable (Q)SARs to be readily available and accessible, including the development of database of accepted (Q)SARs.

The aim of Work Item 1, completed in 2004, was to apply the Setubal principles to selected (Q)SARs, in order to evaluate the principles, and to refine them wherever necessary. The aim of Work Item 2 was to develop guidance documents for the validation of (Q)SARs to assist (Q)SAR practitioners and (Q)SAR end-users in developing and evaluating (Q)SARs with respect to the validation principles. The aim of Work Item 3 was to identify practical approaches to make (Q)SARs readily available and accessible to scientists in regulatory bodies, industry and universities.

17. To manage the OECD work plan on QSARs, the first Expert Group Meeting proposed a subgroup, called the Coordinating Group of the Expert Group on (Q)SARs. In June 2003, the proposed work plan was endorsed by the 35th JM. At the same meeting, the JRC offered to take the lead in coordinating Work Item 1 on the evaluation of the Setubal principles, with the support of the Coordinating Group. The offer was welcomed and accepted by the 35th Joint Meeting.

18. To carry out Work Item 1, a team of experts (the Work Item 1 Team) produced a total of eleven case studies, by applying the Setubal principles to specific (Q)SARs or software models. The models chosen included literature-based models for acute fish toxicity, atmospheric degradation, mutagenicity and carcinogenicity, and the following software models: the Multi-CASE model for *in vitro* chromosomal aberrations; Multi-CASE and MDL models for human NOEL; ECOSAR; BIOWIN; Derek; the Derek skin sensitisation rulebase; the Japanese METI biodegradation model; and the rat oral chronic toxicity models in TOPKAT. These models were considered to collectively provide a representative range of (Q)SAR approaches, covering a variety of physicochemical, environmental, ecological and human health endpoints.

19. To provide guidance on the application of the proposed principles, a check list of considerations (questions) was developed by the Coordinating Group, and this was refined on the basis of experience obtained by carrying out Work Item 1). The refined check list (see Annex B) was presented to the 16th Meeting of the OECD Working Group of National Coordinators of the Test Guidelines Programme (WNT), held on 26-28 May 2004.

20. The report on the outcome of Work Item 1 including the refined check list mentioned above was discussed by the second Expert Group Meeting, hold at the OECD Headquarters in Paris, on 20-21 September 2004. The report consisted of a consolidated report by the Coordinating Group, including a proposal for revision of the Setubal principles, followed by a set of annexes containing the 11 case studies. The Expert Group refined the wording of the consolidated report, which included combining the internal and external validation principles into a single principle (OECD, 2004), which then represented the consensus view of the Expert Group. It was also agreed that the views expressed in the annexes of the

report should be regarded as views of the identified authors, and not necessarily the views of the Expert Group.

21. The final report on the outcome of Work Item 1 (OECD, 2004) and in particular the proposed OECD Principles for (Q)SAR Validation (see para 6 and Annex A), were adopted by the 37th Joint Meeting on 17-19 November 2004. The Joint Meeting supported the Expert Group's proposal that Work Item 1 should be followed up with Work Item 2 in the development of this Guidance Document on the Validation of (Q)SAR Models, which should provide detailed and non-prescriptive guidance to explain and illustrate the application of the OECD Principles for (Q)SAR Validation to different types of models.

22. The 37th Joint Meeting also agreed on some changes in the coordination of the OECD QSAR work programme. In particular, the (Q)SAR Group, often referred to as the "(Q)SAR Expert Group" was changed to "Ad hoc Group on (Q)SARs" and the membership of the Ad hoc Group was re-established to include not only (Q)SAR experts but also those who use (Q)SARs for regulatory purposes. Following receipt of the nominations from the member countries, the 38th Joint Meeting on 8-10 June 2005 agreed to replace the Coordinating Group with a smaller Steering Group, consisting of those members of the Ad hoc Group who are most closely involved in the planning and routine management of the (Q)SAR project.

Definition of Validation for (Q)SAR Models

23. The guidance for (Q)SARs in the present document is consistent with the general guidance given in OECD Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment (OECD, 2005).

24. According to the OECD guidance, the term "validation" is defined as follows:

> *"Validation: The process by which the reliability and relevance of a particular approach, method, process or assessment is established for a defined purpose."*

25. The terms "reliability" and "relevance" are defined as follows:

> *"Reliability: Measures of the extent that a test method can be performed reproducibly within and between laboratories over time, when performed using the same protocol. It is assessed by calculating intra- and inter-laboratory reproducibility and intra-laboratory repeatability."*
>
> *"Relevance: Description of relationship of the test to the effect of interest and whether it is meaningful and useful for a particular purpose. It is the extent to which the test correctly measures or predicts the biological effect of interest. Relevance incorporates consideration of the accuracy (concordance) of a test method."*

Based on these definitions, the term "reliability" refers to the reproducibility of the method, both within and between laboratories, and over time. The term "relevance" refers to the scientific basis for expecting an experimental method to predict a response of important assessment endpoints which cannot be measured directly.

26. The conventional OECD uses of the terms "reliability" and "relevance" can be extended to the validation process for (Q)SAR models. However, because (Q)SAR models are derived from experimental data, the concepts of reliability and relevance for test guideline purposes are necessary but not necessarily sufficient for validation of (Q)SAR models. This guidance document for (Q)SAR validation expands the concepts of reliability in a manner that retains that from a test method as the "maximum reliability" which can be expected from (Q)SAR model. Since few test methods have documented the reproducibility between laboratories for a single chemical, the validation of (Q)SAR models based on experimental data

from different laboratories incorporates this implicit, but not often documented, reproducibility of the experimental test methods along with other important performance elements of the (Q)SAR model. In particular, the assessment of (Q)SAR reliability places greater emphasis on the accuracy of the (Q)SAR predictions with respect to many different chemicals than on the reproducibility of the (Q)SAR within and between laboratories. Moreover, reliability is more often described for an entire group of tested chemicals than as the reproducibility of individual endpoint estimations.

27. Likewise, the term "relevance" must be extended for the validation of (Q)SAR models because biological effects (endpoints) measured by test methods may appear to be similar for different chemicals but result of different molecular interactions and pathways. Consequently, even though the relevance of a test endpoint in regulatory assessments may be established, an additional assessment of the (Q)SAR model relevance must be made with respect to the expected molecular interactions and pathways by which each causes the biological effect. This important distinction between experimental test methods and (Q)SAR models is sometimes expressed by the extent to which each can be applied to the chemicals being regulated. The more reliable test methods tend to be more globally applicable to measuring the same endpoint for many different chemicals whereas the more reliable (Q)SAR models of major toxicity pathways reflected in a given endpoint tend to be relevant for specific classes of chemicals.

28. The extension of the traditional meaning of reliability and relevance for the purposes of (Q)SAR validation is readily captured by placing greater emphasis on the OECD Principle 3 involving the domain of application. The domain of application for a (Q)SAR model describes whether the model will predict an endpoint for a specific chemical with a given reliability. If the domain has well-defined mechanistic requirements, the model may be considered reliable, but only for a small subset of chemicals being regulated. As the domain of application is expanded to include more chemicals, the (Q)SAR model mixes different molecular processes and pathways and the reliability for larger domains decreases. The regulatory decision-maker must balance the global nature of the domain of application for a (Q)SAR model with the reliability needed for specific regulatory constraints and decisions.

29. One promising approach to improve reliability and expand the domain of application for (Q)SAR models is to compile numerous (Q)SAR models for the same regulatory endpoint and create an expert knowledge base which explains which domain of application includes each specific chemical of interest. The expert system for domains may be a simple decision tree involving chemical substructures or a computerized system for molecular similarity analysis. Regardless of the complexity of the expert system, this approach is capable of expanding the overall domain for a given endpoint while maintaining a higher reliability for individual estimates and providing greater transparency for the basis of the final estimate. Efforts to expand the domain of application for (Q)SAR models either by mixing mechanisms or by using expert systems with multiple well-defined domains has led (Q)SAR specialists to use the term "performance" to mean the goodness-of-fit, robustness and predictive ability of the model for a given endpoint.

The (Q)SAR Validation Process

30. For the purposes of this guidance document, a "validation process" refers to any exercise in which the OECD Principles for (Q)SAR validation are applied to a given model or set of models. It is not implied that the validation process should be carried out by any particular organisation, committee or formal validation body. When applying the OECD Principles, the basis for judging validity of a (Q)SAR includes the level of performance, the endpoint and the chemical domain required. Statistical validation of a (Q)SAR for purely scientific purposes is encouraged and is often used in the validation process; however, such scientific assessments of (Q)SAR performance can be misleading unless the relevance to the particular regulatory purpose is also established in the validation process.

31. The outcome of a (Q)SAR validation process should be a dossier providing information on the model performance of a (Q)SAR. The information should be obtained by applying the (Q)SAR Validation Principles, and the dossier should be structured accordingly. Until the scientific community designs (Q)SAR models with the OECD Principles in mind, it may not be possible to fulfil all principles for all models of regulatory interest. Although the transparency of the predictions may be less, reliable models may simply have not have been reported with the details to fulfil the OECD Principles. Therefore, the output of a successful validation exercise is a dossier that is as complete as possible, given the scientific and practical constraints. The output of a successful validation process is not intended to include a formal opinion on the validity of a model, but rather an objective checklist to document the performance and transparency of the model.

32. It follows that each regulatory authority will need to apply flexibility when considering the acceptability of a given (Q)SAR, taking into account the information provided in the (Q)SAR validation dossier, and the needs and constraints of the its particular regulatory programme.

Application of the (Q)SAR Validation Principles

33. The (Q)SAR validation principles are intended to be applicable to a diverse range of models types including SARs, QSARs, decision trees, neural network models, and expert systems which may contain multiple models for a given endpoint. The guidance provided in this document is intended to reflect this diversity of (Q)SAR models. In the case of these "complex models" that are actually based on the use of multiple models, it is important to identify the smallest component that functions independently, and to apply the principles to the individual component. Examples of such models include ECOSAR and Derek for Windows.

34. This guidance document covers the validation of models, but not the verification of computer programmes. It is important to distinguish between the *validation* of a model, and the *verification* of the software programme that executes the prediction. A highly predictive model could be regarded as valid, without considering whether the model has been coded correctly in the computer programme. Conversely, a poorly predictive model which might not be regarded as valid could be accurately translated into a specific programming language for implementation in a specific software package.

35. In principle, any model could be implemented in a variety of computer platforms, however, in practice, for certain types of models, it may be difficult to separate the model from the platform. This is particularly true of commercially available models, where certain components of the model (*e.g*. training sets, algorithms) are hidden for proprietary purposes.

36. A separate document on the regulatory uses and applications in OECD member countries of (Q)SARs in the assessment of new and existing chemicals is being developed as an accompanying document under Work Item 2 (OECD, 2006).

Overview of Chapters 2-6, Annex A-C and Glossary

37. Chapter 2 provides examples how the concepts of a "defined endpoint" can be interpreted in relation to different types of (Q)SARs.

38. Chapter 3 illustrates how the concept of an "unambiguous algorithm" can be interpreted in relation to different types of (Q)SARs.

39. Chapter 4 describes the current state-of-the-art of statistical methods and other approaches for the assessment of the applicability domains of (Q)SARs.

40. Chapter 5 describes the current state-of-the-art of statistical methods and other approaches for assessing the goodness-of-fit, robustness and predictivity of (Q)SARs, and explains the concepts of internal and external validation. This chapter also illustrates, by means of flow charts, logical sequences of steps that could be taken during the validation of (Q)SARs.

41. Chapter 6 provides examples to illustrate how the concept of "mechanistic interpretation" can be applied to (Q)SARs, where feasible. The mechanistic interpretation of a (Q)SAR includes two considerations: a) the interpretation of the (Q)SAR descriptors and consequently their relevance for the prediction of the endpoint; b) the relevance of the mathematical form of the relationship between the descriptors and the endpoint being modelled.

42. Annex A provides the OECD Principles for the Validation, for Regulatory Purposes, of (Q)SAR Models and explanatory notes for each of principles agreed at the 37th Joint Meeting in November 2004.

43. Annex B provides a check list of considerations that can be used to summarise which validation principles have been applied, and which pieces of information have been obtained.

44. Annex C provides a template for the reporting of (Q)SAR models, which could be included in dossiers that are submitted to regulatory authorities, or in (Q)SAR databases and decision support systems.

45. The Glossary provides a glossary of commonly-used terms in the (Q)SAR literature.

CHAPTER 2. GUIDANCE ON PRINCIPLE OF DEFINED ENDPOINTS

Summary of Chapter 2

46. This chapter introduces the rationale behind the first OECD Validation Principle which emphasises that a (Q)SAR should be associated with a "defined endpoint" (Principle 1). Guidance is provided on the interpretation of this principle, by describing what constitutes a defined endpoint. Following an introduction to the principle (paras 47-50), the concept of the defined endpoint is discussed (paras 51-68). It is emphasised that what constitutes a defined endpoint in the context of test guidelines does not necessarily constitute a defined endpoint for the purpose of (Q)SAR development. Difficulties in applying Principle 1 are illustrated with reference to the endpoints of acute fish toxicity, acute mammalian toxicity and biodegradation.

Introduction

47. (Q)SARs are relationships between the many different measures of chemical activity and measures of chemical structure. Measures of activity for chemicals made under specific conditions are called "endpoints". Because there are many different conditions under which the activity (especially toxicity) of chemicals can be measured, and there are different uses for endpoint data, it is important that the specific endpoint associated with a (Q)SAR be well described so that the user can judge whether the intended use is appropriate. For many regulatory applications, the endpoints of interest are often defined by a specific test guideline, and a (Q)SAR based on such endpoints would be intended to estimate the results of that specific test guideline.

48. OECD Principle 1 encourages (Q)SARs to be associated with a "defined endpoint", where endpoint refers to any physicochemical property, biological effect or environmental parameter related to chemical structure that can be measured and modelled. The intent of this OECD Validation Principle is to ensure transparency in the endpoint being predicted by a given model, since a given endpoint could be determined by different experimental protocols and under different experimental conditions. When comparing (Q)SAR predictions with experimental data, it is important to know that the model was developed and is intended to generate estimates of endpoints which are the same as the experimental data.

49. It should be noted that the definition of the term 'endpoint' might vary depending on the scientific or regulatory context and also on the user's professional focus. In fact, the term 'endpoint' may be used to designate information requirements on very different levels of specificity:

- A toxicologist will most likely define 'endpoint' as a specific biological effect which is precisely defined in terms of biological target structure and associated changes in tissue structure(s) and/or other parameters, *e.g.* induction of certain cytochromes, hypertrophy of hepatocytes, increased serum levels of aminotransferases, increased relative kidney weight etc.
- However, when toxicological data are generated for regulatory purposes, standardised and harmonised toxicological test protocols (such as those laid down in the OECD Test Guidelines) are used which generally cover a range of toxicological endpoints, which, in the case of *e.g.* repeat-dose experiments, may range from observations such as clinical signs, body weight or food consumption changes, to clinical chemistry and haematology parameters and macro- and

microscopical pathology findings. In this context, the term 'endpoint' has also been used for 'test protocol endpoints', such as 'skin irritation/corrosion' or 'acute oral toxicity in rats'.

- On a yet higher organisational level of regulatory work, especially in the context of classification and labelling, the results of several tests performed according to different protocols (or by similar protocols in different species/organisms) are often grouped into one 'regulatory endpoint', *e.g.* when a battery of different *in vitro* and *in vivo* test protocols are used to determine a chemical's potential with regard to the 'regulatory endpoint' genotoxicity, or when a relevant No-Observed-Adverse-Effect-Level (NOAEL) for 'repeat-dose toxicity' might be determined from a variety of studies in dogs, rats, or mice which are performed for periods from three weeks up to a year or longer, and using different modes of substance administration (feed/gavage/capsules).

50. The central issue addressed by OECD Principle 1 is that the 'endpoint' which a (Q)SAR model was built to predict might in fact be any one of the types of endpoints of different specificity contained in the above listing.

A Defined Endpoint

51. In the original liner notes to the OECD Validation Principles (*cf.* Annex A to this document), it was therefore stated that 'the intent of Principle 1 was to ensure clarity in the endpoint being predicted by a given model'. This transparency is indeed an essential requirement for regulators in order to assess the validity of a given QSAR prediction for application to a particular regulatory problem. Generally, the more congruent the endpoint predicted by a (Q)SAR model is with the regulatory endpoint under question, the more reliable will the prediction for a particular chemical by that particular model be for this regulatory purpose. *E.g.*, a (Q)SAR model built from a training set of Salmonella mutagenicity data, all obtained in accordance with OECD Test Guideline 471 (OECD, 1997), will probably provide useful predictions for the endpoint 'Salmonella mutagenicity'. On the other end of the scale, global prediction of a 'repeat dose LOAEL' from a training set of LOAEL data based on a variety of mechanisms/mode of actions and obtained in different species by different test protocols will hardly be seen as equally reliable. Of course, the question of the reliability of a particular (Q)SAR prediction in the context of a given regulatory problem will have to be answered on a case-by-case basis, however the essence of OECD Principle 1 is that detailed information is required as a sound scientific basis for that decision. The most important components of the information needed are described in more detail in the subsequent paragraphs.

52. It is also important to know whether the experimental data used to develop the model were generated according to a single experimental protocol, or whether data representing different protocols were merged in the training set. Consideration of the experimental protocol includes the design and conduct of the laboratory test and the methods employed for assay acceptance and assay evaluation criteria. Among the details of the experimental protocol which might influence the variability of the original training set data and the quality of the (Q)SAR model itself. Ideally, all (Q)SARs should be developed by using data generated by a single protocol, but this is not feasible unless the databases were designed as part of the (Q)SAR modelling process. Moreover, in many proprietary and regulatory models, information on the variation of protocols in the training sets is not always made publicly available.

53. In some instances, a single (Q)SAR model derived from a training set may not accommodate those Agencies which employ individualized regulatory paradigms. For example, Agencies often employ different procedures to evaluate and to regulate carcinogens: (1) pathology data (benign verses malignant), (2) statistical methods (trends verses pair wise comparisons), (3) regulatory philosophy (weight of evidence verses any evidence), and/or (4) analyses of historical tumour background frequencies verses concomitant experiment controls. For these complex endpoints it may be necessary to adjust the scoring and weighting of the endpoint data to develop more specialized QSAR models to meet each Agency's regulatory needs.

Examples of Defined Endpoints for Regulatory Assessment

54. In the context of this guidance document, the discussion is focused on those endpoints needed for the regulatory assessment of chemicals in OECD member countries (*e.g*. under REACH, in SIDS dossiers and in the GHS classification scheme). To be helpful in a regulatory context, (Q)SARs models are often grouped according to the defined endpoints, or the same toxicity effect, associated with OECD Test Guidelines. For example, the OECD Screening Information Data Set (SIDS) used in assessing existing chemicals includes endpoints from a wide variety of OECD Test Guidelines (Table 2.1). Some well-established endpoints might be estimated from dozens of QSAR models, some common toxicity effects are estimated using general (Q)SARs based upon non-congeneric training data sets, whereas other still evolving endpoints may have fewer or no (Q)SAR models yet capable of predicting the endpoint.

Table 2.1 Most Common Regulatory Endpoints associated with OECD Test Guidelines

Physicochemical Properties	Melting Point Boiling Point Vapour Pressure K octanol/water K organic C/water* Water Solubility
Environmental Fate	Biodegradation Hydrolysis Atmospheric Oxidation Bioaccumulation*
Ecological Effects	Acute Fish Toxicity Acute Daphnid Toxicity Alga Toxicity Long-term Aquatic Toxicity Terrestrial Effects
Human Health Effects	Acute Oral Toxicity Acute Inhalation Toxicity Acute Dermal Toxicity Skin Irritation /Corrosion* Eye Irritation/Corrosion * Skin Sensitisation * Repeated Dose Genotoxicity (in vitro) Genotoxicity (in vitro, non bacterial) Genotoxicity (in vivo) Reproductive Toxicity Developmental Toxicity Carcinogenicity* Organ Toxicity (*e.g.*, hepatotoxicity, cardiotoxicity, nephrotoxicity, etc.)

* non-SIDS endpoints

55. The principle for having a defined endpoint takes on a much greater importance in the validation of (Q)SAR models than simply reflecting the need for reproducibility of the endpoint as is often the case in validation of test guidelines. Test guidelines can often be applied to a broad array of chemicals limited only by physical properties, by common toxicity effects, or other factors that limit the experiment. Consequently, the test guidelines tend to have a global application domain wherein the test endpoint is a

reproducible measurement for broad classes of chemicals. However, the observed endpoints may be the result of a number of different toxicity mechanisms for the different classes of chemicals, all resulting in the same observed toxic effect.

56. For endpoints which can arise from numerous different chemical mechanisms, (Q)SAR models must either be developed for each mechanism which are applied to narrowly-specified classes of chemicals (see domain of application in Chapter 4), or a more general (Q)SAR relationship based upon the same observed toxic effect associated with non-congeneric classes of chemicals and different toxicity mechanisms. The general (Q)SAR relationship blurs the distinction of chemical classes at the expense of more explicit explanation of why a particular chemical produced the endpoint estimate. Either multiple (Q)SAR models for the same endpoint but each for different domains of application must be combined to yield a global estimation capability, or a statistical method capable of more global modelling across multiple mechanisms simultaneously must be used to gain a more global domain of application.

57. The OECD Principle of defined endpoint cannot be viewed in isolation from the other OECD Principles for (Q)SAR validation: *e.g.*, the nature of the defined endpoints in OECD Test Guidelines, therefore, creates a dynamic relationship between the trade-offs in mechanistic interpretation (Chapter 6), domain of application (Chapter 4), the algorithm used in the (Q)SAR model (Chapter 3) and the approach needed to understand the validity a (Q)SAR model for specific purpose (Chapter 5). Chemicals selected outside the domain of a mechanistic model do not make the (Q)SAR model invalid if the prediction does not agree with the measurement. The validation of a (Q)SAR model does not imply that a test guideline can be reliably estimated for all chemicals considered within the regulatory context.

58. For each defined endpoint used in a regulatory application, the domain of application of the available (Q)SAR models can be compared to the domain of chemicals being regulated to identify gaps. Multiple (Q)SAR models, or a general (Q)SAR model, may be required to predict a single endpoint for a heterogeneous list of non-congeneric chemicals. Families of (Q)SARs for a defined endpoint, each with a different applicability domain, can be integrated into a more global model if the domains of application are appropriately defined. Alternatively, if the regulatory context permits latitude in terms of the precision of the estimates, a general (Q)SAR model with a larger domain of application may be found reliable for the specific regulatory application.

Importance of Quality of Measured Endpoint Data

59. In addition to multiple underlying mechanisms by which chemical activity is observed, the endpoints in OECD Test Guidelines vary considerably in terms of their ability to measure the variation of test responses which can be related to intrinsic variation in chemical structure in a training set. Comparison of (Q)SAR models to predict the acute lethality to aquatic organisms and mammals is an example. Acute tests with fish are exposures directly across the respiratory surface wherein a steady-state between blood and external exposure concentration is quickly attained. A steady-state blood concentration yields a highly reproducible incipient lethal threshold which is linked directly to the exposure concentration. The steady-state conditions hold over 6-7 orders of magnitude in water solubility, so the variation in 96-hour LC_{50} endpoints is comparable for a wide variety of chemicals

60. Oral exposures with mammals are defined endpoints in the test guidelines but the measurement of the effects of the chemical on the animal is much more influenced by kinetic factors than intrinsic thermodymnamic variations in chemical structure. As chemical structure varies, the exposure itself is no longer comparable from one chemical to another. The shifting exposure regime with different chemicals will mean that, for some chemicals, a response near the true lethal potency of the chemical is measured whereas for another chemical only 10% of the true lethal potency is measured. In such cases, the endpoint is as much a measure of the toxicity of the chemical as it is an artifact of the way the chemical is tested.

61. Another example of an endpoint where care is needed in the development of the model and in the interpretation of the result is that of biodegradation. The definition of the biodegradation endpoint is very dependent on the experimental method used is also very variable even when the same method is used.

62. Biodegradability can be defined as the molecular degradation of a substance, resulting from the complex action of micro-organisms. It is one of the most important processes determining the fate of organic chemicals in the environment. Hence, biodegradation rates play an important role in the estimation of the fate of organic chemicals in the environment. However, as discussed below, one problem in the development of QSARs is that many regulatory studies are that they measure extent of biodegradation not rates.

63. In general, two types of biodegradation processes can be distinguished. Primary biodegradability occurs when an initial small alteration is made to the molecule, changing its physicochemical properties and integrity. It is quantified by measuring the disappearance of the parent compound with a specific analytical method or by the disappearance of a physico-chemical effect. Information on kinetics of primary degradation is warranted for chemicals whose toxic or inhibitory effects are lost as a result of the first enzymatic or abiotic reaction. Although there are few QSARs based on primary biodegradation, this is probably not unreasonable. Thus for example, in risk assessments, the uncertainties created by the need to assess unknown metabolites, would certainly limit the value of a primary biodegradation prediction.

64. Ultimate biodegradability occurs when a chemical substance is broken down and all the organic carbon is converted into carbon dioxide, methane and/or incorporated into biomass materials. This leads to a complete conversion of the organic carbon with extensive mineralisation and transient metabolites. Methods providing evidence of ultimate biodegradability are based on endpoints that are directly or indirectly related to the measurement of organic carbon oxidation.

65. A number of different tests, associated with a variety of endpoints, attempt to measure biodegradation. For example, the ready biodegradability test yields a value, frequently expressed as % biodegradability, and a term, (not) readily biodegradable. The former indicates the extent to which the substance degraded, while the latter is a legal or regulatory term that indicates whether a chemical passes or fails the OECD ready biodegradability test.

66. Many factors reflecting the extreme difference in biodegradation mechanisms to the very specific environmental conditions of each phenomenon, affect the biodegradability of a substance in the environment. Structural features such as molecular weight, types of bonds and substitutions affect biodegradation rate of organic compounds (Alexander, 1981; Kelcka, 1985). Environmental factors affecting biodegradability include microbial activity and growth as determined by temperature, pH, availability of nutrients, moisture level and residence time of the microbial population in the environmental compartment of interest. Processes such as microbial adaptation and co-metabolism add to the complexity of biodegradation.

67. Lack of uniform endpoints, substrate to biomass ratio, and time allowed for acclimation across the tests are responsible for the limited size of available training sets (compared to those used in toxicology) for Quantitative Structure-Biodegradability Relationships (QSBRs). Within a specific test, intra-laboratory and inter-laboratory variability in endpoint measured add to the difficulties in selecting a training set. For a specific standard biodegradation test method carried out at different laboratories or within a single laboratory discrepancies and large variability can be observed in the results (King and Painter, 1983; Kitano and Takatsuki, 1988). Although biodegradation tests have been standardised by the OECD, a deviation of 20% is considered acceptable when a test is repeated within the same laboratory (OECD, 1993). This limits the development of QSARs for biodegradation and biodegradation kinetics. Therefore, success of future developments in the QSAR area with respect to biodegradability will be dependent on the

availability of high quality experimental data (Peijnenburg *et al.*, 1995) and our ability to use these data and extrapolate them to the real environment.

68. From the discussions above, it follows that a defined endpoint, or the same observed toxic effect, should contain a number of desirable attributes. The list below is not intended to be exhaustive, nor should it mean that an endpoint that lacks one or more of these elements is ambiguous. However, developers and users of (Q)SARs should be aware of the extent to which the endpoint being modelled can be described in the following terms:

1. The endpoint should be defined by providing detailed information on the test protocols which were used to generate the training set data, especially with respect to factors which impact variability, knowledge of uncertainties, and possible deviations from standardised test guidelines.
2. Differences in the protocols that experimentally measure the described endpoint should not lead to markedly different values of the endpoint.
3. Differences within a protocol (*e.g.* media, reagents) should not lead to differences that cannot be rationalised (*e.g.* impact of hardness within a fish LC_{50} study).
4. The chemical domain of the (Q)SAR should be contained within the chemical domain of the test protocol.
5. The endpoint being predicted by a (Q)SAR should be the same as the endpoint measured by a defined test protocol that is relevant for the purposes of the chemical assessment.
6. A well-defined endpoint should reflect differences between chemical structures.

Concluding Remarks

69. Data-driven (Q)SAR models can most often be developed for measures of chemical activity (endpoints) for which numerous chemicals have been tested using comparable protocols. The resulting (Q)SAR models can only be expected to predict the specific endpoint used in the training set. To facilitate the use of (Q)SAR models in regulatory decisions for untested chemicals, many of the endpoints used in (Q)SAR development and validation will be identical to those already in use in the regulatory activities such as the OECD Test Guidelines. At the same time, the users of (Q)SAR models are cautioned that the reproducibility of measured endpoint values from a standardized test protocol is a primary determinant of the reproducibility of a (Q)SAR model for that same endpoint. Consequently, (Q)SAR models based on well defined endpoints in harmonized test protocols have the greatest potential for concordance between measured and estimated values.

CHAPTER 3. GUIDANCE ON PRINCIPLE OF UNAMBIGUOUS ALGORITHMS

Summary of Chapter 3

70. This chapter introduces the rationale behind the second OECD Validation Principle which emphasises the importance for (Q)SARs to be associated with an "unambiguous algorithm" (Principle 2). Guidance is provided on the interpretation of this principle by describing what constitutes an unambiguous algorithm. Following an introduction to the principle (paras 71-72), the concept of the defined algorithm is discussed (paras 73-78). The need for a defined algorithm is discussed in terms of the elements that are needed for an algorithm to be fully transparent, with particular emphasis on the need for information concerning the descriptors used to model the endpoint of interest and the mathematical methods used to develop an algorithm that is based on these descriptors (paras 79-90).

Introduction

71. (Q)SAR models are relationships between the behaviour of chemicals as defined by the model endpoints (Chapter 2) and different descriptors of chemical structure. The form of the relationship between the descriptors of chemical structure and the "activity" endpoint in a (Q)SAR model is called the "algorithm" of the model which may be a mathematical model or a knowledge-based rule developed by one or more experts. This chapter provides guidance on the importance of using (Q)SAR models with unambiguous algorithms and the importance of the algorithm in the (Q)SAR validation process.

72. According to Principle 2, a (Q)SAR model should be expressed in the form of an "unambiguous algorithm". The intent of this OECD Validation Principle is to ensure transparency in the description of the model algorithm so that others can reproduce the model and explain how (Q)SAR estimates are derived. One important contributor to transparency of (Q)SAR estimates is the ability of users to explain how the endpoint estimates of the model were produced. Most models have algorithms comprised of unambiguous statistical methods and/or process models which have been evaluated by scientific peer review. Other more exploratory algorithms are not capable of explaining how the model estimate was derived, nor independently reproduced by others. In many proprietary models, the algorithm may not be publicly available so that the results may be reproduced by others but not explained.

Unambiguous Algorithms

73. The algorithms used in (Q)SAR modelling should be described thoroughly so that the user will understand exactly how the estimated value was produced and can reproduce the calculations, if desired. Important regulatory endpoints are estimated for chemicals by selecting the proper (Q)SAR for the specific class of chemical (see domain of application in Chapter 4), or a proper general (Q)SAR model(s) based upon a common toxic effect, computing the chemical-specific molecular descriptors required by the (Q)SAR model, and using those molecular descriptors in a mathematical algorithm to create an estimate of the endpoint for the chemical. The ability to reproducibly complete all three steps producing an estimate is an important part of the acceptance of (Q)SAR models. All three steps in producing estimated values may involve individual algorithms as is the case of mechanistic estimates of dose-response endpoints. For many binary endpoints where the (Q)SAR model is primarily a classification model, the algorithms may be an association with the presence or absence of important chemical substructures.

74. OECD Principle 2 simply states that (Q)SAR should be comprised of unambiguous algorithms. Because there may be multiple steps in the estimate of an endpoint for a chemical, the unambiguous nature of all algorithms used is important. In practice, the algorithms for chemical classification, molecular description and computing endpoint estimates may involve expert knowledge of statistics, physical chemistry and toxicology. OECD Principle 2 is not intended to suggest that the application of algorithms in (Q)SAR modelling requires expertise in these fields of science. In many cases, it is possible to offer transparent description of algorithms without necessarily delving into the mathematical or statistical methods used to develop the algorithm. For example, a regression-based QSAR can be defined explicitly without any particular discussion of the regression approach. An expert rule can be stated explicitly without the need re-derive the consensus of a panel of experts.

75. Some exploratory algorithms used to understand variation in data sets are inherently ambiguous and are not recommended for regulatory applications. In cases where the definition of the algorithm is more closely associated with the way in which it was derived (*e.g.* a neural network model which includes both a learning process and a prediction process), users must relay on the validation process to determine if an ambiguous algorithm can produce reliable results for a regulatory application. In all cases, it is recommended that (Q)SAR validation exercises should seek as much information as possible on both the method used to develop the algorithm, and the algorithm itself.

76. It is important to distinguish between the transparency of the algorithm and the ability to interpret the algorithm as a cause-effect relationship. For example, a statistically-based QSAR can be transparent in terms of its predictor variables and coefficients, but the descriptor variables themselves may not have an obvious physicochemical meaning or plausible causal link with the endpoint being modelled. Only where a mechanism is known or assumed will the relationship expressed by the QSAR be based on (possible) causality rather than correlation. In such cases any conclusions based on the model are related to and defined by the original dataset used to develop the model. This is especially true for small datasets where there are a limited number of chemicals and a large variability in the data. Thus mechanistic interpretation becomes an additional issue to address when considering the algorithm (see Chapter 6).

77. The following elements should therefore be considered when assessing the algorithm:

1. The dataset of chemicals, end-point values and descriptor values.
2. A clear description of the derivation of the descriptors and how they were measured.
3. A clear description of the test and training sets and, if outliers were removed a clear justification for this.
4. The mathematical model(s) used to explore the descriptor and end-point relationship needs describing.
5. Statistical parameters describing how the model performs (see Chapter 5).
6. The parameters and their values which constitute the (Q)SAR.

78. A brief overview of algorithms commonly used in (Q)SAR are presented in this guidance to distinguish methods which generally fulfill OECD Principle 2 and those where special precautions are needed in the validation process.

Univariate regression (ULR)

79. Univariate regression involves only one dependent response variable *(y)* and one independent variable *(x*) which model a simple relationship between a molecular descriptor and an endpoint. Univariate linear regression (ULR) assumes that the relationship being modelled is a straight line. The normal method of determining the regression coefficients is minimising the sum of the squares of the residuals using the

least squares method. Non-linear univariate regression such as the use of exponential functions are less frequently used due, in part, to the fact that the non-linear model is less constrained than the corresponding linear model. It is vital when assessing a non-linear model to examine the residuals across the full extent of the model to ensure that it has not been overfitted (Draper and Smith, 1991). It should be noted that the linear model is generally considered as unambiguous algorithm.

Multiple Linear Regression (MLR)

80. When the endpoint needs to be modelled using more than one descriptor (selected by different approaches) then multivariate techniques are applied. The technique of multiple linear regression (MLR) is discussed in Chapter 5 and is extensively covered in Draper and Smith (1991). MLR, in particular OLS (Ordinary Least Squares), is the most popular regression method, it produces a transparent and easily reproducible algorithm. As it can suffer of the use of correlated variables, this correlation must be carefully controlled by the methods discussed in Chapter 5. The problem of possible overfitting, common also to other modelling methods, must be also verified by statistical validations methods for predictivity (see Chapter 5). The selection of descriptors in MLR can be performed *a priori* by the model developer on mechanistic basis or by evolutionary techniques such as Genetic Algorithms as well as methods like Principal Component Analysis (PCA) or Factor Analysis (FA). In the latter approach the selected modeling descriptors can be mechanistically interpreted after the model development.

81. There are many examples of the use of MLR, including the baseline toxicity model of Koneman (1981) and, for instance, the studies by de Bruijn and Hermens (1991) and Govers *et al.* (1984, 1991). Successful examples of more novel applications of MLR are in the papers: Netzeva *et al*., 2005; Ren S., 2003; Ghafourian and Cronin, 2005. In some recent publications, particular attention is devoted to model validation for predictivity and chemical domain of applicability, as well as to the descriptor interpretation, thus to the model development according to all the OECD Principles (Gramatica *et al*., 2004; Gramatica and Papa, 2005; Pavan *et al*., 2006).

Principal Component Analysis (PCA) and Principal Component Regression (PCR)

82. Principal Component Analysis (PCA) is a technique used for dimension reduction and is based on linear combinations of the variables. In Principal Component Regression (PCR), one obtains a reduced-order model by neglecting some components of the PCA modelling of the independent variables (X-matrix) and relating the maintained principal components to the dependent variables (Y-variables). The neglected components are usually dominated by non-relevant information in the data. But it is possible that the neglected components contain some relevant information, this information is lost when the higher components are neglected. It is also possible that some noise is maintained in the model. In this situation a good model for the training set (dataset used to construct the model) is obtained but the model has a poor ability to predict the test set. This effect is known as the overtraining effect. When applying PCA it is difficult to identify outliers and hence the model will give undue weighting to these data points. Robust methods (Walczak and Massart, 1995) have been proposed in an attempt to overcome this problem (Niemi, 1990; Kaiser and Esterby, 1991).

Partial Least Squares (PLS)

83. Partial least squares (PLS) is a combination of MLR and PCR. It attempts to explain the variance in the independent variables and also tries to obtain a good correlation between the dependent and the independent variables. One major advantage of PLS is that it is very useful when co-linearity in the descriptors exists. To reduce the overtraining effect, a cross validation can be performed during the model constructing phase. A*s* with PCA, outlier identification are also a problem for PLS and again robust methods have been proposed (Wakeling and Macfie, 1992; Griep *et al.*, 1995). The final model is affected

by the introduction of outliers. A better model will be obtained when the outliers are left out or become less important for the final model, *i.e*. when the model is made more robust.

Artificial Neural Nets (ANN)

84. Neural nets are used in many areas, such as pattern recognition, process analysis and non-linear modelling. An advantage of neural nets is that the neural net model is very flexible in contrast to the classical statistical models. A significant disadvantage is the amount of data needed and the causal ambiguity of the network. The neural net 'learns' from examples by one of two different approaches, supervised or unsupervised learning. During supervised learning, the system is forced to assign each object in the training set to a specific class, while during unsupervised learning, the clusters are formed without any prior information. One approach commonly used is multi-layer feed-forward (MLF) networks consisting of three or more layers: one input layer, one output layer and one or more intermediate (hidden) layer (Smiths *et al.,* 1994; Xu *et al.,* 1994; De Saint Laumer *et al.,* 1991).

Fuzzy Clustering and Regression

85. In contrast to traditional regression/classification techniques, fuzzy clustering or regression is capable of dealing with *probabilities* of finding objects belonging to certain classes, instead of classifying with hard limitations (yes/no decisions) (Friederichs *et al*., 1996). The limiter functions (which have in most cases a sigmoidal shape), may hold all states or values between two extreme assertions.

K-nearest Neighbour Clustering

86. K-nearest neighbour (KNN) clustering determines the class of an object by assessing the class of a number of the closest neighbours to the object. The majority, sometimes weighted depending on distance, will determine the class of the object being assessed.

Genetic Algorithms (GA)

87. A genetic algorithm (GA) is an artificial intelligence technique based on the theory of evolution that through the process of natural selection, formulae evolve to solve problems or develop control strategies. A brief but thorough introduction to genetic algorithms is provided by Forrest (1993). Goldberg (1989) provides an introductory text on genetic algorithm development, while Koza (1992) supplies a more advanced treatment of genetic algorithms.

88. GA has a number of unique features. First, a GA does not search for a single solution, but in fact maintains a set of perhaps thousands of solutions, referred to as a population. Second, the GA attempts to increase the "fitness" of this population at each generation. Each solution is evaluated as to its "fitness" based on some domain-specific function, then kept or discarded based on that evaluation. If discarded, that member of the population is replaced by a new solution, which is created by a recombination of parts of existing good solutions.

89. This process is repeated thousands, perhaps millions of times, combining different aspects of good solutions, while searching for a combination of solution features that is optimal under the evaluation function imposed. The GA designer provides a function to evaluate the "fitness" of each individual solution; this fitness function is used to propagate "good" individuals into the next generation. A set of these fit individuals are chosen for a crossover operation, which recombines the strings of the parents into new children, trying to construct fitter solutions in the process. The mutation operator randomly alters some element of an individual (solution) in order to further enhance the population.

90. Because genetic algorithms do not use statistical procedures, they are not limited by statistical assumptions. GA performance is, however, strongly influenced by design decisions made by the programmer. With a GA, the designer has more flexibility than is available with many other types of procedures. They can be modified to accommodate important characteristics of a system, explore alternate hypotheses, and elucidate underlying mechanisms simply by adding variables to the program or altering fitness criteria. GA methods readily lend themselves to the exploration of relationships between input and output data (*i.e*. QSAR development).

Concluding Remarks

91. An important premise in scientific integrity is the ability to explain scientific results thoroughly enough so that the scientific community can reproduce the reported results. In QSAR development, providing the training set of data for a defined endpoint is a major part of describing the QSAR in a transparent manner. The algorithm used to relate the endpoint data to descriptors of chemical structure is a second important part of explaining the (Q)SAR model to the scientific community and users. The unambiguous algorithm makes it possible for the model to be tested as well as for the user to develop a worksheet or other explanatory summary of exactly how the (Q)SAR estimate was made. (Q)SAR models which are not transparent with respect to the algorithms used may be as accurate as many validated models, but the lack of explanation of how the estimates were made negatively influences regulatory acceptability.

CHAPTER 4. GUIDANCE ON PRINCIPLE OF A DEFINED DOMAIN OF APPLICABILITY

Summary of Chapter 4

92. This chapter provides guidance on how to interpret OECD Validation Principle 3 that a (Q)SAR should be associated with "a defined domain of applicability" (Principle 3). This principle expresses the need to establish the scope and limitations of a model based on the structural, physicochemical and response information in the model training set. The importance of the principle lies in the fact that a given model can only be expected to give reliable predictions for chemicals that are similar to those used to develop the model. Predictions that fall outside the applicability domain (AD) represent extrapolations, and are less likely to be reliable. When applying a (Q)SAR, it is important to know whether its AD is known, and whether it is being used inside or outside of this boundary. In its simplest form, the assessment of whether a chemical is located in the AD can be expressed categorically (*i.e.* yes or no). For a quantitative assessment, it is possible to associate a confidence interval with the AD, to determine the degree of similarity between the chemical of interest and the model training set. This chapter begins by explaining of the need for defining the AD (paras 93-95), before introducing some basic concepts and definitions (paras 96-98). The chapter then provides a review of different methods that are currently available or under development for identifying and quantifying the applicability domain, with some examples to illustrate their applicability (paras 99-123). It is emphasised that the subject of the (Q)SAR AD is an evolving field of research, and some research needs are presented in the concluding remarks of the chapter (paras 125-129).

Introduction

93. OECD Principle 3 states that "a (Q)SAR should be associated with a defined domain of applicability" and expresses the need to include supporting information with a (Q)SAR which will define the classes of chemicals with which the model performance will satisfy the regulatory requirements. There is no absolute boundary between reliable and unreliable predictions for a given model, but rather a trade-off between the constraints of the applicability domain (AD) and the overall reliability of prediction for numerous chemicals. In general, the less constrained the AD, the more likely chemicals will be included for which the predictions will be less reliable. The more constrained the AD, the more chemicals will be encountered for which the endpoint cannot be predicted with the (Q)SAR. The balance within these trade-offs depends on the requirements and can be determined by the user in the validation process within the specific regulatory context.

94. Information on the AD helps the user of the model to judge whether the prediction for a new chemical is reliable or not. The definition of the AD is based on the assumption that a model is capable of making reliable predictions only within the structural, physicochemical and response space. As a minimum, the AD can be defined by analysis of the training set as will be described in this guidance document. In the more highly developed (Q)SAR models, the AD is defined by mechanistic structural requirements which are derived from interactive hypothesis generation and testing in the design of the training set. Regardless of how explicitly the AD is defined, the model fit, robustness and predictivity determined by statistical methods (see Chapter 5) are meaningful only if they are used for chemicals within the AD. Even within the AD of a model, different degrees of confidence can be associated with different predictions.

95. The determination of whether a chemical falls within the AD of a model is based on an assessment of the similarity between the chemical and the training set. Since there are many different ways of expressing similarity (often defined in physicochemical properties), it follows that many different methods for defining the AD can be developed. This guidance document summarizes a variety of methods in the scientific literature along with their strengths and limitations

Basic Terms and Concepts

96. For the purposes of this guidance document, the definition of the AD is the following (Netzeva *et al*., 2005):

> *"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."*

97. In this definition, chemical structure can be expressed by information on physicochemical properties and/or structural fragments, and the response can be any physicochemical, biological or environmental effect that is being predicted (*i.e.* the defined endpoint, see Chapter 2). The relationship between chemical structure and the response can be expressed by a variety of SARs and QSARs.

98. The AD principle should be applied in a model-specific manner. Thus, every model should be associated with its own AD derived not only on the chemicals in the training set but also on the descriptors and (statistical) approach used to develop the model. Ideally, the AD should be defined and documented by the model developer. This information should include: a statement of the unambiguous model algorithm (see Chapter 3), details of the training set (chemical identification, descriptors and endpoint values), details of the (statistical) method to derive the model, and structural requirements determined during model development.

Recommendations for Deriving Applicability Domains

99. Ideally, the AD should define the structural, physicochemical and response space of the model. This is because the best assurance that a chemical is predicted reliably is to have confirmation that the chemical is not an outlier in terms of its structural fragments (structural domain), its descriptor values (physicochemical domain) or its response values (response domain). When the AD is defined in more mechanistic terms, the (Q)SAR can predict reliably beyond the physicochemical and response space of the training set.

100. Even though a well-defined AD helps the user of the model to assess the reliability of predictions made by the model, it should not automatically be assumed that all predictions within the defined AD are necessarily reliable. In practice, a prediction could still be unreliable even though the chemical lies within the established structural and physicochemical domains of the model. This could occur in cases where the chemical of interest acts by a different mechanism of action, not captured by the model. If more than one such chemical is discovered, the QSAR practitioner could either try to refine the model, to accurately predict the outliers, or could try to define an exclusion rule. The need to account for such outliers has also led to the concept of the mechanistic domain. Thus, for some models, the application of OECD Validation Principle 3 is linked with the application of Principle 5 on mechanistic interpretation.

101. Historically, the first QSAR models were developed for homologous series of chemicals. Although these models may have limited use today, they are helpful to illustrate how the concept for the AD can be applied. For example, if one knows the narcotic effects of the primary alcohols ethanol, propanol, butanol, hexanol and heptanol, then one can predict the narcotic effect of pentanol by the linear relationship between the narcotic effect and molecular weight (MW). Pentanol is in the AD of this simple model because it is a structural homologue of the other alcohols and has a MW intermediate to two other

alcohols. The alcohols methanol and the n-octanol, however, would not be considered in the AD of the model, because while they are structural homologues of the other alcohols, they have MW values lower than ethanol and greater than heptanol, respectively (Figure 4.1).

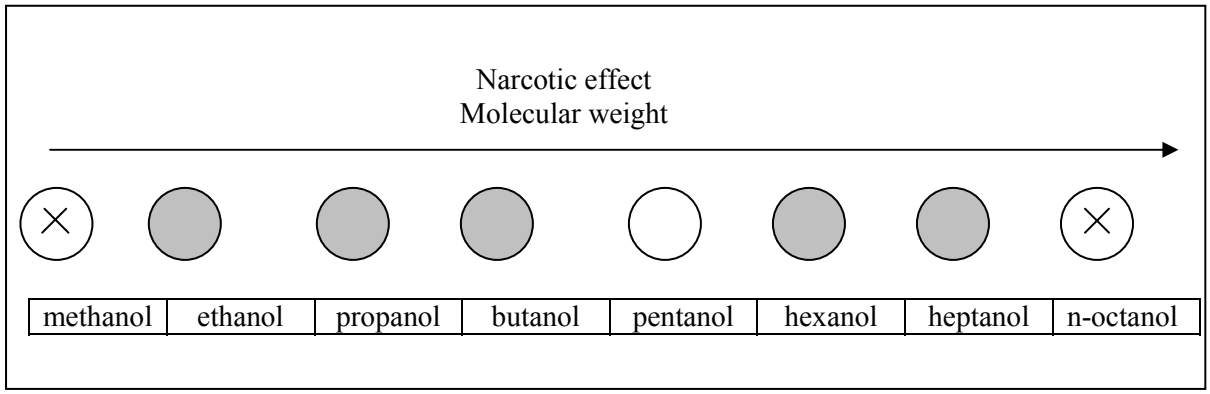

Figure 4.1

102. Other examples support the same reasoning. For example, Deneer *et al.* (1988) have shown that increasing the number of carbon atoms in a homologue series of aldehydes above 10 leads to a change of the mechanism of action. The consequence is that the relationship between the toxicity and the octanol-water partition coefficient (log K_{ow}), found for lower members of the series, does not hold true for the higher members. In another example, Schultz and Cronin (1999) showed that acrolein, the first chemical in the series of α,β-unsaturated aldehydes, was considerably more toxic than predicted by the relationship between log K_{ow} and toxicity for the other α,β-unsaturated aldehydes.

103. In addition to the physicochemical and structural domains, an additional useful element in the AD definition is an understanding of the mechanism of action (MOA) of the chemicals used to develop a model (*i.e*. the mechanistic domain). For example, the phenols and the anilines (if not complicated by more reactive moieties) demonstrate polar narcosis in aquatic organisms (Verhaar *et al*., 1995) even though they belong to different chemical classes. Thus, the effects of chemicals belonging to both chemical classes can be predicted by a single model provided the chemical does not go beyond the range of physicochemical parameters used to develop the model. The grouping of chemical classes into single QSARs is endpoint-specific because the different classes might not behave in the same way for a different endpoint (*e.g.* mutagenicity). In fact, aromatic amines have considerable potential to cause mutations whereas phenols do not.

104. Chemicals that contain multiple functional groups deserve special attention. Such chemicals might exhibit enhanced effects as a result of synergism or even exhibit a different MOA. Such chemicals are likely to be outliers to well established relationships. An example is provided by the α-halogenated esters Schultz *et al.* (2002), in which the presence of a halogen atom on an aliphatic hydrocarbon chain does not alter the narcosis MOA for aquatic toxicity. Aliphatic esters also act as narcotics in aquatic organisms. However, the presence of a halogen atom at the α-position to the carbonyl group of an aliphatic ester results in a drastic increase of toxicity due to the fact that this arrangement of atoms undergoes an SN2 reaction (the halogen atom being the leaving group) with macromolecules.

105. The identification of special atom arrangements (toxicophores) that cause certain types of toxicity provides a way of defining mechanistic domains. Expert judgement is required since the expected toxicological profile could be modulated by the presence of additional functional groups (modulators), which may increase or decrease the toxicity. For example, the methyl groups usually increase the toxicity due to increased lipophilicity without changing the MOA. Thus, the methylphenols are slightly more toxic

to fish than the parent phenol (Russom *et al.*, 1997). However, methyl groups can also block completely the toxicophore; for example the methyl groups in the *tert*-butyl group decrease the toxicity of *tert*-butyl acrylate (Schultz *et al.*, 2005). The presence of a bulky substituent next to a reactive group is one reason why a chemical might fall outside the expected mechanistic domain. The properties of such chemicals or are usually overestimated.

106. Inaccuracy of prediction can appear also if a chemical undergoes metabolic transformation. Such chemicals appear outliers from many different (Q)SAR models irrespectively of whether the model was developed on a mechanistic basis or statistically. The reason for miss-prediction in this case is that the chemical that causes the effect is different from the chemical that was introduced to the biologic system and these out-of-the-domain chemicals are usually most difficult to identify *a priori*. An example could be given with 1,2- and 1,4-dihydroxybenzenes that exhibit enhanced toxicity because of transformation to 1,2- and 1,4-quinones with strong electrophilic potential, or formation of free-radical species (O'Brien, 1991).

107. At present, the identification of mechanistic domains relies heavily on expert judgement. There are, however, some software tools that can assist in the identification of potential toxicophores and modulators. An example is the Derek software (Lhasa Ltd. (Logic and Heuristics Applied to Synthetic Analysis)), an expert system that applies knowledge-based rules for toxicity prediction. A similar functionality is available in HazardExpert (Compudrug.Inc.), which issues an alert if a toxic fragment is found in the query molecule. Another program for toxicity prediction, MULTICASE (Multicase Inc.), evaluates the structural features of molecules from non-congeneric training data sets and identifies substructural molecular fragments that are significantly correlated with specified toxicological activities, and substructural molecular fragments and molecular descriptors that modulate specified toxicological activities. The MDL-QSAR software (MDL Information Systems) evaluates E-state and other molecular descriptor features of molecules from non-congeneric training data sets and identifies descriptors that are significantly correlated with specified types of toxicological activities. The TOPKAT software (Accelrys Inc.) uses an initial classification into chemical classes before applying quantitative models for toxicity prediction. Various software products incorporate knowledge about metabolism and can therefore be used to anticipate the metabolites of the chemical of interest. These systems include CATABOL (Laboratory of Mathematical Chemistry, Bulgaria), META (Multicase, Inc.), MetabolExpert (CompuDrug Inc.) and METEOR (Lhasa Ltd.).

108. If a (Q)SAR is based on physicochemical descriptors, the interpolation space (*i.e.* its coverage), defined by its descriptors, should be characterised. The interpolation space of a one-descriptor model is simply the range between the minimum and the maximum value of that descriptor, as observed in the training set of the model. The interpolation space of multi-descriptor models is more complex. Several statistical methods can be applied to characterise the interpolation space, as described below.

109. The simplest method for describing the AD is to consider the ranges of the individual descriptors. This approach is based on the assumption that the descriptor values follow a normal distribution, and could therefore be unreliable if this assumption is violated. A limitation of this approach is that the AD may include internal empty spaces, *i.e.* interpolation regions where the relationship is not proved (Figure 4.2). Another possible limitation is the fact that intercorrelation between the descriptors is not taken into account, unless the individual descriptors are replaced by their principal components.

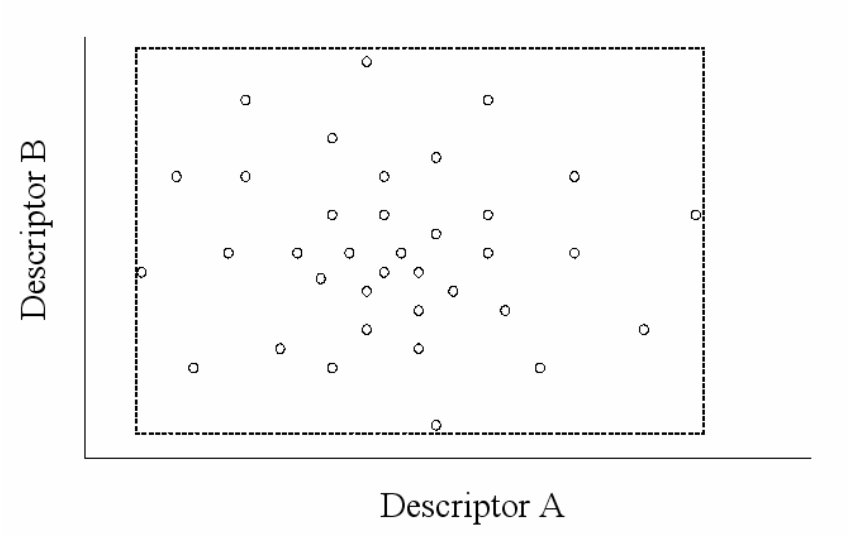

Figure 4.2

110. A more advanced method for defining the interpolation space of a model is to define the smallest convex area that contains the descriptors of the training set. However, this method does not solve completely the problem of empty spaces between the chemicals of the training set. In addition, for models that contain many descriptors, the calculation of the convex area becomes a time-consuming computational problem (see Figure 4.3).

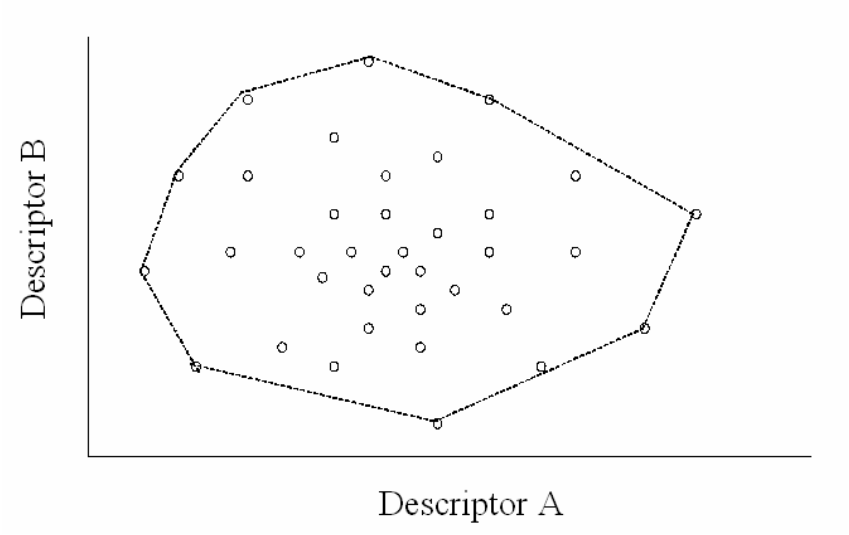

Figure 4.3

111. A different approach to defining the AD is based on a calculation of the distance between a query chemical and a defined point in the descriptor space of the model (typically, the centroid of the training set). A detailed review of methods is given by Jaworska *et al.* (2005). Different methods following this approach can be applied (*e.g.* Euclidean, Mahalanobis, Manhattan distance). The advantage of the distance (also called geometric) approach is that confidence levels can be associated with the AD by drawing iso-distance contours in the interpolation space. The disadvantage is again the assumption of a normal distribution for the underlying data. This means that the contours are drawn solely on the basis of the

distance from the centroid, and the population of the regions between two iso-distance contours is not taken into account.

112. A common approach to distance analysis is to use the Hotelling's test and the associated leverage statistics. The leverage of a chemical provides a measure of the distance of the chemical from the centroid of its training set. Chemicals in the training set have leverage values between 0 and 1. A "warning leverage" (h^*) is generally fixed at $3p/n$, where n is the number of training chemicals, and p the number of descriptors plus one. A leverage value greater than the warning leverage is considered large.

113. The leverage is a useful statistic in both QSAR development and application. During QSAR development, chemicals with high leverage unduly influence the regression parameters of the model, and yet do not appear as statistical outliers (the regression line is forced near the high leverage chemical). It may therefore be appropriate to exclude such chemicals from the training set. During the application of a QSAR, chemicals with high leverage are likely to be outside the descriptor space of the model, and therefore the predictions for such chemicals could be unreliable. The leverage approach is illustrated in Gramatica *et al.* (2004) and Pavan *et al.* (2005).

114. As with all statistical methods based on physicochemical descriptors, the leverage approach needs to be applied with care. While the observation that a chemical has a large leverage indicates that it is outside the descriptor coverage of the model, a chemical with small leverage can also be outside the AD for other reasons (*e.g.* a presence of a toxicophore that is not present in the training set). The inability to discriminate unequivocally between chemicals that are inside and outside the AD is common to all statistical methods based on physicochemical descriptors, and this should be taken into account when applying the concept of the AD.

115. To visualise the outliers in a model, *i.e*. outliers in both the descriptor space and the response space, a plot of standardised residuals (R) *vs.* leverages (or hat values, h), called the Williams graph is sometimes used. An illustration of the Williams plot, taken from Pavan *et al.* (2005), is given in Figure 4.4a. This shows the training set of 86 chemicals for a polar narcosis model of acute toxicity to the fathead minnow (Verhaar *et al*., 1995) as well as a test set of 8 chemicals for which the model was used to make predictions. It can be seen that 6 chemicals in the training set have leverages greater than the warning leverage (0.07), as do 2 of the test chemicals. The corresponding regression line for the model is shown in Figure 4.4b.

116. The most advanced statistical methods that are currently applied for identifying the (Q)SAR AD are probability density distribution-based methods. The probability density function of a data set can be estimated by parametric and non-parametric methods. The parametric methods assume a standard distribution (*e.g.* Gaussian or Poisson distribution) while the non-parametric methods (*e.g.* kernel density estimation function) make no assumptions about the data distribution. An advantage of non-parametric methods is the ability to identify internal empty spaces and, if necessary, to generate concave regions around the borders of the interpolation space to reflect the actual data distribution. It has been argued that the probability density approach is more robust than the range, distance and leverage approaches (Jaworska, 2005). However, it is also more restrictive in terms of the chemical space that falls in the AD of a model.

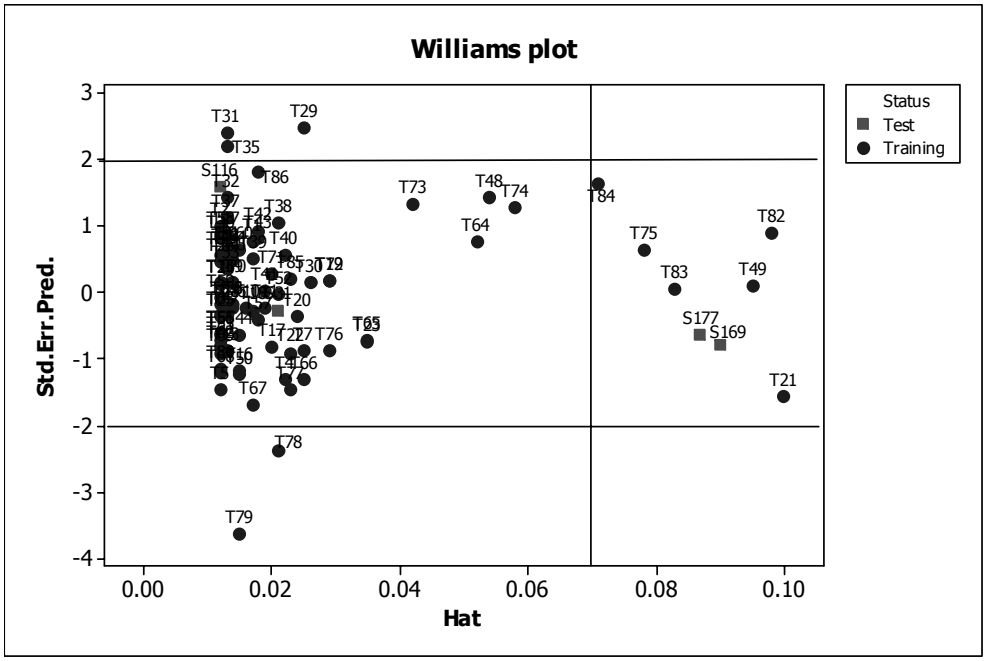

Figure 4.4a

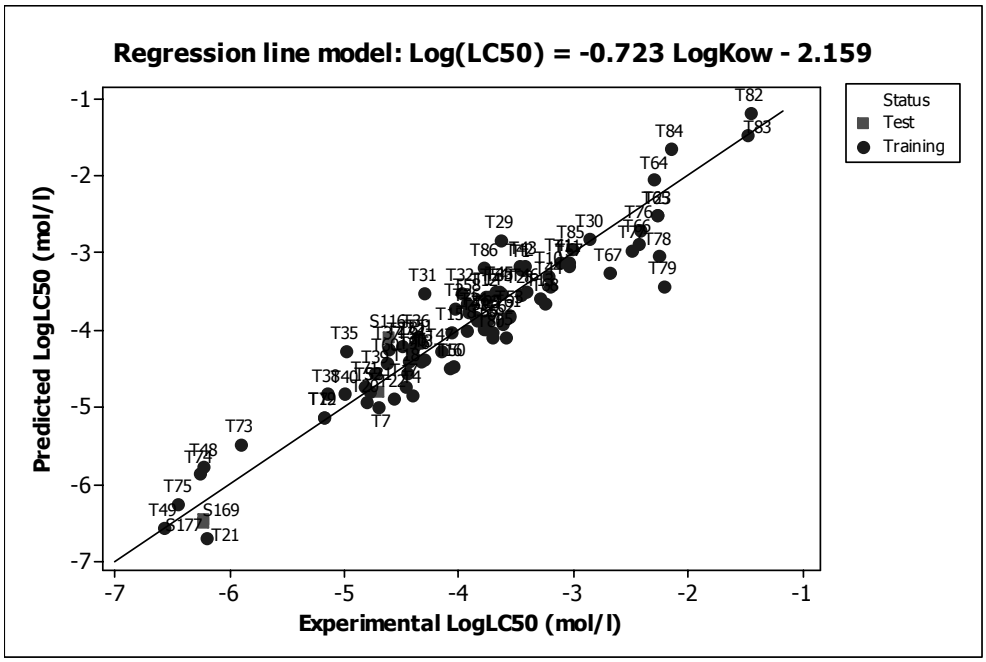

Figure 4.4b

117. While some of the described statistical methods are available in a standard statistical packages (*e.g.* MINITAB, STATISTICA, SYSTAT), they are not adapted to meet the needs of (Q)SAR developers and users. In contrast, a user-friendly software package called Ambit Disclosure being developed under the auspices of CEFIC LRI can be used to calculate the interpolation space by knowing the values of the dependent (endpoint) and independent (descriptors) variables used in a given model. The AD methods

incorporated in the software are independent of the modelling technique and require only transparency of the training set. A free download is available on the internet (Ambit Disclosure Software developed by Jaworska, J.S. and N. Nikolova, accessible: http://ambit.acad.bg, last accessed 6 February 2007).

118. Ideally, the coverage of the training set should be accompanied by information on the structural or physicochemical similarity between the query molecule and the (Q)SAR training set. The similarity can be expressed in a qualitative or quantitative manner. Preferably, some mechanistic rationale should be given of whether the query chemical represents a mechanism common to a group of chemicals in the (Q)SAR training set. However, when such an assessment is not possible, a statistical expression of similarity can be obtained.

119. One possible approach is to split the query molecule in molecular fragments and to check whether all the fragments are represented in the training set of the model. The higher the occurrence of the query fragments in the training set, the higher the confidence that the query chemical of interest can be predicted reliably. This approach is adopted in the MultiCASE software and in the Leadscope Inc. Prediction Model Builder software. These programs issue a warning message that a chemical is outside of the AD of the model if it encounters an unknown fragment.

120. A quantitative expression of similarity can be obtained by using ISIS molecular keys and calculating molecular proximity parameters such as the Tanimoto coefficient, cosine coefficient, etc. The Tanimoto coefficient is the ratio of shared substructures to the number of all substructures that appear in the reference chemical in the training set. The Tanimoto coefficient varies between 0 (total lack of similarity) to 1 (the query chemical has an identical constitution to the reference chemical). It is important to remember that the Tanimoto coefficient does not provide a unique measure of similarity - its meaning is based on how structural fragments are defined for the purposes of the comparison. Thus, two chemicals that are similar with a Tanimoto coefficient of 0.8 on the basis of one set of fragments may not be similar when compared by using a different set of fragments. Algorithms for calculating Tanimoto similarity coefficients are incorporated in several software products, including Ambit disclosure software, the Leadscope Inc., software, and the MDL Information Systems MDL-QSAR software which has several measures of similarity. Another possible approach to measure the domain of applicability and statistical confidence in predictions is the assessment of membership in a class statistics such as that used in MDL-QSAR nonparametric discriminant analysis.

121. Two different approaches may be adopted when multiple (Q)SAR models are being used for the prediction of the same endpoint or same toxic effect (consensus and battery QSAR models). In the first consensus approach if a query chemical falls within the intersection of the ADs of the different models, the confidence of the overall prediction may be obtained by averaging (or other transformation) of the individual predictions. In this situation the confidence in predictions which are the same for two or more models should be greater than the confidence associated with the prediction of a single model. However, it is expected that the common AD will be narrower for multiple models, thus restricting the number of potential chemicals that could be predicted. An example of the use of multiple models is provided by Tong *et al.* (2003), who used a decision forest (*i.e*. multiple comparable and heterogeneous decision trees).

122. In the second battery QSAR modelling approach, the battery of predictions from multiple models using different logic paradigms and algorithms are added (Votano *et al.,* 2004). This approach is used to expand our knowledge and give added insights into molecular properties significantly correlated with the same endpoint or same toxic effect (*e.g.*, chemicals having fragment structure alerts and specific E-state descripters). For example, a (Q)SAR model battery could be selected that utilizes molecular fragments or molecular descriptors which correlate with the same endpoint. Each (Q)SAR is chosen based upon validation experiments in which the (Q)SAR exhibited high specificity for the endpoint. In this situation the collective AD could theoretically be increased because the different (Q)SAR models could have

different and non-overlapping ADs for heterogeneous, non-congeneric test chemicals. The confidence and the handling of discordant predictions by the (Q)SAR models could be addressed by the investigator's application and need to achieve the highest possible sensitivity or specificity for the predictions. For example, the addition of results of two (Q)SARs having high specificity and low or moderate sensitivity would result is a high overall specificity and may also result in an increased sensitivity.

123. Recently, a stepwise approach for determining the model AD has been proposed by Dimitrov *et al.* (2005). It consists of four stages. The first stage identifies whether a chemical falls in the range of variation of physicochemical properties of the model. The second step defines the structural similarity between the query chemical and chemicals correctly predicted by the model. The third stage comprises a mechanistic check by assessing whether the chemical contains the specific reactive groups hypothesised to cause the effect. The fourth and final stage is a metabolic check, based on an assessment of the probability that the chemical undergoes metabolic activation. The four stages are applied in a sequential manner. The advantage of processing query chemicals through all four stages is the increased reliability of prediction for those chemicals that satisfy to all four conditions for inclusion in the AD. The cost of applying this rigorous, multiple AD approach is that the number of chemicals for which reliable predictions are eventually made is reduced compared to the use of a single AD method.

Comparing applicability domains with the spaces of regulatory inventories

124. Defining the AD of a model not only provides a means of increasing the confidence associated with predictions inside the domain, but also of assessing the applicability of the model to a given regulatory inventory of chemicals. A model that gives highly accurate predictions for narrow chemical classes that are not covered by the regulatory inventory of interest would be of questionable value. A number of investigations have addressed the need to screen and prioritise chemical inventories established under different legislations in OECD member countries (Cunningham and Rosenkranz, 2001; Klopman *et al*., 2003; Hong *et al*., 2002). Among the most commonly screened regulatory inventories are those of the High Production Volume Chemicals, Existing Substances, and inventories of pesticides and biocides. Less information is publicly available regarding the inventories of New Substances, mainly because of confidentiality considerations. In addition, these inventories are periodically updated with new chemicals, which implies the need for iterative development of (Q)SAR models (Schmieder *et al*., 2003, Tunkle *et al*., 2005) to expand their domains and adapt them to the regulatory domains of concern. An approach for comparing the AD with a regulatory domain is illustrated in a study (Netzeva *et al*., 2006) in which the AD of a QSAR for estrogenic potential is compared with the domain of the EINECS inventory (the list of Existing Substances in the EU). In this study, the physicochemical space of the EINECS inventory is characterised by using the descriptors in the QSAR model.

Concluding remarks

125. OECD Principle 3 should be considered in combination with the fourth OECD Principle on the need to characterise the statistical validity of a model, since an understanding of the AD can increase or decrease the confidence in a given (Q)SAR estimate. It should be noted, however, that the use of AD methods will never provide absolute certainty in the (Q)SAR estimates: a query chemical may appear to be within the defined AD, and yet the prediction could still be unreliable; conversely, the query chemical may appear to be outside the defined AD, and yet the prediction could be reliable.

126. The model user should therefore be aware that AD methods, like other (statistical) methods discussed in this Guidance Document, provide a useful means of supporting decisions based on the additional use of expert judgement, but they cannot in themselves make the decisions.

127. Numerous AD methods have been proposed based on the following considerations: structural features, physicochemical descriptor values, response values, mechanistic understanding, and metabolism. On this basis, it is useful to conceptualise the AD of a model as the combination of one or more elements relating to the structural, physicochemical, response, mechanistic and metabolic domains. While these different types of domains provide useful distinctions, they should not be assumed to be mutually exclusive. For example, the structural fragments present in a molecule will affect its physicochemical descriptors, its response value, and its mechanistic behaviour.

128. The different AD methods should not be seen as in competition with one another, since the combined use of multiple AD methods should give a higher assurance that query chemicals are predicted accurately by a (Q)SAR model. Inevitably, there is a trade-off between the breadth of applicability of a model and the reliability of the predictions within the domain: the broader the scope of the model, the lower the overall reliability of prediction. The user of a model therefore needs to strike an appropriate balance between the level of confidence in the predictions resulting from AD considerations and the number of reliable predictions that are determined.

129. Attempts to formalise and quantify the concept of the AD are relatively recent, which means that it is still a difficult concept to apply in regulatory practice. Thus, a considerable amount of research and development is still needed to further develop AD methods, as well as an understanding of the applicability of these methods. For example, the following research needs can be identified:

1. the development of confidence limits associated with the AD;

2. the development of AD methods for structural alerts and fragment-based QSAR methods;

3. the assessment of AD methods with a view to better understand their strengths, limitations and applicabilities;

4. the development of automated tools that facilitate the application of AD methods in an integrated manner with traditional statistical methods.

CHAPTER 5. GUIDANCE ON THE PRINCIPLE OF MEASURES OF GOODNESS-OF- FIT, ROBUSTNESS AND PREDICTIVITY

Summary of Chapter 5

130. This chapter provides guidance on how to interpret OECD Validation Principle 4 that "a (Q)SAR should be associated with appropriate measures of goodness-of–fit, robustness and predictivity" (Principle 4). This principle expresses the need to perform statistical validation to establish the performance of the model, which consists of internal model performance (goodness-of-fit and robustness) and external model performance (predictivity), taking into account any knowledge about the applicability domain of the model (Chapter 4). The chapter starts with a brief introduction to Principle 4 and statistical validation (paras 131-134), followed by an explanation of some key terms and concepts (paras 135-141). In paras 142-212, commonly used techniques for model development are then described and illustrated (multiple linear regression, partial least squares, classification modelling, neural network modeling) along with well-established statistical validation techniques for assessing goodness-of-fit, robustness and predictivity (cross-validation, bootstrapping, response randomisation test, training/test splitting, external validation). In the context of these different techniques, the statistics that are commonly used to describe model performance are explained.

Introduction

131. The need for information on the performance of (Q)SAR models is expressed by OECD Principle 4 which states that models should be associated with appropriate measures of goodness-of–fit and robustness (internal performance) and predictivity (external performance). The assessment of model performance is sometimes called statistical validation within the context of the assessment.

132. Statistical validation techniques are used during (Q)SAR development to find a suitable balance between the two extremes of overfitted and underfitted models. The optimal model complexity is a trade-off between models that are "too simple" and lacking in useful information and models that are too "complex" and provide modelling noise (Jouan-Rimbaud *et al*., 1996; Hawkins, 2004). Statistical validation techniques provide various "fitness" functions that can be used by the QSAR practitioner to compare the quality of different models, and to avoid models that are too simplistic or too complex.

133. Statistical validation techniques also provide a means of identifying "spurious" models based on chance correlations, *i.e*. situations in which an apparent relationship is established between the predictor and response variables, but which is not meaningful and not predictive (Topliss and Edwards, 1979; Wold and Dunn, 1983; Clark and Cramer, 1993) .

134. The statistical validation techniques described in this chapter should be considered in combination with any knowledge about the applicability domain (AD) of the model, since the choice of chemicals during model development and validation affects the assessment of performance. In particular, chemicals that are outside the AD during model development may unduly influence the regression parameters of the model, thereby affecting its robustness. Chemicals that are outside the AD during model validation are unlikely to be predicted with the desired level of reliability.

Basic Terms and Concepts

135. This section provides an explanation of some key terms and concepts that are needed to understand the remainder of the chapter. These concepts are also explained in the glossary of QSAR terminology attached to this document.

136. One approach to developing (Q)SAR models begins with the compilation of available endpoint data sets for a variety of representative chemicals likely to be encountered in regulatory programs. If endpoint data are available for a sufficient number of chemicals, the data set is often divided into a *training set*, used to derive the model through the application of a statistical method, and a *test set*, containing chemicals not used in the derivation of the model but used to evaluate the model. The variables in the model, referred to as *predictors,* are chosen to optimise model complexity. In many (Q)SAR models, the predictors are (molecular) descriptors which can be chosen to test hypotheses regarding mechanisms or to explore date sets for models.

137. The model derived from the training set is used to predict of the response data in both the training and the test sets. The *accuracy* of prediction for a given chemical is the closeness of an estimate/prediction to a reference value. Models with greater proportions of accurate predictions are the more *reliable* models.

138. Predictions for chemicals in the training set are used to assess the *goodness-of-fit* of the model, which is a measure of how well the model accounts for the variance of the response in the training set. The generation of predictions within the range of predictor values in the training set is called *interpolation*, whereas *extrapolation* is the generation of a prediction outside the range of values of the predictor in the sample used to generate the model. The more removed the predicted value from the range of values used to fit the model, the more unreliable the prediction becomes, because it is not certain whether the model continues to hold.

139. The *robustness* of model refers to the stability of its parameters (predictor coefficients) and consequently the stability of its predictions when a perturbation (deletion of one or more chemicals) is applied to the training set, and the model is regenerated from the "perturbed" training set.

140. Predictions for chemicals in the test set are used to assess the *predictive ability* of the model, which is a measure of how well the model can predict of new data, which not used in model development. In this document, predictive ability is used synonymously with *predictive capacity*, *predictive power* and *predictivity*.

141. The data-driven approach described above is predominant when the available data sets are large enough and representative of the chemicals being regulated, *i.e*. regulatory domain. The data-driven approach may not always group chemicals according to important mechanisms, which causes some groups in the data set to be outliers in the data-driven (Q)SAR models. An alternative approach to developing (Q)SAR-based predictive capacity begins with the regulatory domain, itself, and groups the chemicals according to expected consistent trends for important endpoints with each group. A variety of techniques are employed to extrapolate measured endpoint data to the untested chemicals in the group. This approach, similar to many of the knowledge-based rules in structural alerts and SAR, is a knowledge-driven approach to (Q)SAR modelling in which the statistical demand for data is replaced with expert knowledge of chemistry and toxicology for grouping chemicals. The determination of predictive capacity is determined by periodic retrospective evaluations of the predictions across all groups of chemicals in the regulatory domain.

Recommendations for Practitioners

142. This section offers guidance on how the robustness and predictivity of (Q)SAR models can be evaluated for a variety of the more common algorithms.

Multiple Linear Regression (MLR)

143. Multiple linear regression (MLR) is the traditional statistical approach for deriving QSAR models. It relates the dependent variable **y** (biological activity) to a number of independent (predictor) variables $\mathbf{x}_i$ (molecular descriptors) by using linear equations (Eq. 1, Table 5.1).

144. **Estimating the regression coefficients**. Regression coefficients $\mathbf{b}_j$ in MLR model can be estimated using the least squares procedure by minimizing the sum of the squared residuals. The aim of this procedure is to give the smallest possible sum of squared differences between the true dependent variable values and the values calculated by the regression model.

145. **Assessing the relative importance of descriptors**. If the variables are standardized to have mean of zero and standard deviation of one, then the regression coefficients in the model are called *beta* coefficients. The advantage of *beta* coefficients (as compared to regression coefficients that are not standardised) is that the magnitude of these *beta* coefficients allows the comparison of the relative contribution of each independent variable in the prediction of the dependent variable. Thus, independent variables with higher absolute value of their *beta* coefficients explain greater part from the variance of the dependent variable.

146. **Assessing goodness-of-fit**. To assess goodness-of-fit, the coefficient of multiple determination (R^2) is used (Eq. 2, Table 5.1). R^2 estimates the proportion of the variation of **y** that is explained by the regression (Massart, 1997a). If there is no linear relationship between the dependent and the independent variables $R^2 = 0$; if there is a perfect fit $R^2 = 1$. R^2 value higher than 0.5 means that the explained variance by the model is higher than the unexplained one. The end-user(s) of a QSAR model should decide what value of R^2 is sufficient for the specific application of the model. One author has recommended that $R \geq 0.9$ for in vitro data and $R \geq 0.8$ for in vivo data can be regarded as good (Kubinyi, 1993).

147. **Avoiding overfitting**. The value of R^2 can generally be increased by adding additional predictor variables to the model, even if the added variable does not contribute to reduce the unexplained variance of the dependent variable. It follows that R^2 should be used with caution. This could be avoided by using another statistical parameter – the so-called adjusted R^2 (R^2_{adj}) (Eq. 3, Table 5.1). R^2_{adj} is interpreted similarly to the R^2 value except it takes into consideration the number of degrees of freedom. It is adjusted by dividing the residual sum of squares and total sum of squares by their respective degrees of freedom. The value of R^2_{adj} decreases if an added variable to the equation does not reduce the unexplained variance.

148. From the calculated and observed dependent variable values the standard error of estimate *s* could be obtained (Eq. 4, Table 5.1). The standard error of estimate measures the dispersion of the observed values about the regression line. The smaller the value of *s* means higher reliability of the prediction. However it is not recommended to have standard error of estimate smaller than the experimental error of the biological data, because it is an indication for an overfitted model (Wold *et al*., 1984).

149. The statistical significance of the regression model can be assessed by means of *F*-value (Eq. 5, Table 5.1). The *F*-value is the ratio between explained and unexplained variance for a given number of degrees of freedom. The higher the *F*-value the greater the probability is that the equation is significant. The regression equation is considered to be statistically significant if the observed *F*-value is greater than a tabulated value for the chosen level of significance (typically, the 95% level) and the corresponding

degrees of freedom of *F*. The degrees of freedom of *F*-value are equal to *p* and *n-p*-1. Significance of the equation at the 95% level means that there is only a 5% probability that the dependence found is obtained due to chance correlations between the variables.

150. The statistical significance of the regression coefficients can be obtained from a *t*-test (Eq. 6, Table 5.1). It is used to test the hypothesis that the regression coefficient is zero. If the hypothesis is true, than the predictor variable does not contribute to explain the dependent variable. Higher *t*-values of a regression coefficient correspond to a greater statistical significance. The statistical significance of a regression coefficient using its *t*-value is determined again from tables for a given level of significance (similar to the use of *F*-value). The degrees of freedom of *t* are equal to *n-p*-1 (corresponding to the degrees of freedom of the residual mean square). Statistical significance at the 95% level means there is only a 5% probability that the regression coefficient of a given variable is not significantly different from zero. The *t*-values are used to calculate the confidence intervals for the true regression parameters. These confidence intervals can also be used to check the significance of the corresponding regression coefficients. In practice the confidential intervals should be smaller than the absolute values of the regression coefficients in order to have statistically significant independent variables (Wold *et al*., 1984).

Partial Least Squares regression (PLS)

151. Partial Least Squares (PLS), introduced by Wold *et al*. (1984, 1993), is a MLR method that allows relationships to be sought between an **X**-block of *p* predictors and a single **y** response (PLS1) or a **Y**-block of *r* responses (PLS2). Thus several activity variables, **Y**, *i.e*. profiles of activity, can be modelled simultaneously. An advantage of PLS is that it tolerates a certain amount of missing data. For instance, in the case of data set containing 20 compounds, 10-20% missing data can be tolerated (Wold, 1995).

152. **Information provided by PLS**. The purpose of PLS is to find a small number of relevant factors (A) that are predictive of **Y** and utilize **X** efficiently (Massart *et al*., 1997b). The PLS model is expressed by a matrix of scores (**T**) that summarizes the **X** variables, a matrix of scores (**U**) that summarizes the **Y** variables, a matrix of weights (**W**) expressing the correlation between **X** and **U**(**Y**), a matrix of weights (**C**) expressing the correlation between **Y** and **T**(**X**), and a matrix of residuals (the part of data that are not explained by the model). For the interpretation of the PLS model a number of informative parameters can be used. The scores **t** and **u** contain information about the compounds and their similarities/dissimilarities with respect to the given problem. The weights **w** and **c** provide information about how the variables can be combined to form a quantitative relation between **X** and **Y**. Hence they are essential for understanding which **X** variables are important and which **X** variables provide the same information. The residuals are of diagnostic interest – large residuals of **Y** indicate that the model is poor and the outliers should be identified (Wold, 1995). PLS regression coefficients can be obtained after re-expression of the PLS solution as a regression model. When **X** values are scaled and centered and **Y** values are scaled, the resulting coefficients are useful for interpreting the influence of the variables **X** on **Y** (Eriksson *et al*., 2001; Netzeva *et al*., 2003).

153. **Assessing Goodness-of-fit.** The quantitative measure of the goodness of fit is given by the parameter R^2 (= the explained variation) analogous to MLR. PLS model is characterized by the following R^2 parameters:

- R^2(Y) – cumulative sum of squares of all dependent values (Y) explained by all extracted components
- R^2(X) – cumulative sum of squares of all descriptor values (X) explained by all extracted components
- $R^2(Y)_{adj}$, $R^2(X)_{adj}$ – cumulative R^2(Y) and R^2(X) respectively adjusted for the degrees of freedom

154. **Avoiding overfitting**. Depending on the number of components, near perfect correlations are often obtained in PLS analysis, due to the usually large number of included **X** variables. Therefore, the high R^2(Y) is not a sufficient criterion for the validity of a PLS model. A cross-validation procedure must be used and Q^2(Y) parameter must be calculated to select the model having the highest predictive ability (Kubinyi, 1993). In contrast to R^2(Y), Q^2(Y) does not increase after a certain degree of model complexity. Hence, there is a zone, where there is a balance between predictive power and reasonable fit (Massart *et al.*, 1997b). According to the proposed reference criteria the difference between R^2(Y) and Q^2(Y) should not exceed 0.3. A substantially larger difference is indication for an overfitted model, presence of irrelevant **X**-values or outliers in the data (Eriksson *et al*., 2003).

155. **Identification of outliers**. As a measure of the statistical fit of the PLS model also the residual standard deviation (RSD) can be used, which corresponds to the standard deviation in the MLR. It should be similar in size to the known or expected noise in the system under investigation. The RSD can be calculated for the responses and predictor variables. The RSD of an **X** variable is indication for its relevance to the PLS model. The RSD of a **Y** variable is a measure of how well this response is explained by the PLS model. The RSD of an observation in the X or **Y** space is proportional to the observation distance in the hyper plane of the PLS-model in the corresponding space (DModX and DModY). The last ones give information about the outliers in **X**- and **Y**-data (Massart *et al*., 1997b; Netzeva *et al*., 2003).

Classification Models (CMs)

156. Chemicals are sometimes classified into two (*e.g*. active/inactive) or more pre-defined categories, for scientific or regulatory purposes. For scientific purposes, the biological variability of certain endpoints is sometimes too large to enable reasonable quantitative predictions, so that the data is converted into one or more categories of toxic effect. Otherwise, in regulatory toxicology, binary classification systems are commonly used to provide a convenient means of labelling chemicals, according to their hazard.

157. Classification-based QSARs, also referred as classification models (CMs), can be developed using a variety of statistical methods. Among the methods appropriate for the development of linear CMs, multivariate discriminant analysis (MDA), logistic regression (LR), and decision or classification trees (CT), among others, have been extensively described in the literature (Worth and Cronin, 2003). Also, rule-based models expressed in symbolic "if… then" decision rules, can be derived from the CMs. For the models associated with non-linear boundaries, embedded cluster modelling (ECM) (Worth and Cronin, 2000), neural networks (NN), and k-nearest neighbour (k-NN) classifiers can be used.

158. **Assessing Goodness-of-fit.** The goodness-of-fit of a CM can be assessed in terms of its Cooper statistics, which were introduced in the late seventies to describe the validity of carcinogen screening tests (Cooper *et al*., 1979). Cooper statistics, based on a Bayesian approach (Feinstein, 1975; Sullivan, 2003) has been extensively applied to assess the results of classification (Q)SAR models (Eriksson *et al*., 2003; McDowell and Jaworska, 2002). Bayesian-based methods can also be used to combine results from different cases, so that judgments are rarely based only on the results of a single study but they rather synthesize evidence from multiple sources. These methods can be developed in an iterative manner, so that they allow successive updating of battery interpretation.

159. In a CM, the results of the classification can be arranged in the so-called *confusion* or *contingency matrix* (Frank and Friedman, 1989), where the rows represent the reference classes (*Ag*), while the columns represent the predicted classes assigned by the CM (*Ag'*). Table 5.2 illustrates the general form of a contingency matrix for the general case of *G* classes.

160. **Interpreting the contingency matrix**. The main diagonal ($c_{gg'}$) represents the cases where the true class coincides with the assigned class, that is, the number of objects correctly classified in each class,

while the non-diagonal cells represent the misclassifications. Overpredictions are to the right and above the diagonal, whereas underpredictions are to the left and below the diagonal. The right-hand column reports the number of objects belonging to each class (n_g), whereas the bottom row reports the total number of objects assigned to each class according to the CM ($n_{g'}$).

161. **Setting the importance of misclassifications**. Depending on the intended use of the CM, some classification errors may be considered "worse" than others. In order to quantify such error, the *loss matrix* (**L**), which the same structure as the contingency matrix, can be used (Table 5.3). It can be considered as a matrix of weights for the different types of classification errors, where the non-diagonal elements quantify the type of error in the classification.

162. According to this matrix of weights, the classification errors that for example confuse class A_1 with class A_3 and class A_G are more significant (loss value of 2) than the ones than confuse class A_1 with class A_2 (loss value of 1). The main diagonal corresponds to the correct classification, so that the loss value is set to zero. This matrix can be defined in an arbitrary way, according to the situation. If it is not explicit all the errors can be assigned to have the same weight of 1.

163. The most commonly used goodness-of-fit parameters for a CM are defined in Table 5.4. When evaluating the results of a CM, the reference situation is generally taken to be the one in which all of the objects are assigned to the class that is most represented. This reference condition corresponds to the absence of a model, and is therefore called the *No-Model*. Goodness-of-fit values close to the ones of the *No-Model* condition give evidence of a poor result of the classification method. The *No-Model* value is unique and independent from the classification method adopted. Other statistics have been proposed, like kappa (k) statistic (Kraemer, 1982).

164. In the particular case of a two-group CM, which evaluates the presence or absence of activity, Cooper statistics can be calculated from a 2x2 contingency table (see Table 5.5).

165. The statistics in Table 5.6 collectively express the performance of a CM, provided they measure its ability to detect known active compounds (sensitivity), non-active compounds (specificity), and all chemicals in general (concordance or accuracy). The false positive and false negative rates can be calculated from the complement of specificity and sensitivity, respectively. The positive and negative "predictivities" focus on the effects of individual chemicals, since they act as conditional probabilities. Thus, the positive "predictivity" is the probability that a chemical classified as active is really active, while the negative "predictivity" gives the probability that a classified non-active chemical is really non-active.

166. The sensitivity is the ability to detect known active compounds, that is to say the percent of the chemicals tested positive that are correctly identified as positive by the QSAR model. Therefore, a high value of sensitivity is associated with a high true positive rate. In addition, a high value of sensitivity is also associated with a low false negative rate. The specificity is the ability to detect known inactive compounds, that is to say the percent of the chemicals tested negative that are correctly identified as negative by the QSAR model. A high specificity is associated with a high true negative rate and a low false positive rate. Given a fixed sensitivity and specificity, the positive and negative predictivities vary according to the prevalence or proportion of active chemicals in a population, *i.e.* $(a+b)/N$. Furthermore, the accuracy is influenced by the performance of the most numerous class. Therefore, CMs should not be judged according to these statistics alone.

167. For the assessment of the predictive performance of two-group CMs, the maximal classification performance achievable should be assessed on the basis of the quality of the predictor and response data and taking also into account the purpose of the model. Thus, for stand-alone classification models, the

Cooper statistics should be significantly greater than 50%, whereas for a CM that identifies active or inactive chemicals in a battery of models, a lower performance could still be useful.

168. The classification ability of a CM depends on the particular data set of chemicals used. It is therefore useful to report some measure of the variability associated with the classifications. This indicates whether the classification performance of the CM would vary significantly if it had been assessed with a different set of chemicals. To estimate the confidence intervals (CI) for the Cooper statistics, the bootstrap re-sampling technique can be used (Wehrens *et al*., 2000; Worth and Cronin, 2001).

169. To compare the performances of a number of classification models, the Receiver Operating Characteristic (ROC) curve can be used. ROC curves are so-named because they were first used for the detection of radio signals in the presence of noise in the 1940s (Lusted, 1971). In the ROC graph, the X-axis is 1-specificity (false positive rate) and the Y-axis is the sensitivity (true positive rate). The best possible CM would yield a point located in the upper left corner of the ROC space, *i.e*. high true positive rate and low false positive rate. A CM with no discriminating power would give a straight line at an angle of 45 degrees from the horizontal, *i.e*. equal rates of true and false positives (Hanley, 1989; Provost and Fawcett, 2001). An index of the goodness of the CM is the area under the curve: a perfect CM has area of 1.0, whilst a non-discriminating test (one which falls on the diagonal) has an area of 0.5.

170. In the case of a CM based on continuous predictors, *i.e*. predictors expressed by continuous values, the ROC curve allows us to explore the relationship between the sensitivity and specificity resulting from different thresholds (cut points), thus allowing an optimal threshold to be determined (Figure 5.1). The threshold is an arbitrary cut-off value which determines when the prediction is considered as positive or negative. Ideally, both sensitivity and specificity would be equal to one, but changing the threshold to increase one statistic usually results in a decrease in the other. In Figure 5.1, points greater than (to the right of) the threshold are classified positive, whereas points less than (to the left of) the threshold are classified negative. The dark blue curve represents the distribution of true negatives, and the dark red curve represents the distribution of true positives. If the threshold (green line) is increased (movement from left to right), the false positive rate (light blue area) decreases. However, as the false positive rate decreases, the true positive rate (red area) also decreases; this corresponds to points in the bottom left of the ROC curve. Otherwise, if the threshold is decreased (movement of green line from right to left), the proportion of true positives (Y axis) increases, rather dramatically initially, and then more gradually; this corresponds to points in the top right of the ROC curve.

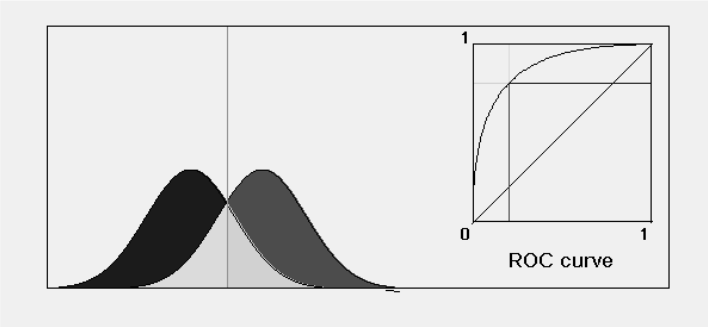

Figure 5.1. ROC curve for a model that produces a continuous output, as a function of the classification threshold marked with a green line

Image taken from http://www.anaesthetist.com/mnm/stats/roc/ (last accessed 6 February 2007). The coordinates are indicative of the performance of the models corresponding to: (0,0) high threshold, (1,1) low threshold, (0,1) perfect classification, y=x, model with no discriminatory power.

171. **Setting the importance of misclassifications**. The assessment of classification accuracy often assumes equal costs of false positives and false negative errors. However, in real applications, the minimisation of costs should be considered alongside the maximization of accuracy. The problems of unequal error costs and uneven class distributions are related, so that high-cost cases can be compensated by modifying their prevalence in the set (Breiman *et al*., 1984).

172. The robustness of a CM can be evaluated by the total number of misclassifications, estimated with the *leave-one-out* method (Hand, 1981). Alternatively, the above-mentioned set of optimal loss factors (*i.e.* weights for different kinds of misclassifications that are minimised in the process of fitting a model) can be defined to reflect that some classification errors are more detrimental than others. The loss function represents a selected measure of the discrepancy between the observed data and data predicted by the fitted function. It can be empirically estimated and employed in a minimum risk decision rule rather than a minimum error probability rule. Also, by combining different predictions, the resulting models are more robust and accurate than single model solutions.

Artificial Neural Networks (ANNs)

173. An Artificial Neural Network (ANN) is a mathematical model that "learns" from data in a manner that emulates the learning pattern in the human brain. The calculations in a neural network model occurs as a result of the "activation" of a series of neurons, which are situated in different layers, from the input layer through one or more hidden layers to the output layer. The neural network learns by repeatedly passing through the data and adjusting its connection weights to minimise the error.

174. There are two main groups of ANN, which differ in architecture and in learning strategy: (i) unsupervised and supervised self organizing maps; and (ii) supervised back-propagation ANN (Lek and Guegan, 1999). The terms "unsupervised" and "supervised" indicate whether only descriptors (input variables), or both descriptors and biological activities (output variables), participate in the training of ANN.

175. ANNs are especially suitable for modelling non-linear relationships and trends and have been used to tackle a variety of mathematical problems, including data exploration, pattern recognition, the modelling of continuous and categorized responses, and the modelling of multiple responses (Anzali *et al*., 1998; Zupan and Gasteiger, 1999), the classification of objects toxicological classes or modes of toxic action (Spycher *et al*., 2005), selection of relevant descriptors, and division of the original data set into clusters (Vracko, 2005).

176. **Assessing Goodness-of-fit.** Several tests for assessing the goodness-of-fit of NN models (based on the training set) are recommended. In the recall ability test (Guha and Jurs, 2005; Mazzatorta *et al*., 2003; Vracko and Gasteiger, 2002; Devillers and Domine, 1999), the activity values are calculated for the objects of training set, to provide an indication of how well the model recognises the objects of training set. The test results are usually reported as the standard deviation and the parameters of the regression line between reference values and predicted values. Since the recall ability test is a test for goodness-of-fit only, it is recommended additional tests are also used, such as leave-one-out, leave-many-out, Y-Scrambling, and assessment with independent test set.

177. **Measures of robustness.** The aim of validation techniques is thus to find a model which represents the best trade-off between the model simplicity and its variability, in order to minimize the Mean Squared Error (MSE) (Table 5.7), minimising the bias as well as the unexplained variance.

178. A necessary condition for the validity of a regression model is that the multiple correlation coefficient R^2 is as close as possible to one and the standard error of the estimate s small. However, this

condition (*fitting ability*), which measures how well the model is able to mathematically reproduce the end point data of the training set, is an insufficient condition for model validity. In fact, models that give a high fit (smaller s and larger R^2) tend to have a large number of predictor variables (Eriksson *et al*., 2003). These parameters are measures of the quality of the fit between predicted and experimental values, and do not express the ability of the model to make reliable predictions on new data.

179. It is well known that increasing the model complexity always increases the multiple correlation coefficient (R^2), *i.e.* the explained variance in fitting, but if model complexity is not well supervised then the predictive power of the model, *i.e.* the explained variance in prediction (Q^2) decreases. The differing trends of R^2 and Q^2 with an increasing number of predictor variables is illustrated in Figure 5.2.

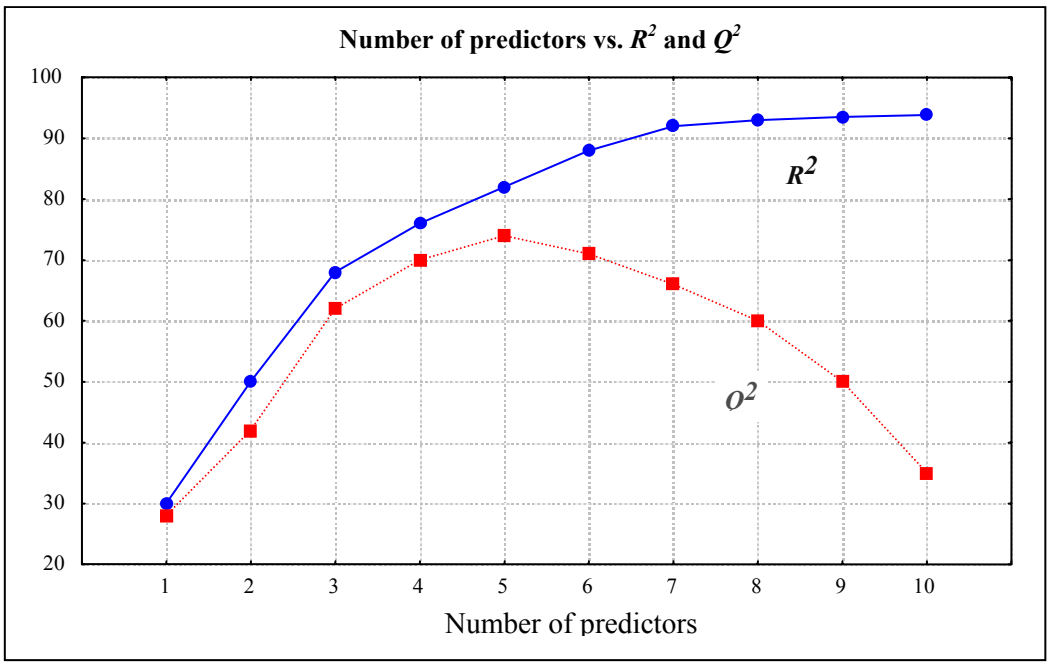

Figure 5.2. Comparison of the explained variance in fitting with the explained variance in prediction

180. In Figure 5.2, it can be seen that increasing the number of predictors improves the explained variance in fitting (R^2). On the other hand, the explained variance in prediction (Q^2) only up to 5 predictors (which represents the maximum predictive power in this case) but adding further statistically insignificant predictors decreases the model performance in prediction.

181. The first condition for model validity deals with the ratio of the number of objects (*i.e.* chemicals) over the number of selected variables. This is called the Topliss ratio. As a rule-of-thumb, it is recommended that the Topliss ratio should have a value of at least 5.

182. The quality of multivariate regression models is usually evaluated by different fitness functions (*e.g.* adjusted R^2, Q^2) (Table 5.7) able to find the optimal model complexity and useful to compare the quality of different QSAR models.

183. For this reason, the structure of a QSAR model (number of predictors, number of PCs, number of classes) should always be inspected by validation techniques, able to detect overfitting due to variable multicollinearity, noise, sample specificity, and unjustified model complexity.

184. Model validation can be performed by internal validation techniques and external validation techniques. As illustrated in Figure 5.3, in case of internal validation a number of modified data sets are

created by deleting, in each case, one or a small group of objects and each reduced data set is used to estimate the predictive capability of the final model built by using the whole data set. This means that the model predictivity is estimated by compounds (the test set) which took part in the model development, thus the information of these compounds is included in the final model. On the other hand a more demanding evaluation is the one provided by an external validation where the model predictivity is estimated by new experimentally tested compounds (external test set) which did not take part in the model development.

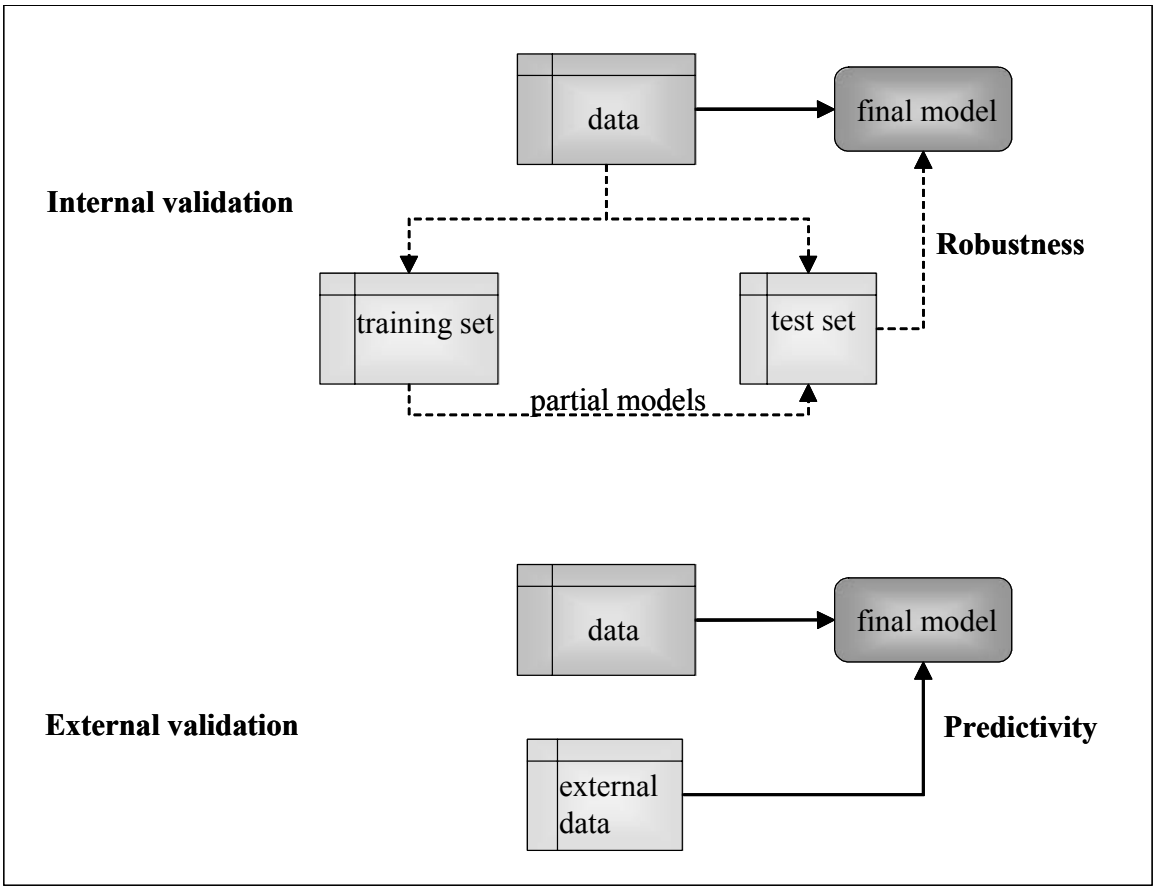

Figure 5.3. Internal and external validation.

185. A number of internal validation techniques can be used to simulate the predictive ability of a model (Diaconis and Efron, 1983; Cramer *et al.*, 1988). The most popular validation ones are listed below:

- Cross validation (leave-one-out (LOO) and leave-many-out (LMO)).
- Bootstrapping
- Y-scrambling or response permutation testing
- Training/test set splitting

186. *Cross validation* is the most common validation technique where a number of modified data sets are created by deleting, in each case, one or a small group of compounds from the data in such a way that each object is removed away once and only once. From the original data set, a reduced data set (training set) is used to develop a partial model, while the remaining data (test set) are used to evaluate the model predictivity (Efron, 1983; Osten, 1988). For each reduced data set, the model is calculated and responses for the deleted compounds are predicted from the model. The squared differences between the true response and the predicted response for each compound left out are added to the predictive residual sum of

squares (*PRESS*). From the final predictive residual sum of squares, the Q^2 (or R^2_{CV}) and *SDEP* (*standard deviation error of prediction*) values are calculated (Cruciani *et al*., 1992) (Table 5.7).

187. The simplest cross validation procedure is the *leave-one-out* (LOO) technique, where each compound is removed, one at a time. In this case, given *n* compounds, *n* reduced models are calculated, each of these models is developed with the remaining *n*-1 compounds and used to predict the response of the deleted compound. The model predictive power is then calculated as the sum of squared differences between the observed and estimated response. This technique is particularly important as this deletion scheme is unique and the predictive ability of the different models can be compared accurately. However, the predictive ability obtained is often too optimistic, particularly with larger datasets compounds, because the perturbation in the dataset is small and often insignificant when only one compound is omitted.

188. To obtain more realistic estimates of the predictive ability, it is often necessary to remove more than one compound at each step. In the *leave-many-out* (LMO) cross-validation procedure, the data set is divided into a number of blocks (cancellation groups) defined by the user. At each step, all the compounds belonging to a block are left out from the derivation of the model. The cancellation groups *G* range from 2 to *n*. For example, given 120 compounds ($n = 120$), for 2, 3, 5, 10 cancellation groups *G*, at each time *m* (= *n/G*) objects are left in the test sets, *i.e*. 60, 40, 24, and 12 compounds, respectively. Rules for selecting the group of compounds for the test set at each step must be adopted in order to leave out each compound only one time. The LOO method is equivalent to a LMO method with $G = n$, *i.e.* with a number of cancellation groups equal to the number of compounds. By introducing a larger perturbation in the data set, the predictive ability estimated by LMO is more realistic than the one by LOO.

189. *Bootstrap resampling* is another technique to perform internal validation (Wehrens *et al*., 2000). The basic premise of bootstrap resampling is that the data set should be representative of the population from which it was drawn. Since there is only one data set, bootstrapping simulates what would happen if the samples were selected randomly. In a typical bootstrap validation, *K* groups of size *n* are generated by a repeated random selection of *n* compounds from the original data set. Some of these compounds can be included in the same random sample several times, while other compounds will never be selected. In this validation technique, the original size of the data set (*n*) is preserved by the selection of *n* compounds with repetition. In this way, the training set usually consists of repeated compounds and the test set of the compounds left out (Efron and Tibshirani, 1993). The model is derived by using the training set and responses are predicted by using the test set. All the squared differences between the true response and the predicted response of the compounds of the test set are expressed in the *PRESS* statistic. This procedure of building training sets and test sets is repeated thousands of times. As with the LMO technique, a high average Q^2 in bootstrap validation is indicative of model robustness and what is sometimes referred to as "internal predictivity".

190. *Y-scrambling* or response permutation testing is another widely used technique to check the robustness of a QSAR model, and to identify models based on chance correlation, *i.e*. models where the independent variables are randomly correlated to the response variables. The test is performed by calculating the quality of the model (usually R^2 or, better, Q^2) randomly modifying the sequence of the response vector $\boldsymbol{y}$, *i.e*. by assigning to each compound a response randomly selected from the true set of responses (Figure 5.4) (Lindgren *et al*., 1996). If the original model has no chance correlation, there is a significant difference in the quality of the original model and that associated with a model obtained with random responses (Figure 5.5). The procedure is repeated several hundred of times.

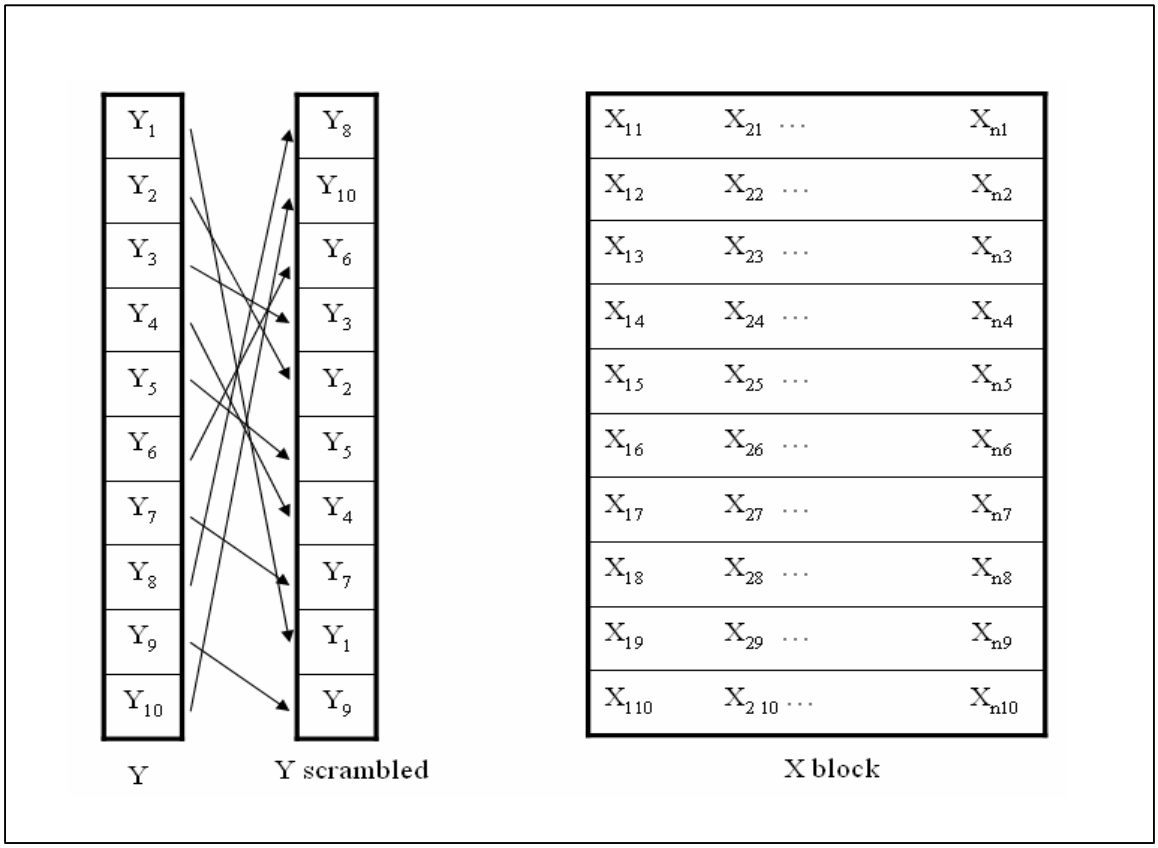

Figure 5.4. Y-scrambling by random permutations of activity values (Y);

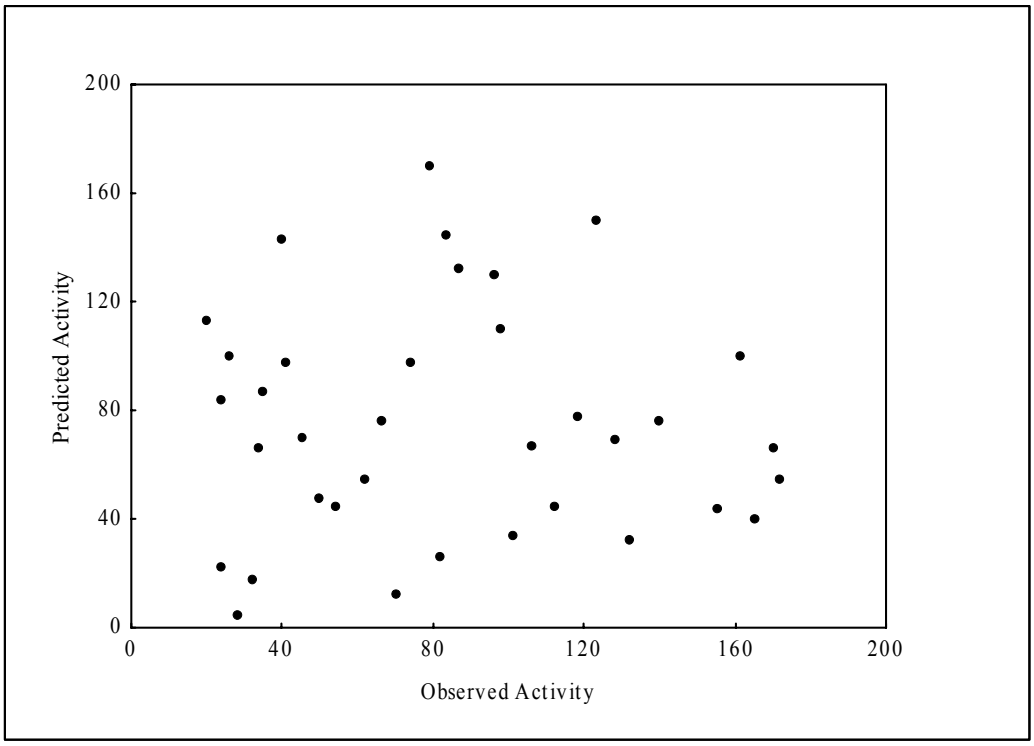

Figure 5.5. Plot of predicted versus observed activity values (Y) – random scatter plot indicates that the model is not probably due to chance correlations.

191. Models based on chance correlation can be detected by using the QUIK rule. Proposed in 1998 (Todeschini *et al*., 1999), the QUIK rule is a simple criterion that allows the rejection of models with high predictor collinearity, which can lead to chance correlation (Todeschini *et al*., 2004). The QUIK rule is based on the *K* multivariate correlation index (Table 5.7) that measures the total correlation of a set of variables. The rule is derived from the evident assumption that the total correlation in the set given by the model predictors ***X*** plus the response ***Y*** (K_{XY}) should always be greater than that measured only in the set of predictors (K_X). Therefore, according to the QUIK rule only models with the K_{XY} correlation among the [X + Y] variables greater than the K_X correlation among the [X] variables can be accepted. The QUIK rule has been demonstrated to be very effective in avoiding models with multi-collinearity without prediction power.

192. An example of the application of the QUIK rule in QSAR studies is provided (Todeschini *et al*., 1999) by a series of 11 3-quinuclidinyl benzylates represented by three physicochemical descriptors: Norrington's lipophilic substituent constant π_N (x_1), its squared values π^2_N (x_2), and the Taft steric constant E_s (x_3). The *y* response was the apparent equilibrium constant K_{app}. This data set has been extensively discussed by Stone and Jonathan (1993) and by Mager (1995), who concluded that the model has multicollinearity without prediction power. The regression model obtained by Ordinary Least Squares regression (OLS) was:

$$y = -8.40 + 8.35\ x_1 - 1.70\ x_2 + 1.43\ x_3 \qquad \text{(Eq 1)}$$

with the following statistics:

$$R^2 = 91.8 \quad Q^2_{LOO} = 81.5 \quad Q^2_{LMO} = 67.0$$

where R^2, Q^2_{LOO} and Q^2_{LMO} are the explained variances in fitting, by leave-one-out cross validation and by leave-many-out cross validation (two objects left out at each step), respectively. The large decrease in the predictive performance of the model was already suspect. The same conclusions were reached applying the *QUIK* rule. In fact, for the proposed model, the *K* values were:

$$K_{xy} = 47.91 < K_x = 54.87$$

According to the QUIK rule, the model would be rejected, the X correlation being greater than the X+Y-correlation.

193. Another method to check chance correlation is to add a percentage of artificial noisy variables to the set of available variables. This approach allows the detection of optimal model size, *i.e*. the size for which no noisy variable is present in models of this size and an example of its capability tested on a spectral matrix was extensively illustrated in Jouan-Rimbaud *et al*. (1996). In fact, when simulated noisy variables start to appear in the evolving model population it means that the allowed maximum model size can no longer be increased since optimal complexity has been reached. However, this approach does not account for the likely correlation between a generated noisy variable and the Y response. In fact, there is a high probability that on generating a number of noisy predictors some will be significantly correlated with the y response. While chance correlation is considered explicitly in the Y scrambling procedure, by response randomisation, a noisy predictor could play an important role in modelling in the latter approach, contributing in the same way as a true predictor with a small, but significant, correlation with response. For this reason, noisy variables should only be used if a check on their correlation with the y response is performed first, excluding all the noisy variables with correlation greater than a fixed threshold value (Todeschini *et al*., 2004). An optimal value of this threshold can be chosen only if the experimental error of the response is known *a priori*.

194. The training/test set splitting is a validation technique based on the splitting of the data set into a training set and a test set. The model is derived from the training set and the predictive power is estimated by applying the model to the test set. The splitting is performed by randomly selecting the objects belonging to the two sets. As the results are strongly dependent on the splitting of the data, this technique is better used by repeating the splitting several hundred of times and averaging the predictive capabilities, *i.e*. using the repeated test set technique (Boggia *et al*., 1997).

Evaluating Predicitive Capacity for Individual (Q)SAR Models

195. One of the most important characteristics of a (Q)SAR model is its predictive power, *i.e*. the ability of a model to predict accurately the (biological) activity of compounds that were not used for model development. While the internal validation techniques described above can be used to establish model robustness, they do not directly assess model predictivity.

196. In principle, external validation is the only way to "determine" the true predictive power of a QSAR model. This type of assessment requires the use of an external test set, *i.e*. compounds not used for the model development. It is generally considered the most rigorous validation procedure, because the compounds in the external test set do not affect the model development. In fact, the test set is often constituted of new experimentally tested compounds used to check the predictive power of the model.

197. External validation should be regarded as a supplementary procedure to internal validation, rather than as a (superior) alternative. This is because a model that is externally predictive should also be robust, although a robust model is not necessarily predictive (of independent data). Indeed, a high value of the leave-one-out cross-validated correlation coefficient, Q^2, can be regarded as a necessary, but insufficient, condition for a model to have a high predictive power (Golbraikh and Tropsha, 2002a; Gramatica *et al*., 2004; Gramatica and Papa, 2005; Papa, *et al*., 2005).

198. The predictivity of a regression model is estimated by comparing the predicted and observed values of a *sufficiently large and representative* external test set of compounds that were not used in the model development. By using the selected model, the values of the response for the test objects are calculated and the quality of these predictions is defined in terms of external explained variance Q^2ext (Table 5.7). Unlike the cross-validated correlation coefficient, Q^2, in the external explained variance Q^2ext the sum of the predictive residual sum of squares on the numerator runs over the external test chemicals and the reference total sum of squares on the denominator is calculated comparing the predicted response of the external test chemicals with the average response of the training set.

199. Analogously, the predictivity of a classification model is estimated by comparing the predicted and observed classes of a *sufficiently large and representative* test set of compounds that were not used in the model development. The parameters described in Table 5.4, but derived by using the external test set, are used to quantify the CM predictivity.

200. In practice, for reasons of cost, time and animal welfare, it is often difficult or impossible to obtain new experimentally tested compounds to check model predictivity, and for this reason a common practice is to split the available dataset into training set, used to develop the (Q)SAR model and an external test set, containing compounds not present in the training set and used to assess the predictive capability of a (Q)SAR model. This technique can be used reliably only if the splitting is performed by partitioning the compounds according to a pre-defined and suitable criterion, such as a criterion based on experimental design or cluster analysis.

201. When performing statistically designed external validation, the goal is to ensure that: a) the training and test sets separately span the whole descriptor space occupied by the entire data set; and b) the

structural domains in the two sets are not too dissimilar. It is important that the training set contains compounds that are informative and good representatives of many other similar compounds. Thus, the following criteria were recently proposed for training and test selection (Golbraikh and Tropsha, 2002b; Gramatica *et al*., 2004; Gramatica and Papa, 2005; Papa *et al*., 2005): a) representative points of the test set must be close to points in of the training set; b) representative points of the training set must be close to points in the test set; and c) the training set must be diverse. These criteria were proposed to ensure that the similarity principle can be adopted when predicting the test set.

202. To accomplish a well-planned selection, some approach to statistical experimental design is needed (Box *et al*., 1978). An ideal splitting leads to a test set such that each of its members is close to at least one member of the training set (Tropsha *et al*., 2003). Developing rational approaches for the selection of training and test sets is an active area of research. These approaches range from the straightforward random selection (Yasri and Hartsough, 2001) through activity sampling and various systematic clustering techniques (Potter and Matter, 1998; Taylor, 1995), to the methods of self-organising maps (Gastaiger and Zupan, 1993), Kennard and Stone (1969), formal statistical experimental design (factorial and D-Optimal) (Eriksson and Johansson, 1996), and recently proposed modified sphere exclusion algorithm (Golbraikh *et al*., 2003). These methods help achieve desirable statistical characteristics of the training and test sets.

203. A frequently used approach is *activity sampling* (Kauffman and Jurs, 2001), according to which the choice of training and test sets is made by binning the range of experimental values and randomly selecting an even distribution of compounds from each bin. This guarantees that members of the test set span the entire range of the experimental measurements and are numerically representative of the data set. However, because the binning is based on the response, it does not guarantee that the training set represents the entire descriptor space of the original dataset and that each compound of the test set is close to at least one of the training set.

204. In several applications, the training/test splitting is performed by using clustering techniques. *K*-means algorithm is often used, and from each cluster one compound for the training set is randomly selected. Given that all compounds are represented in a multidimensional descriptor space, the clustering algorithm can be performed on the descriptor values (X values), on response values (Y values), or on the descriptor/response values (X/Y values). Clustering on X/Y values allows clustering the compounds according to all of the given information (Burden, 1999). An alternative clustering approach to select representative subset of compounds is the one based on the maximum dissimilarity method (Potter and Matter, 1998; Taylor, 1995). The method starts with the random selection of a seed compound, then every new compound is successively selected such that it is maximally dissimilar from all the other compounds of the dataset. The process ends either when a maximum number of compounds has been selected or when no other compound can be selected without being too similar to one already selected. Since this method is based on a random starting point, the variance of the results is normally checked by comparing various selections. Hierarchical clustering provides a more specific control by assigning every single compound to a cluster of compounds. It does not require any prior assumption about the number of clusters, and after the clustering process the compound closest to the centre of a cluster is selected as representative compound.

205. Another way to perform a statistically planned training/test selection is by using the Kohonen's Self-Organising Maps (Loukas, 2001). The main goal of the neural network is to map compounds from *n*-dimensional into two-dimensional space. Representative compounds falling in the same areas of the map are randomly selected for the training and test sets. This approach preserves the closeness between compounds: compounds which are similar in the original multidimensional space are close to each other on the map.

206. Similarly to the maximum dissimilarity method, the Kennard Stone algorithm can be used to perform data splitting (Bourguignon *et al*., 1994). It is sequential and consists in maximizing the Euclidean distances between the newly selected compounds and the ones already selected. An additional compound is selected by calculating for each compound, which is not selected, the distance to each selected compound and by maximizing the distance to the closest compound already selected. Both the maximum dissimilarity and the Kennard Stone methods guarantee that the training set compounds are distributed more or less evenly within the whole area of the representative points, and the condition of closeness of the test set to the training set is satisfied.

207. Another data splitting strategy makes use of fractional factorial design (FFD) and D-Optimal design (factorial and D-Optimal) (Kennard and Stone, 1969). A common practice is to process the original data using principal component analysis (PCA) and subsequently to use the principal components (PCs) as design variables in a design selecting a small number of informative and representative training data. These principal components are suitable for experimental design purposes since they are orthogonal and limited in number, reducing the extent of collinearity in the training set. In fractional factorial design all the principal components are explored at two, three or five levels. The training set includes one representative for each combination of components. The drawback of this approach is that it does not guarantee the closeness of the test set to the training set in the descriptor space. D-Optimal design is often performed whenever the classical symmetrical design cannot be applied, because the experimental region is not regular in shape or the number of compounds is selected by a classical design is too large. The basic principle of this method is to select compounds to maximize the determinant of the information (variance-covariance) matrix $|\mathbf{X'X}|$ of independent variables. The determinant of this matrix is maximal when the selected compounds span the space of the whole data, *i.e*. when the most influential compounds (maximal spread) are selected. (Gramatica *et al*., 2004; Gramatica and Papa, 2005; Papa *et al*., 2005)

208. Sphere Exclusion is a dissimilarity-based compound selection method first described by Hudson *et al*. (1996) and then later adapted by various groups (Golbraikh *et al*., 2003; Snarey *et al*., 1997; Nilakatan *et al*., 1997). The algorithm consists in selecting molecules, whose similarities with each of the other selected molecules are not higher than the defined threshold (Gobbi and Lee, 2003). Therefore, each selected molecule creates a (hyper) sphere around itself, so that any candidate molecules inside the sphere are excluded from the selection. The radius of the sphere is an adjustable parameter, determining the number of compounds selected and the diversity between them. The original method starts with the "most descriptive compound" and in each cycle identifies the compound most similar to the centroid of the already selected compounds. This was considered to be very computer intensive, so variations from the original algorithm have been implemented to reduce the computer time required by selecting the next compound quicker.

Evaluating Reliability of Knowledge-Driven (Q)SAR Models

209. Knowledge-driven (Q)SAR models are distinguished from data-driven (Q)SAR models by the initial importance given to assigning chemicals to classes or groups before attempting to predict the specific endpoint values. Because these (Q)SAR models are limited to specific classes, the overall domain of this approach is determined not by a single (Q)SAR model, but rather by combining (Q)SAR models for the classes of chemicals within the purview of the regulatory program. The advantages of this approach are twofold. First, the delineation of chemical classes offers a transparent method of incorporating expert knowledge on the chemical and toxicological properties of chemicals within classes so that (Q)SAR predictions can be more explainable and reliable. Second, because the universe of chemicals under regulatory purview is partitioned into classes, the domain of application for the combined classes can be broader and more closely aligned to the regulatory domain than individual data-driven (Q)SAR models, usually with the trade-off of having fewer measured data for each class of chemicals.

210. For example, the ECOSAR system used by the U.S. EPA has grouped chemicals into more than 100 classes of chemicals and corresponding (Q)SAR models based on a mixture of public and proprietary data (Zeeman *et al*., 1995). The scientific challenge is to assess the predictive capacity of knowledge-based systems is to weigh the significance of information on mechanism on a comparable basis with conventional statistical methods outlined previously. If a chemical is in a class of chemicals for which the toxicity mechanism is known, that information is extremely influential on the ultimate prediction of toxic effects and, if accurate, may often provide a reliable estimate of toxicity within the class. Although there may be exceptions, perhaps the best way to assess the performance of a knowledge-based (Q)SAR modelling system is to conduct periodic retrospective evaluations of the entire system within the regulatory constraints.

211. Moreover, regulatory/decision making bodies may use a set of preliminary classification criteria to make decisions regarding the potential fate and effects of chemicals and may not actually require the use of the discrete experimental or estimated values themselves. These classification schemes typically define ranges to allow the assessors to make more qualitative judgements regarding the chemical of interest. (Q)SARs and classification schemes are used in screening and priority setting to identify potentially hazardous chemicals of concern from the universe of industrial chemicals

212. The results of retrospective evaluations for ECOSAR indicated that the (Q)SAR system could perform with acceptable reliability even though the program was not accompanied by full disclosure of the internal performance parameters (Hulzebos and Postumus, 2003). In other words, regulatory acceptance had been accomplished by years of experience with the (Q)SAR models even though full transparency of the confidential data could not be provided to the scientific community. The study was designed to determine if all (Q)SARs within ECOSAR conformed to the recommended acceptability criteria for (Q)SAR application within Dutch risk assessment. Even though 96 of the 123 (Q)SAR models were found lacking in regard to the OECD Validation Principles, the results indicated that ECOSAR was capable of making accurate and useful predictions.

Concluding remarks

213. Ideally, QSAR modelling should lead to statistically robust models capable of making reliable predictions for new compounds. In this guidance document, reference is made to the reliability, rather than the correctness, of model predictions. This is because from a philosophical viewpoint, it is questionable whether a prediction can ever be correct, or whether a model can ever truly represent reality. As famously quoted by the chemist and statistician, George Box, "all models are wrong, but some are useful" (Box *et al*., 1978).

214. In order for a statistical model to be useful for predictive purposes, it should be built on a sufficiently large and representative amount of information regarding the modelled activity and should contain only relevant variables. As discussed in this chapter, a variety of statistical methods are available for assessing the goodness-of-fit, robustness and predictive ability of QSAR models, and a variety of statistics are routinely used to express these aspects of model performance. Modern statistical software packages provide convenient and automated means of applying these methods and generating a plethora of statistics. The users of (Q)SAR models, such as regulators, need a sufficient understanding of these statistics and the underlying methods in order to interpret the statistics according to their own purposes.

215. The model user should be aware that the performance of a model, while being expressed in quantitative terms and on the basis of well-established techniques, is dependent on the choices by the (Q)SAR modeller. Different types of statistics are generated by different methods, and different values of the same statistics can be generated by altering the compositions of the training and test sets, or by altering

the resampling routine in a cross-validation procedure. This is why transparency in the statistical validation process is needed to form the basis of sound decision-making.

216. Internal validation refers to the assessment of goodness-of-fit and robustness. The goodness-of-fit of a model to its training set can be regarded as the absolute minimum of information needed to assess model performance. It expresses the extent to which the model descriptors "account for" the variation in the training set, and most importantly whether the model is statistically significant. If the model is not statistically significant, or if it is significant but of poor fit, it cannot be expected to be useful for predictive purposes.

217. The robustness of the model provides an indication of how sensitive the model parameters (and therefore predictions) are to changes in the training set. If the model is not robust to small perturbations in the training set, it is unlikely to be useful for predictive purposes. In practice, robustness can be a difficult concept to apply, because there are numerous ways of resampling the data, which affect the statistics generated.

218. The distinction between internal and external validation has important practical implications. Models that are too complex (*i.e*. overfitted) are unlikely to predict independent data as reliably as their internal validation statistics may imply. This problem is increasingly relevant as modern QSAR methods become more powerful and capable of handling large amounts of correlated information and a large number of noisy variables.

219. Predictivity is perhaps the most difficult concept to apply. From a philosophical standpoint, it can be argued that it is impossible to determine an absolute measure of predictivity, since it is highly dependent on the choice of statistical method and test set. Nevertheless, external validation, when performed judiciously, is generally regarded as the most rigorous assessment of predictivity, since predictions are made for chemicals not used in the model development.

220. External validation should be seen as a useful supplement to internal validation, rather than as a substitute. External validation can be difficult to apply in a meaningful when data of sufficient quality are scarce. The model user should therefore be aware that the statistics derived by external validation could be less meaningful than those provided by internal validation, if the external test set is not carefully designed.

221. It is not the aim of this document to define acceptability criteria for the regulatory use of QSAR models, since the use of data in decision-making is highly context-dependent. However, it is possible to identify features of models that are likely to contribute to a high or low performance.

222. A model with high statistical performance is likely to have one or more of the following characteristics:

1. the highest possible prediction power is achieved with the minimum number of variables;
2. there is a low correlation between the predictor variables.

223. A model with low statistical performance is likely to have one or more of the following characteristics:

1. it is lacking one or more relevant variables, *i.e*. has insufficient fitting capability;
2. there is a marked difference between goodness-of-fit and prediction power;
3. one or more (noisy) variables are correlated with the response by chance;
4. there is a high correlation between the predictor variables (multi-collinearity) resulting in redundancy in descriptor information.

Table 5.1. Basic equations and parameters of goodness of fit in MLR

N.	Definition	Equation and terms
1	MLR equation	$\hat{\mathbf{y}} = \mathbf{b}_0 + \mathbf{b}_1\mathbf{x}_1 + \mathbf{b}_2\mathbf{x}_2 + ...\mathbf{b}_p\mathbf{x}_p$ • $\hat{\mathbf{y}}$ = calculated dependent variable • $\mathbf{x}_j$ = predictor variable • $\mathbf{b}_j$ = regression coefficient
2	Coefficient of multiple determination (Multiple correlation coefficient)	$R^2 = \frac{SS_{Reg}}{SS_T} = \frac{(SS_T - SS_{Res})}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$ • $SS_T = \sum_i (y_i - \bar{y})^2$ = total sum of squares • $SS_{Res} = \sum_i (y_i - \hat{y})^2$ = residual sum of squares • $SS_{Reg} = \sum_i (\hat{y} - \bar{y})^2$ = sum of squares due to the regression • y_i = observed dependent variable • $\bar{y}$ = mean value of the dependent variable • $\hat{y}$ = calculated dependent variable
3	Adjusted R^2	$R^2_{adj} = 1 - \frac{SS_{Res}/(n-p-1)}{SS_T/(n-1)}$ $= 1 - (1-R^2)\cdot\left(\frac{n-1}{n-p-1}\right)$ • n = number of observations • p = number of predictor variables
4	Standard error of estimate	$s = \sqrt{\frac{\sum_i (y_i - \hat{y})^2}{(n-p-1)}}$
5	*F*-value	$F = \frac{EMS}{RMS} = \frac{SS_{Reg}/p}{SS_{Res}/(n-p-1)}$ • *RMS = SSRes / (n-p-1)* = residual mean square • *EMS = SSReg / p* = explained mean square
6	*t*-test	$t = \frac{\mathbf{b}_j}{s_{b_j}}$ • $s_{b_j} = \frac{s}{\sqrt{\sum_i (x_i - \bar{x})^2}} = \frac{MS_R}{\sqrt{\sum_i (x_i - \bar{x})^2}}$ = standard deviation of the estimated regression coefficient **b**j • $\bar{x}$ = mean value of the predictor variable

Table 5.2. Confusion or contingency matrix {c_{GG}} for a general case with *G* classes

		Assigned class				
		A1'	***A2'***	***A3'***	***Ag'***	**Marginal totals**
True class	***A1***	$c_{11'}$	$c_{12'}$	$c_{13'}$	$c_{1g'}$	n_1
	A2	$c_{21'}$	$c_{22'}$	$c_{23'}$	$c_{2g'}$	n_2
	A3	$c_{31'}$	$c_{32'}$	$c_{33'}$	$c_{3g'}$	n_3
	Ag	$c_{G1'}$	$c_{G2'}$	$c_{k3'}$	$C_{gg'}$	n_g
	Marginal totals	$n_{1'}$	$n_{2'}$	$n_{3'}$	$n_{g'}$	

Table 5.3. Example of loss matrix {$l_{GG'}$} where the loss function has been arbitrarily defined in an integer scale

		Assigned class			
		A1'	***A2'***	***A3'***	***Ag'***
True class	***A1***	0	1	2	2
	A2	1	0	1	1
	A3	2	1	0	2
	Ag	2	1	2	0

Table 5.4. Definitions of the goodness-of-fit parameters

Statistic	Formula	Definition
Concordance or Accuracy (Non-error Rate)	$\frac{\sum_g c_{gg'}}{n} \times 100$	total fraction of objects correctly classified. $c_{gg'}$ = number of objects correctly classified to each class n = total number of objects
Error Rate	$\frac{n - \sum_g c_{gg'}}{n}$ 1-concordance	total fraction of objects misclassified $c_{gg'}$ = number of objects correctly classified to each class n = total number of objects
NO-Model Error Rate, NOMER%)	$NOMER\% = \frac{n - n_M}{n} \times 100$	Error provided in absence of model n_M = number of objects of the most represented class n = total number of objects
Prior probability of a class	$P_g = \frac{1}{G}$	probability that an object belongs to a class supposing that every class has the same probability (independently of the number of objects of the class). G = number of classes
Prior proportional probability of a class	$P_g = \frac{n_g}{n}$	probability that an object belongs to a class taking into account the number of objects of the class n_g = total number of objects belonging to class g n = total number of objects
Sensitivity of a class	$\frac{C_A}{n_A} \times 100$	percentage of active compounds correctly classified as active compounds. C_A = number of correctly classified active compounds n_A = total number of active compounds
Specificity of a class	$\frac{C_{NA}}{n_{NA}} \times 100$	percentage of non active compounds correctly classified as non active compounds. C_{NA} = number of correctly classified non active compounds n_{NA} = total number of non active compounds.
Misclassification risk	$\sum_g \frac{\left(\sum_{g'} l_{gg'} c_{gg'}\right) P_g}{n_g} \times 100$	risk of incorrect classification (takes into account the number of missclassifications, and their importance) $l_{gg'}$ = diagonal element of the loss matrix

		$c_{gg'}$ = number of objects correctly classified to each class n_g = total number of objects belonging to class g P_g = prior probability class

Footnote: g=1,…, G (G = number of classes)

Table 5.5. 2×2 **contingency table**

		Assigned class		
		Toxic	**Non-toxic**	**Marginal totals**
Observed (*in vivo*) class	**Active**	a	b	$a+b$
	Non-active	c	d	$c+d$
	Marginal totals	$a+c$	$b+d$	$a+b+c+d$

Table 5.6. Definitions of the Cooper statistics

Statistic	Formula	Definition
Sensitivity (True Positive rate)	***a/(a+b)***	fraction of active chemicals correctly assigned
Specificity (True Negative rate)	***d/(c+d)***	fraction of non-active chemicals correctly assigned
Concordance or Accuracy	$\frac{(a+d)}{(a+b+c+d)}$	fraction of chemicals correctly assigned
Positive Predictivity	***a/(a+c)***	fraction of chemicals correctly assigned as active out of the active assigned chemicals
Negative Predictivty	***d/(b+d)***	fraction of chemicals correctly assigned as non-active out of the non-active assigned chemicals
False Positive (over-classification) rate	***c/(c+d)*** **1-specificity**	fraction of non-active chemicals that are falsely assigned to be active
False Negative (under-classification) rate	***b/(a+b)*** **1-sensitivity**	fraction of active chemicals that are falsely assigned to be non-active

Table 5.7. Definitions of the robustness and predictive parameters

Statistic	Definition	Formula
MSE	Mean Squared Error	$\sum_1^n (y_i - \hat{y}_i)^2 \Big/ n$ y_i = observed response for the *i-th* object $\hat{y}_{i/i}$ = response of the *i-th* object estimated by using a model obtained without using the *i-th* object
PRESS	Predictive Residual Sum of Squares	$PRESS = \sum_i \left(y_i - \hat{y}_{i/i}\right)^2$ y_i = observed response for the *i-th* object $\hat{y}_{i/i}$ = response of the *i-th* object estimated by using a model obtained without using the *i-th* object
Q^2	Explained variance in prediction	$Q^2 = 1 - \frac{PRESS}{SS_T} = 1 - \frac{\sum_i \left(y_i - \hat{y}_{i/i}\right)^2}{\sum_i \left(y_i - \bar{y}\right)^2}$ SS_T = total sum of squares y_i = observed response for the *i-th* object $\hat{y}_{i/i}$ = response of the *i-th* object estimated by using a model obtained without using the *i-th* object $\bar{y}$ = average response value of the training set
SDEP	Standard Deviation Error of Prediction	$SDEP = \sqrt{\frac{\sum_i \left(y_i - \hat{y}_{i/i}\right)^2}{n}}$ y_i = observed response for the *i-th* object $\hat{y}_{i/i}$ = response of the *i-th* object estimated by using a model obtained without using the *i-th* object n = the number of training objects
K	Multivariate correlation index	$K\% = \frac{\sum_m \left\lvert \frac{\lambda_m}{\sum_m \lambda_m} - \frac{1}{p} \right\rvert}{\frac{2(p-1)}{p}} \times 100$ λm = eigenvalues obtained from the correlation matrix of the data set **X**(n, p), n = number of objects p = number of variables.

Q^2_{ext}	External explained variance	$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{n_{ext}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{ext}} (y_i - \bar{y})^2}$ y_i = observed response for the *i-th* object $\hat{y}_i$ = predicted response for the *i-th* object $\bar{y}$ = average response value of the training set

CHAPTER 6. GUIDANCE ON THE PRINCIPLE OF MECHANISTIC INTERPRETATION

Summary of Chapter 6

224. This chapter provides guidance on the application of the Principle, "a (Q)SAR should be associated with a mechanistic interpretation, if possible" (Principle 5). The chapter begins with a historical perspective citing several early examples of congeneric (Q)SAR models where the notion of mechanistic interpretation first began. It then goes on to describe examples of more recent (Q)SARs where mechanistic interpretations have been provided. The difference between what is meant by a mechanistic basis and a mechanistic interpretation is clarified through the use of these examples. The chapter also makes raises several discussion points and proposes potential areas for further research.

Introduction

225. OECD Principle 5 for validation of (Q)SARs calls for "a mechanistic interpretation, if possible". Statistical methods used to describe relationships between chemical structure and activity are not intended to replace other knowledge from chemistry and toxicology if such knowledge exists. Any effort in the validation process to show that the (Q)SAR model is consistent with other knowledge of fundamental processes in chemistry and toxicology adds to credibility and acceptance of the predictions from the model. The interpretation of a (Q)SAR model in the context of the molecular descriptors and the endpoint data, both included in and excluded from the training set, is the basis for discovery of underlying causal relationships. When the interpretation of a (Q)SAR model is consistent with existing theories and knowledge of mechanisms, the ability to explain how and why an estimated value from the model was produced increases. Adding that transparency to model performance is the goal of including a mechanistic interpretation of the model.

226. The clause, "if possible," is added to OECD Principle 5 for a very special reason. The evolution of a (Q)SAR model is an iterative process involving the statistical exploration of data, hypothesis generation, and hypothesis testing. The iterative process generally leads to a series of refinements to the training set, both in terms of chemicals included and molecular descriptors for those chemicals. The record of the sequence of hypotheses tested and the mechanistic purpose for refining the training set are not often captured by data mining activities even if they are reported in the literature. Consequently, a useful (Q)SAR model may lack mechanistic interpretation because the model is in the early stages of evolution, or because the mechanistic elements of the application domain have not been compiled from the literature. OECD Principle 5 encourages the validation process to find mechanistic interpretations which can add to the understanding of the statistical validity and the domain of application.

Mechanistic Interpretation

227. Mechanistic interpretations of (Q)SARs begin with the number and the nature of the molecular descriptors used in the model. A molecular descriptor can be any parameter which is a formal mathematical representation of structural attributes of chemicals. The number of molecular descriptors used to quantify structure has proliferated with high speed computing to include many hundreds of parameters, many of which have not been causally linked to intrinsic chemical interactions. The variation in endpoints with the variation in chemical structure is generally attributed to changes in the hydrophobic, electronic and steric attributes of different chemicals. Unless the molecular descriptors can be associated

with these three inherent attributes, or if dozens of molecular descriptors are needed to produce a single (Q)SAR prediction, the likelihood that the models is in the early stages of evolution and of there being a mechanistic basis for the model is small.

228. The basis for mechanistic interpretation is the underlying principle of (Q)SARs which states that the properties and biological interactions of a chemical are inherent to its molecular structure. For example, Hammett (1937) reasoned that similar changes in structure for different chemicals produced similar changes in relative reactivity. He postulated that the effect of substituents on the structure of benzoic acids could be used as a model system to estimate the electronic effects of substituents on similar molecules. The more electron attracting the substituent, the more rapid the reaction. Hammett defined a parameter, σ. Positive values of σ represented electron withdrawal by the substituent from the aromatic ring and negative values indicated electron release.

229. Although the Hammett equation has been modified and extended, σ constants still remain the most general means for estimating the electronic effects of substituents on reaction centres. The power of these simple σ values is that they often take into account solution effects on substituents such as hydrogen bonding, dipole interactions and so on that are still difficult to calculate.

230. Hammett's reasoning was subsequently extended to the development of steric and hydrophobic parameters. These extensions have enabled all kinds of structure-activity relationships of chemical reactions to be tackled.

231. Fifty years ago, Hansch proposed a mathematical model which correlated biological activity such as plant growth regulatory activity of phenoxyacetic acids to Hammett constants and partition coefficients (Hansch *et al*., 1962). In 1964, Hansch and Fujita (1964) showed that the biological activity could be correlated linearly by free-energy related parameters. This approach became known as a Linear Free Energy Relationship (LFER) and expressed in the following equation:

$$\log 1/C = a\pi + b\sigma + cE_S + \ldots\ldots\ldots + \text{constant} \qquad \text{(Eq 1)}$$

when C is the molar concentration of the compound to produce a defined biological response, π is the hydrophobic contribution of the substituent and represented by $\log P_X/P_H$, σ is the Hammett electronic descriptor of the substituents (Hammett, 1970), represented by $\log K_X/K_H$, E_S is Taft's steric parameter (Taft, 1956a) and a, b and c are the appropriate coefficients. In these expressions P_X and P_H are the octanol/water partition coefficients of the substituted and unsubstituted compounds, respectively, and K_X and K_H are the ionization constants of the meta- or para-substituted and unsubstituted benzoic acids at 25 ºC, respectively.

232. The work of Hansch provided perhaps the first example of how a (Q)SAR could give information concerning mechanism. He and his workers (Hansch *et al*., 1977) demonstrated the following relationship for a set of esters binding to the enzyme papain.

$$\text{Log } 1/\text{Km} = 1.03\pi 3' + 0.57\sigma + 0.61\text{MR4} + 3.8 \qquad \text{(Eq 2)}$$
$$N = 25,\ r = 0.907,\ s = 0.208$$

Mechanistic interpretation included the observation that the positive sigma term implied that electron withdrawing substituents favoured formation of the enzyme substrate complex. This made biological sense since the binding to papain involves the electron rich thiol group of a cysteine residue. The positive molar refractivity term suggested that bulkier substituents in the 4 position favoured binding. The two parameters π4 and MR4 are orthogonal to each other in the dataset implying that a bulk effect rather than a hydrophobic effect was important at position 4. The prime sign associated with the π parameter for position 3 indicated that in cases where there were two meta substituents, the π value of more hydrophobic

substituent was used, the other π 3 value being ignored. The rationale for this was that binding of one meta substituent to the enzyme placed the other into an aqueous region and therefore outside the enzyme binding site (Livingstone, 1995).

233. Hansch's early work demonstrated how attempts to rationalize the statistical importance of molecular descriptors could be used to generate and test many hypotheses about the mechanisms. The Hansch approach to interpreting mechanisms became one of the early methods of describing the active sites on proteins by using chemicals as probes into the hydrophobic, electronic, steric nature of receptor sites. Whilst the Hansch approach is mechanistically simple, it is somewhat limited in its breadth of application. Typically, Hansch-type QSARs are limited to biological activity involving molecular initiating events and causal chain in the toxicity pathway.

234. The drive to expand the applicability domain of an individual (Q)SAR to be more responsive to the number and diversity of chemicals being assessed will generally trade greater coverage of the model at the expense of mechanistic relevance. (Q)SAR experts in regulatory agencies are in the best position to balance the uncertainty of models and the need to explain the predictions. Combining families of individual (Q)SAR models using expert systems is a long-term solution to improving the overall performance of (Q)SAR predictions; however, before discussion of expert systems, a general explanation of the common molecular descriptors is needed.

Molecular Descriptors

235. There are two types of parameters used in (Q)SAR models summarized in Table 6.1. Firstly, there are those that are derived from a measurable property of the molecule, *e.g*. the octanol-water partition coefficient, vapour pressure, dissociation constants. The experimental methods and data obtained should be made available for users. Secondly, there are those parameters used to quantify important attributes of chemical structure, or molecular descriptors. In general, each of the molecular descriptors will be computed for a given chemicals structure using a formal computational method. Since the molecular descriptors make up the important starting point in (Q)SAR predictions, it is essential that the method of calculating the molecular descriptors is available to the user and that it can be uniformly applied to the chemical structures without ambiguity. It is important that when selecting descriptors upon which to base a QSAR, the role that these descriptors play, either in the way the chemical behaves or the way the endpoint is expressed, should be known. This is increasingly important now that complex descriptors, based on molecular, electronic or quantum mechanical properties of a molecule are becoming easily available.

236. Descriptors based on measured properties have historically been the most favoured approach when generating QSARs. A number of reviews are available which describe many of these approaches including those by Verhaar *et al.* (1995) and by Clements *et al.* (1993).

237. The most frequently used parameter, especially in effect QSARs, is log K_{ow}. This is probably because log K_{ow} (a measure of hydrophobicity) is considered to reflect the ability of organic substances to partition and accumulate in organisms. However, the use of log K_{ow} assumes that the behaviour of chemicals under consideration is properly modelled by this parameter. Hence if they partition in some other manner rather than passive diffusion or there is significant metabolism or the chemical has a specific mode of action, then log K_{ow} is not a reasonable descriptor.

Presence of Substructures

238. Early attempts to assess the environmental impact of chemicals by SARs used a limited approach based on analogues and chemical class similarities. More recently, QSARs based on the presence of

substructures that indicate the potential for biological activity or for expressing a physicochemical property have been developed.

239. Parameters may be calculated for whole molecules or for well-defined substructures, which may be either functional units, *e.g*. a hydroxyl group, or a clearly defined part of the molecular structure.

240. This method has the distinct advantage that very large databases containing structures, are available allowing for the assessment of an extensive number of substructures and may also reduce the error of predictions based on one result per chemical. However, there are problems that the approach has difficulty in handling. For example, electronic interactions between substructures may vary and cannot always be anticipated. It follows that substructures which were not present in the original database may not be properly assessed. Therefore the approach is best applicable to chemicals containing substructures that have previously been evaluated.

Connectivity Indices

241. The use of molecular connectivity indices (MCI) is extensively discussed (Kier and Hall, 1986) as these are the most successful of all such approaches based on topological information. They can be summarised as follows:

- Path MCI: These are calculated from the non-hydrogen part of a molecule, and can be further divided into zero, first, second and higher order MCIs. The first group are assumed to relate to the bulk properties of a chemical, *e.g.* molecular volume and surface area. Thus Protic and Sabljic (1989) described a zero order valence MCI, ${}^{0}\chi^{v}$, which was used in the development of a QSAR for estimating the toxicity of some chemicals to fathead minnows and which they suggested was a good approximation for molecular volume. This is also supported by Govers *et al.* (1984), who found excellent correlation with molecular weight for a series of PAHs. The higher order MCIs tend to become more related to local structural features and are then normally best used in combination with other parameters (Sabljic, 1991).

- Cluster and path/cluster MCIs: These are strongly associated with branching in a molecule and may have some potential for QSARs requiring steric hindrance descriptors. This was noted by Kuenemann *et al.* (1990) in the assessment of QSARs for biodegradation.

- Chain MCIs: These are associated with rings and their substituents. However, although potentially useful for describing local properties and effects there have been few attempts to date, to use these in QSARs for ecotoxicity.

242. The principal advantages of MCIs are that they are relatively easy to obtain and can be calculated quickly, being based on structure. They are also very flexible, since there are several MCIs available, capable of combining and thus incorporating local, as well as bulk properties of a chemical. However, it is this very flexibility that also tends to be used as a criticism of QSARs based on MCIs. It is often difficult to know what property or feature a particular MCI actually corresponds to in a chemical. Hence it is difficult to propose a relationship based on possible behaviour and then relate that to a certain MCI or group of MCIs.

Calculated Structural and Electronic Descriptors

243. As the speed and availability of computers and software have increased, so has the use of calculated electronic descriptors. There are semi-empirical models now available that can calculate many electronic descriptors in minutes and even the more powerful and precise *ab initio* programs take only

hours now instead of months for modest sized chemicals. These models calculate the electronic nature of chemicals and descriptors that can be measured and some that cannot be directly measured. The following is a partial list of descriptors that can be calculated: LUMO (lowest unoccupied molecular orbital) energy, HOMO (highest occupied molecular orbital) energy, dipole moment, molecular polarisability, solvent accessible surface area, atomic charge on an atom, nucleophilic and electrophilic superdelocalisabilities of bonds, atoms and molecules, heat of formation, and the change in free energy of reactions. Many of these descriptors are useful in predicting reactivity and since some chemicals are toxic because they react with cellular biochemicals to denature them, the descriptors can be used to predict toxicity (Verhaar *et al*., 1996; Purdy, 1991; Lewis, 1992). These descriptors have recently started to be commonly used and so there are not yet many QSARs based on them, but the descriptors appear to provide tools to lump chemicals into larger classes than the traditional classes based on substructures. It is possible to use electronic descriptors to classify chemicals as to the mechanism by which they are toxic and in so doing allow the elimination of some testing. An advantage of this type of QSAR for chemicals with previously untested substructures is that the electronic or structural descriptors for those substructures can be obtained. However, when using a QSAR in this way it is important to remember that this is extrapolating the QSAR and may give rise to unreliable values due to unexpected interactions.

Examples of Mechanistic Interpretations

244. Benigni *et al.* (1994) aimed to study some molecular determinants to discriminate between mutagenic and inactive compounds for aromatic and heteroaromatic amines and nitroarenes. Using a selection of data from the literature (both Ames and SOS repair), he investigated the feasibility of developing (Q)SARs. He found a dramatic difference between those (Q)SARs derived for estimating potency and those derived for predicting the absence or presence of activity. Hydrophobicity was found to play a major role in determining the potency of the active compounds whereas mainly electronic factors differentiated the actives from the inactives. The electronic factors were those expected on the basis of hypothesised metabolic pathways of the chemicals. Electronic factors together with size/shape appeared to determine the minimum requirement for the chemicals to be metabolised whereas hydrophobicity determined the extent of activity.

245. Debnath *et al.* (1992a) modelled mutagenic potency in the TA98 strain of Salmonella typhimurium (+ S9 activation system) and derived the following equation for a set of aminoarenes:

$$\log \text{TA98} = 1.08 \log P + 1.28\ \text{HOMO} - 0.73\ \text{LUMO} + 1.46\ IL + 7.20 \qquad \text{(Eq 3)}$$
$$n = 88,\ r = 0.898\ (r^2 = 0.806),\ s = 0.860,\ F_{1,83} = 12.6$$

The mutagenic potency (log TA98) was expressed as log (revertants/nmol). *IL* in the equation was an indicator variable that assumed a value of 1 for compounds with three or more fused rings. Overall, the principal factor affecting the relative mutagenicity of the aminoarenes was their hydrophobicity (logP). Mutagenicity increased with increasing HOMO values; this positive correlation seemed reasonable since compounds with higher HOMO values are easier to oxidize and should be readily bioactivated. For the negative correlation with LUMO, no simple explanation could be offered.

246. Barratt (1995) proposed a mechanism-based model for predicting the eye irritation potential of neural organic chemicals, as measured in the rabbit draize eye test. A substance which is classified as irritating to eyes according to EC criteria is one which causes a defined degree of trauma in the Draize rabbit eye test following the instillation of 0.1ml (or equivalent weight) as defined in the EC Annex V method (EEC, 1984) and the OECD Test Guideline 405 (OECD, 2002). Neutral organics were described as uncharged, carbon-based chemicals which did not possess the potential to react covalently with or to ionize under the conditions prevalent in biological systems. Common chemical classes covered by this definition were hydrocarbons, alcohols, ethers, esters, ketones, amides, unreactive halogenated compounds,

unreactive aromatic compounds and aprotic polar chemicals. Data on 38 neutral organics taken from the reference databank of eye irritation data published by ECETOC (European Centre for Ecotoxicology and Toxicology) (Bagley *et al*., 1992) together with 8 chemicals drawn from work by Jacob and Martens (1989) was analysed using principal components analysis (PCA). The mechanistic hypothesis underlying this (Q)SAR was summarized as follows. Neutral organic chemicals were irritant as a result of the perturbation of ion transport across cell membranes. These perturbations arise from changes in the electrical properties of the membrane and are related to dipole moments of the perturbing chemicals. In order to affect these electrical properties, a chemical must be able to partition into the membrane and hence possess the appropriate hydrophobic/hydrophilic properties. An appropriately small cross sectional area allowing it to fit easily between lipid components of the membrane was also a requirement. Log P was used as a measure of hydrophobicity. The minor principal inertial axes Ry and Rz were used to represent the cross-sectional area and the dipole moment was used to model the reactivity. Plots of the first two principal components of these parameters showed that PCA was able to discriminate well between the irritant and non-irritant chemicals in the dataset.

247. Abraham and his workers followed a similar mechanistic based approach. In this example a collection of data on the Draize rabbit eye test was analyzed (Abraham *et al*., 1998) using the set of Abraham descriptors (Abraham, 1994). These descriptors included R_2, excess molar refraction, π_2^H polarisability/dipolarity, $\sum\alpha_2^H$ and $\sum\beta_2^H$ effective hydrogen bond acidity and basicity and *Log* L^{16} a descriptor where L^{16} is the vapour-hexadecane solubility at 25°C. A possible model process would be that of transfer of a pure organic liquid to a dilute solution in an organic solvent phase. The equilibrium constant governing such a model process is known as the activity coefficient, γ°, which may be defined for a sparingly soluble liquid as the reciprocal of the solubility of the liquid in the organic solvent phase. Abraham defined the solubility of a vapour into a solvent phase as *L*, where $L = (1/\gamma^{o})/P^{o}$. If the Draize eye score (*DES*) were related to a transport driven mechanism, the transfer process would be from the pure organic liquid into an initial biophase that will be the tear film and cell membranes on the surface of the eye. The more soluble the organic liquid in the initial phase, the larger the *DES* and hence greater irritation. Thus *DES* values would be proportional to $1/\gamma^{\infty}$, the physicochemical solubility and hence $Log(DES/P^{o}) = Log\ L$ where P^{o} is the saturated vapour pressure in ppm at 25°C. A general equation for the correlation and prediction of a series of *Log L* values for solutes into a given condensed phases had already been established.

$$\text{Log SP} = c + r R_2, + s\pi_2^H + a\sum\alpha_2^H + b\sum\beta_2^H + 1.\ \text{Log}\ L^{16} \qquad \text{(Eq 4)}$$

Application of Eq 4 to Log(DES) values yielded an extremely poor correlation but when Log(DES/P°) was used as the dependent variable, a strong relationship (Eq 5) was found.

$$\text{Log(DES/P°)} = -6.955 + 0.1046\pi_2^H + 4.437\sum\alpha_2^H + 1.350\sum\beta_2^H + 0.754\ \text{Log}\ L^{16} \qquad \text{(Eq 5)}$$
$$n = 37,\ r^2 = 0.951,\ SD = 0.32,\ F = 155.9$$

On transforming the calculated Log(DES/P°) values back to calculated DES values, there was good agreement with the original DES values (Eq 6).

$$\text{Log(DES)obs} = 0.022 + 0.979\ \text{Log(DES)calc} \qquad \text{(Eq 6)}$$
$$n = 37,\ r^2 = 0.771,\ SD = 0.3,\ F = 117.6$$

It was suggested that the DES/ P° values referred to the transfer of the irritants from the vapour phase to the biophase and hence that a major factor in the Draize eye test was simply the transfer of the liquid (or the vapour) to the biological system.

248. Models for skin sensitisation have varied from those based on an *a priori* approach to those interpreted *a posteriori*. An example of both is described here. The first physicochemical mathematical model for skin sensitisation was the RAI (Relative Alkylation Index) model (Roberts and Williams, 1982). This index quantifies the relative extent of sensitiser binding to the skin protein as a function of the dose given, the chemical reactivity (which could be expressed in the terms of the measured rate constants for reaction with a model nucleophile, in terms of Taft or Hammett substituent constants or in terms of computed molecular orbital indices) and hydrophobicity expressed as the octanol/water partition coefficient. The general form of the RAI expression is:

$$RAI = \text{Log } D + a \text{ Log } k + b \text{ Log } P \qquad \text{(Eq 7)}$$

where D is dose, k is the relative rate constant and P is the octanol/water partition coefficient. Log P here models both penetration and lipid/polar fluid partitioning.

249. Topological indices are often thought of as being different to interpret. In this example a model for skin sensitisation was developed relating the potency of a set of 93 diverse chemicals to a range of topological indices (Estrada *et al*., 2003). The indices used in the final model accounted for hydrophobicity (H), polar surface area (PS), molar refractivity (MR), polarisability (PSR), charges (GM), van der Waals radii (VDW). Such parameters can be assigned as relevant in the context of skin sensitisation in that partition could be modelled by hydrophobicity, polar surface area, molar refractivity, van der Waals radii as bulk parameters and the reactivity accounted for by polarisability and charges. The Topological Sub-Structural Molecular Design (TOPS-MODE) approach used in this example is based on the method of moments (Estrada, 1996, 1997, 1998). The approach consists of using the topological bond matrix (edge adjacency matrix) of the molecular graph. Bond weights in the main diagonal entries of the bond matrix are used to account for effects that could be involved in biological processes. An advantage with this approach is that a structural interpretation of TOPS-MODE results can be carried out by using the bond contributions to skin sensitization. These are calculated on the basis of the local moments which are defined as the diagonal entries of the different powers of the weighted bond matrix. This provides a mechanistic interpretation at a bond level and enables the generation of new hypotheses such as structural alerts.

250. The following (Q)SAR, taken from the European Technical Guidance Document for chemical risk assessment (European Commission, 1996), predicts the acute toxicity of organic chemicals to the fathead minnow (*Pimephales promelas*). The equation developed was:

$$\text{Log } (LC50) = -0.846 \text{ log Kow} - 1.39 \qquad \text{(Eq 8)}$$

where LC_{50} is the concentration (in moles per litre) causing 50% lethality in *Pimephales promelas*, after an exposure of 96 hours; and Kow is the octanol-water partition coefficient.

251. The (Q)SAR was developed for chemicals considered to act by a single mechanism of toxic action, non-polar narcosis, as defined by Verhaar *et al.* (1992), and therefore has a clear mechanistic basis. In fact, non-polar narcosis is one of the most established mechanisms of toxic action. Non-polar narcosis has been established experimentally by using the Fish Acute Toxicity Syndrome methodology (McKim *et al*., 1987). The (Q)SAR is based on a descriptor for hydrophobicity (log Kow), which is relevant to the mechanism of action, *i.e*. toxicity results from the accumulation of molecules in biological membranes.

Expert Systems

252. An expert system for predicting toxicity is considered to be any formalised system not necessarily computer based, which enables a user to obtain rational predictions about the toxicity of chemicals. All expert systems for the prediction of chemical toxicity are built upon experimental data

representing one or more manifestations of chemicals in biological systems (the database) and/or rules derived from such data (the rulebase). Individual rules within the rulebase are generally of two main types. Some rules are based on mathematical induction whereas other rules are based on existing knowledge and expert judgement. Typically induced rules are QSARs whereas expert rules are often based on knowledge about reactive chemistry. Expert systems are sometimes characterized according to the nature of the rules in their rulebase. An expert system based primarily on statistically induced rules is sometimes called an "automated rule-induction system", whereas a system based primarily on expert rules is referred to as a "knowledge based system" (Dearden *et al*., 1997). The following two examples, referring to ECOSAR and Derek for Windows, outline the mechanistic interpretation for these two types of expert system.

253. As part of the work by the OECD (Q)SAR Group, the ECOSAR tool was evaluated with respect to the OECD Principles (OECD, 2004; ECOSAR, 1996, 1998, 2000, 2005) predicts defined endpoints as required by the US EPA regulatory framework, such as acute $L(E)C_{50}$ and long-term NOECs for fish, daphnids and algae. The (Q)SAR equations are based on linear regression analysis, using log Kow as the sole descriptor for predicting the $L(E)C_{50}$ values (except for the class of surfactants). There is no explicit description of the chemical classes or the exclusion rules. The (Q)SAR for neutral organics is based on the assumption that all chemicals have a minimal toxicity based on the interference of the chemical with biological membranes, which can be modelled by the octanol-water partition coefficient (Kow). All other chemical classes show excess toxicity compared to the neutral organics.

254. Derek for Windows is a knowledge-based expert system created with knowledge of structure-toxicity relationships and an emphasis on the need to understand mechanisms of action and metabolism. The Derek knowledge base covers a broad range of toxicological endpoints, including mutagenicity, carcinogenicity and skin sensitisation.

255. The expert knowledge incorporated into the Derek for Windows system originated from Sanderson and Earnshaw (Sanderson *et al*., 1991). These workers identified a series of 'structural alerts' associated with certain types of toxic activity. The Derek knowledge base was written, developed and continues to be enhanced by Lhasa Ltd and its members at the School of Chemistry, University of Leeds, UK. Lhasa Ltd is a non-profit making collaboration consisting of the University of Leeds and various other educational and commercial institutions (including agrochemical, pharmaceutical and regulatory organizations) created to oversee the development of the Derek for Windows system and the evolution of its toxicity knowledge base.

256. Derek for Windows provides an explicit description of the substructure and substituents. When a query structure is processed, the alerts that match are displayed in a hierarchy called the prediction tree and are highlighted in bold in the query structure. The prediction tree includes the endpoint, the species and reasoning outcome, the number and name of the alert, and the example from the knowledge base if it exactly matches the query structure. The alert description provides a description depicting the structural requirement for the toxicophore detected and a reference to show the bibliographic references used. Some rules are extremely general with substructures only taking into account the immediate environment of a functional group. This means that remote fragments that may modulate a toxicity are not always taken into consideration. In other cases, the descriptions are much more specific.

257. All the rules in Derek are based on either hypotheses relating to mechanisms of action of a chemical class or observed empirical relationships, the ideas for which come from a variety of sources, including published data or suggestions from the Derek collaborative group. This group consists of toxicologists who represent Lhasa Ltd and customers who meet at regular intervals to give advice and guidance on the development of the databases and rulebases. The hypotheses underpinning each alert are documented in the alert descriptions as comments. These comments often include descriptions of features

acting as electrophiles or nucleophiles. However, the detail depends on the specific alert. Some alerts contain no comments, apart from the modulating factors of skin penetration.

Artificial Intelligence systems

258. Many of the models so far discussed involve the use of transparent algorithms, typically regression equations where the mechanistic interpretation is achieved by interpreting the descriptors, the size of their coefficients, and perhaps the mathematical form of the equation. In contrast, AI-based models are sometimes considered to be non-transparent, since the algorithms are deeply embedded.

259. For example, Kohonen networks are specific types of networks that can provide mechanistic insights. Graphical representations of individual layers may indicate the roles of individual descriptors in the model. When a new compound is presented to the model it will be located on a defined position in the Kohonen network. Its mechanism of activity may be deduced from the mechanisms of neighbouring compounds.

Concluding remarks

260. There are many types of different types of modelling approaches. In this chapter, guidance is presented through the use of examples, to illustrate how to consider mechanism in the context of different types of model.

261. The mechanistic rationale of a (Q)SAR can be established *a priori*, in which case the descriptors are selected before modelling on the basis of their known or anticipated role in driving the response, or *a posteriori*, in which case the descriptors are selected on the basis of statistical fit alone, with their mechanistic rationale being rationalised after modelling. Models can also be developed by a combination of these two approaches.

262. In the case of a QSAR with continuous descriptors, a mechanistic interpretation can be based on the physicochemical interpretation of each descriptor and its association with a mode or mechanism of action. The magnitudes of the model coefficients and model structure might also be taken into consideration.

263. In the case of a SAR, a mechanistic interpretation can be based on the chemical reactivity or molecular interaction of the substructure.

264. In the case of expert systems, it is not possible to generalise how a mechanistic interpretation could be assigned, due to the variety of such systems. Some systems are based primarily on expert knowledge, whereas others are based primarily on learned rules. For example, Derek for Windows is based on the use of multiple structural alerts, each of which has its own scientific supporting evidence; whereas METEOR and CATABOL incorporate a significant amount of information on known metabolic pathways.

265. The architecture of neural network models does not generally correspond in any obvious way with underlying mechanisms of action.

Table 6.1. Commonly used molecular descriptors in QSAR studies

Molecular descriptor	Physicochemical interpretation	Examples of QSAR applications
Logarithm of the Partition coefficient: **log P = log (C_{org} /C_{water})** C_{org} = concentration of the non-ionised solute in the organic phase C_{water} = concentration of the non-ionised solute in the water phase	Describes the distribution of a compound between organic (usually n-octanol) and water phase logP>0 – greater solubility in the organic phase; logP<0 – greater solubility in the aqueous phase. Measure of hydrophobicity / lipophilicity	Many applications in QSAR analysis of toxicological data sets (Cronin *et al*., 2002)
Hydrophobic substituent constant (π) : $\pi_X = \log P_{R-X} - \log P_{R-H}$ $\log P_{R-H}$ = logP of the parent compound $\log P_{R-X}$ = logP of *X* substituted derivative	Describes the contribution of a substituent to the lipophilicity of a compound.	QSAR for mutagenicitiy of substituted N-nitroso-N-benzylmethylamines (Singer *et al.*, 1986; Benigni, 2005)
Hammett electronic substituent constant (σ): $\log(Ka_x/Ka_H) = \rho\sigma$ Ka_H = acid dissociation constant of benzoic acid Ka_x = acid dissociation constant of *X* substituted derivative of benzoic acid ρ = a series constant	Describes the electron-donating or -accepting properties of an aromatic substituent, in the *ortho*, *meta* and *para* positions.	QSARs of the relative toxicities of monoalkylated or monohalogenated benzyl alcohols (Schultz *et al*., 1988)
Taft steric parameter (E_S) : $\log k = \log k_0 + \rho^* \sigma^* + \delta E_S$ σ* = polar substituent constant ρ = constant	Steric substituent constant. Describes the intramolecular steric effects on the rate of a reaction.	Original reference of the formulation of Taft steric parameter (Taft, 1956b)
Aqueous solubility (S_{aq}) : The maximum concentration of the compound that will dissolve in pure water at a certain temperature, at equilibrium	Measures the hydrophilicity of a compound	QSARs for fish bioconcentration factor (Dearden and Shinnawei, 2004)
Molecular refractivity (MR): $MR = [(n^2-1)/(n^2+2)]*M/\rho$ n = refractive index M = relative molecular mass ρ = density	Describes the size and polarizability of a fragment or molecule. It could be considered as both an electronic and a steric parameter.	QSARs for binding of tetrahydroisoquinoline derivatives with dstrogen receptors (Hansch *et al*., 2003)
Dissociation Constant (pKa)	Describes extent of ionization of a compound. Reflects electron-directing effects of substitutents.	QSARs for relative toxicity of monosubstituted phenols (Schultz *et al*., 1992)

Table 6.1. Commonly used molecular descriptors in QSAR studies *(cont.)*

Dipole moment Determined via experimental measurement of dielectric constant, refractive index and density, or calculated using molecular orbital theory	Describes separation of charge (polarity) in a molecule, and also considered as measure of hydrophilicity. Hypothesised to reflect the influence of electrostatic interactions with biological macromolecules (Dearden, 1990)	QSARs for eye irritation of neutral organic chemicals (Barratt, 1995)
Atomic charge Calculated by different molecular orbital methods	Descriptor that determines the electrostatic potential around a molecule, thus influencing intermolecular interactions with electrostatic nature.	QSARs for mutagenicity of quinolines (Debnath *et al*., 1992b)
HOMO (Highest Occupied Molecular Orbital) and **LUMO** (Lowest Unoccupied Molecular Orbital) reactivity indices. Calculated using molecular orbital theory.	Descriptors of molecular orbital energies. The HOMO energy describes the nucleophilicity of a molecule, whereas the LUMO energy describes electrophilicity.	Mutagenicity of aromatic and heteroaromatic amines (Debnath *et al*., l992a)
Hydrogen bonding Various measures have been proposed.	Descriptors of chemical reactivity (electrostatic interactions between molecules). Hydrogen-bond donors are proton donors (electronegative atoms or groups) and hydrogen-bond acceptors are groups with the capacity to donate a lone electron pair.	Modelling of aquatic toxicity of environmental pollutants (Raevsky and Dearden, 2004)
Molecular weight (MW) and Molecular volume (MV): $MV = MW/\rho$ ρ - density	Simple molecular size descriptors.	QSPR models for *in vivo* blood-brain partitioning of diverse organic compounds (Hou and Xu, 2003) QSARs of a series of xanthates as inhibitors and inactivators of cytochrome P450 2B1 (Lesigiarska *et al*., 2002)
Molecular surface area (MSA)	Size descriptor defined on the basis of the van der Waals surface of an energy minimised molecule by excluding gaps and crevices	Prediction of blood-brain partitioning for structurally diverse molecules (Kaznessis *et al*., 2001)
Topological Descriptors Numerous types have been proposed, *e.g*. Wiener, Randić, Zagreb, Hosoya, Balaban, Kier and Hall molecular connectivity indices, kappa indices	Descriptors based on chemical graph theory, calculated from the connectivity tables of molecules. Used to express different aspects of the shape and size of molecules, including degree of branching, and flexibility.	Modelling structural determinants of skin sensitisation (Estrada, *et al*., 2003) QSAR of Phenol Toxicity (Hall and Vaugh, 1997)

Table 6.1. Commonly used molecular descriptors in QSAR studies*(cont.)*

Electrotopological descriptors	Atom-based topological descriptors that encode information about the topological environment and electronic interactions of the atom.	QSAR Models for Antileukemic Potency of Carboquinones (Gough and Hall ,1999)
Electronic Density Function (ρ) Obtained from Quantum Chemical Calculations.	Descriptors of molecular similarity, based on electrostatic and steric interactions of the molecule	QSAR of antimycobacterial benzoxazines (Gallegos *et al*., 2004)

REFERENCES

Chapter 1

Cronin, M.T.D., *et al*. (2003a), "Use of Quantitative Structure-Activity Relationships in International Decision-making Frameworks to Predict Ecologic Effects and Environmental Fate of Chemical Substances", *Environmental Health Perspectives*, 111, 1376-1390.

Cronin, M.T.D., *et al.* (2003b), "Use of Quantitative Structure-Activity Relationships in International Decision-making Frameworks to Predict Health Effects of Chemical Substances", *Environmental Health Perspectives,* 111, 1391-1401.

Eriksson, L., *et al*. (2003), "Methods for Reliability, Uncertainty Assessment, and Applicability Evaluations of Classification and Regression Based QSARs", *Environmental Health Perspectives,* 111, 1361-1375.

Jaworska, J.S., *et al*. (2003), "Summary of a Workshop on Regulatory Acceptance of (Q)SARs for Human Health and Environmental Endpoints", *Environmental Health Perspectives,* 111, 1358-1360.

OECD (2004), *Report from the Expert Group on (Quantitative) Structure-Activity Relationships [(Q)SARs] on the Principles for the Validation of (Q)SARs*, Series on Testing and Assessment, No. 49, OECD, Paris, 206pp, http://www.oecd.org/document/30/0,2340,en_2649_34365_1916638_1_1_1_1,00.html, accessed 6 February 2007

OECD (2005), *Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment*, Series on Testing and Assessment, No. 34, OECD, Paris, 96pp, http://www.oecd.org/document/30/0,2340,en_2649_34365_1916638_1_1_1_1,00.html, accessed 6 February 2007.

OECD (2006), *Report on the Regulatory Uses and Applications in OECD Member Countries of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models in the Assessment of New and Existing Chemicals*, Series on Testing and Assessment, No. 58, OECD, Paris, 79pp., http://www.oecd.org/document/30/0,2340,en_2649_34365_1916638_1_1_1_1,00.html, accessed 6 February 2007.

Chapter 2

Alexander, M. (1981), "Biodegradation of Chemicals of Environmental Concern", *Science,* 211, 132-138.

Kelcka, G.M. (1985), "Biodegradation", in W.B. Neely and W.E. Blau (eds.), *Environmental Exposure from Chemicals,* Boca Raton, FI, USA, 109-155.

King, E.F. and H.A. Painter (1983), *Ring-test Programme 1981-82. Assessment of Biodegradability of Chemicals in Water by Manometric Respirometry,* Report No. EUR 8631 EN. European Commission, DG XI, Brussels.

Kitano, M. and M. Takatsuki (1988), *Evaluation of the 1988 OECD-Ring Test of Ready Biodegradability*, 12th Sept. 1988, Chem. Inspect. and Test. Inst., Japan.

OECD (1993), *Structure-Activity Relationships for Biodegradation,* OECD Environment Monograph No. 68, OECD, Paris, http://www.oecd.org/document/30/0,2340,en_2649_34365_1916638_1_1_1_1,00.html, accessed 6 February 2007.

OECD (1997), *OECD Guidelines for the Testing of Chemicals, Test Guideline 471, Bacterial Reverse Mutation Test,* OECD, Paris, http://www.oecd.org/document/40/0,2340,en_2649_34365_37051368_1_1_1_1,00.html, accessed 7 February 2007

Peijnenburg, W.J.G.M. and W. Karcher (1995), *Proceedings of the Workshop "Quantitative Structure Activity Relationships for Biodegradation"*, September 1994, Belgirate Italy. RIVM Report No. 719101021. National Institute of Public Health and Environmental Protection, Bilthoven, The Netherlands.

Chapter3

De Bruijn, J. and J. Hermens (1991), "Qualitative and Quantitative Modeling to Toxic Effects of Organophosphorous Compounds to Fish", in J.L.M. Hermens and A. Opperhuizen (eds.), *QSAR in Environmental Toxicology – IV*, pp. 441-455. Elsevier, Amsterdam, The Netherlands.

De Saint Laumer, J.Y., M. Chastrette and J. Devillers (1991), "Multilayer Neural Networks Applied to Structure-Activity Relationships", in W. Karcher and Devillers (eds.), *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Applied Multivariate Analysis in SAR and Environmental Studies*, pp. 153-169. Kluwer, Dordrecht, The Netherlands.

Draper, N.R. and H. Smith (1991), *Applied Regression Analysis*, John Wiley and Sons, New York, USA.

Forrest, S. (1993), "Genetic Algorithms: Principles of Natural Selection Applied to Computation", *Science*, 261, 872-885.

Friederichs, M., O. Fränzle and A. Salski (1996), "Fuzzy Clustering of Existing Chemicals According to Their Ecotoxicological Properties", *Ecological Modelling*, 85, 27-40.

Ghafourian, T. and M.T.D. Cronin (2005), "The Impact of Variable Selection on the Modelling of Oestrogenicity", *SAR and QSAR in Environmental Research*, 16, 171-190.

Goldberg, D.E. (1989), *Genetic Algorithms in Search of Optimization, and Machine Learning*, Addison Wesley Longman Inc., Reading, MA.

Govers, H., C. Ruepert and H. Aiking (1984), "Quantitative Structure-Activity Relationships for Polycyclic Aromatic Hydrocarbons: Correlation between Molecular Connectivity, Physico-chemical Properties, Bioconcentration and Toxicity in *Daphnia pulex*", *Chemosphere*, 13, 227-236.

Govers, H.A.J., R. Luijk and E.H.G. Evers (1991), "Descriptors for Isomer Resolution of (Bio)-distribution of Chlorinated Aromatic Compounds", in J.L.M. Hermens and A. Opperhuizen (eds.), *QSAR in Environmental Toxicology – IV*, pp. 105-119, Elsevier, Amsterdam, The Netherlands.

Gramatica, P., P. Pilutti and E. Papa (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.

Gramatica, P. and E. Papa (2005), "An Update of the BCF QSAR Model Based on Theoretical Molecular Descriptors", *QSAR and Comb. Sci., 24,* 953-960.

Griep, M.I., I.N.Wakeling, P. Vankeerberghen, and D.L. Massart (1995), "Comparison of Semirobust and Robust Partial Least Squares Procedures", *Chemometrics and Intelligent Laboratory Systems*, 29, 37-50.

Kaiser, K.L.E. and S.R. Esterby (1991), "Regression and Cluster Analysis of the Acute Toxicity of 267 Chemicals to Six Species of Biota and the Octanol/Water Partition Coefficient", *Science of the Total Environment*, 109/110, 499-514.

Koneman, H. (1981), "Quantitative Structure-Activity Relationships in Fish Toxicity Studies. Relationships for 50 Industrial Pollutants", *Toxicology*, 19, 209-221.

Koza, J.R. (1992), *Genetic Programmimg: on the Programming of Computers by Means of Natural Selection*, The MIT Press, Cambridge, MA.

Netzeva T.I., *et al.* (2005), "Description of the Electronic Structure of Organic Chemicals Using Semiempirical and Ab Initio Methods for Development of Toxicological QSARs", *Journal of Chemical Information and Modelling*, 45: 106-114

Niemi, G.J. (1990), "Multivariate Analysis and QSAR: Application of Principal Components Analysis", in W. Karcher and Devillers (eds.), *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*, pp. 153-169. Kluwer, Dordrecht, The Netherlands.

Pavan, M., T.I. Netzeva and A.P. Worth (2006), "Validation of a QSAR Model for Acute Toxicity", *SAR and QSAR in Environmental Research,* 17, 147-171.

Ren, S. (2003), "Ecotoxicity Prediction Using Mechanism- and Non-mechanism-based QSARs: a Preliminary Study", *Chemosphere*, 53, 1053-1065.

Smiths, J.R.M., *et al*. (1994), Using Artificial Neural Networks for Solving Chemical Oroblems. Part I. Multi-layer Feed-forward Networks", *Chemometrics and Intelligent Laboratory Systems*, 22, 165-189.

Wakeling, I.N. and H.J.H Macfie (1992), "A Robust PLS Procedure", *Journal of Chemometrics,* 6, 189-198.

Walczak, B. and D.L. Massart (1995), "Robust Principal Components Regression as a Detection Tool for Outliers", *Chemometrics and Intelligent Labortory Systems*, 27, 41-54.

Xu, L., *et al.* (1994), "Quantitative Structure-Activity Relationships for Toxicity of Phenols Using Regression Analysis and Computational Neural Networks", *Environmental Toxicology and Chemistry*, 13, 841-851.

Chapter 4

Cunningham, A.R. and H.S Rosenkranz (2001), "Estimating the Extent of the Health Hazard Posed by High-Production Volume Chemicals", *Environmental Health Perspectives*, 110, 953−956.

Deneer, J.W., W. Seinen and J.L.M Hermens (1988), "The Acucte Toxicity of Aldehydes to Guppy", *Aquatic Toxicology*, 12, 185−192.

Dimitrov, S., *et al*. (2005), "A Stepwise Approach for Defining the Applicability Domain of SAR and QSAR Models", *Journal of Chemical Information and Modeling*, 45, 839 −849.

Gramatica, P., P. Pilutti and E. Papa (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.

Hong, H., *et al*. (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.

Jaworska, J., N. Nikolova-Jeliazkova and T. Aldenberg (2005), "QSAR Applicability Domain Estimation by Projection of the Training Set in Descriptor Space: A Review", *Alternatives to Laboratory Animals*, 33, 445–459.

Klopman, G., *et al*. (2003), "In-silico Screening of High Production Volume Chemicals for Mutagenicity using the MCASE QSAR Expert System", *SAR and QSAR in Environmental Research*, 14, 165−180.

Netzeva, T.I., *et al*. (2005), "Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure-Activity Relationships", *Alternatives to Laboratory Animals*, 33, 155−173.

Netzeva, T.I., A. Gallegos Saliner and A.P. Worth (2006), "Comparison of the Applicability Domain of a QSAR for Estrogenicity with a Large Chemical Inventory", *Environmental Toxicology and Chemistry*, 25, 1223-1230.

O'Brien, P.J, (1991), "Molecular Mechanisms of Quinone Toxicity", *Chemico-Biological Interactions*, 80, 1−41.

Pavan, M., A. Worth and T. Netzeva (2005), *Preliminary Analysis of an Aquatic Toxicity Dataset and Assessment of QSAR Models for Narcosis*, JRC Report EUR 21479 EN, European Commission, Joint Research Centre, Ispra, Italy.

Russom, C.L., *et al*. (1997), "Predicting Modes of Toxic Action from Chemical Structure: Acute Toxicity in the Fathead Minnow (*Pimephales promelas*)", *Environmental Toxicology and Chemistry*, 16, 948−967.

Schmieder, P., *et al*. (2003), "QSAR Prioritisation of Chemical Inventories for Endocrine Disruptor Testing", *Pure and Applied Chemistry*, 75, 2389−2396.

Schultz, T.W. and M.T.D. Cronin (1999), "Response-surface Analysis for Toxicity to *Tetrahymena pyriformis*: Reactive Carbonyl-containing Chemicals", *Journal of Chemical Information and Computer Sciences*, 39, 304−309.

Schultz, T.W., *et al*. (2002), "Structure-Toxicity Relationships for Aliphatic Chemicals Evaluated with *Tetrahymena pyriformis*", *Chemical Research in Toxicology*, 15, 1602–1609.

Schultz, T.W., *et al*. (2005), "Structure-Toxicity Relationships for the Effects to *Tetrahymena pyriformis* of Aliphatic, Carbonyl-containing α,β-unsaturated Chemicals", *Chemical Research in Toxicology*, 18, 330−341.

Tong, W., *et al.* (2003), "Decision Forest: Combining the Predictions of Multiple Independent Decision Tree Models", *Journal of Chemical Information and Computer Sciences*, 43, 525−531.

Tunkel, J., *et al*. (2005), "Practical Considerations on the Use of Predictive Models for Regulatory Purposes", *Environmental Science and Technology*, 39, 2188−2199.

Verhaar, H.J.M., W. Mulder and J.L.M. Hermens (1995), "QSARs for Ecotoxicity", in J.L.M. Hermens (ed.), *Overview of Structure-Activity Relationships for Environmental Endpoints. Part 1: General Outline and Procedure*, Report Prepared within the Framework of the Project "QSAR for Prediction of Fate and Effects of Chemicals in the Environment", Contract with the European Commission EV5V-CT92-0211.

Votano, J.R., *et al*. (2004), "Three New Consensus QSAR Models for the Prediction of Ames Genotoxicity", *Mutagenesis*, 19, 365-377.

Chapter 5

Anzali, S., *et al*. (1998), "The Use of Self-Organizing Neural Networks in Drug Design", *Perspectives in Drug Discovery and Design*, Vol. 9-11, pp. 273-299.

Boggia, R., *et al*. (1997), "Chemometric Study and Validation Strategies in the Structure-Activity Relationship of New Cardiotonic Agents", *Quantitative Structure-Activity Relationships*, 16, 201-213.

Bourguignon, B., *et al*. (1994), "Optimization in Irregularly Shaped Regions: pH and Solvent Strength in Reversed Phase High-Performance Liquid Chromatography Separations", *Analytical Chemistry*, 66, 893-904.

Box, G.E.P., W.G. Hunter and J.S. Hunter (1978), *Statistics for Experimenters. An Introduction to Design, Data Analysis and Model Building,* John Wiley & Sons, New York, NY.

Breiman, L., *et al*. (1984), *Classification and Regression Trees*, Wadsworth International Group, Belmont, CA, USA.

Burden, F.R. (1999), "Robust QSAR Models Using Bayesian Regularized Neural Networks", *Journal of Medicinal Chemistry*, 42, 3183-3187.

Clark, M. and R.D. Cramer III (1993), "The Probability of Chance Correlation Using Partial Least Squares (PLS)", *Quantitative Structure-Activity Relationships*, 12, 137-145.

Cooper, J.A., R. Saracci and P. Cole (1979), "Describing the Validity of Carcinogen Screening Tests", *British Journal of Cancer*, 39, 87–89.

Cramer III, R.D., *et al*. (1988), "Cross Validation, Bootstrapping and Partial Least Squares Compared with Multiple Regression in Conventional QSARs Studies", *Quantitative Structure-Activity Relationships*, 7, 18-25.

Cruciani, G., *et al*. (1992), "Predictive Ability of Regression Models. Part I: Standard Deviation of Prediction Errors (SDEP)", *Journal of Chemometrics*, 6, 335-346.

Devillers, J. and D. Domine (1999), "A Noncongeneric Model for Predicting Toxicity of Organic Molecules to Vibrio fischeri", *SAR and QSAR in Environmental Research*, 10, 61-70.

Diaconis, P. and B. Efron (1983), "Computer Intensive Methods in Statistics", *Scientific American*, 248, 96-108.

Efron, B. (1983), "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation", *Journal of American Statistical Association,* 78, 316-331.

Efron, B. and R.J. Tibshirani (1993), *An Introduction to the Bootstrap*, Chapman & Hall, London.

Eriksson, L. and E. Johansson (1996), "Multivariate Design and Modeling in QSAR. Tutorial", *Chemometrics and Intelligent Laboratory Systems*, 34, 1-19.

Eriksson, L., *et al*. (2001), *Multi- and Megavariate Data Analysis. Principles and Applicationsm*, Umetrics AB, Umeä, Sweden.

Eriksson, L., *et al*. (2003), "Methods for Reliability, Uncertainty Assessment, and Applicability Evaluations of Regression Based and Classification QSARs", *Environmental Health Perspectives*, 111, 1361-1375.

Feinstein, A.R. (1975), "Clinical Biostatistics XXXI. On the Sensitivity, Specificity, and Discrimination of Diagnostic Tests", *Clinical Pharmacology and Therapeutics*, 17, 104–116.

Frank, I.E. and J.H. Friedman (1989), "Classification: Oldtimers and Newcomers", *Journal of Chemometrics*, 3, 463-475.

Gastaiger, J. and J. Zupan (1993), "Neural Networks in Chemistry", *Angewandte Chemie International Edition*, 32, 503-527.

Gobbi, A. and M-L. Lee (2003), "Database DISE: Directed Sphere Exclusion", *Journal of Chemical Information & Computer Sciences*, 43, 317-323.

Golbraikh, A. and A. Tropsha (2002a), "Beware of q^2!", *Journal of Molecular Graphics and Modelling*, 20, 269-276.

Golbraikh, A. and A. Tropsha (2002b), "Predictive QSAR Modeling Based on Diversity Sampling of Experimental Datasets for the Training and Test Set Selection", *Journal of Computer Aided Molecular Design*, 16, 357-369.

Golbraikh, A., *et al*. (2003), "Rational Selection of Training and Test Sets for the Development of Validated QSAR Models", *Journal of Computer Aided Molecular Design*, 17, 241-253.

Gramatica, P., P. Pilutti and E. Papa (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.

Gramatica, P. and E. Papa (2005), "An Update of the BCF QSAR Model Based on Theoretical Molecular Descriptors", *QSAR andComb. Sci*., 24*,* 953-960.

Guha, R. and P.C. Jurs (2005), "Determining the Validity of a QSAR Model – a Classification Approach", *Journal of Chemical Information and Modeling*, 43, 65-73.

Hand, D. (1981), *Discrimination and Classification*, Wiley & Sons, New York.

Hanley, J.A. (1989), "Receiver Operating Characteristic (ROC) Methodology: The State of the Art", *Critical Reviews in Diagnostic Imaging*, 29, 307-335.

Hawkins, D.M. (2004), "The Problem of Overfitting", *Journal of Chemical Information & Computer Sciences*, 44, 1-12.

Hudson, B.D., *et al.* (1996), "Parameter Based Methods for Compounds Selection from Chemical Databases", *Quantitative Structure-Activity Relationships*, 15, 285-289.

Hulzebos, E.M. and R. Posthumus (2003), "(Q)SARs: Gatekeepers against Risk on Chemicals?", *SAR and QSAR in Environmental Research*, 14(4): 285-316.

Jouan-Rimbaud, D., D.L. Massart and O.E. de Noord (1996), "Random Correlation in Variable Selection for Multivariate Calibration with a Genetic Algorithm", *Chemometrics and Intelligent Laboratory Systems*, 35, 213-220.

Kauffman, G.W. and P.C. Jurs (2001), "QSAR and *k*-Nearest Neigbor Classification Analysis of Selective Cyclooxygenase-2 Inhibitors Using Topologically-based Numerical Descriptors", *Journal of Chemical Information and Computer Sciences*, 41, 1553-1560.

Kennard, R.W. and L.A. Stone (1969), "Computer Aided Design of Experiments", *Technometrics*, 11, 137-148.

Kraemer, H.C. (1982), "Kappa Coefficient", in S. Kotz and N.L. Johnson (eds.), *Encyclopedia of Statistical Sciences*, John Wiley & Sons, New York.

Kubinyi, H. (1993), "QSAR: Hansch Analysis and Related Approaches. Methods and Principles" in R. Mannhold, P. Kroogsgard-Larsen and H. Timmerman (eds.), *Medicinal Chemistry, Vol. 1*., VCH, Weinheim.

Lek, S. and J.F. Guegan (1999), "Artificial Neural Networks as a Tool in Ecological Modeling, an Introduction", *Ecological Modelling,* 120, 65-73.

Lindgren, F., *et al*. (1996), "Model Validation by Permutation Tests: Applications to Variable Selection", *Journal of Chemometrics*, 10, 421-532.

Loukas, Y.L. (2001), "Adaptive Neuro-fuzzy Inference System: an Instant and Architecture-free Predictor for Improved QSAR Studies", *Journal of Medicinal Chemistry*, 44, 2772-2783.

Lusted, L.B. (1971), "Signal Detectability and Medical Decision-making", *Science*, 171, 1217-1219.

Mager, P.P. (1995), "Diagnostics Statistics in QSAR", *Journal of Chemometrics*, 9, 211-221.

Massart, D.L., *et al*. (1997a), *Handbook of Chemometrics and Qualimetrics: Part A*, Elsevier Science.

Massart, D.L., *et al.* (1997b), *Handbook of Chemometrics and Qualimetrics: Part B*, Elsevier Science, Amsterdam, The Netherlands.

Mazzatorta, P., *et al*. (2003), "Modeling Toxicity by Using Supervised Kohonen Neural Networks", *Journal of Chemical Information and Modeling*, 43, 485-492.

McDowell, R.M. and J. Jaworska (2002), "Bayesian Analysis and Inference of QSAR Predictive Model Results", *SAR and QSAR in Environmental Research*, 13, 111–125.

Netzeva, T.I., *et al.* (2003), "Partial Least Squares Modelling of the Acute Toxicity of Aliphatic Compounds to Tetrahymena pyriformis", *SAR and QSAR in Environmental Research*, 14, 265-283.

Nilakatan, R., N. Bauman and K.S. Haraki (1997), "Database Diversity Assessment: New Ideas, Concepts and Tools", *Journal of Computer Aided Molecular Design*, 11, 447-452.

Osten, D.W. (1988), "Selection of Optimal Regression Models via Cross-validation", *Journal of Chemometrics*, 2, 39-48.

Papa, E., F. Villa and P. Gramatica (2005), "Statistically Validated QSARs and Theoretical Descriptors for the Modelling of the Aquatic Toxicity of Organic Chemicals in *Pimephales promelas* (Fathead Minnow)", *J. Chem. Inf. Model*, 45, 1256-1266.

Potter, T. and H. Matter (1998), "Random or Rational Design?", Evaluation of Diverse Compound Subsets from Chemical Structure Databases", *Medicinal Chemistry*, 41, 478-488.

Provost, F. and T. Fawcett (2001), "Robust Classification for Imprecise Environments", *Machine Learning Journal*, 42, 203-231.

Snarey, M., *et al.* (1997). "Comparison of Algorithms for Dissimilarity-based Compound Selection", *Journal of Molecular Graphics and Modeling*, 15, 373-385.

Spycher, S., E. Pellegrini and J. Gasteiger (2005), "Use of Structure Descriptors to Discriminate between Modes of Toxic Action of Phenols", *Journal of Chemical Information and Modeling*, 45, 200-208.

Stone, M. and P. Jonathan (1993), "Statistical Thinking and Technique for QSAR and Related Studies. Part I: General Theory", *Journal of Chemometrics*, 7, 455-475.

Sullivan Pepe, M. (2003), *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford Statistical Science Series 28, Oxford University Press.

Taylor, R. (1995), "Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals", *Journal of Chemical Information & Computer Sciences*, 35, 59-67.

Todeschini, R., V. Consonni and A. Maiocchi (1999), "The K Correlation Index: Theory Development and Its Applications in Chemometrics", *Chemometrics and Intelligent Laboratory Systems*, 46, 13-29.

Todeschini, R., *et al*. (2004), "Detecting “Bad” Regression Models: Multicriteria Fitness Functions in Regression Analysis", *Analytica Chimica Acta*, 515, 199-208.

Topliss, J.G. and R.P. Edwards (1979), "Chance Factors in Studies of Quantitative Structure-Activity Relationships", *Journal of Medicinal Chemistry*, 22, 1238-1244.

Tropsha, A., P. Gramatica and V.K. Gombar (2003), "The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models", *QSAR & Combinatorial Science*, 22, 69-77.

Vracko, M. and J. Gasteiger (2002), "A QSAR Study on a Set of 105 Flavonoid Derivatives Using Descriptors Derived from 3D Structures", *Internet Electronic Journal of Molecular Design*, 1, 527-544.

Vracko, M. (2005), "Kohonen Artificial Neural Network and Counter Propagation Neural Network in Molecular Structure-Toxicity Studies", *Current Computer-Aided Drug Design,* 1, 73-78.

Wehrens, R., H. Putter and L.M.C. Buydens (2000), "Bootstrap: a Tutorial", *Chemometrics and Intelligent Laboratory Systems*, 54, 35-52.

Wold, S. and W.J. Dunn III (1983), "Multivariate Quantitative Structure-Activity Relationships (QSAR): Conditions for their Applicability", *Journal of Chemical Information & Computer Sciences*, 23, 6-13.

Wold, S., *et al*. (1984), "The Collinearity Problem in Linear Regression, the Partial Least Squares (PLS) Approach to Generalized Inverses", *SIAM Journal of Science Statistics and Computer*, 5, 735-743.

Wold, S., E. Johansson and M. Cocchi (1993), "PLS: Partial Least Squares Projections to Latent Structures", in H. Kubinyi (ed.), *3D-QSAR in Drug Design: Theory, Methods and Applications*, pp523-550. ESCOM Science, Leiden, The Netherlands.

Wold, S. (1995), "PLS for Multivariate Linear Modeling", in H. van de Waterbeemd (ed.), *Chemometric Methods in Molecular Design*, VCH, Weinheim, Germany.

Worth, A.P. and M.T.D. Cronin (2000), "Embedded Cluster Modelling: A Novel Quantitative Structure-Activity Relationship for Generating Elliptic Models of Biological Activity", in M. Balls, A.M. van Zeller and M.E. Halder (eds.), *Progress in the Reduction, Refinement and Replacement of Animal Experimentation*, pp. 479-491, Elsevier Science, Amsterdam, The Netherlands.

Worth, A.P. and M.T.D. Cronin (2001), "The Use of Bootstrap Resampling to Assess the Uncertainty of Cooper Statistics", *Alternatives to Laboratory Animals*, 29, 447-459.

Worth, A.P and M.T.D. Cronin (2003), "The Use of Discriminant Analysis, Logistic Regression and Classification Tree Analysis in the Development of Classification Models for Human Health Effects", *Journal of Molecular Structure (Theochem)*, 622, 97-111.

Yasri, A. and D. Hartsough (2001), "Toward an Optimal Procedure for Variable Selection and QSAR Model Building", *Journal of Chemical Information and Computer Sciences*, 41, 1218-1227.

Zeeman, *et al*. (1995), "U.S. EPA Regulatory Perspectives on the Use of QSAR for New and Existing Chemical Evaluations", *SAR and QSAR in Environmental Research*, 3, 179-201.

Zupan, J. and J. Gasteiger (1999), *Neural Networks for Chemistry and Drug Design*, Wiley-VCH, Weinheim, Germany.

Chapter 6

Abraham, M.H. (1994), "Scales of Solute Hydrogen-bonding: Their Construction and Application to Physicochemical and Biochemical Processes", *Chemical Society Reviews*, 22, 73-83.

Abraham, M.H., *et al*. (1998), "A (Q)SAR for a Draize Eye Irritation Database", *Toxicology in Vitro*, 12, 201-207.

Bagley, D.M., *et al*. (1992), "Eye Irritation: Reference Chemicals Databank", *Toxicology in Vitro*, 6, 487-491.

Barratt, M.D. (1995), "A Quantitative Structure-Activity Relationship for the Eye Irritation Potential of Neutral Organic Chemicals", *Toxicology Letters*, 80, 69-74.

Benigni, R. (2005), "Structure-Activity Relationship Studies of Chemical Mutagens and Carcinogens: Mechanistic Investigations and Prediction Approaches", *Chemical Reviews*, 105, 1767-1800.

Benigni, R., C. Andreoli and A. Giuliani (1994), "(Q)SAR Models for Both Mutagenic Potency and Activity: Application to Nitroarenes and Aromatic Amines", *Environmental and Molecular Mutagenesis*, 24, 208-219.

Clements, R.G., *et al*. (1993), "The Use and Application of QSARs in the Office of Toxic Substances for Ecological Hazard Assessment of New Chemicals", in W.G. Landis, J.S. Hughes and M.A. Lewis (eds.), *Environmental Toxicology and Risk Assessment – 1st Volume*, ASTM STP 1179, American Society for Testing and Materials, Philadelphia, USA, pp. 56-64.

Cronin, M.T.D, *et al.* (2002), "The Importance of Hydrophobicity and Electrophilicity Descriptors in Mechanistically-based QSARs for Toxicological Endpoints", *SAR and QSAR in Environmental Research*, 13, 167-176.

Dearden, J.C. (1990), "Physico-Chemical Descriptors", in W. Karcher and J. Devillers (eds.), *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*, pp 25-61, Kluwer Academic Publishers.

Dearden, J.C., *et al*. (1997), "The Development and Validation of Expert Systems for Predicting Toxicity", *Alternatives to Laboratory Animals*, 25, 223-252.

Dearden, J.C. and N.M. Shinnawei (2004), "Improved Prediction of Fish Bioconcentration Factor of Hydrophobic Chemicals", *SAR and QSAR in Environmental Research*, 15, 449-455.

Debnath, A. K., *et al.* (1992a), "A QSAR Investigation of the Role of Hydrophobicity in Regulating Mutagenicity in the Ames Test: 1. Mutagenicity of Aromatic and Heteroaromatic Amines in Salmonella Typhimurium TA98 and TA100", *Environmental and Molecular Mutagenesis*, 19, 37-52.

Debnath, A.K., *et al*. (1992b), "Mutagenicity of Quinolines in Salmonella typhimurium TA100. A QSAR Study Based on Hydrophobicity and Molecular Orbital Determinants", *Mutation Research*, 280, 55-65.

Ecological Structure Activity Relationships (ECOSAR) (1996), *Estimating Toxicity of Industrial Chemicals to Aquatic Organisms Using Structure-Activity Relationships (ECOSAR Technical Reference Manual)*, http://www.epa.gov/oppt/newchems/tools/sarman.pdf, accessed 7 February 2007.

ECOSAR (1998), *User's Guide for the ECOSAR Class Program (ECOSAR User Manual),* http://www.epa.gov/oppt/newchems/tools/manual.pdf, accessed 7 February 2007.

ECOSAR (2000), *Ecological Structure Activity Relationships, v.0.99g,* January 2000, U.S.EPA website, http://www.epa.gov/opptintr/newchems/tools/21ecosar.htm, accessed 7 February 2007.

ECOSAR (2005), *Ecological Structure Activity Relationships, v. 0.99h,* December 2005, U.S.EPA website, http:www.epa.gov/opptintr/exposure/pubs/episuite.htm, accessed 7 February 2007. [Note: This URL is the site of EPI (Estimation Programs Interface) Suite Version 3.12 (8 December 2005) which includes ECOSAR v.0.99h. Once EPI 3.12 is downloaded, ECOSAR v.0.99h can be run separetely.]

EEC (1984), 84/449/EEC. Commission Directive of 25 April 1984 Adapting to Technical Progress for the Sixth Time Council Directive 67/548/EEC on the Approximation of Laws, Regulations and Administrative Procedures Relating to the Classification, Packaging and Labelling of Dangerous Substances, *Official Journal of European Communities*, L25, 106-108.

Estrada, E. (1996), "Spectral Moments of the Edge Adjacency Matrix in Molecular Graphs. 1. Definition and Applications to the Prediction of Physical Properties of Alkanes", *Journal of Chemical Information and Computer Science*, 36, 844-849.

Estrada, E. (1997), "Spectral Moments of the Edge-Adjacency Matrix of Molecular Graphs. 2. Molecules Containing Heteroatoms and (Q)SAR Applications", *Journal of Chemical Information and Computer Science*, 37, 320-328.

Estrada, E. (1998), "Spectral Moments of the Edge-Adjacency Matrix of Molecular Graphs. 3. Molecules Containing Cycles", *Journal of Chemical Information and Computer Science*, 38, 23-27.

Estrada, E., *et al.* (2003), "Computer-aided Knowledge Generation for Understanding Skin Sensitization Mechanisms: the TOPS-MODE Approach", *Chemical Research in Toxicology*, 16, 1226-1235.

European Commission (1996), *Technical Guidance Document 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*, Luxembourg, European Commission, Office for Official Publications of the European Communities.

Gallegos,A., *et al.* (2004), "Similarity Approach to QSAR. Application to Antimycobacterial Benzoxazines", *International Journal of Pharmaceutics*, 269, 51-60.

Gough, J. and L.H. Hall (1999), "QSAR Models of the Antileukemic Potency of Carboquinones: Electrotopological State and Chi Indices", *Journal of Chemical Information and Computer Science*, 39, 356-361.

Govers, H., C. Ruepert and H. Aiking (1984), "Quantitative Structure-Activity Relationships for Polycyclic Aromatic Hydrocarbons: Correlation between Molecular Connectivity, Physico-chemical Properties, Bioconcentration and Toxicity in *Daphnia pulex*", *Chemosphere*, 13, 227-236.

Hall, L. H. and T.A. Vaughn (1997), "QSAR of Phenol Toxicity Using Electrotopological State and Kappa Shape Indices", *Medicinal Chemistry Research,* 7, 407-416.

Hammett, L.P. (1937), "The Effect of Structure on the Reaction of Organic Compounds. Benzene Derivatives", *Journal of the American Chemical Society*, 59, 96-103.

Hammett, L.P. (1970), *Physical Organic Chemistry*, Mc Graw Hill, New York.

Hansch, C. and T. Fujita (1964), "Rho-sigma-pi Analysis. A Method for the Correlation of Biological Activity with Chemical Structure", *Journal of the American Chemical Society*, 86, 1616-1626.

Hansch, C, *et al*. (1962), "Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients", *Nature*, 194, 178-180.

Hansch, C., *et al*. (1977), "Structure-Activity Relationships in Papain and Bromelain Ligand Interactions", *Archives of Biochemistry and Biophysics*, 183, 383-392.

Hansch, C., *et al*. (2003), "Quantitative Structure-Activity Relationships of Phenolic Compounds Causing Apoptosis", *Bioorganic and Medicinal Chemistry*, 11, 617-620.

Hou, T.J. and X.J. Xu (2003), "ADME Evaluation in Drug Discovery. 3. Modeling Blood-Brain Barrier Partitioning Using Simple Molecular Descriptors", *Journal of Chemical Information and Computer Sciences*, 43, 2137-2152.

Jacobs, G.A. and M.A.Martens (1989), "An Objective Method for the Evaluation of Eye Irritation in vivo", *Food and Chemical Toxicology*, 27, 255-258.

Kaznessis, Y.N., M.E. Snow and C.J. Blankley (2001), "Prediction of Blood-Brain Partitioning Using Monte Carlo Simulations of Molecules in Water", *Journal of Computer-Aided Molecular Design,* 15, 697-708.

Kier, L.B. and L.H. Hall (1986), *Molecular Connectivity in Structure-Activity Analysis*, Res. Studio Press Ltd, Letchworth, UK.

Kuenemann, P., P. Vasseur and J. Devillers (1990), "Structure-Biodegradability Relationships", in W. Karcher and J Devillers (eds.), *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*, pp. 343-370, ECSC, EEC, EAEC, Brussels and Luxembourg.

Lesigiarska, I., I. Pajeva and S. Yanev (2002), "Quantitative Structure-Activity Relationship (QSAR) and Three-Dimensional QSAR Analysis of a Series of Xanthates as Inhibitors and Inactivators of Cytochrome P450 2B1", *Xenobiotica,* 32, 1063-1077.

Lewis, D.F.V. (1992), "Computer-assisted Methods in the Evaluation of Chemical Toxicity" in K.B. Lipkowitz and D.B. Boyd (eds.), *Reviews in Computational Chemistry*, *Vol. III*, pp. 173-222. VHC Publishers, New York.

Livingstone, D.L. (1995), *Data Analysis for Chemists*, Oxford Science Publications.

McKim, J.M., *et al*. (1987), "Use of Respiratorycardiovascular Responses of Rainbow-trout (*Salmo gairdneri*) in Identifying Acute Toxicity Syndromes in Fish. 1. Pentachlorophenol, 2,4-

Dinitrophenol, Tricaine Methanesulfonate and 1-Octanol", *Environmental Toxicology and Chemistry,* 6, 295-312.

OECD (2002), *OECD Guidelines for the Testing of Chemicals, Test Guideline 405, Acute Eye Irritation/ Corrosion* (adopted in 1981, first revised in 1987, second revised in 2002), OECD, Paris, http://www.oecd.org/document/40/0,2340,en_2649_34365_37051368_1_1_1_1,00.html, accessed 7 February 2007

OECD (2004), *Report from the Expert Group on (Quantitative) Structure-Activity Relationships [(Q)SARs] on the Principles for the Validation of (Q)SARs*, Series on Testing and Assessment, No. 49, OECD, Paris, 206pp, http://www.oecd.org/document/30/0,2340,en_2649_34365_1916638_1_1_1_1,00.html, accessed 6 February 2007

Protic M. and A. Sabljic (1989), "Quantitative Structure-Activity Relationships for Acute Toxicity of Commercial Chemicals on Fathead Minnows: Effect of Molecular Size", *Aquatic Toxicology*, 14, 47-64

Purdy, R. (1991), "The Utility of Computed Superdelocalizability for Predicting the LC_{50} Values of Epoxides to Guppies", *Science of the Total Environment*, 109/110, 553-556.

Raevsky, O.A. and J.C. Dearden (2004), "Creation of Predictive Models of Aquatic Toxicity of Environmental Pollutants with Different Mechanisms of Action on the Basis of Molecular Similarity and HYBOT Descriptors", *SAR and QSAR in Environmental Research*, 15, 433-448.

Roberts, D.W. and D.L. Williams (1982), "The Derivation of Quantitative Correlations between Skin Sensitisation and Physico-Chemical Parameters for Alkylating Agents and their Application to Experimental Data for Sulfones", *Journal of Theoretical Biology*, 99, 807-825.

Sabljic, A. (1991), "Chemical Topology and Ecotoxicology", *Science of the Total Environment*, 109/110, 197-220.

Sanderson, D.M. and C.G. Earnshaw (1991), "Computer Prediction of Possible Toxic Action from Chemical Structure. The DEREK system", *Human and Experimental Toxicology*, 10, 261-273.

Schultz, T.W., *et al*. (1988), "Structure-Toxicity Relationships for Selected Benzyl Alcohols and the Polar Narcosis Mechanism of Toxicity", *Ecotoxicology and Environmental Safety*, 16, 57-64.

Schultz, T.W., D.T. Lin and S.K. Wesley (1992), "QSARs for Monosubstituted Phenols and the Polar Narcosis Mechanism of Toxicity", *Quality Assurance*, 1, 132-143.

Singer, G.M., A.W. Andrews and S.M. Guo (1986), "Quantitative Structure-Activity Relationship of the Mutagenicity of Substituted N-Nitroso-N-benzylmethylamines: Possible Implications for Carcinogenicity", *Journal of Medicinal Chemistry*, 29, 40-44.

Taft, R.W. (1956a), *Steric effects in Organic Chemistry*, Wiley, New York.

Taft, R.W. (1956b), "Separation of Polar, Steric and Resonance Effects in Reactivity", in M.S. Newman (ed.), *Steric Effects in Organic Chemistry*, pp. 556-675, Wiley, New York.

Verhaar, H.J.M., C.J. van Leeuwen and J.L.M. Hermens (1992), "Classifying Environmental Pollutants. 1. Structure-Activity Relationships for Prediction of Aquatic Toxicity", *Chemosphere*, 25, 471-491.

Verhaar, H.J.M., W. Mulder and J.L.M. Hermens (1995), "QSARs for Ecotoxicity", in J.L.M. Hermens (ed.), *Overview of Structure-Activity Relationships for Environmental Endpoints. Part 1: General Outline and Procedure*, Report Prepared within the Framework of the Project "QSAR for Prediction of Fate and Effects of Chemicals in the Environment", Contract with the European Commission EV5V-CT92-0211.

Verhaar, H.J.M., *et al*. (1996), "Modelling the Nucleophilic Reactivity of Organochlorine Electrophiles: a Mechanistically-based Quantitative Structure-Activity Relationship", *Environmental Toxicology and Chemistry*, 16*,* 1011-1018.

ANNEX A. OECD PRINCIPLES FOR THE VALIDATION, FOR REGULATORY PURPOSES, OF (QUANTITATIVE) STRUCTURE-ACTIVITY RELATIONSHIP MODELS

These principles were agreed by OECD member countries at the 37th Joint Meeting of the Chemicals Committee and Working Party on Chemicals, Pesticides and Biotechnology in November 2004. The principles are intended to be read in conjunction with the associated explanatory notes which were also agreed at the 37th Joint Meeting.

To facilitate the consideration of a (Q)SAR model for regulatory purposes, it should be associated with the following information:

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

Notes

1. The intent of Principle 1 (defined endpoint) is to ensure clarity in the endpoint being predicted by a given model, since a given endpoint could be determined by different experimental protocols and under different experimental conditions. It is therefore important to identify the experimental system that is being modeled by the (Q)SAR. Further guidance is being developed regarding the interpretation of "defined endpoint". For example, a no-observed-effect level might be considered to be a defined endpoint in the sense that it is a defined information requirement of a given regulatory guideline, but cannot be regarded as a defined endpoint in the scientific sense of referring to a specific effect within a specific tissue/organ under specified conditions.

2. The intent of Principle 2 (unambiguous algorithm) is to ensure transparency in the model algorithm that generates predictions of an endpoint from information on chemical structure and/or physicochemical properties. It is recognized that, in the case of commercially-developed models, this information is not always made publicly available. However, without this information, the performance of a model cannot be independently established, which is likely to represent a barrier for regulatory acceptance. The issue of reproducibility of the predictions is covered by this Principle, and will be explained further in the guidance material.

3. The need to define an applicability domain (Principle 3) expresses 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 the models can generate reliable predictions. Further work is recommended to define what types of information are needed to define (Q)SAR applicability domains, and to develop appropriate methods for obtaining this information.

4. The revised Principle 4 (appropriate measures of goodness-of–fit, robustness and predictivity) includes the intent of the original Setubal Principles 5 and 6. The wording of the principle is intended to simplify the overall set of principles, but not to lose the distinction between the internal performance of a model (as represented by goodness-of-fit and robustness) and the predictivity of a model (as determined by external validation). It is recommended that detailed guidance be developed on the approaches that could be used to provide appropriate measures of internal performance and predictivity. Further work is recommended to determine what constitutes external validation of (Q)SAR models.

5. It is recognised that it is not always possible, from a scientific viewpoint, to provide a mechanistic interpretation of a given (Q)SAR (Principle 5), or that there even be multiple mechanistic interpretations of a given model. The absence of a mechanistic interpretation for a model does not mean that a model is not potentially useful in the regulatory context. The intent of Principle 5 is not to reject models that have no apparent mechanistic basis, but to ensure that some consideration is given to the possibility of a mechanistic association between the descriptors used in a model and the endpoint being predicted, and to ensure that this association is documented.

ANNEX B. CHECK LIST FOR THE OECD PRINCIPLES FOR (Q)SAR VALIDATION

The OECD Principles for (Q)SAR validation encourage (Q)SARs to be associated with the following information:

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

This annex provides a series of questions associated with each principle, intended to provide an overview of the main considerations associated with the application of each principle. The questions are neither intended to be definitive, nor equally relevant for a given type of model.

CHECK LIST FOR PROVIDING GUIDANCE ON THE INTERPRETATION OF THE OECD PRINCIPLES FOR (Q)SAR VALIDATION

PRINCIPLE	CONSIDERATIONS Is the following information available for the model?	Yes/No/NA
1) Defined endpoint		
1.1	A clear definition of the scientific purpose of the model (*i.e*. does it make predictions of a clearly defined physicochemical, biological or environmental endpoint)?	
1.2	The potential of the model to address (or partially address) a clearly defined regulatory need (*i.e*. does it make predictions of a specific endpoint associated with a specific test method or test guideline)?	
1.3	Important experimental conditions that affect the measurement and therefore the prediction (*e.g*. sex, species, temperature, exposure period, protocol)?	
1.4	The units of measurement of the endpoint?	
2) Defined algorithm		
2.1	In the case of a SAR, an explicit description of the substructure, including an explicit identification of its substituents?	
2.2	In the case of a QSAR, an explicit definition of the equation, including definitions of all descriptors?	
3) Defined domain of applicability		
3.1	In the case of a SAR, a description of any limits on its applicability (*e.g*. inclusion and/or exclusion rules regarding the chemical classes to which the substructure is applicable)?	
3.2	In the case of a SAR, rules describing the modulatory effects of the substructure's molecular environment?	
3.3	In the case of a QSAR, inclusion and/or exclusion rules that define the following variable ranges for which the QSAR is applicable (*i.e*. makes reliable estimates): a) descriptor variables? b) response variables?	
3.4	A (graphical) expression of how the descriptor values of the chemicals in the training set are distributed in relation to the endpoint values predicted by the model?	

4A) Internal performance		
4.1	Full details of the training set given, including details of: a) number of training structures b) chemical names c) structural formulae d) CAS numbers e) data for all descriptor variables f) data for all response variables g) an indication of the quality of the training data?	
4.2	a) An indication whether the data used to the develop the model were based upon the processing of raw data (*e.g.* the averaging of replicate values) b) If yes to a), are the raw data provided? c) If yes to a), is the data processing method described?	
4.3	An explanation of the approach used to select the descriptors, including: a) the approach used to select the initial set of descriptors b) the initial number of descriptors considered c) the approach used to select a smaller, final set of descriptors from a larger, initial set d) the final number of descriptors included in the model ?	
4.4	a) A specification of the statistical method(s) used to develop the model (including details of any software packages used) b) If yes to a), an indication whether the model has been independently confirmed (*i.e.* that the independent application of the described statistical method to the training set results in the same model)?	
4.5	Basic statistics for the goodness-of-fit of the model to its training set (*e.g.* r^2 values and standard error of the estimate in the case of regression models)?	
4.6	a) An indication whether cross-validation or resampling was performed b) If yes to a), are cross-validated statistics provided, and by which method? c) If yes to a), is the resampling method described?	
4.7	An assessment of the internal performance of the model in relation to the quality of the training set, and/or the known variability in the response?	

4B) Predictivity		
4.8	An indication whether the model has been validated by using a test set that is independent of the training set?	
4.9	If an external validation has been performed (yes to 4.8), full details of the test set, including details of: a) number of test structures b) chemical names c) structural formulae d) CAS numbers e) data for all descriptor variables f) data for all response variables g) an indication of the quality of the test data?	
4.10	If an external validation has been performed (yes to 4.8): a) an explanation of the approach used to select the test structures, including a specification of how the applicability domain of the model is represented by the test set ? b) was the external set *sufficiently large and representative* of the training data set? c) a specification of the statistical method(s) used to assess the predictive performance of the model (including details of any software packages used) d) a statistical analysis of the predictive performance of the model (*e.g.* including sensitivity, specificity, and positive and negative predictivities for classification models) e) an evaluation of the predictive performance of the model that takes into account the quality of the training and test sets, and/or the known variability in the response f) a comparison of the predictive performance of the model against previously-defined quantitative performance criteria?	
5) Mechanistic interpretation		
5.1	In the case of a SAR, a description of the molecular events that underlie the properties of molecules containing the substructure (*e.g.* a description of how substructural features could act as nucleophiles or electrophiles, or form part or all of a receptor-binding region)?	
5.2	In the case of a QSAR, a physicochemical interpretation of the descriptors that is consistent with a known mechanism of (biological) action?	

5.3	Literature references that support the (purported) mechanistic basis?	
5.4	An indication whether the mechanistic basis of the model was determined *a priori* (*i.e*. before modelling, by ensuring that the initial set of training structures and/or descriptors were selected to fit a pre-defined mechanism of action) or *a posteriori* (*i.e*. after the modelling, by interpretation of the final set of training structures and/or descriptors) ?	

ANNEX C. REPORTING FORMATS FOR (Q)SARS VALIDATION

Introduction

A (Q)SAR Model Reporting Format (QMRF) is a framework for structuring and summarising key information about a model, to provide the end-user with details on: a) the source of the model (including the developer, where known); b) model type; c) model definition; d) the development of model; e) the validation of the model; and f) possible applications of the model.

The QMRF should be regarded as a communication tool. It draws on some of the information provided by applying the OECD Principles for QSAR validation, but it is not intended in itself to be a complete characterisation of the model.

The QMRF should involve an input from the developer(s) and/or proponent of the model, as well as information from any evaluation studies performed with the model.

QMRFs will need to include specific information associated with particular kinds of models. It therefore needs to be investigated, depending on the level of resolution desired in the format, whether a single format can be applied to all models, or whether certain kinds of models (*e.g.* MultiCase models) will need additional fields to capture model-specific information.

The QMRF should not be confused with the reporting formats used to provide QSAR estimates for chemicals that are registered/notified within a given regulatory programme, even though such formats are likely contain similar information fields.

Case Studies

In this Annex, case studies for reporting QSAR validation using the QMRF are given for the following models:

1. Derek for Windows Model for Skin Sensitisation
2. MULTICASE Model for In Vitro Chromosomal Aberrations in Mammalian Cells
3. Fish Acute Neutral Organics 96-hour (Q)SAR, a constituent of ECOSAR
4. CATABOL for Biodegradation
5. BIOWIN for Biodegradation

QPRF and TERF

The (Q)SAR Prediction Reporting Format (QPRF) will explain how an estimate has been derived by applying a specific model or method to a specific substance. This should include information on the model prediction(s), including the endpoint, a precise identification of the substance modelled, the relationship between the modelled substance and the defined applicability domain, and the identities of close analogues.

In the overall assessment of a given chemical, it will often be necessary to integrate the QSAR estimates with other sources of information (*e.g*. in vitro and in vivo test data). This data integration should be based on "weight-of-evidence" considerations, which could be perhaps better thought of as "totality-of-evidence" considerations, because it is not necessarily the case that weights will be attached to individual pieces of information. It is proposed that this level of integration should be documented in detail in a Totality of Evidence Reporting Format (TERF). The reasoning and assessments for a given substance and given endpoint included in the QMRF and QPRF (or multiple QMRFs and QPRFs) could be carried over to (or referenced in) the TERF. Collectively, these three levels of reporting formats would provide an comprehensive description of the use of the (Q)SAR and other approaches applied during the risk assessment of a given substance for a specific endpoint.

No definitive formats for QPRF and TERF are proposed in this document. These types of formats are likely to evolve over time. Draft versions have been developed by the European Chemicals Bureau. Details and developments could be found on the website of the European Chemicals Bureau's QSAR Action [http://ecb.jrc.it/QSAR/, accessed 7 February 2007].

CASE STUDY 1: Derek for Windows Model for Skin Sensitisation

1. QSAR identifier

Derek for Windows skin sensitisation rulebase. Version No 9

Note: Version 9 is the latest version. The reporting format is written largely independent of the version in the case of Derek.

2. Source

2.1 Reference(s) to scientific papers and/or software package:

- Barratt, M.D., *et al*. (1994), "An Expert System Rulebase for Identifying Contact Allergens", *Toxicology in Vitro*, 8, 1053-1060.
- Barratt, M.D. and J.J. Langowski (1999), "Validation and Subsequent Development of the Derek Skin Sensitisation Rulebase by Analysis of the BgVV List of Contact Allergens", *Journal of Chemical Information and Computer Science*, 39, 294-298.
- Greene, N. (2002), "Computer Systems for the Prediction of Toxicity: an Update", *Advanced Drug Delivery Reviews*, 54, 417-431.
- Greene, N., *et al*. (1999), "Knowledge-based Expert Systems for Toxicity and Metabolism Prediction: DEREKfW, StAR and METEOR", *SAR and QSAR in Environmental Research*, 10, 299-314.
- Sanderson, D.M. and C.G. Earnshaw (1991), "Computer Prediction of Possible Toxic Action from Chemical Structure; The DEREK System", *Human and Experimental Toxicology*, 10, 261-273.
- Zinke, S., I. Gerner and E. Schlede (2002), "Evaluation of a Rule Base for Identifying Contact Allergens by Using a Regulatory Database: Comparison of Data on Chemicals Notified in the European Union with 'Structural Alerts' Used in the DEREKfw Expert System", *ATLA* 30, 285-298.

2.2 Date of publication:

A number of publications though key dates are notably 1986 when the first Derek system was created at Schering Agrochemicals in the UK and 1989 when Lhasa Ltd. adopted the Derek system and began coordinating the main development of the structure-toxicity knowledge base.

2.3 Identification of the model developer(s)/authors:

Lhasa Limited
LHASA is the acronym for Logic and Heuristics Applied to Synthetic Analysis.

2.4 Contact details of the model developer(s)/authors:

22-23 Blenheim Terrace,
Woodhouse Lane,
Leeds LS2 9HD
UK

Tel: +44 (0)113 394 6020
Fax: +44 (0)113 394 6099
Email: info@lhasalimited.org
Web: www.lhasalimited.org

2.5 Indication of whether the model is proprietary or non-proprietary:

Model is proprietary, the datasets within are taken from both public and proprietary sources.

3. Type of model

3.1 ☑ 2D SAR
3.2 ☐ 3D SAR (*e.g*. pharmacophore)
3.3 ☐ Regression-based QSAR
3.4 ☐ 3D QSAR
3.5 ☐ Battery of (Q)SARs
(overall prediction depends on application of multiple models/rules)
3.6 ☑ Expert system
(overall prediction depends on application of multiple models/rules and use of data in a knowledge base)
3.7 ☐ Neural network
3.8 ☐ Other

4. Definition of the model

4.1 Dependent variable:

4.1.1 Species

The relevant test guideline determines which species underpins the toxicity information used. In the case of skin sensitisation, this will be predominantly the guinea pig and the mouse from Guinea Pig Maximisation Tests (GPMT)/Buehler and Local Lymph Node Assay (LLNA) tests. More detailed information on these test protocols can be found in OECD Guidelines 406 and 429 respectively. In addition there will be some rules that are based on human data (*e.g.* from the Human Repeat Insult Test (HRIPT) or maximisation test).

4.1.2 Endpoint (including exposure time)

The endpoint modelled is the overall endpoint without specific reference to a given test guideline. Derek makes qualitative predictions of skin sensitisation using a range of different and available data principally that from the public domain but additionally proprietary data from its members where feasible. The data might be categorical in nature in providing a yes/no answer of whether a compound is a sensitiser *e.g*. a R43 classification or it might be quantitative providing a measure of relative potency *e.g*. an EC3 from the LLNA. In all cases, Derek provides an estimate of the presence of a potential skin sensitisation

hazard in the form of structural alerts and related supporting information. It does not provide information on the relative potency of a skin sensitiser *i.e*. a compound might be identified as a sensitiser but Derek will not discriminate between a weak, extreme or moderate sensitiser.

4.1.3 Units of measurement

Qualitative predictions are made which do not incorporate any specific unit of measurement.

4.1.4 Reference to specific experimental protocol(s)

The skin sensitisation data encoded within Derek includes both public and proprietary data generated through a number of different test methods including GPMT, Buehler, LLNA, Mouse Ear Swelling test, human maximisation test as well as the HRIPT. The earlier alerts were based on GPMT, more recent alerts have been based on LLNA data. Information about the experimental conditions is only provided in the references associated with a given alert. Since only a subset of these is fully referenced, the quality of the data used in the derivation of an alert can not be fully verified. However where possible and practically feasible - the data is evaluated by Lhasa Ltd for its quality, robustness and suitability of use within an alert.

4.2 Number of descriptors used as independent variables:

There are no independent variables, and no mathematical equation as this is a SAR model. Some of the structural alerts are published (see Barratt *et al*. (1994, 1999) of Section 2.1 for examples).

4.3 Identification of descriptors (names, symbols):

Not applicable

4.4 Explicit algorithm for generating predictions from the descriptors:

Derek provides an explicit description of the substructure and substituents. When a query structure is processed, the alerts that match are displayed in a hierarchy called the prediction tree and are highlighted in bold in the query structure. The prediction tree includes the endpoint, the species and reasoning outcome, the number and name of the alert, and the example from the knowledge base if it exactly matches the query structure. The alert description provides a description depicting the structural requirement for the toxicophore detected and a reference to show the bibliographic references used. Some rules are extremely general with substructures only taking into account the immediate environment of a functional group. In other cases, the descriptions are much more specific. This means that remote fragments that may modulate sensitisation are not always taken into consideration in the assessment.

4.5 Goodness-of-fit statistics

Derek does not provide the full details of the training data used to develop an alert. Only a subset of the references and example chemicals used to develop the alert are provided for illustrative purposes.

4.6 Information on the applicability domain of the model

Derek includes some inclusion/exclusion rules associated with an alert. These are documented in the alert description as particular substituents. For some sensitisation rules there are very clear descriptions of what is covered by a specific substructure, in other cases the rules are extremely general, *e.g*. alpha,beta-unsaturated carbonyls vs. alkyl halides. Physicochemical parameters namely Log Kow and Molecular Weight are used to limit the domain by accounting for skin penetration. Whilst the domain has not been

defined as such, Derek is able to make reasonable estimates for many organic compounds and metals. It can not make predictions for polymers. There are no negative alerts for skin sensitisation.

4.7 Information on the mechanistic basis/interpretation of the model

All the rules in Derek are based on either hypotheses relating to mechanisms of action of a chemical class or observed empirical relationships, the ideas for which come from a variety of sources, including published data or suggestions from the Derek collaborative group. This group consists of toxicologists who represent Lhasa Ltd. and members who meet at regular intervals to give advice and guidance on the rule development work and predictions made by the program. The hypotheses underpinning each alert are documented in the alert descriptions as comments. These comments often include descriptions of features acting as electrophiles or nucleophiles. However, the detail depends on the specific alert. Some alerts contain no comments, aside from the modulating factors of skin penetration.

5. Development of the model

5.1 Explanation of the method (approach) used to generate each descriptor

Any information would be found in the comments section of the alert but this is not systemically provided.

5.2 Selection of descriptors

5.2.1 Indication of initial number of descriptors screened

Not applicable

5.2.2 Explanation of the method (approach) used to select the descriptors and develop the model from them

Not applicable.

5.2.3 Indication of final number of descriptors included in the model:

Not applicable

5.3 Information on experimental design for data splitting into training and validation sets.

Not applicable

5.4 Availability of the training set

5.4.1 ☐ Chemical names (common names and/or IUPAC names)
5.4.2 ☐ CAS numbers
5.4.3 ☐ 1D representation of chemical structure (*e.g.* SMILES)
5.4.4 ☐ 2D representation of chemical structure (*e.g.* ISIS sketch file)
5.4.5 ☐ 3D representation of chemical structure (*e.g.* MOL file)
5.4.6 ☐ Data for each descriptor variable
5.4.7 ☐ Data for the dependent variable

Derek rules describe generalised structure-activity relationships and do not record internally the specific chemical structures on which they are based. Derek is a knowledge base as opposed to a database.

This means it is possible to use data from confidential sources as a basis for new rules without revealing exact chemicals to end-users. This provides a means by which proprietary data can be used without revealing potentially sensitive information. This is a clear advantage for the purposes of securing business confidentially, but reduces the transparency of the system.

The training set information visible to the enduser is limited to a few key example compounds that illustrate the scope of the alert. The number of examples is dependent on the sensitisation alert, some alerts may have no examples. Where examples are provided, the CAS#, name, test result (summary data), bibliographic reference and 2D structural representation are provided.

6. Validation of the model

6.1 Statistics obtained by leave-one-out cross-validation

None

6.2 Statistics obtained by leave-many-out cross-validation

None

6.3 Statistics obtained by Y-scrambling

None

6.4 Statistics obtained by external validation

Some "external" validation studies have been performed to evaluate the performance of Derek. "External", as in most cases the aim has been to take a dataset of chemicals of interest to a company/organisation etc and evaluate how well Derek performs for those specific chemicals. In this way, the evaluation has not been designed to consider the real applicability domain of Derek *i.e*. the scope of the training sets within Derek. Some recent exercises are described below. In some of these exercises, part or all the testset used is provided.

Validation by Zinke et al.(2002)

An external validation for skin sensitisation, using the BgVV database, was performed by Zinke *et al*. (2002). The BgVV database includes 1039 chemicals that have reliable data for the assessment of skin sensitising potential. The results indicated a concordance of 67%, a sensitivity of 37% (*i.e*. a false negative rate of 63%) and a specificity of 85% (*i.e*. a false positive rate of 15%). Zinke *et al*. gave comments on which structural alerts worked well, which needed to be adapted, and which might need to be left out.

Validation by Seaman et al. (2001)

A total of 78 chemicals which underwent testing using the LLNA to identify moderate and severe skin sensitisers were also evaluated with Derek by Seaman *et al*. (2001).They obtained a concordance of 59%, a sensitivity of 79% and a specificity of 47%. A total of 39 of the 49 LLNA negatives were then examined in the Guinea Pig maximisation test (GMPT). The LLNA missed 15 GPMT positives. By excluding the LLNA negatives that were Derek positive, the number of false negatives was decreased by 10 to 5/39 (15%) although this addition introduced 11 false positives.

Validation using 89 chemicals from Henkel

A total of 89 compounds (mostly aromatic amines) taken from Henkel were evaluated using Derek v 3.6.0 (Delbanco, 2002). Previously these chemicals had undergone experimental testing using the guinea pig maximisation test (GPMT) and/or Buehler test (BT). Overall the predictions of Derek were in concordance with about 42% of the sensitisers and non-sensitisers when compared to the results of both test types or to the results of each test system. The Derek software was over predictive for skin sensitisation, which was shown by many false positive predictions.

Validation using 80 chemicals from IUCLID

The application of Derek v 5.01 for predicting skin sensitisation potential has also been examined using a set of 80 substances from the IUCLID database for which guinea pig maximisation test results have been published (ECETOC, 2003). The results indicated a concordance of 62.5%, a sensitivity of 62.5%, and a specificity of 62.5%.

6.5 Definition of the applicability domain of the model

Approach for establishing the applicability domain of the model is yet to be defined. Some principles/approaches were discussed in an ECVAM workshop on applicability domains (Netzeva *et al*., 2005).

6.6 Availability of the external validation set

6.6.1 ☐ Chemical names (common names and/or IUPAC names)
6.6.2 ☐ CAS numbers
6.6.3 ☐ 1D representation of chemical structure (*e.g*. SMILES)
6.6.4 ☐ 2D representation of chemical structure (*e.g*. ISIS sketch file)
6.6.5 ☐ 3D representation of chemical structure (*e.g*. MOL file)
6.6.6 ☐ Data for each descriptor variable
6.6.7 ☐ Data for the dependent variable

Variable access to all or part of the data in each case.

7. Applications of the model

Suggestions for possible applications for the model:

Predicting likely skin sensitisation hazard on a case by case or HTS basis, provide mechanistic insights; *i.e*. to screen out undesirable chemicals on the basis of sensitisation and to examine potential mechanisms of action to explain why a given query chemical was potentially sensitising.

8. Miscellaneous information

- Derek is essentially a knowledge archive of structure-toxicity relationships.

- Derek is limited in that it identifies only ‘activating’ fragments, meaning the negative prediction is based solely on the lack of structural alerts. Only qualitative outcomes are provided, no measure of potency is provided. Training sets of chemicals containing these structural alerts are not provided. Derek does not provide a comprehensive list of references used in the development

of each alert. Insufficient information is provided about the quality of the data used in the development of each alert.

- No clear explanation of the domain of applicability is provided that would alert the user as to when a query structure was within or outside the chemical domain of Derek.

- Some of the alerts within Derek are very general, explaining the high number of false positives in the external validation studies.

- Derek covers a small subset of chemical space, a huge number of rules would need to be developed in order to account for each chemical class. Development of Derek is incremental, focusing on each chemical class in turn. Derek would improve from adding more information about the modulating factors in the environment of an alert such as remote groups or by calculation of other physiochemical descriptors.

9. References

Delbanco, E.H. (2002), "Use of the Prediction Software DEREK in the Hazard Assessment of Raw Materials", *Naunyn Schmiedeberg's Archive Pharmacology,* Suppl 365, R 639.

ECETOC (2003), *QSARs: Evaluation of the Commercially Available Software for Human Health and Environmental Endpoints with Respect to Chemical Management Applications,* ECETOC Technical Report No. 89.

Netzeva, T.I., *et al*. (2005), "Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure-Activity Relationships. The Report and Recommendations of ECVAM Workshop 52", *Alternatives To Laboratory Animals*, 33, 155-173.

Seaman, C.W., F.J. Guerriero and G.L. Sprague (2001), "The Use of DEREK (a Structure/Toxicity Prediction Program) in the Identification of Skin Sensitisers", *Toxicologist*, 60, 1452.

Zinke, S., I. Gerner and E. Schlede (2002), "Evaluation of a Rule Base for Identifying Contact Allergens by Using a Regulatory Database: Comparison of Data on Chemicals Notified in the European Union with 'Structural Alerts' Used in the DEREKfW Expert System", *Alternatives To Laboratory Animals* 30, 285-298.

CASE STUDY 2: MULTICASE Model for In Vitro Chromosomal Aberrations in Mammalian Cells

1. QSAR identifier

MultiCASE MC4PC Release 2003

Chromosomal Aberration Test In Vitro

2. Source

2.1 Reference(s) to scientific papers and/or software package:

- Klopman, G. (1992), "Multicase, 1. A Hierarchial Computer Automated Structure Evaluation Program", *Quant. Struct. Act. Relat*., 11, 176 – 184.
- Kusakabe, H., *et al*. (2002), "Relevence of Chemical Structure and Cytotoxicity to the Induction of Chromosome Aberrations Based on Testing of 98 High Production Volume Industrial Chemicals", *Mutation Research*, 517, 187-198.
- Niemelä, J. and E. Wedeby (2004), "Evaluation of the Setubal Principl for Establishing the Status of Development and Validation of (Q)SARs, Annex 4, A "Global" MULTI-CASE Model for in vitro Chromosomal Aberrations in Mammalian Cells", in OECD, *Report from the Expert Group on (Quantitative) Structure-Activity Relationships [(Q)SARs] on the Principles for the Validation of (Q)SARs*, Series on Testing and Assessment, No. 49, OECD, Paris, pp113-133, http://www.oecd.org/document/30/0,2340,en_2649_34365_1916638_1_1_1_1,00.html, accessed 6 February 2007.
- OECD (1997), *OECD Guidelines for the Testing of Chemicals, Test Guideline 473, In Vitro Mammalian Chromosome Aberration Test,* http://www.oecd.org/document/40/0,2340,en_2649_34365_37051368_1_1_1_1,00.html, accessed 7 February 2007
- Sofuni, T. (ed.) (1998), *Data Book of Chromosomal Aberration Test In Vitro, Revised Edition*. Life-Science Information Center, Tokyo, Japan.
- Ishidate, M. Jr. (ed.) (1988), *Data Book of Chromosomal Aberration Test In Vitro, Revised Edition,* Elsevier, Amsterdam, New York, Oxford.

2.2 Date of publication:

Publication of the model: In 2004: Niemelä, J. and Wedebye,E.

2.3 Identification of the model developer(s)/authors:

Jay Niemelä,
Danish Institute for Food and Veterinary Research, Toxicology and Risk Assessment Division

Eva Bay Wedebye,
Danish Institute for Food and Veterinary Research, Toxicology and Risk Assessment Division

2.4 Contact details of the model developer(s)/authors:

Mørkhøj Bygade 19,
DK – 2860 Søborg,
Denmark

Tel: +45 (0) 72 34 75 92 , Jay Niemelä
Email: jayni@dfvf.dk

Tel: +45 (0) 72 34 76 04, Eva Bay Wedebye
Email: ebawe@dfvf.dk

Fax: +45 (0) 72 34 70 01
Web: www.dfvf.dk

2.5 Indication of whether the model is proprietary or non-proprietary:

Proprietary

3. Type of model

3.1 ☐ 2D SAR
3.2 ☐ 3D SAR (*e.g*. pharmacophore)
3.3 ☐ Regression-based QSAR
3.4 ☐ 3D QSAR
3.5 ☐ Battery of (Q)SARs
(overall prediction depends on application of multiple models/rules)
3.6 ☑ Expert system
(overall prediction depends on application of multiple models/rules and use of data in a knowledge base)
3.7 ☐ Neural network
3.8 ☐ Other

4. Definition of the model

4.1 Dependent variable:

4.1.1 Species

All tests were performed using a Chinese Hamster Lung Cell (CHL) fibroblast cell line, which has been kept as a single cell sub-clone since 1973 (Sofuni, 1998).

4.1.2 Endpoint (including exposure time)

The endpoint used was *Chromosomal Aberration Test In Vitro* in order to identify agents that cause structural chromosome aberrations in cultured mammalian cells, visible in light microscopy. The test system and its purpose are described in OECD Guideline for the Testing of Chemicals, No. 473. The current Test Guideline does not specify testing for a length of time (OECD, 1997).

4.1.3 Units of measurement

In the Data Book (Ishidate, 1988; Sofuni, 1998) results in the experimental studies were indicated as positive (active) or negative (inactive).

In MULTICASE during breaking down the structures of the training set all fragments "produced" are assigned a MULTICASE activity score according to the activity of the parent structure (Klopmann, 1992). If the parent compound is "inactive it is assigned a score of 10, while fragments from active parents are given a score of 45.

4.1.4 Reference to specific experimental protocol(s)

- OECD Guidelines for the Testing of Chemicals, Test Guideline 473, In Vitro Mammalian Chromosome Aberration Test (OECD, 1997), describing guidelines for the experimental studies
- Data Book of Chromosomal Aberration Test in Vitro (Ishidate, 1988; Sofuni, 1998), reporting the experimental results
- A "global" MULTICASE model for in vitro chromosomal aberrations in mammalian cells (Niemelä and Wedebye, 2004), the selection of data for modelling is described

4.2 Number of descriptors used as independent variables:

Not applicable

4.3 Identification of descriptors (names, symbols):

Not applicable

4.4 Explicit algorithm for generating predictions from the descriptors:

MULTICASE is a fragment-based statistical model system. The methodology involves breaking down the structures of the training set into all possible fragments from 2 to 10 heavy (non-hydrogen) atoms in length.

Fragments from the entire training set are combined into gross activity categories. A structural fragment is considered as a "biophore" if it has a statistical association with chemicals in the active category. It is considered a "biophobe" if it has a statistically significant relation with the inactive category.

4.5 Goodness-of-fit statistics

Internal performance for predictions within the domain

	Active	**Inactive**	**Total**	**Accuracy**
Predicted +	241	2	243	99.2%
Predicted -	1	238	239	99.6%
Total	242	240	482	
Percentage	99.6 (sensitivity)	99.2 (specificity)		

Results excluding the 31 inconclusive values

Chi square = 470.082; Phi square = 0.975

Expected Correct Predictions (ECP) = 50.00 %
Observed Correct Predictions (OCP) = 99.38 %

Footnote
Phi-square is the Pearson Chi-square, divided by the number of cases. It has value 0 if there is no association, and a value of 1 of there is a perfect association.

4.6 Information on the applicability domain of the model

During the prediction process for a substance for chromosomal aberration MULTICASE provide warnings if the substance is outside the domain of the model. Warnings may be due to presence of fragments not present in the training set and not covered by the model, or the presence of inactivating fragments associated with an active prediction (or the opposite). It is up to the user to take account of these warnings or not, we consider any MULTICASE warning to be an indication that that the molecule being predicted is outside of the model domain.

4.7 Information on the mechanistic basis/interpretation of the model

The exact mechanism of action of the chemicals causing chromosomal aberration is not known, but it is assumed that a covalent reaction with a biological macromolecule (*e.g.* DNA) may be involved. Many resulting predictions have mode of action that are obvious for the person with expert knowledge for the endpoint in question. Knowledge to mode of action is extremely desirable in the final evaluation of predictions.

5. Development of the model

5.1 Explanation of the method (approach) used to generate each descriptor

Not appropriate.

5.2 Selection of descriptors

5.2.1 Indication of initial number of descriptors screened

Not applicable

5.2.2 Explanation of the method (approach) used to select the descriptors and develop the model from them

See 4.4

5.2.3 Indication of final number of descriptors included in the model:

Not applicable

5.3 Information on experimental design for data splitting into training and validation sets.

Not applicable

5.4 Availability of the training set

5.4.1 ☑ Chemical names (common names and/or IUPAC names)
5.4.2 ☑ CAS numbers
5.4.3 ☑ 1D representation of chemical structure (*e.g.* SMILES)
5.4.4 ☐ 2D representation of chemical structure (*e.g.* ISIS sketch file)
5.4.5 ☐ 3D representation of chemical structure (*e.g.* MOL file)
5.4.6 ☐ Data for each descriptor variable
5.4.7 ☑ Data for the dependent variable

Out of 911 substances from the Data Book (Sofuni, 1998), 513 were used to establish the model. The exclusion criteria used include inorganic status, inadequate smile code, etc. A decision was made to include chemicals as being positive if they were active in inducing either aberrations or polyploidy (Niemelä and Wedebye, 2004). Polyploidy is not included in the current Test Guideline (OECD, 1997).

6. Validation of the model

6.1 Statistics obtained by leave-one-out cross-validation

None

6.2 Statistics obtained by leave-many-out cross-validation

10x10% cross-validation

Taking account of the model's ability to identify the domain the following results were obtained:
Sensitivity = (98/155) x 100 = 63.23%
Specificity = (155/180) x 100 = 86.11%
Concordance = (253/335) x 100 = 75.52%

100x50% cross-validation

Taking the domain into account, we obtained, for 14619 predictions within the model domain as defined above:
Sensitivity = (4431/6934) x 100 = 63.90%
Specificity = (6410/7684) x 100 = 83.42%
Concordance = (10841/14618) x 100 = 74.16%

6.3 Statistics obtained by Y-scrambling

As a further check on model performance, we randomly scrambled the toxicity scores in our training set of 513 chemicals, and performed 10 cross-validations, leaving out 50% of the chemicals in each cross-validation. None of the resulting validations was statistically significant. The Chi Square value averaged 0.7126 (probability = ca. 0.4). For chemicals, estimated as being within the domain, concordance was 49.69%.

6.4 Statistics obtained by external validation

The statistical analysis for specificity, sensitivity and concordance gave results that were broadly similar to the cross-validations.

The initial data (see 6.6) comprised 98 substances which was reduced to 62 due different reasons such as some of the chemicals were included in the training set, some were only active at very high

concentrations, or chromosomal aberrations were only induced under non-physiological culture conditions (ex. pH<6).

The results for the 62 chemicals within the domain are:
Sensitivity = (10/17) x 100 = 58.82%
Specificity = (37/45) x 100 = 82.22%
Concordance = (47/62) x 100 = 75.81%

6.5 Definition of the applicability domain of the model

From MULTICASE warnings the domain of the model is defined; see 4.6.

6.6 Availability of the external validation set

6.6.1 ☑ Chemical names (common names and/or IUPAC names)
6.6.2 ☑ CAS numbers
6.6.3 ☑ 1D representation of chemical structure (*e.g.* SMILES)
6.6.4 ☐ 2D representation of chemical structure (*e.g.* ISIS sketch file)
6.6.5 ☐ 3D representation of chemical structure (*e.g.* MOL file)
6.6.6 ☐ Data for each descriptor variable
6.6.7 ☑ Data for the dependent variable

For external validation, we used data generated over a six-year period (1991-1996) for chromosomal aberration testing of high production volume (HPV) industrial chemicals that had been conducted using Chinese hamster lung (CHL/IU) cells according to the OECD HPV testing program and the national program in Japan (Kusakabe *et al*., 2002).

7. Applications of the model

Suggestions for possible applications for the model:

Predicting for Chromosomal Aberrations in mammalian cells in vitro.

8. Miscellaneous information

9. References

See 2.1

CASE STUDY 3: Fish Acute Neutral Organics 96-hour (Q)SAR, a constituent of ECOSAR

1. QSAR identifier

ECOlogical Structure Activity Relationships (ECOSAR)

Acute Fish 96-hour LC_{50} – Neutral Organics

MS-Windows – Version 0.99h.

The new updated version of ECOSAR (Version 1.00) is scheduled for release in 2007.

PLEASE NOTE: The (Q)SAR under evaluation in this case study is only one of many available in the ECOSAR program. The evaluation, statistics, and data presented are only applicable to the acute fish 96-hour LC_{50} (Q)SAR, and not other (Q)SARs available within the program.

2. Source

2.1 Reference(s) to scientific papers and/or software package:

- **ECOSAR Program:**

  Publicly available for download at: http:www.epa.gov/opptintr/exposure/pubs/episuite.htm, accessed 7 February 2007. [Note: This URL is the site of EPI (Estimation Programs Interface) Suite Version 3.12 (released 8 December 2005) which includes ECOSAR v.0.99h. Once EPI 3.12 is downloaded, ECOSAR v.0.99h can be run separetely.]

- **User's Manual:**

  User's Guide for the ECOSAR Class Program (1998), Risk Assessment Division (7403), Office of Pollution Prevention and Toxics, U.S. Environmental Protection Agency, 1200 Pennsylvania Ave., N.W., Washington, DC 20460.
  Available at: http://www.epa.gov/oppt/newchems/tools/manual.pdf (accessed 7 February 2007)

- **Technical Reference Manual:**

  Estimating Toxicity of Industrial Chemicals to Aquatic Organisms Using Structure Activity Relationships (1996), Environmental Effects Branch, Health and Environmental Review Division, Office of Pollution Prevention and Toxics, U.S. Environmental Protection Agency, Washington, DC 20460.
  Available at: http://www.epa.gov/oppt/newchems/tools/sarman.pdf, accessed 7 February 2007

2.2 Date of publication:

Publication of Model: 8 December 2005

2.3 Identification of the model developer(s)/authors:

J. Vincent Nabholz
U.S. Environmental Protection Agency, OPPT Risk Assessment Division

Gordon G. Cash
U.S. Environmental Protection Agency, OPPT Risk Assessment Division

Bill Meylan
Syracuse Research Corporation

Phil Howard
Syracuse Research Corporation

2.4 Contact details of the model developer(s)/authors:

2.4.1 Technical Contacts

J. Vincent Nabholz
Risk Assessment Division
U.S. Environmental Protection Agency
Ariel Rios Building, Mail Code: 7403M
1200 Pennsylvania Avenue, N.W.
Washington, DC 20460
Phone: 202 564-8909
Email: nabholz.joe@epa.gov

Gordon G. Cash
Risk Assessment Division
Ariel Rios Building, Mail Code: 7403M
1200 Pennsylvania Avenue, N.W.
Washington, DC 20460
Phone: 202 564-8923
Email: cash.gordon@epa.gov

2.4.2 Model Contacts

Bill Meylan
Syracuse Research Corporation
Environmental Science Center
301 Plainfield Road, Suite 350
Syracuse, NY 13212
Phone: 315 452-8421
Fax: 315 452-8440

Philip H. Howard
Syracuse Research Corporation
Environmental Science Center
301 Plainfield Road, Suite 350
Syracuse, NY 13212
Phone: 315 452-8417
Fax: 315 452-8440

2.5 Indication of whether the model is proprietary or non-proprietary:

Non-proprietary

3. Type of model

3.1 ☐ 2D SAR
3.2 ☐ 3D SAR (*e.g*. pharmacophore)
3.3 ☑ Regression-based QSAR
3.4 ☐ 3D QSAR
3.5 ☐ Battery of (Q)SARs
(overall prediction depends on application of multiple models/rules)
3.6 ☑ Expert system
(overall prediction depends on application of multiple models/rules and use of data in a knowledge base)
3.7 ☐ Neural network
3.8 ☐ Other

4. Definition of the model

4.1 Dependent variable:

4.1.1 Species

Standard test species were used when developing this model. (Q)SARs are specific for the effect modeled, but not specific with respect to a single species.

4.1.2 Endpoint (including exposure time)

Acute Fish 96-hour LC_{50}: the aqueous concentration predicted to kill 50% of a population following a 96-hour exposure period.

4.1.3 Units of measurement

LC_{50} values are in presented in mg/L.

4.1.4 Reference to specific experimental protocol(s):

OPPTS850.1075 Fish acute toxicity test, freshwater and marine
40CFR797.1400 Fish acute toxicity test, freshwater and marine
OECD TG 203: Fish, acute toxicity test
These guidelines are preferred but not obligatory.
All test data are validated prior to inclusion regardless of test protocol.

4.2 Number of descriptors used as independent variables:

Two independent variables, the Log of the octanol-water partition coefficient (Log Kow) and the molecular weight (MW) are required to predict the acute fish 96-hour LC_{50}. The descriptors are calculated from the chemical structure, obtained through input of CAS RN or SMILES notation (Refer to Section 5.1.2) entered in the initial data entry screen.

A searchable database of CAS RNs and corresponding SMILES structures are provided within the ECOSAR program. CAS RNs are available for approximately 103,000 discrete organic chemicals. If a CAS RN is not available, a SMILES notation can be directly entered by the user. The encoding rules for SMILES are located in the help menu of the ECOSAR data entry page, as well as at
http://www.syrres.com/esc/smilecas.htm (accessed 8 February 2007)

Additional batch mode data entry options are available as well for importing structural information.

4.3 Identification of descriptors (names, symbols):

- Log of Octanol/water partition coefficient, Log K_{ow} (or Log P)
- Molecular weight, MW

4.4 Explicit algorithm for generating predictions from the descriptors:

For chemicals with a Log Kow of less than 5.0:

- Neutral Organics, Acute Fish 96-hour (Q)SAR: Log LC_{50} = -0.862 (Log Kow) + 1.6108
- The LC_{50} predictions from this equation are presented in millimoles per liter (mM/L). ECOSAR then converts the LC_{50} from mM/L to mg/L, by multiplying LC_{50} value by the molecular weight of the compound.
- Neutral Organics, Acute Fish 96-hour (Q)SAR: Log (LC_{50}/MW) = -0.862 (Log Kow) +1.6108

For chemical with a Log Kow of greater than 5.0:

- Neutral Organics, Acute Fish 96-hour (Q)SAR: Log LC_{50} = No-toxic-effect-at-saturation, or "*"

4.5 Goodness-of-fit statistics

The Correlation Coefficient (r^2) for the Neutral Organics Fish 96-hour (Q)SAR equals 0.886, obtained from standard statistical regression software.

4.6 Information on the applicability domain of the model

This (Q)SAR may be used to obtain ***quantitative*** acute LC_{50} estimates for toxicity of neutral organic compounds (solvents, non-reactive, non-ionizable) with log Kow values of less than 5.0. However, the method may be used to estimate toxic effects equal to "no-toxic-effect-at-saturation or "*" " for chemicals exceeding Log Kow values of 5.0. Therefore, the domain of the model is much larger than the values covered in the regression equation and covers all Log Kow ranges.

This model was derived from data on 337 neutral organic compounds (*e.g.*, solvents, non-reactive, non-ionizable). For chemicals with a Log Kow of less than 5.0, the 96-hour model is sufficient. Data used in the regression equation are for compounds with Log Kow values of 5.0 or less. Compounds with a molecular weight of greater than 1000 g/mol are considered too large to present any significant toxicity. Also, if the predicted toxicity exceeds the water solubility, no acute toxicity is expected to be observed in the absence of an organic carrier solvent.

To obtain ***quantitative*** acute LC_{50} estimates for toxicity of neutral organic compounds with log Kow greater than 5.0 and less than 7.0, use the fish 14-day LC_{50} for neutral organics.

4.7 Information on the mechanistic basis/interpretation of the model

ECOSAR classes are grouped based on similar relationships between toxicity and the various types of pharmacologic properties. Neutral organic compounds have a narcotic effect on aquatic organisms, which is a reversible state of arrested activity of protoplasmic structures. (Veith *et al*., 1983)

5. Development of the model

5.1 Explanation of the method (approach) used to generate each descriptor

5.1.1 Calculation of Log Kow

To estimate Log Kow, ECOSAR uses the method KOWWIN, developed by Syracuse Research Corporation. KOWWIN uses a "fragment constant" methodology to predict Log Kow and the equation is as follows:

Log Kow = Sum ($f_i n_i$) + Sum ($c_j n_j$) + 0.229, where Sum ($f_i n_i$) is the summation of f_i (the coefficient for each atom/fragment) times n_i (the number of times the atom/fragment occurs in the structure), and ($c_j n_j$) is the summation of c_j (the coefficient for each correction factor) times n_j (the number of times the correction factor occurs (or is applied) in the molecule).

5.1.2 Calculation of MW

MW is determined through summation of atomic weights of each atom in the molecule.

5.2 Selection of descriptors

5.2.1 Indication of initial number of descriptors screened

Not applicable

5.2.2 Explanation of the method (approach) used to select the descriptors and develop the model from them:

Use of Kow and MW to predict acute toxicity was determined experimentally through experience in the U.S. EPA, OPPT New Chemical Program and a need to derive the simplest approach for calculating acute toxicity to fish.

5.2.3 Indication of final number of descriptors included in the model:

Two: Log Kow and MW

5.3 Information on experimental design for data splitting into training and validation sets.

Not applicable

5.4 Availability of the training set

5.4.1 ☑ Chemical names (common names and/or IUPAC names)
5.4.2 ☑ CAS numbers
5.4.3 ☐ 1D representation of chemical structure (*e.g.* SMILES)
5.4.4 ☐ 2D representation of chemical structure (*e.g.* ISIS sketch file)
5.4.5 ☐ 3D representation of chemical structure (*e.g.* MOL file)
5.4.6 ☐ Data for each descriptor variable
5.4.7 ☑ Data for the dependent variable

A list of training set chemicals for the Neutral Organic Fish 96-hour (Q)SAR will be available in 2006 through an update to the ECOSAR Technical Reference Guide. The previous training set for version 0.99g can be found in the current version listed at the beginning of the case study document.

A total of 376 toxicity data points for various neutral organic compounds were used for the development of the model. Only 337 of these data points were incorporated in development of the regression equation, as 39 of these data points were for chemicals that exceeded cut-off criteria (*i.e*., water solubility for solids, or Log Kow for liquids) and presented no effects at saturation. These chemicals were not used in development of the regression equation, but are included in the neutral organics chemical class to support justification of neutral organics solubility and Log Kow cut-off criteria indicating no effects at saturation.

Of the 337 data points used in development of the neutral organics acute fish 96-hour regression equation, 21 of the chemicals (6%) represented confidential studies which U.S. EPA is restricted from disclosing to the public. For the 39 chemicals that exceeded cut-off criteria, 28 of those chemicals (76%) were confidential studies. For all CBI chemicals, only molecular weight and the predicted Log Kow are available in the reference manual.

The measured toxicity values used to create the algorithm were the discrete (*e.g*., no ranges or inequalities) dose levels that were determined to produce 50% lethality (LC_{50}) following a 96-hour exposure of the test compound. The tests were preferably conducted using flow-through systems and measured test concentrations rather than static or static renewal systems and nominal test concentrations. Preferred studies reported water hardness values less than or equal to 150 mg/L $CaCO_3$, and TOC concentrations less than or equal to 2.0 mg TOC/L. Only validated data were used. Criteria for exclusion of study data include an inadequate test duration, inadequate endpoints, and unidentified test substance composition. All endpoint values were adjusted for percent active ingredient.

6. Validation of the model

All available valid data were used by U.S. EPA/OPPT in development of the (Q)SARs within ECOSAR. Subsequent validation studies have been completed on ECOSAR by multiple stakeholders, and the results of those external validation studies and/or peer reviews of ECOSAR can be found at the following locations:

- Hulzebos, E.M. and R. Posthumus (2003), "(Q)SARs: Gatekeepers against Risk on Chemicals?", *SAR and QSAR in Environmental Research*, 14(4): 285-316.

- Kaiser, K.L.E, *et al.* (1997), "On Simple Linear Regression, Multiple Linear Regression, and Elementary Probabilistic Neural Network with Gaussian Kernel's Performance in Modeling Toxicity Values to Fathead Minnow Based on Microtox Data, Octanol/Water Partition Coefficient, and Various Structural Descriptors for a 419-Compound Dataset", in F. Chen and G Schuumann (eds.), *Quantitative Structure-Activity Relationships in Environmental Sciences-VII*, SETAC Press, Pensacola, FL, pp. 285-297.

- Kaiser, K.L.E., *et al*. (1999), "A Note of Caution to Users of ECOSAR", *Water Quality Res. J. Canada*, 34(1): 179-182.

- Moore, D.R.J., R.L. Breton and D.B. MacDonald (2003), "A Comparison of Model Performance for six QSAR Packages that Predict Acute Toxicity to Fish", *Environmental Toxicology and Chemistry*, 22(8): 1799-1809.

- Nabholz, J.V., *et al.* (1993), "Validation of Structure Activity Relationships used by the Office of Pollution Prevention and Toxics for the Environmental Hazard Assessment of Industrial Chemicals", in W. Joseph *et al*. (eds), *Environmental Toxicology and Risk Assessment,* 2nd Volume, ASTM STP 1216, American Society for Testing and Materials, Philadelphia, PA, pp. 571-590.

- OECD (1994), *US EPA/EC Joint Project on the Evaluation of (Quantitative) Structure Activity Relationships*, Environment Monographs No. 88, OECD, Paris, 366 pp. http://www.oecd.org/document/30/0,2340,en_2649_34365_1916638_1_1_1_1,00.html, accessed 8 February 2007.

- Posthumus, R. and W. Sloof (2001), "Implementation of QSARS in Ecotoxicological Risk Assessments", RIVM (Research for Man and Environment/National Institute of Public Health and the Environment), Bilthoven, The Netherlands, RIVM report 601516003, 93 pp.

- USEPA (1994), *U.S. EPA/EC Joint Project on the Evaluation of (Quantitative) Structure Activity Relationships*, Washington, DC: U.S. EPA's Office of Pollution Prevention & Toxics, EPA Report #EPA-743-R-94-001. Available from National Technical Information Service (NTIS), U.S. Department of Commerce, 5285 Port Royal Road, Springfield, Virginia 22161, Tel: 703-487-4650 and at http://www.epa.gov/oppt/newchems/tools/21ecosar.htm, accessed 8 February 2007.

6.1 Statistics obtained by leave-one-out cross-validation

Not applicable.

6.2 Statistics obtained by leave-many-out cross-validation

Not applicable.

6.3 Statistics obtained by Y-scrambling

Not applicable.

6.4 Statistics obtained by external validation

Not applicable.

6.5 Definition of the applicability domain of the model

Not applicable.

6.6 Availability of the external validation set

6.6.1 ☐ Chemical names (common names and/or IUPAC names)
6.6.2 ☐ CAS numbers
6.6.3 ☐ 1D representation of chemical structure (*e.g*. SMILES)
6.6.4 ☐ 2D representation of chemical structure (*e.g*. ISIS sketch file)
6.6.5 ☐ 3D representation of chemical structure (*e.g*. MOL file)
6.6.6 ☐ Data for each descriptor variable
6.6.7 ☐ Data for the dependent variable

7. Applications of the model

The fish acute toxicity (Q)SAR for neutral organic chemicals has been used to predict the fish 96-h LC_{50} for industrial chemicals under the Toxic Substance Control Act (TSCA). This SAR has also been used to predict the toxicity of some pesticide active ingredient and pesticide inert ingredients, chemicals found in hazardous waste, chemicals found in water, and chemicals found in air. The Office of Pollution Prevention and Toxics (OPPT) has used this SAR to assist in the validation of measured toxicity test data for fish 96-h LC_{50} values. This SAR has been used by testing laboratories to select test concentrations in lieu of doing a range-finding test. This SAR has been used to predict the toxicity of some pharmaceuticals (Sanderson *et al*., 2003, 2004). For organic chemicals which have a more specific mode to toxic action, this SAR will only predict baseline toxicity or the toxicity just associated with narcosis.

8. Miscellaneous information

The ECOSAR Class Program is a computerized version of the methods employed by the OPPT to assess the environmental toxicity of new chemicals under TSCA. It has been developed within the regulatory constraints of the TSCA and is a pragmatic approach to (Q)SAR, initiated and refined based on experience with chemical under TSCA.

The QSARs presented in this program are used to predict the aquatic toxicity of chemicals based upon their similarity of structure to chemicals for which the aquatic toxicity has been previously measured. Most (Q)SAR calculations in the ECOSAR Class Program are based upon the octanol/water partition coefficient (K_{ow}). Various surfactant (Q)SAR calculations are based upon the average length of carbon chains or the number of ethoxylate units (User's Guide for the ECOSAR Class Program; Meylan, W.M and P.H. Howard, 1998).

Additional information on the ECOSAR program can be found at the following references:

- Auer, C.M., J.V. Nabholz and K.P. Baetcke (1990), "Mode of Action and the Assessment of Chemical Hazards in the Presence of Limited Data: Use of Structure-Activity Relationships (SAR) under TSCA, Section 5", *Environ. Health Perspect*, 87: 183-197.
- Meylan, W.M. and P.H. Howard (1998), *User's Guide for the ECOSAR Class Program. MS-Windows Version 0.99d,* Prepared for the U.S. EPA, OPPT, Risk Assessment Division (RAD). 26 pp, http://www.epa.gov/oppt/newchems/tools/manual.pdf (accessed 7 February 2007)
- Wagner, P.M., J.V. Nabholz and R.J. Kent (1995), "The New Chemicals Process at the Environmental Protection Agency (EPA): Structure-Activity Relationships for Hazard Identification and Risk Assessment", *Toxicology Letters*, 79: 67-73.

Additional information via the internet can be found at the follow sites:

- ECOSAR (Ecological Structure Activity Relationships), http://www.epa.gov/oppt/newchems/tools/21ecosar.htm, accessed 7 February 2007.

9. References

Bailey *et al*. (1985), "Time/Toxicity Relationships in Short-Term Static, Dynamic, and Plug-Flow Bioassays", in R.C. Bahner and D.J. Hansen (eds), *Aquatic Toxicology and Hazard Assessment: 8th Volume*, ASTM STP 891, American Society for Testing and Materials, Philadelphia, PA, pp.193-212.

Broderius *et al*. (2005), "A Comparison of the Lethal and Sublethal Toxicity of Organic Chemical Mixtures to the Fathead minnow", *Environ. Toxicol. Chem*, 24(12): 3117-3127.

Brooke, L.T., *et al*. (eds.), (1984-1990), Acute Toxicities of Organic Chemicals to Fathead minnows (Pimephales promelas), Superior, WI: Center for Lake Superior Environmental Studies, Univ. of Wisconsin-Superior. Vols. 1-5. [Note: the five volumes have different author sequence and publication years]

Chui, Y.C., R.F. Addison and F.C.P Law (1990), "Acute Toxicity and Toxicokinetics of Chlorinated Diphenyl Ethers in Trout", *Xenobiotica*, 20(5): 489-499.

Edsall, C.C. (1991), "Acute Toxicities to Larval Rainbow Trout of Representative Compounds Detected in Great Lake Fish", *Bull. Environ. Contam. Toxicol*., 46(2): 173-178.

Mount, D.I. and C.E. Stephen (1967), "A Method for Establishing Acceptable Toxicant Limits for Fish - Malathion and the Butoxyethanol Ester of 2,4-D", *Trans. Amer. Fish. Soc.*, 96(2): 185-193.

Sanderson, H., *et al*. (2003), "Probabilistic Hazard Assessment of Environmentally Occurring Pharmaceuticals Toxicity to Fish, Daphnids and Algae by ECOSAR Screening", *Toxicol. Lett.,* 144(3): 383-395.

Sanderson, H., *et al*. (2004), "Ranking and Prioritization of Environmental Risks of Pharmaceuticals in Surface Waters", *Reg. Toxicol. & Pharmacol.,* 39(2):158-183.

Shell Oil Co (1984), Unpublished information on the production, uses, and toxicity of sulfolane and Shell Technical Bulletin IC:71-20 (October, 1971, 20 pp) and Material Safety Data Sheet No. 5,620-4 (January, 1083; 4 pp) submitted by JP Sepesi, Shell Oil Co, to M Greif, TSCA Interagency Testing Committee, January 12, 1984 in CRCS Inc. 1984. Sulfolame. IR-434. Rockville MD: CRCS, Inc, 11426 Rockville Pike.

SIDS Initial Assessment Report for 13th SIAM (2001), N,N-Dimethylacetamide, UNEP Publication (Kennedy, 86).

U.S. Environmental Protection Agency (USEPA) (1990), Summary of Structure-Activity Data Files: University of Wisconsin-Superior (UWS) and ORD Environmental Research Laboratory, Duluth, MN (ERL-D) Research Team. Computer printout from Environmental Effects Branch, HERD, U.S. EPA, Washington, DC.

USEPA (1992), *Environmental Toxicity Fact Sheet (ETFS)*, Washington DC: Office of Water, USEPA, 1400 Pennsylvania Ave., N.W.

USEPA (2006), Database of Environmental Toxicity Data from Premanufacture Notifications (PMN), Washington DC: Risk Assessment Division (RAD), OPPT, USEPA, 1400 Pennsylvania Ave., N.W. (unpublished test data).

USEPA (2006), Database of Environmental Toxicity Data from Data Submitted under the “Toxic Substance Control Act” (TSCA), Public Law 94-469, 90 Stat. 2003, October 11, 1976. Washington DC: OPPT, USEPA, 1400 Pennsylvania Ave., N.W.

Veith, G.D., D.J. Cal and L.T. Brooke (1983), "Structure-Toxicity Relationships for the Fathead minnow, Pimephales promelas: Narcotic Industrial Chemicals", *Canadian Journal of Fisheries and Aquatic Sciences*, 40: 743-748.

CASE STUDY 4: CATABOL for Biodegradation

1. QSAR identifier

CATABOL M v5.082

2. Source

2.1 Reference(s) to scientific papers and/or software package:

- Dimitrov, S., *et al*. (2002), "Quantitative Prediction of Biodegradability, Metabolite Distribution and Toxicity of Stable Metabolites", *SAR and QSAR in Environmental Research*, 13, 445-455.
- Dimitrov S., *et al*. (2004), "Predicting the Biodegradation Products of Perfluorinated Chemicals using CATABOL", *SAR and QSAR in Environmental Research*, 15, 69-82.
- Dimitrov, S., *et al*. (2005), "A Stepwise Approach for Defining the Applicability Domain of SAR and QSAR Models", *Journal of Chemical Information and Modelling*, 45, 839-849.
- Jaworska, J. S., *et al.* (2002), "Probabilistic Assessment of Biodegradability Based on Metabolic Pathways: CATABOL System", *SAR and QSAR in Environmental Research*, 13, 307-323.

2.2 Date of publication:

In 1996, a pilot version of metabolism simulator (METABOL) was created. In 1998, METABOL was adapted for prediction of biodegradation. In 1999, prototype CATABOL with MITI data as a training set was developed. This version (v5.082) was released in 2005.

2.3 Identification of the model developer(s)/authors:

Laboratory of Mathematical Chemistry, Bourgas Prof. Assen Zlatarov University

2.4 Contact details of the model developer(s)/authors:

Prof. Ovanes Mekenyan
Laboratory of Mathematical Chemistry, Head
Bourgas "Prof. As. Zlatarov" University,
"Yakimov" St. #1,
8010 Bourgas,
Bulgaria

Tel: +359 56 858 343
Fax: +359 56 880249
Email: omekenya@btu.bg
Web: http://www.oasis-lmc.org/

2.5 Indication of whether the model is proprietary or non-proprietary:

Proprietary

3. Type of model

3.1 ☐ 2D SAR
3.2 ☐ 3D SAR (*e.g*. pharmacophore)
3.3 ☐ Regression-based QSAR
3.4 ☐ 3D QSAR
3.5 ☐ Battery of (Q)SARs
(overall prediction depends on application of multiple models/rules)
3.6 ☑ Expert system
(overall prediction depends on application of multiple models/rules and use of data in a knowledge base)
3.7 ☐ Neural network
3.8 ☐ Other

4. Definition of the model

4.1 Dependent variable:

4.1.1 Species

CATABOL predicts biodegradation endpoints of organic chemicals in the presence of mixed population of environmental micro-organism, which is defined in OECD 301C method.

4.1.2 Endpoint (including exposure time)

CATABOL predicts degradation pathways of a chemical until mineralize under OECD 301C test conditions and, quantitatively predicts Biochemical Oxygen Demand (BOD) degradability and the amount of residuals (both parent and degradants) at 28 days under OECD 301C test conditions.

4.1.3 Units of measurement

The unit of the predicted BOD degradability and the amount of residuals are the ratio of oxygen consumed in 28 days to TOD and the ratio of the amount of residuals to the amount of the parent chemical, respectively.

4.1.4 Reference to specific experimental protocol(s):

CATABOL is modelled to predict biodegradation of a chemical under OECD 301C test method.

4.2 Number of descriptors used as independent variables:

Not applicable

4.3 Identification of descriptors (names, symbols):

Not applicable

4.4 Explicit algorithm for generating predictions from the descriptors:

CATABOL has a predictive engine that consists of 613 hierarchically ordered metabolic transformations. The biodegradation pathway of a chemical substance is generated by sequentially matching its substructure with transformations in the hierarchy and thus indicates the path of degradation to become inorganic substances. Each metabolic transformation in the hierarchy is assigned a probability. BOD and residual amounts are calculated using these probabilities. The transformation probabilities in CATABOL are defined according to the following equation:

$$BOD = \left[\frac{\Delta k_1}{k_{TOD}} P_1 + \frac{\Delta k_2}{k_{TOD}} P_1 P_2 + \frac{\Delta k_3}{k_{TOD}} P_1 P_2 P_3 + \cdots + \frac{\Delta k_l}{k_{TOD}} P_1 P_2 P_3 \cdots P_l \right] \times 100 \quad (1)$$

where, the theoretical oxygen demand is $k_{TOD} = \Sigma \Delta k_i$, and P_i is the probability of initiation of the i-th transformation. In each layer, it is defined the inhibited transformation. The chemicals possessing such inhibited transformation are not matched in the layer and proceed to the next layer.

4.5 Goodness-of-fit statistics

A data set for 745 chemicals based on OECD301C method taken from MITI biodegradation database is used as the training set. The structures, experimental BOD values and BOD values calculated by CATABOL for each chemical in the training set are shown in the software. The coefficient of correlation between the experimental and calculated BOD values was 0.85. The percentage of correctly classified not readily biodegradable (NRB) chemicals was 91% (485/532) and the percentage of correctly classified readily biodegradable (RB) chemicals was 86% (183/213).

4.6 Information on the applicability domain of the model

CATABOL has a function to evaluate whether a target chemical is in the applicability domain or not by comparing a target chemical and chemicals in the training set. The applicability domain is defined by LogP, molecular weight, water solubility and substructures.

4.7 Information on the mechanistic basis/interpretation of the model

The CATABOL model is based on metabolic biodegradation paths and most of the reactions are interpreted in the help files of this software.

5. Development of the model

5.1 Explanation of the method (approach) used to generate each descriptor

In order to establish the prediction engine, the degradation paths of those 745 training set chemicals were defined by experts and were not disclosed. For the defined degradation path of each chemical, experimental BOD values were inserted into the equation (1) and the probability (P) of each reaction was calculated by using least square method. In this process, detailed pathways were generalized by merging sequences of elementary reaction steps into principal metabolic transformations. The reactions were categorized into 44 spontaneous reactions and 72 catabolic reactions, and all spontaneous reactions were ascribed the highest probability value (one). All reaction groups and their reaction probabilities are shown in the help file of the software. The CATABOL prediction engine based on 613 reaction schemes was constructed using the reaction probabilities by experts. The probabilities and their order in the prediction engine are further optimized by comparing calculated pathways and experimental pathways for 207 chemicals.

5.2 Selection of descriptors

5.2.1 Indication of initial number of descriptors screened

Not applicable

5.2.2 Explanation of the method (approach) used to select the descriptors and develop the model from them

Not applicable

5.2.3 Indication of final number of descriptors included in the model:

Not applicable

5.3 Information on experimental design for data splitting into training and validation sets.

Not applicable

5.4 Availability of the training set

The structures, experimental BOD values and BOD values calculated by CATABOL for each chemical in the training set are shown in the software. All the parameters and the rules used for making prediction are shown in the software.

5.5.1 ☑ Chemical names (common names and/or IUPAC names)
5.5.2 ☑ CAS numbers
5.5.3 ☑ 1D representation of chemical structure (*e.g.* SMILES)
5.5.4 ☐ 2D representation of chemical structure (*e.g.* ISIS sketch file)
5.5.5 ☐ 3D representation of chemical structure (*e.g.* MOL file)
5.5.6 ☑ Data for each descriptor variable
5.5.7 ☑ Data for the dependent variable

6. Validation of the model

6.1 Statistics obtained by leave-one-out cross-validation

None

6.2 Statistics obtained by leave-many-out cross-validation

The robustness of the present version of CATABOL is not reported. The robustness of the previous version, which had a training set of 532 chemicals, is reported as Q^2=0.88 for 4 times of Leave 25% out by Jaworska *et al*. (2002).

6.3 Statistics obtained by Y-scrambling

None

6.4 Statistics obtained by external validation

Validation by Mcdowell et al. (2002)

External validations of a pervious version of CATABOL were performed using 77 chemicals from a P&G database. The percentage of correctly classified NRB chemicals were 90% (19/21) and the percentage of correctly classified RB chemicals were 92% (49/53).

Validation by Sakuratani et al. (2005)

External validation of the CATABOL v.4.562 was conducted using test data of 338 existing chemicals and 1123 new chemicals under the Japanese Chemical Substances Control Law (CSCL). The percentage of correctly classified NRB chemicals were 88% (925/1055) and the percentage of correctly classified RB chemicals were 58% (234/406). The features of chemical structures affecting CATABOL predictability were described.

6.5 Definition of the applicability domain of the model

None

6.6 Availability of the external validation set

It is available the data set of existing chemicals used as external validation set. (Sakuratani *et al.*, 2005)

6.6.1 ☑ Chemical names (common names and/or IUPAC names)
6.6.2 ☑ CAS numbers
6.6.3 ☐ 1D representation of chemical structure (*e.g.* SMILES)
6.6.4 ☐ 2D representation of chemical structure (*e.g.* ISIS sketch file)
6.6.5 ☐ 3D representation of chemical structure (*e.g*. MOL file)
6.6.6 ☐ Data for each descriptor variable
6.6.7 ☐ Data for the dependent variable

7. Applications of the model

Suggestions for possible applications for the model:

CATABOL can be used for classifying chemicals by the probability of persistent for screening purpose. And, CATBOL can be used as a supporting tool for risk assessor to predict stable degradants in environment.

8. Miscellaneous information

9. References

Mcdowell, R.M., and J.S. Jaworska (2002), "Bayesian Analysis and Inference from QSAR Predictive Model Results", *SAR and QSAR in Environmental Research*, 13, 111-125.

Sakuratani, Y., *et al.* (2005), "External Validation of the Biodegradability Prediction Model CATABOL Using Data Sets of Existing and New Chemicals under the Japanese Chemical Substances Control Law", *SAR and QSAR in Environmental Research,* 16, 403-431.

CASE STUDY 5: BIOWIN for Biodegradation

1. QSAR identifier

BIOWIN v4.02: containing the following six separate models.

- BIOWIN1 (linear probability model)
- BIOWIN2 (nonlinear probability model)
- BIOWIN3 (expert survey ultimate biodegradation model)
- BIOWIN4 (expert survey primary biodegradation model)
- BIOWIN5 (Japanese MITI linear model)
- BIOWIN6 (Japanese MITI nonlinear model)

2. Source

2.1 Reference(s) to scientific papers and/or software package:

- Boethling, R.S., *et al*. (1994), "Group Contribution Method for Predicting Probability and Rate of Aerobic Biodegradation", *Environmental Science and Technology*, 28, 459-465.
- Howard, P.H., A.E. Hueber and R.S. Boethling (1987), "Biodegradation Data Evaluation for Structure/Biodegradability Relations", *Environmental. Toxicology and Chemistr.* 6, 1-10.
- Howard, P.H., *et al*. (1992), "Predictive Model for Aerobic Biodegradability Developed from a File of Evaluated Biodegradation Data", *Environmental. Toxicology and Chemistry*, 11, 593-603.
- Tunkel, J., *et al*. (2000), "Predicting Ready Biodegradability in the MITI Test", *Environmental. Toxicology and Chemistry*, 19, 2478-2485.

2.2 Date of publication:

The publication dates of each model are as follows:

- BIOWIN1, 2: 1992.
- BIOWIN3, 4: 1994.
- BIOWIN5, 6: 2000.

2.3 Identification of the model developer(s)/authors:

U.S. Environmental Protection Agency
Syracuse Research Corporation

2.4 Contact details of the model developer(s)/authors:

Dr. Robert Boethling
U.S. Environmental Protection Agency
1200 Pennsylvania Ave., N.W. (Mail Code 7406M)
Washington, DC 20460
Tel: +1 202 564 8533
Email: boethling.bob@epa.gov
Web: http://www.epa.gov/

Dr. Philip Howard
Syracuse Research Corporation
6225 running Ridge Road
North Syracuse, NY 13212
Tel: +1 315 452 8417
Email: howardp@syrres.com
Web: http://www.syrres.com

2.5 Indication of whether the model is proprietary or non-proprietary:

Non-proprietary

3. Type of model

3.1 ☐ 2D SAR
3.2 ☐ 3D SAR (*e.g.* pharmacophore)
3.3 ☑ Regression-based QSAR
3.4 ☐ 3D QSAR
3.5 ☑ Battery of (Q)SARs
(overall prediction depends on application of multiple models/rules)
3.6 ☐ Expert system
(overall prediction depends on application of multiple models/rules and use of data in a knowledge base)
3.7 ☐ Neural network
3.8 ☐ Other

4. Definition of the model

4.1 Dependent variable:

4.1.1 Species

BIOWIN models predict biodegradation endpoints of organic chemicals in the presence of mixed population of environmental micro-organism.

4.1.2 Endpoint (including exposure time)

The endpoints of each model are as follows:

- BIOWIN1, 2: The probability that a chemical is easily biodegradable in the typical environment (aerobic biodegradation).
- BIOWIN3: The time required for ultimate biodegradation, which is the transformation of a parent compound to carbon dioxide and water, in the typical environment (aerobic biodegradation).
- BIOWIN4: The time required for primary biodegradation, which is the transformation of a parent compound to an initial metabolite, in the typical environment (aerobic biodegradation).
- BIOWIN5, 6: The probability that a chemical is readily biodegradable in the MITI test (OECD301C).

4.1.3 Units of measurement

The units are probability of fast biodegradation for BIOWIN1, BIOWIN2, BIOWIN5 and BIOWIN6; and approximate total time to complete and primary degradation for BIOWIN3 and BIOWIN4 respectively.

4.1.4 Reference to specific experimental protocol(s):

BIOWIN1, BIOWIN2, BIOWIN3 and BIOWIN4 are not based on specific experimental endpoint. BIOWIN1 and BIOWIN2 are based on the summary aerobic biodegradability descriptors from SRC Environ Fate Database "BIODEG Summary" file, which represent weight-of-evidence judgments. Any and all mixed-culture biodegradation data are used to make these summary judgments. For BIOWIN3 and BIOWIN4, the training set data are from a survey of expert judgment for 200 chemicals.

BIOWIN5 and BIOWIN6 were developed to predict biodegradation of a chemical under the MITI test (OECD301C).

4.2 Number of descriptors used as independent variables:

BIOWIN 1-4: 36 descriptors.
BIOWIN 5, 6: 43 descriptors.

4.3 Identification of descriptors (names, symbols):

BIOWIN 1-4: 35 kinds of fragment and molecular weight.
BIOWIN 5, 6: 42 kinds of fragment and molecular weight.
All descriptors are shown in the help file of the software.

4.4 Explicit algorithm for generating predictions from the descriptors:

BIOWIN1, 3, 4, 5

The following type of linear regression equations gives predictions.

$$Y_j = a_0 + a_1 f_1 + a_2 f_2 + \cdots + a_n f_n + a_m MW \quad (1)$$

Here,
Y_j : Probability that chemical j is easily biodegradable (BIOWIN 1), time required for ultimate biodegradation (BIOWIN3), the time required for primary biodegradation (BIOWIN4), probability that chemical j is readily biodegradable (BIOWIN5).
f_n : Number of fragment n in the chemical j
a_n : Regression coefficient for fragment n
MW : Molecular weight of the chemical j
a_m : Regression coefficient for molecular weight of the chemical j
a_0 : Intercept

BIOWIN2, 6

The following type of non-linear regression equations gives predictions.

$$Y_j = \frac{\exp(a_0 + a_1 f_1 + a_2 f_2 + \cdots + a_n f_n + a_m MW)}{1 + \exp(a_0 + a_1 f_1 + a_2 f_2 + \cdots + a_n f_n + a_m MW)} \quad (2)$$

Here,
Y_j : Probability that chemical j is easily biodegradable (BIOWIN2), probability that chemical j is readily biodegradable (BIOWIN6)
f_n : Number of fragment n in the chemical j
a_n : Regression coefficient for fragment n
MW : Molecular weight of the chemical j
a_m : Regression coefficient for molecular weight of the chemical j
a_0 : Intercept

4.5 Goodness-of-fit statistics

The training set of the BIOWIN1 and BIOWIN2 consists of 109 chemical that were critically evaluated as "does not biodegrade fast" and 186 chemicals that were critically evaluated as "biodegrades fast" by weight-of-evidence judgments. The percentage of correctly classified chemicals evaluated as "does not biodegrade fast" were 76% (83/109) for BIOWIN1 and 86% (94/109) for BIOWIN2; The percentage of correctly classified chemicals "biodegrades fast " were 97% (181/186) for BIOWIN1 and 97% (181/186) for BIOWIN2.

The 200 chemicals of the training set of BIOWIN3 and BIOWIN4 were selected from a variety of sources. The ultimate and primary biodegradation of the 200 chemicals were rated on a scale of 1 to 5 by experts. The ratings correspond to the following time units: 5 - hours; 4 - days; 3 - weeks; 2 - months; 1 - longer. For BIOWIN3, the coefficient of correlation between the expert rated time and calculated time required for ultimate biodegradation is 0.85; For BIOWIN4, the coefficient of correlation between the expert rated time and calculated time required for primary biodegradation is 0.84.

Japanese MITI biodegradation data for 589 chemicals are used as the training set of BIOWIN5 and BIOWIN6. The percentage of correctly classified NRB chemicals are 85% (283/335) for BIOWIN5 and 85% (283/335) for BIOWIN6; The percentage of correctly classified RB chemicals were 79% (201/254) for BIOWIN5 and 80% (204/254) for BIOWIN6.

4.6 Information on the applicability domain of the model

All BIOWIN models target general low-molecular weight organic compounds and detail information on the applicability domain of the model is not given.

4.7 Information on the mechanistic basis/interpretation of the model

The fragments were selected based largely on known structural influences on aerobic biodegradability, such as the ester linkage which is hydrolyzed by microorganisms.

5. Development of the model

5.1 Explanation of the method (approach) used to generate each descriptor

Fragments are used as descriptors which were selected by experts.

5.2 Selection of descriptors

5.2.1 Indication of initial number of descriptors screened

The number screened equals the number selected.

5.2.2 Explanation of the method (approach) used to select the descriptors and develop the model from them

The fragments used as descriptors were selected by experts.. The fragment constants are calculated by regression analysis. For linear models (BIOWIN1, BIOWIN3, BIOWIN4 and BIOWIN5), the method of least squares was used and for nonlinear models (BIOWIN2 and BIOWIN6), the maximum likelihood method was used to estimate regression coefficients.

5.2.3 Indication of final number of descriptors included in the model:

BIOWIN 1-4: 35 kinds of fragment and molecular weight.
BIOWIN 5, 6: 42 kinds of fragment and molecular weight.

5.3 Information on experimental design for data splitting into training and validation sets.

The validation set of BIOWIN1 and BIOWIN2 (27 chemicals) were selected from a variety of sources using same criteria as the training set. The validation set of BIOWIN3 and BIOWIN4 (13 chemicals) were selected from the chemicals that had water grab sample data in the SRC Environ Fate Database "BIODEG Summary" file. For BIOWIN5 and BIOWIN6, the complete database of 884 chemicals was divided into training set (two-thirds of the full data set; 589 chemicals) and validation set (one-thirds of the full data set; 265 chemicals). Chemicals in the training set were selected from the electronic file using a visual basic script based on a random number generator.

5.4 Availability of the training set

5.5.1 ☑ Chemical names (common names and/or IUPAC names)
5.5.2 ☑ CAS numbers
5.5.3 ☑ 1D representation of chemical structure (*e.g.* SMILES)
5.5.4 ☐ 2D representation of chemical structure (*e.g.* ISIS sketch file)
5.5.5 ☐ 3D representation of chemical structure (*e.g.* MOL file)
5.5.6 ☑ Data for each descriptor variable
5.5.7 ☑ Data for the dependent variable

6. Validation of the model

6.1 Statistics obtained by leave-one-out cross-validation

None

6.2 Statistics obtained by leave-many-out cross-validation

None

6.3 Statistics obtained by Y-scrambling

None

6.4 Statistics obtained by external validation

Validation by Langenberg et al. (1996)

Previous version (v3.0) of BIOWIN1 and BIOWIN2 were evaluated using MITI data for 488 chemicals. The percentages of correctly classified chemicals were 56% (BIOWIN1) and 63% (BIOWIN2).

Validation by Rorije et al. (1999)

External validation of BIOWIN1 was performed using MITI data for 733 chemicals. The percentage of correctly classified NRB chemicals were 56% (357/635) and the percentage of correctly classified RB chemicals were 68% (179/263).

Validation by Tunkel et al. (2000)

External validation of BIOWIN5 and BIOWIN6 were performed using MITI data for 295 chemicals. The percentage of correctly classified NRB chemicals were 82% (135/164) for BIOWIN5 and 82% (135/164) for BIOWIN6; The percentage of correctly classified RB chemicals were 80% (105/131) for BIOWIN5 and 79% (103/131) for BIOWIN6.

The performances of BIOWIN1 and BIOWIN2 were also evaluated using MITI data for 884 chemicals. The percentages of correctly classified chemicals were 65% (BIOWIN1) and 68% (BIOWIN2).

Validation by Boethling et al. (2003)

The performances of five BIOWIN models were evaluated using a data set for 305 pre-manufacture notice (PMN) substances under the Toxic Substances Control Act (TSCA), which is containing six tests data (OECD301A-F). The percentages of correctly classified chemicals of each model were 54% (BIOWIN1), 67% (BIOWIN2), 88% (BIOWIN3), 77% (BIOWIN5) and 77% (BIOWIN6), respectively.

Validation by Boethling et al. (2004)

The performances of three BIOWIN models were evaluated using data sets for 374 PMN substances under the TSCA and 63 pharmaceuticals. For PMN substances, the percentages of correctly classified chemicals were 86% (BIOWIN3), 81% (BIOWIN5) and 82% (BIOWIN6). For pharmaceuticals, the percentages of correctly classified chemicals were 76% (BIOWIN3), 83% (BIOWIN5) and 87% (BIOWIN6).

Validation by Posthums et al. (2005)

External validations of five BIOWIN models were performed using 110 chemicals which were notified in The Netherlands under EU raw. The performances of each model were compared by two or three pass levels. The percentages of correctly classified chemicals at the best pass level in each model were 64% (BIOWIN1), 70% (BIOWIN2), 88% (BIOWIN3), 77% (BIOWIN5) and 77% (BIOWIN6), respectively.

6.5 Definition of the applicability domain of the model

The definition of the applicability domain is not shown.

6.6 Availability of the external validation set

Te data set used as external validation set for BIOWIN5 and BIOWIN6 is available (Tunkel *et a*l. (2000)).

6.6.1 ☑ Chemical names (common names and/or IUPAC names)
6.6.2 ☑ CAS numbers
6.6.3 ☐ 1D representation of chemical structure (*e.g*. SMILES)
6.6.4 ☐ 2D representation of chemical structure (*e.g*. ISIS sketch file)
6.6.5 ☐ 3D representation of chemical structure (*e.g*. MOL file)
6.6.6 ☑ Data for each descriptor variable
6.6.7 ☑ Data for the dependent variable

7. Applications of the model

Suggestions for possible applications for the model:

BIOWIN can be used for screening readily biodegradable chemicals from wide range of organic compound.

8. Miscellaneous information

9. References

Langenberg, J.H., *et al*. (1996), "On the Usefulness and Reliability of Existing QSBRs for Risk Assessment and Priority Setting", *SAR and QSAR in Environmental Research*, 5, 1-16.

Rorije, E.H., *et al*. (1999), "Evaluation and Application of Models for the Prediction Ready Biodegradability in the MITI-test", *Chemosphere*, 38, 1409-1417.

Boethling, R.S., D.G. Lynch and G.C. Thom (2003), "Predicting Ready Biodegradability of Premanufacture Notice Chemicals", *Environmental Science and Technology*, 22, 837-844.

Boethling, R.S., *et al*. (2004), "Using BIOWIN, Bayes, and Batteries to Predict Ready Biodegradability", *Environmental Science and Technology*, 23, 911-920.

Posthums, R., *et al*. (2005), "External Validation of EPIWIN Biodegradation Models", *SAR and QSAR in Environmental Research*, 16, 135-148.

Tunkel, J., *et al*. (2000), "Predicting Ready Biodegradability in the MITI Test", *Environmental. Toxicology and Chemistry*, 19, 2478-2485.

GLOSSARY

This Glossary provides additional explanation for common scientific terms which are presented in order to enhance communication between (Q)SAR experts and users of (Q)SAR models.

Applicability Domain (AD):

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

The *AD* of a (Q)SAR can be thought of as a theoretical region in multi-dimensional space in which the model is expected to make reliable predictions. Thus, information on the *AD* helps the user of the model to judge whether the prediction for a new chemical is reliable or not. The region depends on the nature of the chemicals in the training set, and the method used to develop the model.

The development and assessment of methods for defining the domain of applicability is an important area of QSAR research.

Acute toxicity:

Acute toxicity refers to the short-term biological effects on an organism of a chemical. A common adverse outcome associated with acute toxicity is lethality; however other effects such as immobilisation (*e.g*. *Daphnia*), reduction in light emission (*e.g*. the Microtox test) and physiological and histological changes are also accepted. Acute lethality is the concentration or dose that produces 50% mortality and is reported as LV_{50} or LD_{50} respectively.

Artificial Neural network (ANN):

Artificial neural networks (ANN) are computational models that make predictions by simulating the functioning of human neurons.

The first step in the development of an *ANN* is to design a specific network architecture that includes a specific number of "layers", each of which consists of a certain number of "neurons". The *ANN* is then subjected to a "training" process, an iterative process in which the neurons apply an iterative process to the number of inputs (variables) to adjust the weights of the network in order to optimally predict the sample data. After the phase of learning from an existing data set, the new network can be used to generate predictions.

The resulting "network" developed in the process of "learning" represents a pattern detected in the data, and is the functional equivalent of a model of relations between variables in the traditional model building approach. However, unlike in the traditional models, the relations in the "network" cannot be articulated in the usual terms used in statistics (*e.g*. "A is positively correlated with B").

ANNs are useful for pattern recognition problems, and for modelling non-linear relationships. They can be fully transparent in terms of being associated with a description of the layers, neurons and connection weights. However, the network architecture is unlikely to correspond with any mechanism or theory underlying the observed phenomena.

Baseline toxicity:

The *baseline toxicity* is the toxicity resulting from the weakest binding forces such as van der Waal or hydrophobic forces between a chemical and cellular targets. The acute toxicity prior to lethality is reversible and is considered non-specific with respect to chemical structure. *Baseline toxicity* is an estimate of the minimum toxicity for a chemical and can be estimated from (Q)SAR for narcotic endpoints. The baseline is a hypothetical reference point for the hazard identification of chemicals involving irreversible and/or selective binding forces with membranes, proteins or DNA.

Bayesian statistics:

The field of statistics is based on two main paradigms: conventional (frequentist) and Bayesian. The Bayesian paradigm is based on an interpretation of probability as a conditional measure of uncertainty, which can be modified in the light of available evidence. This means that Bayesian methods allow for the incorporation of existing knowledge/expectations about what the true relationship might be, before new data become available. This information is expressed in a prior probability distribution, which is subsequently modified to a posterior distribution once data have been obtained and the existing knowledge/expectations have been revised.

In QSAR analyses, *Bayesian statistics* can be applied, for example, to a test battery, in which results of several QSAR models with varying sensitivities and specificities are combined, in order to increase the reliability of QSAR-based predictions. *Bayesian statistics* can also be applied in combination with neural networks (Bayesian Neural Networks).

Bioaccumulation:

Bioaccumulation is the process by which the chemical concentration in an aquatic organism exceeds that in the water as a result of chemical uptake through all possible routes of chemical exposure (*e.g.* dietary absorption, respiratory transport, inhalation).

Bioaccumulation factor (BAF):

The *bioaccumulation factor (BAF)* expresses the extent of chemical bioaccumulation. It is defined as the ratio of the chemical concentration in the organism (C_B) to that in water (C_W). It is a parameter most often defined based on partitioning between water and aquatic organisms, especially fish.

Bioconcentration:

Bioconcentration is the process by which the chemical concentration in an aquatic organism exceeds that in water as a result of exposure to waterborne chemical. *Bioconcentration* refers to a condition usually achieved under laboratory conditions, where the chemical is absorbed only from the water via the respiratory surface and/or skin. *Bioconcentration* can be considered as a transport process in the environment.

Bioconcentration factor (BCF):

The *bioconcentration factor (BCF)* expresses the extent of chemical bioconcentration. It is defined as the ratio of the chemical concentration in an organism to the concentration in water. It is used as a surrogate for the bioaccumulation factor (BAF), but is generally not a good surrogate for chemicals for which significant accumulation occurs via dietary route of exposure.

Biomagnification:

Biomagnification is the process by which the chemical concentration in an organism exceeds that in the organism's diet, due to dietary absorption.

Biomagnification factor:

The *biomagnification factor* (BMF) expresses the extent of chemical biomagnification. It is defined as the ratio of the chemical concentration in the organism to the concentration in the organism's diet.

Bootstrap resampling:

Bootstrap resampling (or bootstrapping) is an approach to internal validation. The basic premise of *bootstrap resampling* is that the data set should be representative of the population from which it was drawn. Since there is only one data set, bootstrapping simulates what would happen if the samples were selected randomly.

In a typical bootstrap validation, K groups of the size n are generated by a repeated random selection (typically, >1000 times) of n objects from the original data set. It is possible for some objects to be included in the same random sample several times, while other objects may never be selected. The model obtained from the data set of n randomly selected objects is used to predict the target properties for the excluded objects. As in the case of LMO validation, a high average q^2 in the bootstrap validation is a demonstration of the model robustness.

Bootstrap resampling also provides non-parametric confidence intervals for the estimated parameters. The resampling process generates a large number of values (>1000) for each parameter, and one then estimates the true values, standard deviations and confidence intervals from these values. Similar methods are sometimes called resampling or jackknife methods.

Chronic toxicity:

Chronic toxicity refers to the long-term biological effects on an organism exposed to a toxicant. The measured endpoints may vary from lethality (LD_{50}) to many sublethal effect concentrations (EC_{50}) and to no observable effect concentrations (NOELs).

Classification:

Classification is the assignment of objects (*e.g.* chemicals) to one of several existing classes based on a classification rule. *Classification* is also called supervised pattern recognition, as opposed to unsupervised pattern recognition.

A class or category is a distinct subspace of the whole measurement space. The classes are defined *a priori* by groups of objects in the training set. The objects of a class have one or more characteristics in common, indicated by the same value of a categorical variable (*e.g.*, biodegradable/not biodegradable).

The goal of a *classification* method is to develop a *classification* rule (by selection of the predictor variables) based on a training set of objects with known classes so that the rule can be applied to a test set of objects with unknown classes. There is a wide range of *classification* methods, including: discriminant analysis (DA), linear DA (LDA), quadratic DA, regularised DA, SIMCA (Soft Independent Modeling of Class Analogy), KNN (K Nearest Neighbours) and CART (Classification And Regression Tree) etc.

The outputs of a *classification* model are the class assignments and the misclassification matrix, which shows how well the classes are separated. The predictive performances of *classification* models can be verified by comparing the cross-validated error rate or risk with the No-Model error rate or risk.

Cluster analysis:

Cluster analysis is the grouping, or clustering, of large data sets on the basis of similarity criteria for appropriately scaled variables that represent the data of interest. Similarity criteria (distance based, associative, correlative, probabilistic) among the several clusters facilitate the recognition of patterns and reveal otherwise hidden structures in the data. Different types of cluster analysis have been developed, referred to as hierarchical or non-hierarchical methods.

In hierarchical cluster analysis (or tree clustering), objects (*e.g*. chemicals) are organised by similarity into a tree, called a dendrogram, similar to the trees seen in phylogenetics. Two objects are next to each other if they are very similar, and increasingly far apart as they become more divergent. The procedure can work bottom up or top down. The bottom-up method starts by joining the two closest objects to form a cluster, then joins the next two closest items (which may be two objects or a object and the newly formed cluster), and continues by joining the two closest items at each step (which may be objects or clusters) until done. The top-down method does the opposite: it starts with all units in one giant cluster, divides the cluster in two, and continues dividing clusters until all objects are separated out.

There are several types of non-hierarchical clustering. An example is *k*-means clustering, in which the researcher defines *a priori* the number of clusters the objects should be arranged. The *k*-means clustering algorithm then produces *k* different clusters, and places the objects in clusters with the goal to minimise the variability within clusters, and to maximise the variability between them.

Coefficient of determination (r^2):

The total variation of any data set is made up of two parts, the part that can be explained by the regression equation and the part that cannot be explained by the regression equation. The *coefficient of determination* is the percent of the variation that can be explained by the regression equation. It represents the explained variance of the model, and is used as a measure of the goodness-of-fit of the model.

The *coefficient of determination* equals the square of the correlation coefficient r between the experimental response (the dependent variable y) and the predictors (the independent variables x), multiplied by 100. It can also be calculated by the formula:

$$r^2 = ESS/TSS = 1-(RSS/TSS)$$

where ESS is the Explained Sum of Squares, RSS is the Residual Sum of Squares and TSS is the Total Sum of Squares.

Collinearity:

Collinearity is a situation where there is a linear relationship between two or more of the independent variables in a regression model. In practical terms, this means there is some degree of redundancy or

overlap the variables. Interpretation of the effects of the independent variables is difficult in this situation, and the standard error of their estimated effects may become very large.

Collinearity should be described by a correlation matrix in the case when more molecular descriptors are involved in the QSAR model. The correlation matrix is formed from correlation coefficients of correlations of all pairs of the descriptors used (even between two descriptors).

Congeneric series:

A group of chemicals with one or more of the following: a common parent structure (*e.g.* aliphatic alcohols), same mechanism of action, and rate-limiting step.

Comparative molecular field analysis (CoMFA):

Comparative Molecular Field Analysis (CoMFA) is a 3D-QSAR method that uses multivariate statistical analysis to quantify the relationship between the biological activities of a set of compounds with a specified alignment, and their three-dimensional electronic and steric properties.

Cooper statistics:

A common problem in QSAR analysis is the prediction of group membership from molecular descriptors. In the simplest case, chemicals are categorised into one of two groups depending on their biological activity: active/inactive or toxic/non-toxic. A variety of statistical methods are available for developing QSARs for two-group classification (*e.g.* discriminant analysis, logistic regression).

The performance of a two-group QSAR is sometimes represented in the form of a 2x2 contingency table:

		Predicted class		
		Active	Inactive	Marginal totals
Known class	Active	a	b	a+b
	Inactive	c	d	c+d
	Marginal totals	a+c	b+d	a+b+c+d

The goodness-of-fit of a two-group QSAR can be summarised in the form of Cooper statistics, which are based on data in the contingency table, and defined as follows:

Statistic	**Definition: "the proportion (or percentage) of the …**	
sensitivity	active chemicals (chemicals that give positive results experimentally) which are predicted to be active."	= a/(a+b)
specificity	inactive chemicals (chemicals that give negative results experimentally) which are predicted to be inactive."	= d/(c+d)
concordance or accuracy	chemicals which are classified correctly."	= (a+d)/(a+b+c+d)
positive predictivity	chemicals predicted to be active that give positive results experimentally."	= a/(a+c)
negative predictivity	chemicals predicted to be inactive that give negative results experimentally."	= d/(b+d)

false positive (over-classification) rate	Inactive chemicals that are falsely predicted to be active."	= c/(c+d) = 1 - specificity
false negative (under-classification) rate	active chemicals that are falsely predicted to be inactive."	= b/(a+b) = 1 - sensitivity

The statistics sensitivity, specificity and concordance provide measures of a two-group QSAR to detect known active (toxic) chemicals (sensitivity), inactive (non-toxic) chemicals (specificity) and all chemicals (accuracy or concordance). The false positive and false negative rates can be calculated from the specificity and sensitivity.

The other two statistics, the positive and negative predictivities, are conditional probabilities: if a chemical is predicted to be active (toxic), the positive predictivity gives the probability that it really is active (toxic); similarly, if a chemical is predicted to be inactive (non-toxic), the negative predictivity gives the probability that it really is inactive (non-toxic). These conditional probabilities can be calculated by Bayesian statistics.

Correlation coefficient (r):

The *correlation coefficient (r)* is a statistical measure of the relationship between a dependent variable *y* (*e.g.* a toxicity endpoint) and one or more independent variable(s) *x*. It is given a value from 0 (for no relationship) to –1 (for a perfect negative correlation) or +1 (for a perfect positive correlation).

In QSAR analysis, it is commonly used as a measure of the statistical fit of a regression-based model, or to describe the relationship and hence potential collinearity between two descriptors.

The variance in the dependent variable is expressed as the total sum of squares (TSS), which can be divided into the variance attributed to the model (the explained sum of squares [ESS]), and the variance attributed to the prediction error (the residual sum of squares [RSS]).

The correlation coefficient (r) id defined by the following equation:

$$r = \sqrt{\frac{ESS}{TSS}}$$

where ESS is the Explained Sum of Squares and TSS is the Total Sum of Squares. The squared correlation coefficient is the **coefficient of determination**.

Cross-validated explained variance (q^2):

The *cross-validated explained variance* or cross-validated correlation coefficient (q^2) is used as a measure of the internal performance, and sometime used to estimate predictivity. It is calculated by the formula:

$$q^2 = 1\text{-PRESS/TSS}$$

where PRESS is the Predictive Error Sum of Squares and TSS is the Total Sum of Squares.

In contrast to r^2, which always increases by adding more descriptors, the value of q^2 increases when useful predictors are added, but decreases otherwise.

Cross-validation:

Cross-validation refers to the use of one or more statistical techniques in which different proportions of chemicals are omitted from the training set (*e.g.* leave-one-out [LOO], leave-many-out [LMO]). The QSAR is developed on the basis of the data for the remaining chemicals, and then used to make predictions for the chemicals that were omitted. This procedure is repeated a number of times, so that a number of statistics can be derived from the comparison of predicted data with the known data.

Cross-validation techniques can be used to assess the robustness of the model (stability of model parameters), and to make estimates of predictivity.

In *k*-fold *cross-validation*, the training set is randomly split into *k* mutually exclusive subsets (called folds) of approximately equal size. The model is trained and tested k times, each time being used to make predictions for chemicals that were left out of the training set.

Cross-validated estimates of accuracy are random numbers that depend on the division into folds. Complete *cross-validation* gives the average of all possibilities of choosing *k* subsets of objects out of a total training set of *n* objects. Except for leave-one-out (LOO) *cross-validation*, which is always complete, *k*-fold *cross-validation* provides an estimate of complete *cross-validation* by using a single split of the training set into folds. To provide a better (Monte-Carlo) estimate of complete *cross-validation*, *k*-fold *cross-validation* can be repeated a number of times.

In stratified *cross-validation*, the folds are stratified so that each fold contains approximately proportions of the classes present in the original training set.

Cross-validation by the Leave-One-Out (LOO) procedure:

Cross-validation by the leave-one-out (LOO) procedure employs *n* training sets in which 1 object has been excluded from the original training set. A total of *n* models are developed by using each training set of n-1 objects. For each model, the value of the excluded object is predicted. In the case of a regression model, q^2 can be computed. In the case of a classification model, cross-validated Cooper statistics can be calculated.

Cross-validation by the Leave-Many-Out (LMO) procedure:

Cross-validation by the leave-many-out (LMO) procedure employs a number of training sets, derived by omitting a fixed proportion (typically, up to 50%) of objects from the original training set. In contrast to LOO cross-validation, which is necessarily complete, LMO cross-validation is generally repeated a number of times, due to the large number of possible combinations of training sets generated by leaving out a fixed proportion of objects from the original training set.

If a QSAR model has a high average q^2 in LMO validation, it is generally concluded that the obtained model is robust.

Data mining:

Data mining is a collective term that refers to all procedures (informatic and statistical) that are applied to heterogeneous data sets, in order to develop a data matrix amenable to statistical methods. For example, there are large databases of toxic effect values, such as the Registry of Toxic Effects of Chemical Substances (RTECS) compilation of rat oral LD_{50} values.

Degradation:

Chemicals that are released in the environment are subject to different (biotic and abiotic) *degradation* processes: biodegradation by microorganisms, photolysis by light, hydrolysis by water, oxidation by different oxidants (for instance, in the atmosphere by hydroxyl and nitrate radicals or by ozone). These degradative processes are usually modelled in terms of the rate constants of the corresponding chemical reactions.

Dependent Variable:

A *dependent variable* (y) is a variable modelled by an equation in which one or more independent variables (x) are used as predictors of the *dependent variable*.

In QSPR and QSAR analysis, the dependent variable generally refers to a physicochemical property, toxicity endpoint, ecotoxicity endpoint or environmental parameter. The independent variables (x) in a QSAR model are generally **molecular descriptors**.

Descriptor: see Molecular descriptor

Domain of applicability: see Applicability domain (AD)

Discriminant analysis:

Discriminant analysis refers to a group of statistical techniques that can be used to find a set of descriptors to detect and rationalise (in terms of a predictive model) the separation between activity classes

Electrophilicity:

Electrophilicity is the molecular or substructural property of having an attraction for electrons or negative charge. Molecular *electrophilicity* is often described by the molecular orbital characteristics: the energy of the lowest unoccupied molecular orbital (E_{LUMO}) and electrophilic superdelocalisability.

Energy of the highest occupied molecular orbital (E_{HOMO}):

The *energy of highest occupied molecular orbital (E_{HOMO})* is the energy of the highest energy level that contains electrons in a molecule. Molecules with high HOMO energy values can donate their electrons more easily compared to molecules with low HOMO energy value and hence are more reactive as nucleophiles. This molecular orbital property is therefore often used as a measure of nucleophilicity in QSAR models. It is equivalent to the negative of the ionisation potential.

Energy of the lowest unoccupied molecular orbital (E_{LUMO}):

The *energy of lowest occupied molecular orbital (E_{LUMO})* is the energy of the lowest energy level that contains no electrons in a molecule. Molecules with low LUMO energy values are more able to accept electrons than molecules with high energy values. This molecular orbital property is often used as a measure of electrophilicity in QSAR models. It is related to the electron affinity.

Expert system:

Any formalised system, not necessarily computer-based, which enables a user to obtain rational predictions about the properties or activities of chemicals. All *expert systems* for the prediction of chemical

properties or activities are built upon experimental data representing one or more effects of chemicals in biological systems (the database), and/or rules derived from such data (the rulebase).

External validation:

External validation refers to a validation exercise in which the chemical structures selected for inclusion in the test set are different to those included in the training set, but which should be representative of the same chemical domain. The QSAR model developed by using the training set chemicals is then applied to the test set chemicals in order to verify the predictive ability of the model.

Many QSAR practitioners regard *external validation* to be the most stringent form of validation, provided that sufficient experimental data are available, and the test structures are selected judiciously, in order to allow for a sufficient coverage of the applicability domain of the model.

In the ideal validation process, the results of *external validation* will be used to supplement the results obtained by internal validation. However, in practice, there may be insufficient data to perform an external validation.

False negative rate: see Cooper statistics

False positive rate: see Cooper statistics

Fisher statistic:

The *Fisher statistic* (F), or variance ratio, is the ratio of two s^2 values (estimates of population variance, based on the information in two or more random samples). In the F test, the obtained value of F is used to test the statistical significance of the observed differences among the means of two or more random samples.

The F test employs the F statistic to test various statistical hypotheses about the mean (or means) of the distributions from which a sample or a set of samples have been drawn.

Fragment analysis:

Fragment analysis refers to the analysis of a dataset that involves breaking down molecular structures into fragments of one or more atoms, in order to identify activitating and inactivating fragments (biophores and biophobes). It can be the basis of fragment-based (or group contribution) methods, in which the properties (activities) of a molecule are estimated by summation of the properties (activities) of the fragments.

Free-Wilson analysis:

Free-Wilson analysis is a regression technique using the presence or absence of substituents or groups (indicator variable) as the only molecular descriptors in correlations with biological activity.

Functional group:

Chemicals can be thought of as consisting of a relatively unreactive backbone and one or more *functional groups*. The *functional group* is an atom, or a group of atoms, which has specific chemical attributes, particularly for interactions with other chemicals. *Functional groups* often be the primary cause for chemical characteristic when only a few functional groups are present. However, for complex

chemicals with many functional groups, the simple interactions associated with individual functional groups are not reliable predictors of chemical behaviour.

Genetic algorithm (GA):

A *genetic algorithm (GA)* is an optimisation method based on evolutionary principles. In GA terminology, a chromosome is a p-dimensional vector (a string of bits) where each position (a gene) corresponds to a variable (1 if included in the model, 0 otherwise). Each chromosome or individual in the population represents a model with a subset of variables. A population of models is obtained which evolves, according to *genetic algorithm* rules, in order to maximise the predictive power of the models (for instance, the explained variance in prediction, q^2).

In the first generation, the variables are chosen randomly. In the next step, reproduction takes place, so that the new individual contains characteristic of both its parents. The next steps are crossovers and mutations, which allow better variable combinations to be found. This reproduction-crossover-mutation process is repeated during the evolution of the population until a desired target fitness score is reached. Only the models producing the highest predictive power are finally retained and further analysed.

*GA*s are used in QSAR analysis as a strategy for variable subset selection (VSS) in multivariate situations where a large number of molecular descriptors are potential x-variables. There are different types of GA analysis, which perform reproduction, crossover and mutation in different ways. An important characteristic of the GA-VSS method is that the result is usually a population of acceptable models.

Half-life:

The *half-life* (commonly denoted as $t_{½}$) is the time required for the concentration of a particular chemical in a medium to be reduced to half of its initial value. Environmental *half-life* data generally reflect the rate of disappearance of a chemical from a medium, without identifying the mechanism of chemical loss. For example, loss from water may be due to a combination of evaporation, biodegradation and photolysis. If the elimination rate involves transport and transformation processes that follow first-order kinetics, the *half-life* time is related to the total elimination rate constant *k* as follows: 0.693/k. In some cases, lifetime is used instead of *half-life*.

Hansch analysis:

Hansch analysis is the investigation of the quantitative relationship between the biological activity of a series of compounds and their physicochemical substituent or global parameters representing hydrophobic, electronic, steric, and other effects, using a multiple regression method.

Henry constant:

The *Henry constant (H)* is an air-water partition coefficient that expresses the tendency of a chemical to volatilise from an aqueous medium. It can be determined by measurement of the solute concentrations in both phases. Due to the difficulty of accurate analytical determination, the H constant is mainly calculated as the ratio of vapour pressure to solubility.

Heterogenous: see Training set

Homogeneous: see Training set

Homologous series:

A *homologous series* is a family of chemicals containing a common functional group and differing only in the length of their carbon chain.

Hydrophilicity:

Hydrophilicity refers to the affinity of a molecule or substitutent for a polar solvent (especially water) or for polar groups. It represents the tendency of a molecule to be solvated by water.

Hydrophobicity:

Hydrophobicity refers to the association of non-polar groups or molecules in an aqueous environment, which arises from the tendency of water to exclude non-polar molecules. It is related to lipophilicity. It represents the tendency of a molecule to partition between a polar and a non-polar phase, and is therefore often measured by a partition coefficient between a polar and non-polar phase (usually, but not always, n-octanol and water).

It is often highly related to biological activity due to its strong relationship with the transport and distribution of a molecule, particularly through phospholipid membranes.

Independent Variable: see Dependent variable

Indicator Variable:

An *indicator variable* is a descriptor that can assume only two values indicating the presence (=1) or absence (=0) of a given condition. In **Free-Wilson analysis**, it is used to indicate the absence or presence of a substituent or substructure.

Internal validation:

Internal validation refers to a validation exercise in which one or more statistical methods are applied to the training set of chemicals. *Internal validation* results in one or more measures of goodness-of-fit, robustness of model parameters, and estimates of predictivity.

Many QSAR practitioners regard *internal validation* to be an essential, but not sufficient, aspect of statistical validation, which should ideally be supplemented by external validation.

Lipophilic:

Lipophilic refers a tendency of a molecule to dissolve in fat-like (*e.g.*, hydrocarbon) solvents.

Lipophilicity:

Lipophilicity refers to the affinity of a molecule or of a substituent for a lipophilic environment. It is commonly measured by its distribution behaviour in a biphasic system (*e.g.*, octanol-water partition coefficient) .

Molecular Descriptor:

A *molecular descriptor* is a structural or physicochemical property of a molecule, or part of a molecule, which characterises a specific aspect of a molecule and is used as an independent variable in a QSAR.

Guidance on the appropriate use of descriptors is provided in Chapter 6, and a list of commonly-used descriptors is provided as Table 6.1.

Molecular modelling:

Molecular modelling refers to the investigation of molecular structures and properties by using computational chemistry and graphical visualisation techniques to provide a plausible three-dimensional (3D) representation of a chemical.

It can refer to the modelling of small organic molecules, macromolecules (*e.g.* proteins, DNA), crystals and inorganic structures. The 3D structure of the molecule is usually obtained by a process of geometry optimisation. The geometry-optimised molecule provides the basis for calculating molecular properties.

Molecular Orbital Properties:

Molecular orbital properties (molecular structure and electronic properties) are estimated by applying quantum chemical calculations to molecular structures.

Molecular orbital properties are usually calculated from semi-empirical rather than *ab initio* methods. Freely available software, such as MOPAC, is available to perform these calculations. A variety of *molecular orbital properties* have been found useful in QSAR analysis, including the energies of the highest occupied and lowest unoccupied molecular orbitals (E_{HOMO} and E_{LUMO} respectively), atomic charges and superdelocalisabilities, dipole moment, and electrostatic potential.

Narcosis:

Narcosis is the non-specific surpression of physiological functions by chemicals which bind reversibly to membranes and proteins. The effect is brought about by non-reactive chemicals and is thought to result from an accumulation of the toxicant in cell membranes, diminishing their functionality. The narcotic effect is reversible, so that an organism will recover when the toxicant is removed.

The potency for narcotic effects are strongly associated with molecular hydrophobicity and vapour pressure, and hence good relationships have been found between the acute toxicity of narcotics and log P (inhalation in fish) and vapour pressure (inhalation in mammals) Within the narcotic mode of toxic action, a number of more selective mechanisms such as non-polar narcosis, polar narcosis, amine narcosis, ester narcosis, anaesthetics, and sensory irritation have been proposed.

Multivariate analysis:

Multivariate analysis is the analysis of multi-dimensional data matrices by using statistical methods. Such data matrices can involve multiple dependent and/or independent variables.

Nucleophilicity:

Nucleophilicity refers to the molecular or substructural property of having a repulsion for electrons or an attraction for positive charge. Molecular *nucleophilicity* is often described by the energy of the highest occupied molecular orbital (E_{HOMO}) and by nucleophilic superdelocalisability.

Outlier:

An *outlier* of a QSAR model refers to a data point (chemical) that falls outside the confidence interval of the regression line. *Outliers* can be defined statistically in various ways. Typically, the *outlier* of a QSAR model has a cross-validated standardised residual greater than three standard deviation units.

The *outliers* of a QSAR model should always be identified, and the reason for their outlying behaviour should be provided.

Parameter space:

The *parameter space* of a model is a multi-dimensional space in which the axes are defined by the descriptors of the model. See domain of applicability.

Pattern recognition:

Pattern recognition is the identification of patterns in (generally large) data sets, using appropriate chemometric methods. Examples are exploratory methods like Principal Component Analysis (PCA), Factor Analysis, Cluster Analysis, Artificial Neural Networks (ANN).

(Model) Performance:

The performance of a (Q)SAR model refers to its goodness-of-fit, robustness and predictive ability in relation to a defined applicability domain.

Model performance is established by using the techniques of statistical validation.

Persistence:

The term persistent is used to characterise chemicals that have long lifetimes in the environment. The *persistence* of a chemical depends on its kinetics or reactivity, as expressed by its rates of degradation. See also Degradation.

Pharmacophore:

The ensemble of steric and electronic features that is necessary to ensure the optimal intermolecular interaction with a specific biological target molecule, which may result in the activation or inhibition of a specific biological response.

Partition coefficient:

A *partition coefficient* is the ratio of the concentrations of a substance between two phases when the heterogeneous system of two phases is in equilibrium. In QSAR analysis, the octanol-water *partition coefficient* (log K) is often used as a descriptor of hydrophobicity, where

$$\text{Log Ko/w} = \text{Log } [\text{chemical}]_{\text{n-octanol}} / [\text{chemical}]_{\text{water}}$$

Predictor: see molecular descriptor

Predictive Error Sum of Squares (PRESS):

Predictive Error Sum of Squares (PRESS) is the sum of the squares of the differences (residuals) between the experimental and predicted responses when predictions are made for objects left out of the training sets, but included in the external test set.

Predictivity:

The *predictivity* (or predictive capacity/ability) of a model is a measure of its ability to make reliable predictions for chemical structures not included in the training set of the model.

For regression models, a measure of *predictivity* is the coefficient of determination. For classification models, measures of *predictivity* include the positive *predictivity* and the negative *predictivity*.

Some (Q)SAR practitioners distinguish between internal and external *predictivity*, depending on whether the estimate or measure of *predictivity* is based on internal or external validation. For other researchers, "*predictivity*" is by definition "external", in which case the term "internal performance" would be used in preference to "internal *predictivity*".

Principal components analysis (PCA):

Principal components analysis (PCA) is a method for reducing data dimensionality by applying mathematical techniques. The main element of this approach consists of the construction of a reduced set of new orthogonal, *i.e.* not correlated, variables, each of which is derived from a linear combination of the original variables. It is an explorative method that is useful for visualising the structure of the data in a complex matrix. In QSAR analysis, it is also used to verify the correlation among the descriptors, thereby supporting the selection of molecular descriptors in models.

Principal components regression (PCR):

Principal components regression (PCR) is the application of regression analysis to a data set in which the descriptors are principal components, derived from more fundamental descriptors.

Quantitative structure-activity relationship (QSAR):

A *Quantitative Structure-Activity Relationship (QSAR)* is a quantitative relationship between a biological activity (*e.g.* toxicity) and one or more molecular descriptors that are used to predict the activity.

Quantitative structure-property relationship (QSPR):

A *Quantitative Structure-Property Relationship (QSPR)* is quantitative relationship between a physicochemical property or environmental parameter (*e.g.* a partition coefficient) and one or more descriptors that are used to predict the property.

Randomisation testing:

Randomisation testing is a technique for checking the robustness of a QSAR model. In this test, the dependent variable vector, y-vector, is randomly shuffled and a new QSAR model is developed using the original independent variable matrix. The process is repeated several times. It is expected that the resulting QSAR models should generally have low r^2 and low q^2 LOO values.

If the new models developed from the data set with randomised responses have significantly lower R^2 and Q^2 than the original model, then strong evidence is provided that the proposed model is well founded, and not just the result of chance correlation.

In contrast, if all QSAR models obtained in the y-randomisation test have relatively high r^2 and q^2 LOO, it implies that an acceptable QSAR model cannot be obtained for the given data set by the current modelling method.

Receiver Operating Characteristics (ROC) Graph:

A *Receiver Operating Characteristics (ROC) Graph* can be used to visually compare the predictive abilities of different two-group classification models. The y axis of the ROC graph is the sensitivity (true positive rate) whereas the x axis is the false positive rate (1-specificity). The diagonal line in the plot represents models with random responses, whereas the top left corner represents the ideal model performance. Therefore, the best classification models are located in the upper left triangle, as close as possible to the corner.

Reliable (Q)SAR and reliability:

A (Q)SAR that is considered to be "reliable" or "valid" for a particular purpose is a model that exhibits an adequate performance for the intended purpose.

The criteria for determining whether the model performance is "adequate" will depend on the particular purpose and are highly context- dependent.

Regression Analysis:

Regression analysis is the use of statistical methods for modelling a dependent variable y in terms of predictors x (independent variables or molecular descriptors).

$$y= b_0+ b_1x1+b_2x2+\ldots\ldots b_n x_n$$

Simple *regression analysis* allows for a line of best fit to be placed between two sets of data. In QSAR analysis, quantitative measures of potency (*e.g.* LD_{50}, EC_{50}) may be used as the dependent variable and the physicochemical and / or structural descriptors of the molecule as independent variables. Thus, for a series of chemicals, the general form of the regression model between the concentrations causing a response (C) and a physicochemical property (PP) is:

$$\text{Log } 1 / C = a\,PP + c$$

Where *a* is the regression coefficient and *c* is the constant.

When more than one independent variable is used, it is termed multiple linear regression (MLR) analysis. This has the general form:

$$\text{Log } 1 / C = a\,PP_1 + b\,PP_2 \ldots. + c$$

There are a number of conditions that must be met for successful use of regression analysis for the development of QSARs. The number of independent variables must be as low as possible, and the ratio of observations to variables must be as high as possible (a ratio of 5:1 is considered an absolute minimum). Care must also be taken to ensure that all variables in a multiple linear regression analysis are significant

(this can be assessed by reference to the t-value for each variable and its associated probability and by the standardised regression coefficient) and preferably that no combinations of independent variables are collinear (unless an appropriate method has been applied to control the collinearity). Methods to check collinearity and to obtain reliable models even in presence of "some" collinearity in the descriptors are described in Chapter 5.

For both simple and multiple linear regression analysis, a number of measures of statistical fit are commonly applied. These include the standard error of the estimate, coefficient of determination, the Fisher statistic (and its associated probability) as well as measures of predictivity. The number of chemicals (data points) should also always be reported.

For large numbers of independent variables (*i.e*. physicochemical and/or structural properties) some form of variable selection technique is commonly applied. This may be an empirical process from the user, *i.e.* the selection of properties known or thought to be important. Alternatively, variable selection may employ stepwise selection techniques (forward or backward), best subsets selection, or the use of genetic algorithms.

Residual Sum of Squares (RSS):

Residual Sum of Squares (RSS) is the sum of the squares of the differences (residuals) between the experimental and estimated responses when predictions are made for objects in the training set.

Sensitivity: see Cooper statistics

Similarity analysis:

Similarity analysis refers to a variety of methods for quantifying the similarity between molecules in terms of their molecular structure (including shape, size, electronic and hydrophobic characteristics. Methods for performing *similarity analysis* generally are generally based on quantum-mechanical calculations, and are therefore implemented by specialised software packages.

Simplified Molecular Line Entry System (SMILES):

Simplified Molecular Line Entry System (SMILES) is a 2D or (very occasionally a 3D) representation of chemical structure. It is in the form a 2D string and has become a standard method for denoting structures in databases, and for inserting chemical structures into models for property calculation. The *SMILES* string is written by following a small number of rules, which are simple to learn and use. Briefly, in the *SMILES* string each non-hydrogen atom (hydrogen is only explicitly included in special circumstances) is denoted by its symbol; double and triple bonds are shown by "=" and "#" symbols, respectively; branches are shown in parentheses; and rings are opened and closed by the use of numbers.

Specificity: see Cooper statistics

Standard Deviation:

The *standard deviation* (s) is the square root of the variance. The variance of a sample (s^2) is given by the following formula:

$$s^2 = 1/(n-1) \sum (x_i - x')$$

where x_i are the values of the objects in the sample, and x' is the sample mean.

Standard Deviation Error in Calculation (SDEC):

The *Standard Deviation Error in Calculation (SDEC)* is similar to the standard error of the estimate. *SDEC* is given by the following formula:

$$SDEC = \sqrt{RSS / n}$$

where RSS is the residual sum of squares (RSS) and *n* is the number of objects in the training set.

Standard Deviation Error in Prediction (SDEP):

The *Standard Deviation Error in Prediction (SDEP)* is similar to SDEC, but the residuals are calculated by using the predicted value of the dependent variable (PRESS: Predictive Error Sum of Squares) when an observation is left out of the training set and put in the test set.

Standard Error of the Estimate (s):

The *standard error of the estimate (s)* is the square root of the residual sum of squares (RSS). The RSS are the sum of the squares of the residuals divided by the corresponding degrees of freedom.

Standardised Regression Coefficient:

The *standardised regression coefficients* are the coefficients of the independent variables (predictors) in a regression model divided by the standard deviation of the corresponding predictor. They provide a measure of the relative importance of the corresponding variable.

Structural alert:

A *structural alert* is a molecular (sub)structure associated with the presence of a biological activity.

Structure-activity relationship (SAR):

A Structure-Activity Relationship (SAR) is qualitative relationship (*i.e.* an association) between a molecular (sub)structure and the presence or absence of a biological activity, or the capacity to modulate a biological activity imparted by another substructure. A substructure associated with the presence of a biological activity is sometimes called a structural alert.

A *SAR* can also be based on the ensemble of steric and electronic features (biophore or toxicophore) considered necessary to ensure the intermolecular interaction with a specific biological target molecule, which results in the manifestation of a specific biological effect.

Similarly, the biophobe (or toxicophobe) refer to the features that are necessary to ensure the optimal intermolecular interaction with a specific biological target molecule, which results in the absence of a specific toxic effect.

Substructure:

A substructure is an atom, or group of adjacently connected atoms, in a molecule.

Superdelocalisability:

Superdelocabilisability is a descriptor, derived by quantum-mechanical calculation, that serves as an index of the reactivity of occupied and unoccupied orbitals in a molecule. A distinction is made between electrophilic and nucleophilic *superdelocalisability* (or acceptor and donor *superdelocalisability*, respectively): the former describes the interactions with an electrophilic centre, whereas the latter describing the interactions with a nucleophilic centre in the second reactant.

Supervised learning:

Supervised learning refers to the development of an algorithm (*e.g*. QSAR model) by a process that uses both the predictor and the response values, whereas in unsupervised learning, only the predictor values are used. Examples of *supervised learning* methods are (multiple) linear regression and discriminant analysis. Examples of unsupervised learning methods are different types of cluster analysis and principal components analysis (PCA).

Test set:

A *test set* is sometimes called an "independent" or "external" *test set* (or validation set), and distinguished from "training set". It is a set of chemicals, not present in the training set, selected for their use in assessing the predictive ability of a (Q)SAR.

For the purpose of (Q)SAR validation, it is important that the *test set* is representative of the training set, and contains a sufficient number of chemical structures.

Theoretical molecular descriptor:

A *theoretical molecular descriptor* is a number, obtained by applying a scientifically-based algorithm, that represents a particular aspect or feature (mono-dimensional, two-dimensional or three-dimensional) of the chemical structure. *Theoretical molecular descriptors* have the advantage that they can be generated for any chemical from a simple representation of its molecular structure (generally by using a specialised software programme). These descriptors can therefore be generated for chemicals that have not been synthesised, and used in QSARs for the purpose of lead identification in drug development.

Three-dimensional (3D) QSAR:

A technique that uses properties or theoretical descriptors derived from the 3D structure of a molecule (*e.g.* related to molecular size and the electric field around the molecule) as the descriptors for QSAR generation.

Topological descriptor:

A *topological descriptor* (or index) is a 2D descriptor of a molecule based on Graph Theory. Topological indices describe the connections between the atoms in a molecule. Typically, they are associated with the size or bulk of a molecule. Specific indices may describe the extent of branching *vs.* linearity in a molecule, or the contribution of rings to a molecule.

There are many different types of topological index, and thousands have been proposed in the QSAR literature. Specialised software packages have been developed to calculate many of these.

Total Sum of Squares (TSS):

The *total sum of squares (TSS)* is the sum of the squares of the differences between the experimental responses and the mean values.

Toxic endpoint:

A *toxic endpoint* is a measure of the deleterious effect to an organism following exposure to a chemical. A large number of *toxic endpoints* are used in regulatory assessments of chemicals. These include lethality, generation of tumours (carcinogenicity), immunological responses, organ effects, development and fertility effects.

It is the purpose of a toxicity test to determine whether a chemical has the potential to exhibit the toxic effect of interest, and in some cases, to determine relative potency. In QSAR analysis, it is important to develop models for individual *toxic endpoints*, and different methods may be required for different endpoints.

Toxicophore:

The ensemble of steric and electronic features that is necessary to ensure the optimal intermolecular interaction with a specific biological target molecule, which results in the manifestation of a specific toxic effect.

Training set:

A *training set* is a set of chemicals used to derive a QSAR. The data in a *training set* are typically organised in the form of a matrix of chemicals and their measured properties or effects in a consistent test method. A homogeneous *training set* is a set of chemicals which belong to a common chemical class, share a common chemical functionality, have a common skeleton, or common mechanism of action. A heterogeneous *training set* is a set of chemicals which belong to multiple chemical classes, or which do not share a common chemical functionality or common mechanism of action.

Unsupervised learning: see supervised learning.

Validation:

According to the OECD Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment, *validation* is defined as the process by which the reliability and relevance of a particular approach, method, process or assessment is established for a defined purpose (http://www.oecd.org/document/30/0,2340,en_2649_34365_1916638_1_1_1_1,00.html, accessed 6 February 2007).

Valid (Q)SAR, validated QSAR and validity:

A *validated (Q)SAR* is a model considered to be reliable for a particular purpose based on the results of the validation process in which the domain of application and the level of uncertainty required is defined.

A *valid (Q)SAR* is a model considered to be adequate for the intended purpose either because reliability has been demonstrated by historical use or by a validation process

The critieria for judging (Q)SAR *validity (reliability)* are determined by specific regulatory constraints in member countries which include the number of chemicals, time required in the descision process and the level of uncertainty acceptable for the regulatory application.

Variance ratio: see Fisher statistic

Y-scrambling: see Randomisation testing: