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

17-Dec-2004

English - Or. English

ENV/JM/MONO(2004)24
Unclassified

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

OECD SERIES ON TESTING AND ASSESSMENT
Number 49

THE REPORT FROM THE EXPERT GROUP ON (QUANTITATIVE) STRUCTURE-ACTIVITY RELATIONSHIPS [(Q)SARs] ON THE PRINCIPLES FOR THE VALIDATION OF (Q)SARs

2nd Meeting of the ad hoc Expert Group on QSARs
OECD Headquarters, 20-21 September, 2004

JT00176183

English - Or. English

OECD Environment Health and Safety Publications

Series on Testing and Assessment

No. 49

REPORT FROM THE EXPERT GROUP ON
(QUANTITATIVE) STRUCTURE-ACTIVITY RELATIONSHIPS [(Q)SARs]
ON THE PRINCIPLES FOR THE VALIDATION OF (Q)SARs

Environment Directorate
ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

Paris
November 2004

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)*

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 (in preparation)*

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)*

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 Pacific, 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 subsidiary 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 subsidiary 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 nine 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; and Emission Scenario Documents.** 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.

TABLE OF CONTENTS

SUMMARY

1. This report presents the outcome of the work under the OECD Work Programme on (Quantitative) Structure-Activity Relationships [(Q)SARs], which had the objective of establishing a set of principles for assessing the state of development and validation of (Q)SARs. The OECD Work Programme on (Q)SARs started in 2003 in order to enhance the use of (Q)SARs in the regulatory assessment of chemicals. The Expert Group on (Q)SARs, composed of experts nominated by OECD member countries to plan and implement the Work Programme, attached the highest priority to this exercise.

2. The Expert Group decided to start with the "Setubal Principles", which were originally proposed at an international workshop held in Setubal, Portugal, in March 2002. A check list was developed to provide guidance on the interpretation of Setubal principles, and to encourage consistency in their application to individual (Q)SARs (Table 3).

3. A team of experts (members listed in Table 2) applied the Setubal Principles to different (Q)SAR models, with a view to covering a diverse range of models. The individual contributions by the members of the team are presented in the 11 annexes to this report. The views expressed in each annex should be regarded as the views of the individual expert(s) who provided the contribution. On the basis of these contributions, the wording of the principles was refined and the check list was further developed.

4. The main conclusions from this exercise, endorsed by the Expert Group and the Joint Meeting of the Chemicals Committee and the Working Party on Chemicals, Pesticides and Biotechnology, are as follows:

- no substantial changes are required to the wording of the Setubal principles, although some rewording is proposed, to emphasise the intent of the principles and to simplify them
- the reworded Setubal principles should be referred to as the OECD Principles for (Q)SAR validation, to avoid confusion with the original principles
- the check list (revised as Table 4) provides useful guidance on the interpretation and application of the principles
- the check list should be updated periodically, to take account of new and useful considerations
- in the case of software programmes and expert systems that are based on the use of multiple (Q)SARs, it is important to identify the smallest component of the programme or expert system that functions independently, and to apply the principles to the individual component
- in the future work of the (Q)SAR Expert Group, detailed and non-prescriptive guidance should be developed, to explain and illustrate the application of the principles to different types of (Q)SAR models

5. Thus, the OECD principles for (Q)SAR validation, which are intended to be read in conjunction with the explanatory comments on paragraphs 22-29, read 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 "

6. It is also emphasised that the principles of (Q)SAR validation and the associated check list only identify the types of information that are considered useful for the regulatory review of (Q)SARs. The definition of criteria for determining the scientific validity and regulatory acceptability of (Q)SARs falls outside the scope of this Work Item, but should eventually be considered by individual national authorities.

7. Member countries should consider the OECD principles for (Q)SAR validation as scientific goals that provide generic base-line guidance for integrating (Q)SARs into regulatory frameworks. Flexibility will be needed in the interpretation and application of each 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.

INTRODUCTION

8. The regulatory use of structure-activity relationships (SARs) and quantitative structure-activity relationships (QSARs), collectively referred to as (Q)SARs, varies considerably between OECD Member Countries, and even between different agencies in the same Member Country. This is partly because different regulatory frameworks impose different requirements, but also because there have been no internationally-harmonised principles for assessing the development and validation status of (Q)SARs, which are needed to assess (Q)SARs according to a consistent conceptual framework.

9. It is therefore considered important to develop an internationally-agreed set of principles for (Q)SAR validation, not only to provide regulatory bodies with a scientific basis for making decisions on the acceptability (or otherwise) of (Q)SAR estimates of regulatory endpoints, but also to promote the mutual acceptance of (Q)SAR models.

10. A number of principles for assessing the validity of (Q)SARs 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 (1-4).

11. According to the so-called "Setubal principles", a (Q)SAR should:

1. be associated with a defined endpoint of regulatory importance

2. take the form of an unambiguous algorithm

3. ideally, have a mechanistic basis

4. be accompanied by a definition of domain of applicability

5. be associated with a measure of goodness-of–fit

6. be assessed in terms of its predictive power by using data not used in the development of the model.

12. The workshop did not produce any guidance on how to interpret and apply these principles.

13. In November 2002, the 34th Joint Meeting (JM) of the Chemicals Committee and the Working Party on Chemicals, Pesticides and Biotechnology decided to start a new OECD activity aimed at increasing the regulatory acceptance of (Q)SARs, and to establish an Expert Group for this work.

14. The 1st 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 1st Expert group Meeting proposed a (two-year) work plan for the OECD work on (Q)SARs. This included a Work Item based on the application of the Setubal principles to selected (Q)SARs, in order to evaluate the principles, and to refine them wherever necessary.

15. In June 2003, the proposed work plan was endorsed by the 35th JM. At the same meeting, the Commission (JRC) offered to take the lead in coordinating the Work Item on the Evaluation of the Setubal Principles. The offer was welcomed by the 35th JM, and was subsequently accepted by the Coordinating Group of the Expert Group on (Q)SARs (Table 1), which is responsible for managing the OECD Work Programme on QSARs.

16. A total of eleven case studies were developed by individual experts (or small groups of experts) by applying the principles to specific (Q)SARs or software models, which were considered to collectively provide a representative range of (Q)SAR approaches, covering a variety of physicochemical, environmental, ecological and human health endpoints (Table 2). To provide guidance on the application of the proposed principles, a check list of considerations was developed by the Coordinating Group (Table 3) and subsequently refined by the Expert Group. The experts were asked to consider the need to refine this check list, based on the experience gained in the case studies. On the basis of this experience, a refined check list was developed (Table 4). The refined check list was presented to the 16th Meeting of the Working Group of National Coordinators of the Test Guidelines Programme, held on 26-28 May.

17. This consolidated report provides an overview of the case studies provided by experts and summarises the progress made in refining the wording of the Setubal principles and the guidance on their application. The individual case studies are attached as Annexes 1-11, and the views expressed therein are the views of the identified authors. The present report, which represents the consensus view of the QSAR Expert Group, makes conclusions regarding the wording and application of the (Q)SAR principles, and discusses the status and future use of the principles by regulatory authorities and validation bodies.

CONCLUSIONS AND RECOMMENDATIONS

Rewording of the Setubal Principles

18. The authors of the case studies generally expressed support for the Setubal principles and noted the usefulness of the initial check list developed to guide the application of the principles.

19. Within the Coordinating Group, there was some initial concern that the principles would be presented as mandatory for the acceptance all types of models, which could result in a low acceptance of models, if the principles were too strict. However, further discussion clarified that the principles were

intended to guide the (Q)SAR validation process, and to facilitate any subsequent regulatory acceptance process, by defining the types of information that regulators would find useful when considering the acceptability of individual (Q)SAR models.

20. Bearing this in mind, the Coordinating Group submitted a set of revised Principles to the Expert Group for their consideration. The Expert Group has agreed to the following rewording of the Setubal principles:

> *"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"*

21. These principles are to be submitted to the Joint Meeting as the OECD principles for (Q)SAR validation. There was consensus that the concepts of both internal and external validation are important to the overall validation process for (Q)SARs. However, there was extensive discussion and a lack of consensus at the Expert Group meeting on Principle 4, appropriate measures of goodness-of-fit, robustness and predictivity. A considerable number of the experts asked that this Principle be reworded as two separate Principles, consistent with the original wording of the Setubal Principles, on the basis that the current approach does not give sufficient emphasis to external validation. Other Members felt that a single Principle was more appropriate to allow flexibility in regulatory acceptance for member countries.

Comments on Principle 1

22. The intent of Principle 1 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 modelled by the (Q)SAR.

23. The Expert Group recommends that further guidance is provided regarding the interpretation of "defined endpoint". For example, a no-observed-effect level could be regarded as 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.

Comments on Principle 2

24. The intent of Principle 2 is to ensure transparency in the model algorithm that generates predictions of an endpoint from information on chemical structure and/or physicochemical properties. The Expert Group recognises 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.

Comments on Principle 3

25. 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.

26. The Expert Group recommends that further work is carried out to define what types of information are needed to define (Q)SAR applicability domains, and to develop appropriate methods for obtaining this information.

Comments on Principle 4

27. This Principle includes the intent of the original Setubal Principles 5 and 6. The wording of the revised 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).

28. The Expert Group recommends the development of detailed guidance on the statistical approaches that could be used to provide appropriate measures of internal performance and predictivity.

Comments on Principle 5

29. The Expert Group recognises that it is not always possible, from a scientific viewpoint, to provide a mechanistic interpretation of a given (Q)SAR, or 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.

Guidance on the application of the (Q)SAR Principles

30. The Expert Group reached the following conclusions on the reworded principles and their application:

1. the reworded principles should be referred to as the “OECD Principles for (Q)SAR Validation”, to avoid confusion with the original Setubal principles.

2. the reworded principles are of sufficient generality to be applicable to a wide range of current and future (Q)SAR models.

3. the check list is a useful summary of considerations that can be applied during the validation of (Q)SARs,

4. the check list should be updated periodically, to include new considerations;

5. in the case of software programs or expert systems based on the use of multiple (Q)SAR models, the principles should be applied to the smallest component of the program/system that functions as an independent unit for the purpose of generating a prediction.

6. a Guidance Document on (Q)SAR validation, as foreseen in Work Item 2, is needed to explain, in a detailed and non-prescriptive manner, how the principles can be applied for different types of (Q)SAR models.

Eventual use of the (Q)SAR Principles

31. The revised principles and check list are intended to define the types of information that should be provided to facilitate the regulatory consideration of (Q)SARs. The principles are not intended to pre-empt how regulatory decisions should be made, since this will depend on the regulatory framework making use of (Q)SARs, and on the specific circumstances surrounding a given decision.

32. Nevertheless, when considering the acceptability of a (Q)SAR estimate, regulatory authorities will need to evaluate not only whether they have sufficient information on the model used to generate the estimate, but also whether they consider the estimate to be reliable. In other words, it will be necessary to evaluate whether the chemical of interest falls within the applicability domain of the model, and whether the model is of sufficiently high quality for estimates within this domain to be considered reliable. This means that criteria will need to be developed to determine: a) whether a given chemical falls within the defined applicability domain of a (Q)SAR; and b) whether the (Q)SAR generates reliable estimates within the defined applicability domain. The definition of such criteria falls outside the scope of this report.

REFERENCES

Cronin, M.T.D., Walker, J.D., Jaworska, J.S., Comber, M.H.I., Watts, C.D. & Worth, A.P. (2003). 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., Jaworska, J.S., Walker, J.D., Comber, M.H.I, Watts, C.D. & Worth, A.P. (2003). 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., Jaworska, J.S., Worth, A.P., Cronin, M.T.D., McDowell, R.M. & Gramatica, P. (2003). Methods for reliability, uncertainty assessment, and applicability evaluations of classification and regression based QSARs. *Environmental Health Perspectives* **111**, 1361-1375.

Jaworska, J.S., Comber, M., Auer, C. & van Leeuwen, C.J. (2003). Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints. *Environmental Health Perspectives* **111**, 1358-1360.

TABLE 1 MEMBERS OF THE (Q)SAR COORDINATING GROUP

Name	**Affiliation**
Richard Becker	Business & Industry Advisory Committee (BIAC)
Romualdo Benigni	Istituto Superiore di Sanita', Rome, Italy
Michael Comber	Business & Industry Advisory Committee (BIAC)
Drew MacDonald	Environment Canada
Betty Hakkert	RIVM, Utrecht, Netherlands
Oscar Hernandez	US EPA, Washington, USA
Joanna Jaworska	Business & Industry Advisory Committee (BIAC)
Peter Howden	HSE, Bootle, UK
Jay Niemelå	Danish EPA, Copenhagen, Denmark
Herbert Rosenkranz	International Council on Animal Protection in OECD Programmes [ICAPO]
Yuki Sakuratani	Japan
Eisaku Toda	OECD, Paris, France
Gilman Veith	OECD, Paris, France
Suzanne Wiandt	UBA, Germany
Andrew Worth	European Commission - Joint Research Centre, Ispra, Italy

TABLE 2 CASE STUDIES ON THE APPLICATION OF THE SETUBAL PRINCIPLES

Case study	Model	Expert(s) involved
1	QSARs for acute fish toxicity	Mark Cronin Liverpool John Moores University, UK
2	QSARs for atmospheric degradation	Paola Gramatica University of Insubria, Varese, Italy
3	QSARs for mutagenicity and carcinogenicity	Romualdo Benigni Istituto Superiore di Sanita', Rome, Italy
4	Multi-CASE model for *in vitro* chromosomal aberrations	Jay Niemelå and Eva Wedebye Danish EPA, Copenhagen, Denmark
5	Multi-CASE and MDL models for human NOEL	Edwin Matthews and Joseph Contrera US FDA, Washington, USA
6	ECOSAR	Etje Hulzebos RIVM, Utrecht, Netherlands
7	BIOWIN	Theo Traas RIVM, Utrecht, Netherlands
8	DEREK	Etje Hulzebos RIVM, Utrecht, Netherlands
9	DEREK skin sensitisation rulebase	Grace Patlewicz Unilever, Bedford, UK
10	Japanese METI biodegradation model	Yuki Sakuratani Ministry of Economy, Trade and Industry (METI), Japan
11	Rat oral chronic toxicity models in TOPKAT	Roger Breton, Ottawa, Canada

TABLE 3: INITIAL CHECK LIST USED TO PROVIDE GUIDANCE ON THE INTERPRETATION OF THE SETUBAL PRINCIPLES

PRINCIPLE	CONSIDERATIONS	YES / NO
1) Defined endpoint	**1.1)** Does the model have a clearly defined scientific purpose? (i.e. does it make predictions of a clearly defined physicochemical, biological or environmental effect?) **1.2)** Does the model have the potential 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)** Is information given about important experimental conditions that affect the measurement and therefore the prediction (e.g. sex, species, temperature, exposure period) **1.4)** Are the units of measurement of the endpoint given?	
2) Defined algorithm	**2.1)** In the case of a SAR, is there an explicit description of the substructure, including an explicit identification of its substituents? **2.2)** In the case of a QSAR, is the equation explicitly defined, including definitions of all descriptors used[3]?	
3) Mechanistic basis	**3.1)** In the case of a SAR, is there a description of the molecular events that underlie the reactivity of the molecule? (e.g. a description of how substructural features could act as nucleophiles or electrophiles, or form part or all of a receptor-binding region) **3.2)** In the case of a QSAR, do the descriptors have a physicochemical interpretation that is consistent with a known mechanism (of biological action)? **3.3)** Are any literature references cited in support of the purported mechanistic basis of the (Q)SAR?	
4) Domain of applicability	**4.1)** In the case of a SAR, is the substructure associated with any inclusion and/or exclusion rules on its applicability to groups of chemicals? **4.2)** In the case of a SAR, is the substructure associated with rules regarding the modulatory effects of the substructure's molecular environment?	

	4.3) In the case of a (Q)SAR, are the descriptor and response variables associated with inclusion and/or exclusion rules that define the variable ranges for which the QSAR is applicable (i.e. makes reliable estimates)?
5) Internal performance	**5.1)** Are full details of the training set given, including details of chemical names, structural formulae, CAS numbers (if available), and data for all descriptor and response variables?
	5.2) If the data used to the develop the model were based upon the processing of raw data (e.g. the averaging of replicate values):
	a) is there an adequate description of the data processing?
	b) are the raw data provided?
	5.3) Is there a specification of the statistical method(s) used to develop the QSAR (including details of any software packages used)?
	5.4) Is the QSAR associated with basic statistics for its goodness-of-fit to the training set?
	(e.g. r^2 values and standard error of the estimate in the case of regression models)
	5.5)
	a) Is the QSAR associated with any statistics based on cross-validation or resampling?
	b) If yes, is the number or samples used indicated?

6) Predictivity (External validation)	**6.1)** Does application of the appropriate statistical method(s) to the training set result in the same (Q)SAR model?
	6.2) Is there any information to indicate that the (Q)SAR has been validated previously, using a test set that is independent of the training set?
	6.3) If an external validation has been performed, is the following information available:
	a) the number of test structures?
	b) the identities of the test structures?
	c) the approach for selecting the test structures?
	d) the statistical analysis of the predictive performance of the model?
	(e.g. including sensitivity, specificity, and positive and negative predictivities for classification models)
	e) a comparison of the predictive performance of the model against previously-defined quantitative performance criteria?

FOOTNOTES

The term “model” refers to a single QSAR equation, or set of SARs (structural alerts) associated with the prediction of a specific physicochemical, biological or environmental effect.

It should be noted that that the term “model” is sometimes also used to refer to a software package or system based on multiple QSAR equations and/or sets of SARs. In such cases, an evaluation should be made, if possible, of the individual QSARs and sets of SARs that form the component parts of the complex system.

TABLE 4: REVISED CHECK LIST FOR PROVIDING GUIDANCE ON THE INTERPRETATION OF THE OECD PRINCIPLES

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?	

4) Internal performance and predictivity

Internal performance

4.1 Full details of the training set given, including details of:

a) chemical names

b) structural formulae

c) CAS numbers

d) data for all descriptor variables

e) data for all response variables

f) 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?

Predictivity

4.8 a) An indication whether the model has been validated by using a test set that is independent of the training set

b) If yes to a), is there an indication of the quality of the data in the test set?

4.9 If an external validation has been performed:

a) the number of test structures

b) the identities of the test structures

c) the approach for selecting the test structures, specifying how the applicability domain of the model is represented by the test set

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 indication of quality of the test data

f) 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

g) 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 interpretation 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)?

ANNEXES

ANNEX 1
QSARS FOR ACUTE FISH TOXICITY

Dr Mark Cronin
School of Pharmacy and Chemistry
Liverpool John Moores University
Liverpool L3 3AF
UK

TABLE OF CONTENTS

1. INTRODUCTION

33. Dr Mark Cronin (Liverpool John Moores University, UK) has retrospectively applied, under the terms of a JRC contract, Setubal Principles 1-6 to the following four QSAR models for acute fish toxicity:

1. QSAR 1 (non-polar narcosis): European Commission (1995*) QSAR for Predicting Fate and Effects of Chemicals in the Environment,* Final report of DG XII contract No. EV5V-CT92-0211.

2. QSAR 2 (polar narcosis): Veith, G.D., Broderius, S.J. (1987) Structure-toxicity relationships for industry chemicals causing type (II) narcosis syndrome. In: Kaiser, K.L.E. (ed.) *QSAR in Environmental Toxicology – II*. D. Reidel, Dordrecht, pp. 385-391.

3. QSAR 3 (mixed mechanism of toxic action): Veith, G.D., Mekenyan, O.G. (1993) A QSAR approach for estimating the aquatic toxicity of soft electrophiles [QSAR for soft electrophiles]. *Quantitative Structure-Activity Relationships* 12: 349-356

4. QSAR 4 (MultiCASE model): Klopman, G., Saiakhov, R., Rosenkranz, H.S. (2000) Multiple Computer-Automated Structure Evaluation study of aquatic toxicity II. Fathead minnow. *Environmental Toxicology and Chemistry* 19: 441-447.

34. For comparative purposes, all QSARs were developed by using the fathead minnow (*Pimphales promelas*) database. The fathead minnow database was developed by the United States Environmental Protection Agency (Mid-Continent Ecology Division, Duluth MN, USA). The database is one of the few publicly available databases for QSAR development of a regulatory endpoint that has been measured by a single protocol and in a single laboratory (key amongst the requirements for high quality data). Further, the data are well supported by mechanistic evaluation and comprehensive, notably being derived from joint binary toxicity studies, and Fish Acute Toxicity Syndrome (FATS) data. The quality, size, and chemical breadth of the fathead minnow dataset developed by the United States Environmental Protection Agency have made it a very suitable data set for modelling. The complete database and mechanistic information is summarised by Russom et al (1).

35. The QSARs considered here represent models for two very well defined mechanisms of action: non-polar narcosis (QSAR 1) and polar narcosis (QSAR 2); a more general model that incorporates potentially bioreactive (electrophilic) compounds (QSAR 3); and a commercially available expert system derived from the fathead minnow database (QSAR 4).

36. For each QSAR, the results of the investigation are provided in a discursive form and in tabular form (Appendices 1-4).

2. APPLICATION OF THE SETUBAL PRINCIPLES TO QSAR 1

2.1. Defined endpoint (Principle 1)

37. The QSAR could potentially fulfil a clear regulatory need, since the 96h LC_{50} in the fathead minnow is one of the endpoints referred to in OECD Test Guideline 203.

2.2. Defined algorithm (Principle 2)

38. The QSAR considered here for predicting the acute toxicity of organic chemicals to the fathead minnow (*Pimephales promelas*) is recommended for use in the European Union Technical Guidance Document (2), and was published originally by the European Commission (3):

$$\text{Log } (LC_{50}) = -0.846 \log K_{ow} - 1.39$$

where LC_{50} is the concentration (in moles per litre) causing 50% lethality in *Pimephales promelas*, after an exposure of 96 hours; and K_{ow} is the octanol-water partition coefficient. It should be noted that the inverse of toxicity (negative logarithm) was not reported in the original form of this equation.

39. The QSAR is a regression model, based on a single parameter. Linear regression analysis is considered to be one of the most transparent methods for the development of QSARs (Cronin and Schultz, 4; Schultz and Cronin, 5), and the use of a single predictor variable makes the QSAR both simple and user-friendly.

2.3. Mechanistic basis (Principle 3)

40. The QSAR was developed for chemicals considered to act by a single mechanism of toxic action, non-polar narcosis, as defined by Verhaar et al. (6), and therefore has a clear mechanistic basis. In fact, non-polar narcosis is one of the most established mechanisms of toxic action.

41. Non-polar narcosis has been established experimentally by using the Fish Acute Toxicity Syndrome methodology (McKim et al., 7). The QSAR is based on a descriptor for hydrophobicity (log K_{ow}), which is relevant to the mechanism of action, i.e. toxicity results from the accumulation of molecules in biological membranes. There are numerous regression models based on log K_{ow} for this mechanism of action, which has enabled the development of such models for inter-species comparisons (Dimitrov et al., 8).

2.4. Domain of applicability (Principle 4)

42. The domain of applicability of the QSAR was well defined by the model developer. The QSAR was stated to be applicable to chemicals having log K_{ow} values in the range from –1.24 to 5.13, and operating by a non-polar narcosis mechanism of action. Such chemicals can be identified on a structural basis (6), or from physicochemical descriptors (Boxall et al., 9).

2.5. Internal performance (Principle 5)

2.5.1 Independent assessment of data quality

43. The training set is given in Table 1. It consists of the biological (96h LC_{50}) data being modelled, and data for a single descriptor (K_{ow}), which is being used as a predictor variable. The biological data can be considered to be of very high quality, since they were obtained by applying a single protocol, and measured in the same laboratory, possibly by the same worker.

44. The descriptor (K_{ow}) data are a mixture of experimental and calculated values. Generally, K_{ow} is considered to be a high quality physicochemical descriptor, and the range of log K_{ow} values (Table 1) is well within that usually considered to provide adequate measured values. However, there is no certainty that the measurements of K_{ow} were made by the same protocol, or in the same laboratory, so this could result in a small amount of variability. Furthermore, using a mixture of calculated and experimental values will also result in some variability.

2.5.2 Goodness-of-fit

45. The following statistics were reported for this QSAR: n = 58, r^2 = 0.937, Q^2 = 0.932 and s.e. = 0.361.

2.6. External validation for predictivity (Principle 6)

46. The application of linear regression to the training set of data enabled the QSAR to be confirmed. However, it was decided to re-express the QSAR as follows:

$$\text{Log } (1/LC_{50}) = 0.846 \log K_{ow} + 1.39$$

47. This is a standard way of expressing models for the prediction of LC_{50} values, and is easy to understand since a numerical increase in the response variable means an increase in toxicity (acute lethality in this case). Furthermore, it is apparent that the QSAR relates an increasing toxicity to an increasing partition coefficient.

48. To validate the "confirmed" QSAR, Worth et al (10) took physicochemical and toxicity data for 20 chemicals from Russom et al. (1). This is also the source of data used to develop the training set (Table 1), so the LC_{50} and K_{ow} data in the test set (Table 2) can be assumed to be of a similar quality.

49. The domain of the external test set falls within that of the training set. The range of log K_{ow} values for the test set is from -0.43 to 4.50. The chemical structures represented by the test set are consistent with those representing non-polar narcosis (Verhaar et al., 6), and are similar to those in the training set. The predicted toxicities of the test set chemicals are reported in Table 2. There is a statistically significant relationship between predicted and observed toxicity:

$$\text{Log } (1/LC_{50})\text{observed} = 0.963 \log K_{ow} - 0.08$$

$$n = 20 \qquad r^2(\text{adj}) = 0.922 \qquad s = 0.329 \qquad F = 225$$

50. This relationship is shown in Figure 1. Investigation of the residuals of predicted toxicity, as well as the standard error from the above equation, suggests that the toxicity of chemicals may be predicted within a 95% confidence interval of ±0.64 log units.

3. APPLICATION OF THE SETUBAL PRINCIPLES TO QSAR 2

3.1. Defined endpoint (Principle 1)

51. The QSAR could potentially fulfil a clear regulatory need, since the 96h LC_{50} in the fathead minnow is one of the endpoints referred to in OECD Test Guideline 203.

3.2 Defined algorithm (Principle 2)

52. The QSAR considered here was published by Veith and Broderius (1987) and is very similar to the model that is recommended by the European Union Technical Guidance Document (European Economic Community (1996)) for polar narcosis.

53. Specifically, the following QSAR for predicting the acute toxicity of organic chemicals to the fathead minnow (*Pimephales promelas*) was reported by Veith and Broderius (11):

$$\text{Log } (LC_{50}) = -0.65 \log K_{ow} - 2.29$$

where LC_{50} is the concentration (in moles per litre) causing 50% lethality in *Pimephales promelas*, after an exposure of 96 hours; and K_{ow} is the octanol-water partition coefficient. It should be noted that the inverse of toxicity (negative logarithm) was not reported in the original form of this equation.

54. The QSAR is a regression model, based on a single parameter. Linear regression analysis is considered to be one of the most transparent methods for the development of QSARs (4, 5), and the use of a single predictor variable makes the QSAR both simple and user-friendly.

3.3. Mechanistic basis (Principle 3)

55. The QSAR was developed for chemicals considered to act by a single mechanism of toxic action, polar narcosis, as defined by Veith and Broderius (11), and therefore has a clear mechanistic basis.

56. Polar narcosis is a defined mechanism of toxic action with potency above that normally associated with non-polar narcosis. This report will use the term “polar narcosis”, this mechanism of toxic action is also termed “Type II Narcosis” by some workers e.g. Russom et al (1).

57. Polar narcosis has been established experimentally by using the Fish Acute Toxicity Syndrome methodology (McKim et al., 7). The QSAR 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. Polar narcotics are typically defined as aromatic molecules that have a polar group (typically an hydroxyl or amine, but also possibly a nitro group). Further they may have a number of substituents such as alkoxy or alkyl groups and three or less halogens. Such molecules are clearly narcotic (in that they cause a reversible effect) however, their toxic effects are well in excess of that elicited by non-polar narcosis, and joint binary toxicity studies indicate different mechanisms of action. In terms of QSAR modelling, it is commonly considered that there is still a strong relationship between toxicity and

hydrophobicity, and QSARs based on log Kow alone should have a lower slope and higher intercept than those for non-polar narcosis.

3.4. Domain of applicability (Principle 4)

58. The domain of applicability was not defined by the model developer. Investigation of the data reveals that it is applicable to chemicals having log K_{ow} values in the range from 0.90 to 6.36, and operating by a polar narcosis mechanism of action. Such chemicals can be identified on a structural basis (Verhaar et al., 6), or from physicochemical descriptors (Boxall et al., 9).

3.5. Internal performance (Principle 5)

3.5.1. Independent assessment of data quality

59. The training set is given in Table 3. It consists of the biological (96h LC_{50}) data being modelled, and data for a single descriptor (K_{ow}), which is being used as a predictor variable. The biological data can be considered to be of very high quality, since they were obtained by applying a single protocol, and measured in the same laboratory, possibly by the same worker.

60. The descriptor (K_{ow}) data are a mixture of experimental and calculated values. Generally, K_{ow} is considered to be a high quality physicochemical descriptor, and the range of log K_{ow} values (Table 3) is well within that usually considered to provide adequate measured values. However, there is no certainty that the measurements of K_{ow} were made by the same protocol, or in the same laboratory, so this could result in a small amount of variability. Furthermore, using a mixture of calculated and experimental values will also result in some variability.

3.5.2. Goodness-of-fit

61. The following statistics were reported for this QSAR: n = 39, r^2 = 0.90, Q^2, s, F not given and s.e. = 0.361.

3.5.3. Cross-validation

62. In order to assess internal predictivity in this study, a leave-many-out (LMO) analysis was performed. Ideally as large a number of compounds as possible should be omitted (upto 50%). However due to the small number of compounds in the model (39), omitting 50% was felt to be too severe a test and may adversely affect the model. Therefore a leave-25%-out approach was performed. This was achieved by ordering the compounds according to toxicity and omitting every 4^{th} compound, then re-calculating the QSAR and making predictions for the 25% omitted. This was performed four times so every compound had been left out once. The toxicity predicted by this analysis is also shown in Table 3. The relationship between observed toxicity and that predicted by LMO was:

$$\text{Log } (1/LC_{50})_{observed} = 0.987 \text{ log } (1/LC_{50})_{LMO} - 0.051$$

n = 39 r^2 = 0.881 s = 0.290 F = 281

63. Toxicity is well predicted by LMO as demonstrated by the high r^2 and low s. Both these values are similar to the original. In addition the slope of the equation is almost unity and intercept almost zero.

3.6. External validation for predictivity (Principle 6)

64. The application of linear regression to the training set of data enabled the QSAR to be confirmed. However, it was decided to re-express the QSAR as follows:

$$\text{Log } (1/LC_{50}) = 0.651 \log K_{ow} + 2.29$$

$$n = 39 \qquad r^2 = 0.898 \quad Q^2 = 0.888 \quad s = 0.267 \quad F = 337$$

65. To validate the "confirmed" QSAR, physicochemical and toxicity data for 11 chemicals were taken from Russom et al. (1). This is also the source of data used to develop the training set (Table 3), so the LC_{50} and K_{ow} data in the test set (Table 4) can be assumed to be of a similar quality.

66. The domain of the external test set falls within that of the training set. The range of log K_{ow} values for the test set is from 0.48 to 3.00. The chemical structures represented by the test set are consistent with those representing polar narcosis (Verhaar et al., 6). However, it should be noted that the range of log K_{ow} values for the test set is about half that of the training set, and this is important in the subsequent statistics. Further, whilst the test set meets the criteria of polar narcosis from a structural basis, some compounds (such as the pyridines) are not well represented by the training set.

67. The predicted toxicities of the test set chemicals are reported in Table 4. There is a statistically significant relationship between predicted and observed toxicity, although the exact fit is not high:

$$\text{Log } (1/LC_{50})\text{observed} = 1.15 \log K_{ow} - 0.56$$

$$n = 11 \qquad r^2(\text{adj}) = 0.551 \qquad s = 0.559 \quad F = 13.5$$

68. This relationship is shown in Figure 2. Investigation of the residuals of predicted toxicity, as well as the standard error from the above equation, suggests that the toxicity of chemicals may be predicted within a 95% confidence interval of ±1.09 log unit.

4. APPLICATION OF THE SETUBAL PRINCIPLES TO QSAR 3

4.1. Defined endpoint (Principle 1)

69. The QSAR could potentially fulfil a clear regulatory need, since the 96h LC_{50} in the fathead minnow is one of the endpoints referred to in OECD Test Guideline 203.

4.2. Defined algorithm (Principle 2)

70. The following QSAR for predicting the acute toxicity of organic chemicals to the fathead minnow (*Pimephales promelas*) was reported by Veith and Mekenyan (12):

$$\text{Log } (1/LC_{50}) = 0.58 \log K_{ow} - 0.42 - E_{LUMO} - 2.41$$

where LC_{50} is the concentration (in moles per litre) causing 50% lethality in *Pimephales promelas*, after an exposure of 96 hours; and K_{ow} is the octanol-water partition coefficient; and E_{LUMO} is the energy of the lowest unoccupied molecular orbital.

71. The QSAR is a regression model, based on two parameters. Linear regression analysis is considered to be one of the most transparent methods for the development of QSARs (4, 5), and the use of a two predictor variables makes the QSAR relatively simple and user-friendly.

4.3. Mechanistic basis (Principle 3)

72. Typically, "mixed" mechanisms of action include compounds acting by narcoses mechanisms as well as those acting by unspecific "bioreactive" mechanisms. Bioreactive mechanisms have been characterised as involving electrophilic (and in some cases nucleophilic) reactions within the cell or organism. They are clearly associated with toxicity in excess of that from narcosis, and joint binary toxicity and fish acute toxicity syndrome studies demonstrate that they are separate from narcosis. Mostly the exact mechanism of action is not known, but it is widely assumed to be a covalent reaction with a biological macromolecule (e.g. a protein or DNA etc). Many reactive toxicants will also be mutagenic and skin sensitisers etc. Due to the reactivity component of the toxicity, a further parameter is required to be included in the modelling process. Typically this has been a molecular orbital property such as the energy of the lowest unoccupied molecular orbital (E_{LUMO}), or a nucleophilic superdelocalisability.

73. The QSAR was developed for aromatic chemicals considered to act by a number of mechanisms of toxic action. These include non-polar and polar narcosis as well as unspecific electrophilicity as defined by Russom et al. (1). Whilst there are a number of mechanisms of action, they are well defined, and so the QSAR can be considered to have a clear mechanistic basis.

74. The QSAR is based on two descriptors. The first for hydrophobicity (log K_{ow}) is relevant to the mechanism of action, i.e. toxicity results from the accumulation of molecules in biological membranes. The second for electrophilicity (E_{LUMO}) also relates to the mechanism of action for those chemicals that are capable of reacting with biological macromolecules. It should be recognised that in the approach described by Veith and Mekenyan (12), some of the chemicals in the data set (the narcotics) are not acting by electrophilic reactivity; some later approaches using these two parameters tended to exclude narcotics, and non-polar narcotics in particular (e.g. Cronin et al., 13).

4.4. Domain of applicability (Principle 4)

75. The domain of applicability was not defined by the model developer. Investigation of the data reveals that it is applicable to chemicals having log K_{ow} values in the range from 0.34 to 7.54, and E_{LUMO} values ranging from -2.52 to 0.73. As noted above, compounds are present in the training set that may operate by a number of mechanisms of action including non-polar and polar narcosis as well as unspecified electrophilicity. All compounds in the training set were aromatic, and the structures represented included alkyl, halogen benzenes, as well as similar substituents on phenols and anilines.

4.5. Internal performance (Principle 5)

4.5.1. Independent assessment of data quality

76. The training set is given in Table 5. It consists of the biological (96h LC_{50}) data being modelled, and data for a two descriptor (K_{ow} and E_{LUMO},) which are used as predictor variables. The biological data can be considered to be of very high quality, since they were obtained by applying a single protocol, and measured in the same laboratory, possibly by the same worker.

77. The descriptor (K_{ow}) data are a mixture of experimental and calculated values. Generally, K_{ow} is considered to be a high quality physicochemical descriptor, and the range of log K_{ow} values (Table 5) is well within that usually considered to provide adequate measured values. However, there is no certainty that the measurements of K_{ow} were made by the same protocol, or in the same laboratory, so this could result in a small amount of variability. Furthermore, using a mixture of calculated and experimental values will also result in some variability.

78. It should be noted that the calculation of E_{LUMO} in the original paper utilised the MNDO calculation method. This may now be viewed as rather limited to newer Hamiltonians such as AM1, PM3, PM5 as well as more accessible *ab initio* methods. In addition, the calculation method applied was in the OASIS software, which is a restricted package in terms of availability. Further application (and any future further validation) may require re-calculation of E_{LUMO} by AM1 using, for instance, the more widely accessible MOPAC package. There is known to be variation in calculated values of molecular orbital properties, however, much of this variation is due to the geometry optimisation step. The molecules in the training are largely unflexible, and so this issue may be minimised.

4.5.2. Goodness-of-fit

79. The following statistics were reported for this QSAR: n = 114, r^2 = 0.81, s^2 = 0.19, F = 245, Q^2 not given.

4.5.3. Cross-validation

80. In order to assess internal predicitivity in this study, a leave-many-out (LMO) analysis was performed. The relatively large number of compounds in the training set enabled a leave-50%-out approach to be applied. This was achieved by ordering the compounds according to toxicity and omitting every 2nd compound, then re-calculating the QSAR and making predictions for the 50% omitted. This was performed twice so every compound had been left out once. The toxicity predicted by this analysis is also shown in Table 5. The relationship between observed toxicity and that predicted by LMO was:

$$\text{Log } (1/LC_{50})_{observed} = 0.965 \text{ log } (1/LC_{50})_{LMO} + 0.150$$

n = 114 r^2 = 0.769 s = 0.487 F = 378

81. Toxicity is well predicted by LMO as demonstrated by the high r^2 and low s. Both these values are similar to the original. In addition the slope of the equation is almost unity and intercept almost zero.

4.6. External validation for predictivity (Principle 6)

82. The application of linear regression to the training set of data revealed a slightly different QSAR in terms of the coefficient on E_{LUMO} (-0.47 as opposed to -0.42 in the original publication), the reason for this discrepancy is unclear:

$$\text{Log } (1/LC_{50}) = 0.579 \text{ log } K_{ow} - 0.473 - E_{LUMO} - 2.41$$

n = 114 r^2 = 0.801 Q^2=0.790 s = 0.453 F = 228

83. At present, it is not possible to validate the external predictivity of the QSAR. This is due to the non-availability of the OASIS software to calculate the required E_{LUMO} values (as noted above). For this reason, in addition to the use of more accurate methods to calculate E_{LUMO} it recommended that further future validation of the QSAR should involved a complete recalculation of E_{LUMO}.

5. BACKGROUND INFORMATION ON QSAR 4

84. The predictive approaches described above are regression-based models for toxicity prediction. For comparative purposes, it was decided to also apply the Setubal principles to an expert system. A number of commercially available expert system models have been developed from the fathead minnow dataset, including the MulitCASE model considered here, and other systems such as ECOSAR, OASIS, TOPKAT and a variety of neural network approaches.

85. The reason MulitCASE was chosen for this study is because the modelling approach has been described elsewhere in the assessment of the Setubal principles, and that a publication (Klopman et al, 14) details the model and its performance. This assessment is based on the findings and discussion of the original article, and so cannot be considered as a true assessment of system. The reader is asked to bear these factors in mind when making any judgement on the model.

86. The Multiple Computer-Automated Structure Evaluation (MultiCASE or M-CASE) approach is well described in a series of publications from the vendors. The expert system contains a large number of toxicity, pharmacological, and physicochemical prediction methods, of which the model to predict fathead minnow toxicity is one. The models are available commercially from MultiCASE Inc., for further details contacts Prof Gilles Klopman, Department of Chemistry, CASE Western Reserve University, Cleveland, OH, USA (e-mail: klopman@multicase.com). MultiCASE and its associated programs are well recognised in the market place and various modules (although not necessarily the one described here) have been utilised by Regulatory Agencies worldwide (e.g. US Food and Drug Administration (FDA) and the Danish Environmental Protection Agency).

87. The MultiCASE methodology is well described elsewhere (14-17), and only a cursory consideration of the methodology is provided here. Briefly the MultiCASE algorithm works best with large amounts of heterogeneous data. The program divides each molecule into all possible fragments (up to, for instance, 8-10 atoms) and then determines if any of the fragments are related positively (biophores, or toxicophore), or negatively (biophobe, or toxicophobe) to activity. Also included into the hierarchical assessment are physicochemical parameters (such as log K_{ow}, E_{LUMO} etc.) although the exact number and type of these parameters is not specified in this study. The modelling process also incorporates the so-called baseline activity identification algorithm (BAIA). This is designed identify baseline activity associated with a specific physical attribute (such as log K_{ow}). The MultiCASE model therefore starts by identifying any baseline effects. This is, in principle, analogous to developing a non-polar narcosis model and utilising fragments to help account for “excess toxicity” i.e. more toxic narcosis and bioreactive mechanisms.

6. APPLICATION OF THE SETUBAL PRINCIPLES TO QSAR 4

6.1. Defined endpoint (Principle 1)

88. The QSAR could potentially fulfil a clear regulatory need, since the 96h LC_{50} in the fathead minnow is one of the endpoints referred to in OECD Test Guideline 203.

6.2. Defined algorithm (Principle 2)

89. Application of the BAIA to the database of toxicity values showed a baseline effect. A QSAR for "baseline" toxicity is presented:

$$\text{Log } (1/LC_{50}) = 0.85 \log K_{ow} - 4.94$$

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

90. This QSAR is similar to that normally assumed for non-polar narcosis for this endpoint (note the difference in intercept with the model assessed above is due to different units of toxicity). It should also noted, however, that the compounds used to develop this QSAR are not listed. In addition the statistical criteria are not valid (see Section 6.5), as the compounds have been selected to fit a line.

91. It is reported that a total of 39 toxicophores were obtained from the MultiCASE analysis. 13 of these, which are most strongly associated with recognisable mechanisms of action are listed in the paper (along with modulators and associated statistics) and are repeated in Table 6 below. The remaining 26 toxicophores are reportedly available from MultiCASE Inc. As with the toxicity data, further efforts to validate this model should include the obtaining of the remaining toxicophores for inspection.

6.3. Mechanistic basis (Principle 3)

92. Many of the fragments listed in Table 6 are recognisable in terms of mechanism of toxic action e.g. phenols and anilines are polar narcotics, aldehydes are reactive chemicals and unsaturated alcohols are non-polar narcotics etc. Much effort is placed by Klopman et al (14) into placing mechanistic interpretation onto the fragment. With this in mind, and combined with the use of a log K_{ow} –based algorithm for the prediction of baseline toxicity, there would appear to be a reasonable, although not wholly transparent, mechanistic basis underpinning the model. However, the complete model is not reported (presumably for reasons of commercial sensitivity), therefore the overall impression given is of a non-transparent model. It must be stressed, however, that this conclusion is drawn from the published report on the model, and utilisation of the actual software may assist in establishing the transparency of the approach.

6.4. Domain of applicability (Principle 4)

93. The domain of applicability was not defined by the model developer. As the data and descriptors applied are not listed, it is not possible to determine the applicability domain at this time. However, it should be recognised that definition of the domain will be complex due to the highly multivariate nature of the model.

6.5. Internal performance (Principle 5)

94. A total of 675 toxicity values to the fathead minnow (*Pimephales promelas*) were utilised for the purposes of modelling. These data are essentially the data set described by Russom et al (1), and also reportedly supplemented by data extracted from Mekenyan et al (18) and Vittozzi and De Angelis (19). Some assessment of the data from different sources is described (i.e. there was a requirement for a high correlation between comparative data).

95. As is common with expert system developers, the data are not recorded in the publication (but may be available from the model), and neither are the chemical structures or names made available. The methods of the paper contains a comment that the complete database is available from the vendors

(MultiCASE Inc.), should future further validation of the model be required, it is highly recommended that the original data should be obtained to assist in that process. It should be recognised that this is not good practice and peer-reviewed journals should be encouraged to reject manuscripts that do not report the original data. Due to the effect of combining data from different sources, and not reporting the data used to develop the model, the quality of the toxicity data must be viewed more circumspectly.

96. All toxicity data were converted to micromolar units (μM). For model building toxicity data are placed into one of three classifications (based on regulatory recommendations):

1. Toxic: LC_{50} less than 100 μM
2. Borderline (or marginal): LC_{50} between 100 μM and 7,000 μM
3. Non-Toxic: LC_{50} greater than 7,000 μM

97. The exact role of the classifications of toxicity (and also the predictions for compounds that fit the non-polar narcosis line), as opposed to the prediction of LC_{50} is not exactly specified in the publication and requires further investigation if future validation of this approach is undertaken.

6.5.1. Independent assessment of data quality

98. The training set consists of the biological (96h LC_{50}) data being modelled, and a large number of undefined physicochemical and structural data that are used as predictor variables. The biological data can be considered to be of relatively high quality, since they were mostly obtained by applying a similar protocol (see previous comments regarding the source of data).

99. The descriptors and structural fragments are calculated by the model developers by using their own software. It is not possible at this time to place assessment of confidence on these values.

6.5.2. Goodness-of-fit

100. For goodness-of-fit, the following statistics were reported for the baseline toxicity QSAR: n = 166, $r^2 = 0.95$, s = 0.32, F = 3285, Q^2 not given.

101. For the overall model, the percentage of correct predictions (presumably for the classifications of toxicity) is 97.0% and for the accuracy of predicted LC_{50} values the error was 0.41, r = 0.94 and F = 4,978.

6.5.3. Cross-validation

102. A leave-many-approach was attempted, this involved removing 10% of the data set three times and calculating the toxicity values from models developed on the remaining 90%. The percentage of correct predictions is 93.1% and for the accuracy of predicted LC_{50} values the error was 0.44, r = 0.973 and F = 737. There is some question over the validity of these latter statistics as whilst, as would be expected, error and F values are lower than for the complete data, r is reported to be higher. Further descriptions of the statistical quality of the models also suggest that there is a worsening of the predictive capability following the leave-many-out approach: for the precise prediction of LC_{50} the proportion of compounds with a predicted value within 0.4 log units of the actual value is 91% for the complete set falling to 79% after the leave-many-out approach.

6.6. External validation for predictivity (Principle 6)

103. At this time no attempt has been made to assess external predictivity.

7. CONCLUSIONS ON THE SETUBAL PRINCIPLES AND THEIR APPLICABILITY

104. Four computational models for the prediction of toxicity, based on two methodologies (regression-based QSAR and expert system) were evaluated in this study. The models have been evaluated according to so-called "Setubal principles". These principles form a sound basis on which to evaluate computational models for the prediction of toxicity. In this particular study they have allowed for the strengths and weaknesses of each model to be identified. Further this assessment allows for recommendations for use to be made, without having to make more complex decisions regarding which may be the best model i.e. a user is provided with guidance regarding the use of a particular QSAR, but may not be obliged to use a particular model, should several be available.

105. The wording of the Setubal principles is clear and unambiguous. Many of the issues covered are, of course, very broad and will require expert application. Issues such as whether a model takes the form of an "unambiguous algorithm" (Principle 2) and has "a clear mechanistic basis" (Principle 3) are open to interpretation, and may require further definition. Specific comments on the principles are made below.

106. Principle 1. Computational models for toxicity are required, first and foremost, for endpoints of regulatory relevance. This should not, of course, discount other acceptable models that may provide useful information for risk assessment purposes. In this study all models considered were based on the same dataset (the Duluth fathead minnow dataset). Acute fish toxicity is a required endpoint for regulatory submission in the European Union, but it should be noted that the fathead minnow assay is not commonly performed in the EU.

107. A further issue concerns the quality of the toxicological data. In this study the data are recognised as being of high quality. For many models, this may not be the case. Indeed the dataset in the 4th model assessed (MultiCASE) may be thought of as being "diluted" by the inclusion of data from outside of the original Duluth data set.

108. Principle 2. It is desirable (although not essential) that computational models take the form of an unambiguous algorithm. It may be possible to clarify this further in guidance for those using the Setubal principles i.e. that there must be a clear and explicit statement of how toxicity could be calculated, given the relevant descriptors, and the actual algorithm presented. A good comparison can be made in this study between the regression-based models (QSARs 1-3) and the MultiCASE approach (QSAR 4). Regression provides a good indication of the algorithm, however these are not provided for MultiCASE, probably as a result both of commercial sensitivity and space within the original article.

109. Principle 3. Ideally all computational models for the prediction of toxicity require a mechanistic basis (again this should not preclude the use of non-mechanistic approaches). Some of the problems that may occur in deciding this are clearly illustrated in this study. QSARs 1 and 2 are clearly defined in terms of a single mechanism. QSAR 3 is described in terms of mechanisms of electrophilicity, but the exact mechanisms are not entirely known. The MultiCASE model has had mechanistic interpretation placed on the results and descriptors. This final point shows a further clear distinction here between approaches that

have been founded on mechanism of action *a priori* (i.e. QSARs 1-3) and those were mechanistic interpretation has been applied *a posteriori* (after the modeling process).

110. Principle 4. There is an obvious need for the domain of applicability of a QSAR to be stated and provided to its user. In the QSARs assessed here, this was easily defined in terms of physicochemical and structural descriptor space, as well as mechanism of action and chemical class for QSARs 1-3. This was considerably more difficult for the MultiCASE approach, where a large, undefined, and structurally and mechanistically diverse data set was assessed. It may be of use to make a further assessment of the approaches for the definition of applicability domain.

111. Principle 5. A measure of the goodness-of-fit is required and essential for any quantitative approach to toxicity prediction, such as the QSARs assessed here. It may be a good idea in future to stipulate the type and nature of statistical measures (i.e. some form of statistical measures that may include, for instance, r^2, s, F, t-values, as well as any other appropriate and easily available terms) and also further desirable, but not essential statistical measures, e.g. probability values, analysis of residuals, randomisation of y-variables etc. Some consideration may also have to be given to the recommendation of particular statistical packages for this purpose.

112. A measure of the goodness-of-fit is required and essential for any quantitative approach to toxicity prediction, such as the QSARs assessed here. It may be a good idea in future to stipulate the type and nature of statistical measures (i.e. some form of Principle 6. An assessment of predictivity is vital for the assessment of a QSAR. The interpretation and application of this principle will probably become the most contentious of all of the principles. For instance, if external validation is recommended, where will the data be sourced? Should no data be available, will cross-validation and / or bootstrapping be acceptable? There is also the possibility for a tier of tests to be applied. A final issue will be whether any further data for the model (obtained for validation) should ultimately be included in the model to assist in its refinement.

113. Commercial Sensitivity. Another contentious and highly vexed issue will be that of commercial sensitivity. This is already becoming apparent in two areas, the use of data for validation and release of "definitive" algorithms from expert systems for assessment. In this study, it was extremely difficult to apply all of the Setubal principles to the MultiCASE model, and it should be noted that the assessment is based on a partial application of the principles. It must be recognized that the publication from which the model was taken was not intended to be suitable for an assessment of the Setubal principles. Further, it should be possible to apply the Setubal principles to the MultiCASE module for the prediction of acute fish toxicity, if the necessary information (such as full data set, toxicophores and internal validation) were made available from the vendors.

114. In conclusion, the six Setubal principles provide a succinct description of the main issues with regard to the assessment of (Q)SARs. Some principles will need guidance (as noted above) to allow their consistent use by evaluators and interpretation by users.

8. REFERENCES

Boxall, A.B.A., Watts, C.D., Dearden, J.C., Bresnen, G.M. and Scoffin, R. (1997) 'Predicting the toxic mode of action for environmental pollutants based on physico-chemical properties', in F. Chen and G. Schüürmann (eds) *Quantitative Structure-Activity Relationships in Environmental Sciences – VII*, SETAC Press: Pensacola FL, pp. 263-275.

Cronin, M.T.D. and Schultz, T.W. (2003) 'Pitfalls in QSAR', *Journal of Molecular Structure – THEOCHEM*, 622: 39-51.

Cronin, M.T.D., Gregory, B.W., Schultz T.W. (1998) Quantitative structure-activity analyses of nitrobenzene toxicity to *Tetrahymena pyriformis*. *Chemical Research in Toxicology* 11: 902-908.

Dearden, J.C., Barratt, M.D., Benigni, R., Bristol, D.W., Combes, R.D., Cronin, M.T.D., Judson, P.M., Payne, M.P., Richard, A.M., Tichy, M., Worth, A.P. and Yourick, J.J. (1997) The development and validation of expert systems for predicting toxicity. The report and recommendations of an ECVAM/ECB workshop (ECVAM workshop 24), *ATLA*, 25: 223-252.

Dimitrov, S.D., Mekenyan, O.G., Sinks, G.D. and Schultz, T.W. (2003) Global modeling of narcotic chemicals: ciliate and fish toxicity, *Journal of Molecular Structure – THEOCHEM*, 622: 63-70.

European Commission (1995). QSAR for Predicting Fate and Effects of Chemicals in the Environment, Final report of DG XII contract No. EV5V-CT92-0211.

European Economic Community (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.

Klopman, G. (1984) Artificial intelligence approach to structure-activity studies. Computer automated structure evaluation of biological activity of organic molecules. *Journal of the American Chemical Society* 106: 7315-7320.

Klopman, G. (1992) M-CASE: A hierarchical computer automated structure evaluation program. *Quantitative Structure-Activity Relationships* 11: 176-184.

Klopman, G., Saiakhov, R., Rosenkranz, H.S. (2000) Multiple Computer-Automated Structure Evaluation study of aquatic toxicity II. Fathead minnow. *Environmental Toxicology and Chemistry* 19: 441-447.

McKim, J.M., Schmieder, P.K., Carlson, R.W., Hunt, E.P. and Niemi, G.I. (1987) 'Use of respiratory-cardiovascular 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.

Mekenyan, O.G., Veith, G.D., Bradbury, S.P., Russom, C.L. (1993) Structure-toxicity relationships for a,b-unsaturated alcohols. *Quantitative Structure-Activity Relationships* 12: 132-136.

Russom, C.L., Bradbury, S.P., Broderius, S.J., Hammermeister, D.E., Drummond, R.A. (1997) Predicting modes of toxic action from chemical structure: acute toxicity in the fathead minnow (*Pimephales promelas*). *Environmental Toxicology and Chemistry* 16: 948-967.

Schultz, T.W. and Cronin, M.T.D. (2003) 'Essential and desirable characteristics of ecotoxicity quantitative structure-activity relationships', *Environmental Toxicological and Chemistry*, 22: 599-607.

Veith, G.D., Broderius, S.J. (1987) Structure-toxicity relationships for industry chemicals causing type (II) narcosis syndrome. In: Kaiser, K.L.E. (ed.) *QSAR in Environmental Toxicology – II*. D. Reidel, Dordrecht, pp. 385-391.

Veith, G.D., Mekenyan, O.G. (1993) A QSAR approach for estimating the aquatic toxicity of soft electrophiles [QSAR for soft electrophiles]. *Quantitative Structure-Activity Relationships* 12: 349-356.

Verhaar, H.J.M., van Leeuwen, C.J. and Hermens, J.L.M. (1992) 'Classifying environmental pollutants. 1. Structure-activity relationships for prediction of aquatic toxicity', *Chemosphere*, 25: 471-491.

Vittozzi, L., De Angelis, G. (1991) A critical review of comparative acute toxicity data on freshwater fish. *Aquatic Toxicology* 19: 167-204.

Worth, A.P., Cronin, M.T.D., van Leeuwen, C.J. (2004) A framework for promoting the acceptance and regulatory use of (quantitative) structure-activity relationships. In: Cronin, M.T.D., Livingstone, M.T.D. (eds) *Predicting Chemical Toxicity and Fate*, CRC Press, Boca Raton FL, USA, *in press.*

Table 1 Training set of data used to develop a QSAR for non-polar narcosis

No.	Chemical	Log K_{ow}	Log (LC_{50}) (mol/L)
1	1-butanol	0.88	-1.63
2	1-decanol	4.57	-4.81
3	1-dodecanol	5.13	-5.26
4	1-hexanol	2.03	-3.02
5	1-nonanol	4.26	-4.40
6	1-octanol	2.97	-3.98
7	1-undecanol	4.52	-5.21
8	1,1,2-trichloroethane	1.89	-3.21
9	1,1,2,2-tetrachloroethane	2.39	-3.91
10	1,2-dichloroethane	1.48	-2.92
11	1,2,3,4-tetrachlorobenzene	4.63	-5.29
12	1,2,4-trichlorobenzene	4.05	-4.79
13	1,3-dichlorobenzene	3.52	-4.27
14	1,4-dichlorobenzene	3.44	-4.56
15	1,4-dimethoxybenzene	2.15	-3.07
16	2-(2-ethoxyethoxy)ethanol	-0.54	-0.70
17	2-butanone	0.29	-1.35
18	2-decanone	3.73	-4.43
19	2-hydroxy-4-methoxyacetophenone	1.98	-3.48
20	2-methyl-1-propanol	0.76	-1.71
21	2-methyl-2,4-pentanediol	-0.67	-1.04
22	2-octanone	2.37	-3.55
23	2-phenoxyethanol	1.16	-2.60
24	2-propanol	0.05	-0.76
25	2,2,2-trichloroethanol	1.42	-2.69
26	2,3,4-trichloroacetophenone	3.57	-5.04
27	2,3,4-trimethoxyacetophenone	1.12	-3.08
28	2,4-dichloroacetophenone	2.84	-4.20
29	2,6-dimethoxytoluene	2.64	-3.87
30	3-furanmethanol	0.30	-2.28
31	3-methyl-2-butanone	0.56	-1.99
32	3-pentanone	0.79	-1.74
33	3,3-dimethyl-2-butanone	0.96	-3.06
34	3,4-dichlorotoluene	4.06	-4.74
35	4-methyl-2-pentanone	1.31	-2.29
36	5-methyl-2-hexanone	1.88	-2.85
37	5-nonanone	2.90	-3.66
38	6-methyl-5-hepten-2-one	1.70	-3.16
39	Acetone	-0.24	-0.85
40	Acetophenone	1.58	-2.87
41	Benzophenone	3.18	-4.07
42	Cyclohexanol	1.23	-2.15
43	Cyclohexanone	0.81	-2.27
44	Dibutyl ether	3.21	-3.60
45	Diisopropyl ether	1.52	-3.04

46	Dipentyl ether	4.04	-4.69
47	Diphenyl ether	4.21	-4.62
48	Ethanol	-0.31	0.51
49	Furan	1.34	-3.04
50	Hexachloroethane	4.14	-5.19
51	Methanol	-0.77	-0.057
52	4-nitrophenyl phenylether	4.28	-4.90
53	Pentachloroethane	3.62	-4.44
54	*tert*-butylmethyl ether	0.94	-2.09
55	Tetrachloroethene	3.40	-4.08
56	Tetrahydrofuran	0.46	-1.52
57	Trichloroethene	2.42	-3.47
58	Triethylene glycol	-1.24	-0.33

The data were taken from the final report of an EU project (EC, 1995).

Table 2 Test set of data used to validate the QSAR for non-polar narcosis

No.	Chemical	Log K_{ow}	MW	LC_{50} (mg/L)	Observed Log(LC_{50}) (mol/L)	Predicted Log(LC_{50}) (mol/L)	Residual (Observed – Predicted)
1	1-pentanol	1.56	88.15	472	-2.27	-2.71	0.44
2	N,N-dimethyl-4-toluidine	2.81	135.2	48.9	-3.44	-3.77	0.33
3	1-heptanol	2.72	116.2	34.5	-3.53	-3.69	0.16
4	N,N-dimethylaniline	2.31	121.2	64.1	-3.28	-3.34	0.06
5	3-bromobenzamide	1.65	200.04	92.7	-3.33	-2.79	-0.54
6	4-(tert-butyl)benzamide	2.51	177.25	31.9	-3.74	-3.51	-0.23
7	urethane	-0.15	89.09	5240	-1.23	-1.26	0.03
8	4’-aminopropiophenone	1.43	149.19	146	-3.01	-2.60	-0.41
9	2-methyl-2-propanol	0.35	74.12	6410	-1.06	-1.69	0.63
10	2-ethylpyridine	1.69	107.16	414	-2.41	-2.82	0.41
11	1-bromopropane	2.10	122.99	67.3	-3.26	-3.17	-0.09
12	2-heptanone	1.98	114.19	131	-2.94	-3.07	0.13
13	1,4-dichlorobutane	2.24	127.01	51.6	-3.39	-3.29	-0.10
14	2-undecanone	4.09	170.3	1.50	-5.06	-4.85	-0.21
15	2-dodecanone	4.49	184.32	1.18	-5.19	-5.19	0.00
16	1,2,4-trimethylbenzene	3.78	120.2	7.72	-4.19	-4.59	0.40
17	ethylbenzene	3.15	106.17	10.5	-4.00	-4.05	0.05
18	phenyl ether	4.21	170.21	4	-4.63	-4.95	0.32
19	pyrrole	0.75	67.09	210	-2.50	-2.02	-0.48
20	1,3,5-trioxane	-0.43	90.08	5950	-1.18	-1.03	-0.15

Table 3 Training set of data used to develop a QSAR for polar narcosis (the data were taken from Veith and Broderius, 1987)

No.	Chemical	Log K_{ow}	Log (LC_{50}) (mol/L)	LMO Toxicity
1	aniline	0.9	2.84	2.88
2	4-toluidine	1.39	2.86	3.22
3	4-nitroaniline	1.31	3.04	3.1
4	4-methoxyphenol	1.34	3.05	3.15
5	3-methoxyphenol	1.58	3.22	3.34
6	4-ethylaniline	1.96	3.22	3.58
7	phenol	1.46	3.41	3.20
8	4-nitrophenol	1.91	3.53	3.54
9	4-bromoaniline	2.26	3.56	3.78
10	4-chloroaniline	1.83	3.59	3.50
11	2-chloro-4-methylphenol	2.58	3.60	3.92
12	pentafluoroaniline	2.22	3.69	3.75
13	a,a,a,a-4-tetrafluoro-3-toluidine	2.62	3.77	4.02
14	a,a,a,a-4-tetrafluoro-2-toluidine	2.62	3.78	3.98
15	4-methylphenol	1.94	3.82	3.51
16	4-ethoxy-2-nitroaniline	1.94	3.85	3.56
17	2,4-dimethylphenol	2.3	3.87	3.80
18	2-chloro-4-nitroaniline	2.17	3.93	3.70
19	2-allylphenol	2.64	3.95	3.96
20	2,6-diisopropylaniline	4.07	4.06	5.01
21	4-ethylphenol	2.58	4.07	3.99
22	4-propylphenol	3.18	4.09	4.32
23	2-chlorophenol	2.15	4.14	3.65
24	4-butylaniline	3.15	4.16	4.38
25	2,4-dichlorophenol	2.92	4.32	4.22
26	3,4-dichloroaniline	2.69	4.33	4.02
27	3-benzoxyaniline	2.79	4.34	4.06
28	4-chloro-3-methylphenol	3.1	4.40	4.35
29	2-phenylphenol	3.36	4.44	4.51
30	4-tert-butylphenol	3.31	4.46	4.41
31	1-naphthol	2.84	4.49	4.09
32	4-phenoxyphenol	3.75	4.58	4.79
33	2,4,6-trichlorophenol	3.69	4.64	4.73
34	2,3,6-trichloroaniline	3.33	4.73	4.42
35	4-hexyloxyaniline	3.66	4.78	4.62
36	4-*tert*-pentylphenol	3.98	4.80	4.94
37	4-nonylphenol	6.36	6.20	6.45
38	4-octylaniline	5.27	6.23	5.61
39	4-decylaniline	6.32	6.58	6.40

Table 4 Test set of data used to validate the QSAR for polar narcosis

Chemical	Log K_{ow}	MW	LC_{50} (mg/L)	Observed Log ($1/LC_{50}$) (mol/L)	Predicted Log ($1/LC_{50}$) (mol/L)	Residual (Observed – Predicted)
2-chloroaniline	1.90	127.57	5.74	4.35	3.53	0.82
2-cyanopyridine	0.50	104.11	726	2.16	2.62	-0.46
pyridine	0.65	79.10	99.8	2.90	2.71	0.19
2-chloro-4-methylaniline	2.58	141.60	35.9	3.60	3.97	-0.37
4-acetylpyridine	0.48	121.14	168	2.86	2.60	0.26
3-trifluoromethyl-4-nitrophenol	3.00	207.11	9.14	4.36	4.24	0.11
2-cresol	2.12	108.14	14.0	3.89	3.67	0.22
2-amino-5-bromopyridine	1.39	173.01	177	2.99	3.19	-0.20
4-chlorophenol	2.48	128.56	6.11	4.32	3.90	0.42
2-bromo-3-pyridinol	1.65	174.00	469	2.57	3.36	-0.79
2-chloro-3-pyridinol	1.50	129.55	622	2.32	3.27	-0.95

Table 5 Training set of data used to develop a QSAR for mixed mechanisms of action

No.	Chemical	Log K_{ow}	E_{LUMO}	Log (LC_{50}) (mol/L)	LMO Toxicity
1	Phenol	1.46	0.24	3.46	2.2
2	o-cresol	2.12	0.13	3.89	2.61
3	p-cresol	1.94	0.18	3.82	3.73
4	2,4-dimethylphenol	2.3	0.1	3.87	3.04
5	2,4,6-trimethylphenol	3.42	0.1	4.02	3.18
6	2,3,6-trimethylphenol	3.42	0.02	4.22	2.69
7	4-ethylphenol	2.58	0.2	4.07	3.08
8	4-propylphenol	3.18	0.19	4.09	3.57
9	2-allylphenol	2.64	0.18	3.95	3.33
10	4-tert-butylphenol	3.31	0.21	4.46	3.44
11	4-tert-pentylphenol	3.98	0.17	4.8	4.23
12	2,6-di-tert-butyl-4-methyl phenol	6.07	-0.05	5.78	4.11
13	2,4,6-tri-tert-butylphenol	7.4	-0.02	6.63	4.14
14	4-nonylphenol	6.36	0.19	6.2	4.17
15	4-phenylphenol	3.36	0.06	4.44	4.12
16	catechol	0.88	-1.47	4.08	4.17
17	resorcinol	0.8	0.19	3.34	4.04
18	3-methoxyphenol	1.58	0.17	3.22	3.91
19	4-methoxyphenol	1.34	0.16	3.05	3.53
20	4-phenoxyphenol	3.75	-0.03	4.58	4.06
21	2-chlorophenol	2.15	-0.22	3.97	3.82
22	4-chlorophenol	2.48	-0.18	4.32	4.11
23	4-chloro-3-methylphenol	3.1	-0.2	4.42	4.09
24	4-chlorocatechol	1.97	-1.81	4.96	3.39
25	2,4-dichlorophenol	2.92	-0.58	4.32	4.15
26	4,5-dichlorocatechol	2.9	-2.1	5.3	4
27	4,5-dichloroguaiacol	3.26	-0.56	4.64	4.56
28	2,4,6-trichlorophenol	3.69	-0.9	4.61	3.89
29	2,3,4,6-tetrachlorophenol	4.45	-1.22	5.35	4.42
30	2,3,4,5-tetrachlorophenol	4.21	-1.23	5.75	3.87
31	tetrachlorocatechol	4.29	-1.25	5.29	4.03
32	pentachlorophenol	5.12	-1.52	6.04	3.4
33	2,4,6-tribromophenol	4.02	-0.72	4.7	4.13
34	pentabromophenol	5.74	-1.26	6.72	4.32
35	2,4,6-triiodophenol	4.8	-0.83	5.59	4.96
36	2-nitrophenol	1.85	-0.92	2.94	4.59
37	4-nitrophenol	1.91	-1.21	3.53	4.5
38	2,6-dinitrophenol	1.91	-1.73	3.67	4.55
39	2,5-dinitrophenol	1.75	-1.87	4.74	5.1
40	2,4-dinitrophenol	1.54	-1.69	4.23	5.11
41	2-sec-butyl-4,6-dinitrophenol	3.69	-1.61	5.65	4.37
42	4-amino-2-nitrophenol	0.96	-0.64	3.63	4.72
43	3-trifluoromethyl-4-nitrophenol	3	-1.84	4.36	4.2

44	4,6-dinitro-o-cresol	2.56	-1.67	5.06	5.4
45	2,2'-methylenebis(4-chloro-phenol)	4.26	-0.45	5.94	4.88
46	2,2'-methylenebis(3,4,6-trichloro-phenol)	7.54	-1.3	7.29	3.89
47	aniline	0.9	0.54	2.91	5.2
48	4-toluidine	1.39	0.55	2.83	6.15
49	4-ethylaniline	1.96	0.55	3.22	5.64
50	4-butylaniline	3.15	0.54	4.16	6.02
51	4-octylaniline	5.27	0.54	6.23	5.2
52	4-decylaniline	6.32	0.54	6.58	5.89
53	4,6-diisopropylaniline	3.18	0.36	4.06	6.02
54	2,4-diaminotoluene	0.34	0.73	1.94	6.1
55	3-benzyloxyaniline	2.79	0.18	4.34	5.94
56	4-hexyloxyaniline	3.66	0.4	4.81	4.72
57	2-chloroaniline	1.9	0.1	4.35	6.21
58	4-chloroaniline	1.83	0.18	3.61	3
59	2-chloro-4-methylaniline	2.58	0.1	3.6	4.08
60	3,4-dichloroaniline	2.69	-0.26	4.33	3.75
61	2,3,4-trichloroaniline	3.33	-0.59	4.73	3.62
62	2,3,5,6-tetrachloroaniline	4.1	-1.04	5.93	3.31
63	4-bromoaniline	2.26	0.24	3.56	3.9
64	4-fluoroaniline	1.15	0.15	3.81	4.27
65	2,3,4,5,6-pentafluoroaniline	2.22	-1.5	3.69	3.9
66	a,a,a-4-tetrafluoro-3-toluidine	2.62	-0.79	3.77	3.43
67	a,a,a-4-tetrafluoro-2-toluidine	2.62	-0.73	3.78	4.56
68	4-nitroaniline	1.31	-0.87	3.04	4.22
69	4-ethoxy-2-nitroaniline	2.47	-0.62	3.85	4.41
70	2-chloro-4-nitroaniline	2.17	-1.14	3.93	4.42
71	2,4-dinitroaniline	1.84	-1.56	4.09	3.09
72	benzene	2.13	0.37	3.65	3.5
73	toluene	2.73	0.25	3.43	4.18
74	o-xylene	3.12	0.19	3.81	3.74
75	m-xylene	3.2	0.19	3.82	4.61
76	p-xylene	3.15	0.13	4.08	4.71
77	ethylbenzene	3.15	0.26	4	4.22
78	isopropylbenzene	3.66	0.26	4.28	3.84
79	1,2,4-trimethylbenzene	3.78	0.1	4.19	4.35
80	butylbenzene	4.26	0.25	4.83	3.84
81	1,3-diethylbenzene	4.5	0.21	4.51	4.19
82	styrene	2.95	-0.1	4.41	4.43
83	4-tert-butylstyrene	4.84	0.13	5.52	4.55
84	amylbenzene	4.91	0.25	4.94	4.4
85	biphenyl	4.09	0.18	4.9	4.81
86	1,4-dimethoxybenzene	2.15	0.17	3.07	4.23
87	2,6-dimethoxytoluene	2.8	0.21	3.88	4.47
88	1-allyl-4-methoxybenzene	3.31	0.21	4.28	3.97
89	chlorobenzene	2.86	-0.13	3.82	5.22
90	1,2-dichlorobenzene	3.38	-0.52	4.19	4.35
91	1,3-dichlorobenzene	3.6	-0.53	4.26	4.24
92	1,4-dichlorobenzene	3.37	-0.6	4.56	4.72
93	3,4-dichlorotoluene	4.22	-0.58	4.74	5.06

94	1,2,4-trichlorobenzene	4.02	-0.93	4.78	5.14
95	1,2,3,4-tetrachlorobenzene	4.99	-1.18	5.29	4.56
96	pentachlorobenzene	5.17	-1.49	6	4.77
97	1,2,4,5-tetrachlorobenzene	4.82	-1.27	5.8	4.62
98	a,a-2,6-tetrachlorotoluene	4.64	-1.21	5.38	4.73
99	pentachloroanisole	5.34	-1.51	5.64	5.09
100	pentachloropyridine	4.34	-1.76	5.73	4.89
101	bromobenzene	2.99	-0.09	3.94	5.94
102	1,2-dibromobenzene	3.64	-0.42	4.76	5.66
103	a,a,a',a'-tetrabromo-o-xylene	5.17	-1.12	5.98	5.76
104	2-fluorotoluene	2.93	-0.16	3.75	5.63
105	nitrobenzene	1.85	-0.82	3.02	5.47
106	3-nitrotoluene	2.45	-0.84	3.73	5.54
107	1-fluoro-4-nitrobenzene	1.8	-1.22	3.7	5.89
108	1-chloro-3-nitrobenzene	2.41	-1.52	3.92	5.13
109	1-chloro-2-nitrobenzene	2.24	-1.13	3.73	6.21
110	1,4-dinitrobenzene	1.46	-1.83	5.37	5.41
111	2,4-dinitrotoluene	2	-1.73	3.88	5.12
112	1,3-dichloror-4,6-dinitrobenzene	2.49	-2.34	6.71	6.6
113	1,3,5-trichloro-2,4-dinitrobenzene	2.65	-2.52	6.09	6.39
114	hexachlorocyclopentadiene	5.04	-2.12	6.95	7.4

The data were taken from Veith and Mekenyan (1993)

Table 6 Reported toxicophores in the MultiCASE model for the prediction of fathead minnow toxicity

Fragment Number	Description
1	Phosphate, thiophosphate and their esters
2	Nitroaromatics and analogues
3	Phenols
4	Anilines
5	Aliphatic amines
6	Aliphatic and aromatic amines
7	Amides
8	Polychlorinated aromatics
9	Polychlorinated alkanes and pesticides
10	Piperazines
11	Unsaturated alcohols
12	Cresols and toluidines
13	Aldehydes

This is a summary of the toxicophores reported in the paper from both the structural aspect and associated class

Figure 1 Observed vs predicted toxicity for QSAR 1

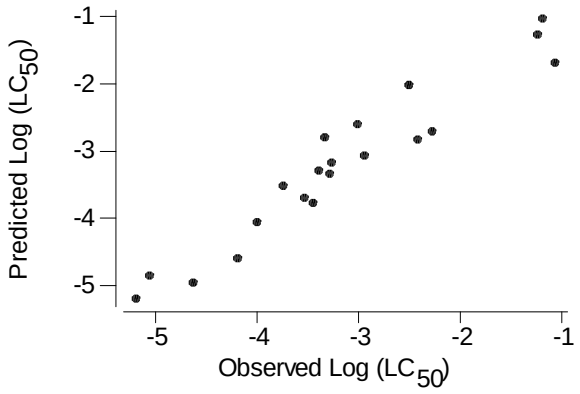
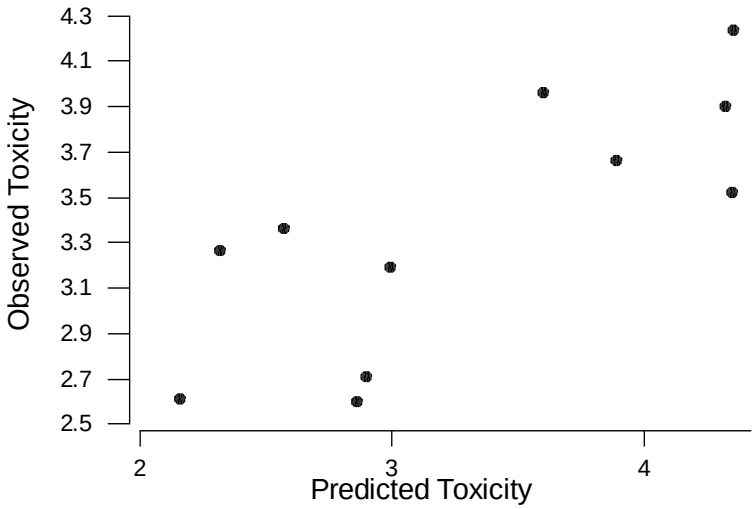


Figure 2 Observed vs predicted toxicity for QSAR 2



9. APPENDICES

APPENDIX 1 SUMMARY REPORT OF THE APPLICATION OF THE SETUBAL PRINCIPLES TO QSAR 1

PRINCIPLE	CONSIDERATIONS	YES / NO
1) Defined endpoint	**1.1)** Does the model have a clearly defined scientific purpose? (i.e does it make predictions of a clearly defined physicochemical, biological or environmental effect?)	Yes
	1.2) Does the model have the potential 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)?	Yes
	1.3) Is information given about important experimental conditions that affect the measurement and therefore the prediction (e.g. sex, species, temperature, exposure period)	Yes
	1.4) Are the units of measurement of the endpoint given?	Yes
2) Defined algorithm	**2.1)** In the case of a SAR, is there an explicit description of the substructure, including an explicit identification of its substituents?	NA
	2.2) In the case of a QSAR, is the equation explicitly defined, including definitions of all descriptors used[3]?	Yes
3) Mechanistic basis	**3.1)** In the case of a SAR, is there a description of the molecular events that underlie the reactivity of the molecule? (e.g. a description of how substructural features could act as nucleophiles or electrophiles, or form part or all of a receptor-binding region)	NA
	3.2) In the case of a QSAR, do the descriptors have a physicochemical interpretation that is consistent with a known mechanism (of biological action)?	Yes
	3.3) Are any literature references cited in support of the purported mechanistic basis of the (Q)SAR?	Yes

4) Domain of applicability	**4.1)** In the case of a SAR, is the substructure associated with any inclusion and/or exclusion rules on its applicability to groups of chemicals?	NA
	4.2) In the case of a SAR, is the substructure associated with rules regarding the modulatory effects of the substructure's molecular environment?	NA
	4.3) In the case of a (Q)SAR, are the descriptor and response variables associated with inclusion and/or exclusion rules that define the variable ranges for which the QSAR is applicable (i.e. makes reliable estimates)?	Yes
5) Internal performance	**5.1)** Are full details of the training set given, including details of chemical names, structural formulae, CAS numbers (if available), and data for all descriptor and response variables?	Yes
	5.2) If the data used to the develop the model were based upon the processing of raw data (e.g. the averaging of replicate values): a) is there an adequate description of the data processing? b) are the raw data provided?	All raw data are available (albeit, not in this publication)
	5.3) Is there a specification of the statistical method(s) used to develop the QSAR (including details of any software packages used)?	Not in this publication
	5.4) Is the QSAR associated with basic statistics for its goodness-of-fit to the training set? (e.g. r^2 values and standard error of the estimate in the case of regression models)	Yes
	5.5) a) Is the QSAR associated with any statistics based on cross-validation or resampling? b) If yes, is the number or samples used indicated?	No
6) Predictivity (External validation)	**6.1)** Does application of the appropriate statistical method(s) to the training set result in the same (Q)SAR model?	Yes
	6.2) Is there any information to indicate that the (Q)SAR has been validated previously, using a test set that is independent of the training set?	No
	6.3) If an external validation has been performed, is the following information available: a) the number of test structures? b) the identities of the test structures? c) the approach for selecting the test structures? d) the statistical analysis of the predictive performance of the model? (e.g. including sensitivity, specificity, and positive and negative predictivities for classification models) e) a comparison of the predictive performance of the model against previously-defined quantitative performance criteria?	Yes, external validation is described above.

APPENDIX 2SUMMARY REPORT OF THE APPLICATION OF THE SETUBAL PRINCIPLES TO QSAR 2

PRINCIPLE	CONSIDERATIONS	YES / NO
1) Defined endpoint	**1.1)** Does the model have a clearly defined scientific purpose? (i.e does it make predictions of a clearly defined physicochemical, biological or environmental effect?)	Yes
	1.2) Does the model have the potential 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)?	Yes
	1.3) Is information given about important experimental conditions that affect the measurement and therefore the prediction (e.g. sex, species, temperature, exposure period)	Yes
	1.4) Are the units of measurement of the endpoint given?	Yes
2) Defined algorithm	**2.1)** In the case of a SAR, is there an explicit description of the substructure, including an explicit identification of its substituents?	NA
	2.2) In the case of a QSAR, is the equation explicitly defined, including definitions of all descriptors used[3]?	Yes
3) Mechanistic basis	**3.1)** In the case of a SAR, is there a description of the molecular events that underlie the reactivity of the molecule? (e.g. a description of how substructural features could act as nucleophiles or electrophiles, or form part or all of a receptor-binding region)	NA
	3.2) In the case of a QSAR, do the descriptors have a physicochemical interpretation that is consistent with a known mechanism (of biological action)?	Yes
	3.3) Are any literature references cited in support of the purported mechanistic basis of the (Q)SAR?	Yes
4) Domain of applicability	**4.1)** In the case of a SAR, is the substructure associated with any inclusion and/or exclusion rules on its applicability to groups of chemicals?	NA
	4.2) In the case of a SAR, is the substructure associated with rules regarding the modulatory effects of the substructure's molecular environment?	NA

	4.3) In the case of a (Q)SAR, are the descriptor and response variables associated with inclusion and/or exclusion rules that define the variable ranges for which the QSAR is applicable (i.e. makes reliable estimates)?	Yes
5) Internal performance	**5.1)** Are full details of the training set given, including details of chemical names, structural formulae, CAS numbers (if available), and data for all descriptor and response variables?	Yes
	5.2) If the data used to the develop the model were based upon the processing of raw data (e.g. the averaging of replicate values): a) is there an adequate description of the data processing? b) are the raw data provided?	Reference is given. All raw data are available (albeit, not in this publication
	5.3) Is there a specification of the statistical method(s) used to develop the QSAR (including details of any software packages used)?	No
	5.4) Is the QSAR associated with basic statistics for its goodness-of-fit to the training set? (e.g. r^2 values and standard error of the estimate in the case of regression models)	No, only basic statistics given in the original publication
	5.5) a) Is the QSAR associated with any statistics based on cross-validation or resampling? b) If yes, is the number or samples used indicated?	No
6) Predictivity (External validation)	**6.1)** Does application of the appropriate statistical method(s) to the training set result in the same (Q)SAR model?	Yes
	6.2) Is there any information to indicate that the (Q)SAR has been validated previously, using a test set that is independent of the training set?	No
	6.3) If an external validation has been performed, is the following information available: a) the number of test structures? b) the identities of the test structures? c) the approach for selecting the test structures? d) the statistical analysis of the predictive performance of the model? (e.g. including sensitivity, specificity, and positive and negative predictivities for classification models) e) a comparison of the predictive performance of the model against previously-defined quantitative performance criteria?	Yes – external validation described above

APPENDIX 3 SUMMARY REPORT OF THE APPLICATION OF THE SETUBAL PRINCIPLES TO QSAR 3

PRINCIPLE	CONSIDERATIONS	YES / NO
1) Defined endpoint	**1.1)** Does the model have a clearly defined scientific purpose? (i.e does it make predictions of a clearly defined physicochemical, biological or environmental effect?)	Yes
	1.2) Does the model have the potential 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)?	Yes
	1.3) Is information given about important experimental conditions that affect the measurement and therefore the prediction (e.g. sex, species, temperature, exposure period)	Yes
	1.4) Are the units of measurement of the endpoint given?	Yes
2) Defined algorithm	**2.1)** In the case of a SAR, is there an explicit description of the substructure, including an explicit identification of its substituents?	NA
	2.2) In the case of a QSAR, is the equation explicitly defined, including definitions of all descriptors used[3]?	Yes
3) Mechanistic basis	**3.1)** In the case of a SAR, is there a description of the molecular events that underlie the reactivity of the molecule? (e.g. a description of how substructural features could act as nucleophiles or electrophiles, or form part or all of a receptor-binding region)	NA
	3.2) In the case of a QSAR, do the descriptors have a physicochemical interpretation that is consistent with a known mechanism (of biological action)?	Yes
	3.3) Are any literature references cited in support of the purported mechanistic basis of the (Q)SAR?	Yes
4) Domain of applicability	**4.1)** In the case of a SAR, is the substructure associated with any inclusion and/or exclusion rules on its applicability to groups of chemicals?	NA
	4.2) In the case of a SAR, is the substructure associated with rules regarding the modulatory effects of the substructure's molecular environment?	NA
	4.3) In the case of a (Q)SAR, are the descriptor and response	Yes

	variables associated with inclusion and/or exclusion rules that define the variable ranges for which the QSAR is applicable (i.e. makes reliable estimates)?	
5) Internal performance	**5.1)** Are full details of the training set given, including details of chemical names, structural formulae, CAS numbers (if available), and data for all descriptor and response variables?	Yes
	5.2) If the data used to the develop the model were based upon the processing of raw data (e.g. the averaging of replicate values): a) is there an adequate description of the data processing? b) are the raw data provided?	Reference is given. All raw data are available (albeit, not in this publication
	5.3) Is there a specification of the statistical method(s) used to develop the QSAR (including details of any software packages used)?	No – see the particular issue with the OASIS software described above.
	5.4) Is the QSAR associated with basic statistics for its goodness-of-fit to the training set? (e.g. r^2 values and standard error of the estimate in the case of regression models)	Yes
	5.5) a) Is the QSAR associated with any statistics based on cross-validation or resampling? b) If yes, is the number or samples used indicated?	No
6) Predictivity (External validation)	**6.1)** Does application of the appropriate statistical method(s) to the training set result in the same (Q)SAR model?	No – see comment regarding coefficient on E_{LUMO}
	6.2) Is there any information to indicate that the (Q)SAR has been validated previously, using a test set that is independent of the training set?	No
	6.3) If an external validation has been performed, is the following information available: a) the number of test structures? b) the identities of the test structures? c) the approach for selecting the test structures? d) the statistical analysis of the predictive performance of the model? (e.g. including sensitivity, specificity, and positive and negative predictivities for classification models) e) a comparison of the predictive performance of the model against previously-defined quantitative performance criteria?	NA

APPENDIX 4 SUMMARY REPORT OF THE APPLICATION OF THE SETUBAL PRINCIPLES TO QSAR 4

PRINCIPLE	CONSIDERATIONS	YES / NO
1) Defined endpoint	**1.1)** Does the model have a clearly defined scientific purpose? (i.e does it make predictions of a clearly defined physicochemical, biological or environmental effect?)	Yes
	1.2) Does the model have the potential 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)?	Yes
	1.3) Is information given about important experimental conditions that affect the measurement and therefore the prediction (e.g. sex, species, temperature, exposure period)	No – data are retrieved from compilations and hence undocumented differences in protocol could potentially occur.
	1.4) Are the units of measurement of the endpoint given?	Yes, but the paper is not totally clear regarding the precise use of qualitative vs. quantitative data.
2) Defined algorithm	**2.1)** In the case of a SAR, is there an explicit description of the substructure, including an explicit identification of its substituents?	NA
	2.2) In the case of a QSAR, is the equation explicitly defined, including definitions of all descriptors used[3]?	No
3) Mechanistic basis	**3.1)** In the case of a SAR, is there a description of the molecular events that underlie the reactivity of the molecule? (e.g. a description of how substructural features could act as nucleophiles or electrophiles, or form part or all of a receptor-binding region)	NA
	3.2) In the case of a QSAR, do the descriptors have a	Yes – as far as

	physicochemical interpretation that is consistent with a known mechanism (of biological action)?	they are listed, although two thirds of toxicophores are not reported in the paper
	3.3) Are any literature references cited in support of the purported mechanistic basis of the (Q)SAR?	Yes – in particular with the baseline algorithm and the toxicophores listed
4) Domain of applicability	**4.1)** In the case of a SAR, is the substructure associated with any inclusion and/or exclusion rules on its applicability to groups of chemicals?	NA
	4.2) In the case of a SAR, is the substructure associated with rules regarding the modulatory effects of the substructure's molecular environment?	Yes, some information on this aspect is provided.
	4.3) In the case of a (Q)SAR, are the descriptor and response variables associated with inclusion and/or exclusion rules that define the variable ranges for which the QSAR is applicable (i.e. makes reliable estimates)?	No
5) Internal performance	**5.1)** Are full details of the training set given, including details of chemical names, structural formulae, CAS numbers (if available), and data for all descriptor and response variables?	No
	5.2) If the data used to the develop the model were based upon the processing of raw data (e.g. the averaging of replicate values): a) is there an adequate description of the data processing? b) are the raw data provided?	Reference is given. All raw data are available (albeit, not in this publication)
	5.3) Is there a specification of the statistical method(s) used to develop the QSAR (including details of any software packages used)?	No – its is not obvious from the descriptions provided.
	5.4) Is the QSAR associated with basic statistics for its goodness-of-fit to the training set? (e.g. r^2 values and standard error of the estimate in the case of regression models)	Yes
	5.5) a) Is the QSAR associated with any statistics based on cross-validation or resampling? b) If yes, is the number or samples used indicated?	Yes - limited

6) Predictivity (External validation)	**6.1)** Does application of the appropriate statistical method(s) to the training set result in the same (Q)SAR model?	Not attempted
	6.2) Is there any information to indicate that the (Q)SAR has been validated previously, using a test set that is independent of the training set?	No
	6.3) If an external validation has been performed, is the following information available: a) the number of test structures? b) the identities of the test structures? c) the approach for selecting the test structures? d) the statistical analysis of the predictive performance of the model? (e.g. including sensitivity, specificity, and positive and negative predictivities for classification models) e) a comparison of the predictive performance of the model against previously-defined quantitative performance criteria?	NA

ANNEX 2
QSARS FOR ATMOSPHERIC DEGRADATION

Prof. Paola Gramatica
QSAR and Environmental Chemistry Research Unit
Department of Structural and Functional Biology (DBSF)
Via J.H. Dunant 3
University of Insubria
21100 Varese

TABLE OF CONTENTS

1. INTRODUCTION

115. Professor Paola Gramatica (Insubria University, Scientific Responsible, QSAR and Environmental Chemistry Research Unit of DBSF) has retrospectively applied, under the terms of a JRC contract, Setubal Principles 1-6 to QSAR atmospheric degradation models published in the following research papers:

1. Gramatica, P., Pilutti, P. and Papa, E. (2003). Predicting the NO_3 tropospheric degradability of organic pollutants by theoretical molecular descriptors. *Atmospheric Environment* **37**, 3115-3124.

2. Gramatica, P., Pilutti, P. and Papa, E. (2003). QSAR Prediction of ozone tropospheric degradation. *QSAR & Combinatorial Science* **22**, 364-373.

3. Gramatica, P., Pilutti, P. and Papa, E. (2002). Ranking of volatile organic compounds for tropospheric degradability by oxidants: a QSPR Approach. *SAR & QSAR in Environmental Research* **13**, 743-753.

116. A summary (in tabular form, Appendix 1) of the extent to which the Setubal principles have been met in all the QSAR models of the above papers is attached.

117. A more-detailed explanation (when necessary) following the items in the Appendix 1 is reported below.

2. APPLICATION OF THE SETUBAL PRINCIPLES

2.1. Defined endpoint (Principle 1)

118. The endpoints, modeled in the above-cited papers, are the rate constants for the degradation of Volatile Organic Chemicals by ozone and nitrate radicals, experimentally measured and collected from Atkinson's papers [4, 5, 6] (cited in the references).

119. In reference [3] a new Index of Atmospheric Persistence is proposed (ATPINdex), based on the linear combination by Principal Component Analysis (PCA) of the most relevant degradative oxidations (experimental data of k_{OH} [4], k_{NO3} [5], k_{O3} [6]): the PC1 score is modeled by theoretical molecular descriptors in order to obtain a chemical ranking according to chemical tropospheric degradability, based only on molecular structure information.

120. For the studied chemicals the experimental conditions are defined in the Atkinson's papers, cited in the reference list [4-6]. The units of measurement of the endpoints are always reported ($cm^3 \cdot sec^{-1} \cdot mol^{-1}$).

121. The experimental data of the defined endpoints are listed in the three papers [1-3], and also in Apendices 2-4.

2.2. Defined algorithm (Principle 2)

122. The QSAR equations of Ordinary Least Square (OLS) regression are explicitly defined with all the statistical parameters. The molecular descriptors are mainly calculated using the *DRAGON* software [7] on the minimal energy conformations determined by the MM+ method, while quantum-chemical descriptors are calculated by MOPAC (PM3 Hamiltonian for geometry optimisation) in the software HYPERCHEM [8].

123. The molecular descriptors in the reported models, selected by Genetic Algorithm-Variable Subset Selection [9], are defined in the text and the software for their calculation reported (DRAGON, available free from the web site http://www.disat.unimib.it/chm). The values for the selected molecular descriptors are reported in paper [1] (here in Appendix 2), the descriptors values for the papers [2] e [3] are here reported in Appendices 3 and 4. It must be noticed that the web site of the free software for their calculation is always indicated in the cited papers.

124. All the QSAR models are statistically validated both internally by different validation techniques (cross-validation LOO, LMO, Y-scrambling) both externally as requested for acceptable QSAR models [10, 11].

125. Standard Deviation Error in Prediction (*SDEP*), Standard Deviation Error in Calculation (*SDEC*), Standard Error of Estimate (*s*), the F-value of the Fisher, inter-correlation of selected descriptors (K_{XX}) and the correlation of the X block with response (K_{XY}) are also reported for each model (see Appendix 5 for formulas).

126. The QSAR equation and statistics parameters for prediction of rate constant of NO_3 radicals [1] is:

$$-\log k(NO_3) = -28.7 - 2.40\ HOMO + 3.41\ nBnz + 20.41\ MATS1m$$

n(training)=77 R^2=91.2% **R^2adj=90.9%** Q^2=90.3% Q^2_{LMO}(50%)=89.6%

s=0.650 F=253.3 SDEP=0.666 SDEC=0.633

127. The QSAR equation and statistics parameters for prediction of rate constant of O_3 [2] is:

$$-\log k(O_3) = -4.73 - 1.70\ (HOMO\text{-}LUMO)Gap + 099\ nAB + 0.39\ AMW + 0.87 nDB + 1.32 MATS7e$$

n(training)= 83 R^2=88.3% **R^2adj=87.5%** Q^2=86.3% Q^2_{LMO}(50%) =84.6%

s= 0.84 F= 115.65 SDEP= 0.87 SDEC=0.80

Note that the coefficients and the statistics of this model are slightly different from the published one, owing to the different version of the software DRAGON for molecular descriptors calculation.

128. The QSAR equation and statistics parameters for prediction of ATmospheric Persistence INdex [3] is:

ATPIN = 43.69 + 1.77 HOMO –2.27 nBnz –27.51 Me + 0.36 DELS

n(training)=44 R^2=92.8% **R^2adj.=92.0%** Q^2=90.3% Q^2_{LMO}(50%)=88.0%

s=0.477 F=125.2 SDEP=0.521 SDEC=0.449

129. The training set for each model has been published in each paper. The chemicals of the data set (training and test) are always reported with chemical names, CAS numbers (the structural formulae are not reported owing to the large number of the chemicals) and response variables. For the molecular descriptor values see above.

130. The studied QSAR models are all reproducible by other labs as both the training and the test sets are available (either X and Y variables).

131. Papers [1-3] give no description of the data processing because the raw experimental data are used directly, as reported in the cited references.

2.3. Mechanistic basis (Principle 3)

132. The molecular descriptors of each QSAR model are explained in the papers as explicative of molecular features related to the known mechanism of the studied degradations. These descriptors support the reaction mechanism of the different degradation and are all highly informative of different aspects of the studied reactions: the energetics of the attack by electrophiles.

133. In particular, quantum-chemical descriptors (HOMO and HOMO-LUMO gap) are the most important descriptors, selected by Genetic Algorithm from among several hundred of possible variables, highlighting their well known and already reported relevance in the modelling of similar degradative reactions. Other descriptors (like, for instance, nBnz, number of aromatic rings, nDB, number of double bonds, and nAB, number of conjugated double bonds) model the different reactivity of various functional groups in relation to the attack sites for oxidants. The electronic distribution is condensed in descriptors like Me (mean atomic Sanderson electronegativity), DELS (molecular electrotopological variation) and by an autocorrelation descriptor (weighted by the atomic electronegativity of Sanderson) like 2D-Moran MATS7e. The information related to the dimension and shape of the molecules is encoded by 2D-Moran MATS1m, weighed by atomic masses.

2.4. Domain of applicability (Principle 4)

134. The domain of applicability is verified by the leverage approach, that allow the definition of a cut-off value: chemicals outside this value are outside the chemical domain of the model, thus their predicted data, being extrapolated, could be unreliable [11, 12].

135. These chemicals are always reported in the original papers [1-3] together with the outliers, verified by the Williams plot. This plot verifies the presence of outliers (i.e. compounds with cross-validated standardized residuals greater than two or three standard deviation units) and chemicals that are too *influential* in the model space as it is identified by the descriptors of the studied model (i.e. compounds with high *leverage* value (h), greater than 3 p'/n, where p' is the number of the model variables plus one, and n the number of the objects used to calculate the model). The leverage *h* of a compound in the original variable space (*x*) is defined as:

$$h_{ii} = x_i^T (X^T X)^{-1} x_i \qquad \text{(i=1,……..,n)}$$

where h_{ii} are the diagonal elements of the Hat Matrix or Influence Matrix ($H=X(X^TX)^{-1}X^T$), X is the variables (here molecular descriptors) matrix.

136. Thus, being the *leverage* values calculated by the X matrix of all the descriptors in the model, the chemical domain cannot be simply defined by the range of single descriptors in the MLR models. Applying this formula by a multivariate analysis package, the leverage values can be calculated also for new chemicals, starting from their descriptors values. This leverage approach for the definition of chemical domain is an *a posteriori* approach: new chemicals can be checked for their leverage value and can be considered into or out the chemical space, defined by the model variables.

137. For completion, in Table 1 we have reported the minimum and maximum values of the descriptors selected in the three models.

2.5. Internal performance (Principle 5)

138. We report in this report the requested missing values of R^2 adjusted for number of descriptors (section 2.2., in **bold**), generally not useful when other internal validation parameters are indicated, as in the studied papers. Any other requested statistics are reported in the original papers and also in this report.

139. The goodness-of-fit to the training set of the studied regression models was always verified by high values of R^2 and small values of the standard errors of the estimate (s and SDEC).

140. This checking is not sufficient to demonstrate the applicability of the model for the data prediction.

141. Thus in addition, several types of cross-validation on each model were performed in order to give an indication of the stabilities of the descriptor coefficients in each models, of the robustness and internal predictivity of the models. Y-scrambling was also applied to exclude chance correlation.

142. Internal cross-validations by *leave-one out* (Q^2) and by *leave-many out* (Q^2_{LMO}) procedure were performed (the stronger perturbation of 50% of the training compounds out was performed, for this reason the values of Q^2_{LMO} leaving out 10% and 30% of the training set are not reported in original papers, being redundant and not useful; in this report we have performed the intermediate perturbations and the respective values are reported in Table 2).

143. The permutation of the responses (Y-scrambling) was also applied several times (300) to verify the reliability of the obtained models. This validation option was performed by permutation testing: models are recalculated for randomly reordered response. If the new models obtained on the set with randomized response 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.

144. This work has been already documented in the original papers [1-3], but the Contractor has re-checked the statistics under this contract work.

2.6. External validation for predictivity (Principle 6)

145. The external validation was always performed by using representative test sets selected by Experimental Design procedure. Predictive capability is calculated from: 1-PRESS/SD, where PRESS is the sum of squared differences between the measured response and the predicted value for each molecule

in the test set and SD is the sum of squared deviations between the measured response for each molecule in the test set and the mean measured value of the training set.

146. No comparison of the predictive performance of the model was performed as quantitative performance criteria had still not been defined.

147. In the original papers [1-3] two different splitting have been performed using molecular descriptors and response: (a) 67% of the chemicals in the training set and 33% in the test set, and (b) 50% of the chemicals in each set.

148. Predictive performances are reported in table 3, but the same parameters are in the texts of the papers.

149. All the published models have high predictive performances Q^2_{EXT} as requested for acceptable QSAR models [10, 11, 12], verified again in this work.

150. It is still not so common to test QSAR model (characterized by a reasonably high Q^2_{LMO}) for their ability to predict accurately activities/properties of compounds from an external test dataset. Although, the low value of Q^2_{LMO} for the training set can indeed serve as indicator of a low predictive ability of a model, the opposite is not necessarily true. Indeed, the high Q^2_{LMO} does not imply automatically a high predictive ability of the model. The only way to estimate the true predictive power of a model is to test it on a sufficiently large collection of compounds from an external test set [10].

3. CONCLUSIONS ON THE SETUBAL PRINCIPLES AND THEIR APPLICABILITY

151. In conclusion, after having checked the application of Setubal principles to the requested QSAR models of atmospheric degradation, the author has verified that:

1. For the investigated QSAR models the Setubal principles have been completely fulfilled; thus, on the basis of this information, the QSAR models could certainly be regarded as sufficiently well developed to undergo an independent, external validation process. In fact, it should be noted that the Setubal principles were originally applied during the development of these models.

2. The Setubal principles are, in the Contractor's opinion and experience, essential for the evaluation of the quality of existing QSARs (regarding the robustness, internal performance, predictive power, domain of applicability and the mechanistic basis). The Contractor, from her wide experience on statistical validation, highlights that:

   a) The checking of the internal performances like goodness of fit (R^2) is absolutely not sufficient to demonstrate the applicability of the model for the data prediction.

   b) Internal cross-validation is essential to verify the robustness of a QSAR model, but validation by leave-one-out (LOO) is too optimistic and more severe perturbation must be performed by leave-many-out (LMO, up to 50%) or bootstrapping. In addition, Y-scrambling of the responses is useful to verify that the models are not by chance.

In conclusion, internal cross-validation, not limited to LOO, must be performed; there should no request like: "if available".

c) Different kinds of internal cross-validation are not sufficient for the definition of true predictive power: in fact, models with high internal validation parameters (even high LMO) can have very low Q2ext (in some cases even negative).

Only externally validated QSAR models can be considered useful for predictive purposes.

d) The ***domain of applicability*** must be defined or by ***a priori*** or ***a posteriori*** approaches.

e) The molecular descriptors in QSAR models, if theoretical, must derive from clearly defined and chemically interpretable algorithms, and interpreted (if possible) on a ***mechanistic basis***.

4. REFERENCES

Atkinson, R. (1989). Kinetics and mechanisms of the gas-phase reactions of the hydroxyl radical with organic compounds. *Journal of Physical Reference Data*, Monograph 1.

Atkinson, R. (1991).Kinetics and mechanisms of the gas-phase reactions of the NO_3 radical with organic compounds. *Journal of Physical Reference Data* **20**, 459-507.

Atkinson, R. and Carter, W.P.L (1984). Kinetics and mechanisms of gas-phase reactions of ozone with organic compounds under atmospheric conditions. *Chemical Review*s **84**, 437-470.

Eriksson, L., Jaworska, J., Worth, A., Cronin, M., McDowell, R.M. and Gramatica, P. (2003). Methods for reliability, uncertainty assessment, and applicability evaluations of regression based and classification QSARs. *Environmental health perspectives* **111** (10), 1361-1375.

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

Gramatica, P., Pilutti, P. and Papa, E. (2002). Ranking of volatile organic compounds for tropospheric degradability by oxidants: a QSPR approach. *SAR & QSAR in Environmental Research* **13**, 743-753.

Gramatica, P., Pilutti, P. and Papa, E. (2003). Predicting the NO_3 tropospheric degradability of organic pollutants by theoretical molecular descriptors. *Atmospheric Environment* **37** (22) , 3115-3124.

Gramatica, P., Pilutti, P. and Papa, E. (2003). QSAR prediction of ozone tropospheric degradation. *QSAR & Combinatorial Science* **22**, 364-373.

HyperChem (2002). Release 7.03 for Windows. Molecular Modeling System. Hypercube, Inc. Gainesville, Florida, USA.

Todeschini, R., Consonni, V. and Pavan, M. (2002). *MOBY DIGS - Software for multilinear regression analysis and variable subset selection by Genetic Algorithm*, Version 1.2 for Windows, Talete srl, Milan, Italy.

Todeschini, R., Consonni, V., Mauri, A. and Pavan, M. (2003). *DRAGON Web version - Software for the calculation of molecular descriptors*, version 3.0 for Windows. Free download available at: http://www.disat.unimib/chm.

Tropsha, A., Gramatica, P. and. Gombar, V.K. (2003). The importance of being Earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR & Combinatorial Science* **22**, 69-76

Table 1 Minimum and maximum values of the descriptors selected in the three models

k(NO_3) [1]	max	min
HOMO	-8.17	-11.98
nBnz	1	0
MATS1m	1.03	0.81
k(O_3) [2]		
(Homo-Lumo)gap	-8.687	-15.868
nAB	6	0
AMW	16.67	3.76
nDB	3	0
MATS7e	0.8	-1
ATPIN [3]		
Homo	-8.62	-11.98
nBnz	1	0
Me	1.07	0.96
DELS	5.84	0

152. In figure 1 we have reported, as an example, the Williams plot of the model for the prediction of rate constant of NO_3 radicals [1], the data are also reported in Appendix 2. This model have five outliers (5, 58, 64, 74, 75) (>2 σ) and no one chemical with an high *leverage* value (h>0.16). The Williams plot of the two other models are not reported here being similar to this one.

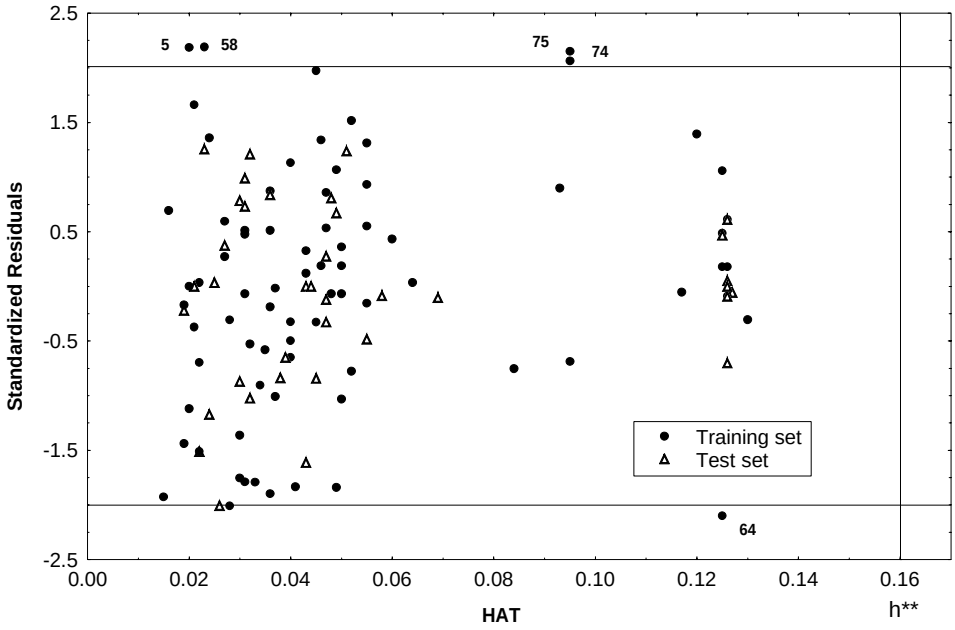
Table 2 Internal cross-validations by leave-one out (Q^2) and by leave-many-out (Q^2_{LMO}) procedures

	Q^2_{LMO}(10%)	Q^2_{LMO}(20%)	Q^2_{LMO}(30%)	Q^2_{LMO}(40%)	Q^2_{LMO}(50%)
k(NO_3) [1]	90.2	90.2	90.0	89.7	89.6
k(O_3) [2]	86.2	86.0	85.7	85.6	84.7
ATPIN [3]	90.0	90.0	89.2	88.5	88.0

Table 3 Predictive performances

	n(test) (a)	Q^2ext (a)	n(test) (b)	Q^2ext (b)
k(NO_3) [1]	37	95.7%	56	94.8%
k(O_3) [2]	42	90.0%	62	89.4%
ATPIN [3]	21	94.3%	32	92.2%

Figure 1 Williams plot of the model for the prediction of rate constant of NO3 radicals



5. APPENDICES

APPENDIX 1 SUMMARY REPORT OF THE APPLICATION OF THE SETUBAL PRINCIPLES

PRINCIPLE	CONSIDERATIONS	YES / NO
1) Defined endpoint	**1.1)** Does the model have a clearly defined scientific purpose? (i.e does it make predictions of a clearly defined physicochemical, biological or environmental effect?)	YES
	1.2) Does the model have the potential 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)?	YES
	1.3) Is information given about important experimental conditions that affect the measurement and therefore the prediction (e.g. sex, species, temperature, exposure period)	YES
	1.4) Are the units of measurement of the endpoint given?	YES
2) Defined algorithm	**2.1)** In the case of a SAR, is there an explicit description of the substructure, including an explicit identification of its substituents?	
	2.2) In the case of a QSAR, is the equation explicitly defined, including definitions of all descriptors used[3]?	YES

3) Mechanistic basis	**3.1)** In the case of a SAR, is there a description of the molecular events that underlie the reactivity of the molecule? (e.g. a description of how substructural features could act as nucleophiles or electrophiles, or form part or all of a receptor-binding region)	
	3.2) In the case of a (Q)SAR, do the descriptors have a physicochemical interpretation that is consistent with a known mechanism (of biological action/environmental endpoint)?	YES The descriptors have been explained as explicative of molecular features related to the known mechanism of the studied degradations
	3.3) Are any literature references cited in support of the purported mechanistic basis of the (Q)SAR?	YES
4) Domain of applicability	**4.1)** In the case of a SAR, is the substructure associated with any inclusion and/or exclusion rules on its applicability to groups of chemicals?	
	4.2) In the case of a SAR, is the substructure associated with rules regarding the modulatory effects of the substructure's molecular environment?	
	4.3) In the case of a (Q)SAR, are the descriptor and response variables associated with inclusion and/or exclusion rules that define the variable ranges for which the QSAR is applicable (i.e. makes reliable estimates)?	YES
5) Internal performance	**5.1)** Are full details of the training set given, including details of chemical names, structural formulae, CAS numbers (if available), and data for all descriptor and response variables?	YES
	5.2) If the data used to the develop the model were based upon the processing of raw data (e.g. the averaging of replicate values): a) is there an adequate description of the data processing?	
	b) are the raw data provided?	YES

	5.3) Is there a specification of the statistical method(s) used to develop the QSAR (including details of any software packages used)?	YES
	5.4) Is the QSAR associated with basic statistics for its goodness-of-fit to the training set?	
	(e.g. r^2 values and standard error of the estimate in the case of regression models)	YES
	5.5)	
	a) Is the QSAR associated with any statistics based on cross-validation or resampling?	YES
	b) If yes, is the number or samples used indicated?	YES
6) Predictivity (External validation)	**6.1)** Does application of the appropriate statistical method(s) to the training set result in the same (Q)SAR model?	YES
	6.2) Is there any information to indicate that the (Q)SAR has been validated previously, using a test set that is independent of the training set?	YES
	6.3) If an external validation has been performed, is the following information available:	
	a) the number of test structures?	YES
	b) the identities of the test structures?	YES
	c) the approach for selecting the test structures?	YES
	d) the statistical analysis of the predictive performance of the model?	YES
	(e.g. including sensitivity, specificity, and positive and negative predictivities for classification models)	
	e) a comparison of the predictive performance of the model against previously-defined quantitative performance criteria?	YES

APPENDIX 2 TABLE OF CHEMICALS STUDIED AND REPORTED IN PAPER [1]

List of studied chemicals (test set in **bold**), experimental, predicted $-\log k_{NO3}$, residuals (r) and descriptor values

ID	Chemicals	CAS	-log k_{NO3} exp.	-log k_{NO3} pred.	r	HOMO[1]	nBnz[2]	MATS1m[3]
1	Formaldehyde	50-00-0	15.23	16.30	1.07	-10.63	0	0.96
2	Methanol	67-56-1	15.68	15.99	0.31	-11.14	0	0.88
3	2-Propanol	67-63-0	14.64	15.44	0.80	-11.04	0	0.87
4	Ethane	74-84-0	17.10	16.59	-0.51	-11.98	0	0.81
5	Ethene	74-85-1	15.91	14.62	-1.29	-10.64	0	0.87
6	Ethyne	74-86-2	16.29	16.89	0.60	-11.01	0	0.94
7	Chloromethane	74-87-3	16.51	16.26	-0.25	-10.48	0	0.97
8	Methanethiol	74-93-1	12.00	11.72	-0.28	-9.21	0	0.90
9	1-Propyne	74-99-7	15.76	15.94	0.18	-10.89	0	0.91
10	1-Chloroethene	75-01-4	15.37	14.83	-0.54	-9.84	0	0.98
11	Acetaldehyde	75-07-0	14.61	15.79	1.18	-10.70	0	0.92
12	Ethanethiol	75-08-1	11.92	11.62	-0.30	-9.25	0	0.89
13	Dichloromethane	75-09-2	17.30	17.33	0.03	-10.58	0	1.01
14	Dimethyl sulfide	75-18-3	12.00	10.85	-1.15	-8.88	0	0.90
15	2-Methylpropane	75-28-5	16.01	15.99	-0.02	-11.59	0	0.83
16	1.1-Dichloroethene	75-35-4	14.91	15.30	0.39	-9.74	0	1.01
17	**2-Methylbutane**	78-78-4	15.80	15.85	0.05	-11.50	0	0.83
18	**2-Methyl-1.3-butadiene**	78-79-5	12.23	11.65	-0.58	-9.31	0	0.88
19	1.1.2-Trichloroethene	79-01-6	15.55	14.77	-0.78	-9.38	0	1.03
20	2.3-Dimethylbutane	79-29-8	15.39	15.43	0.04	-11.30	0	0.84
21	Camphene	79-92-5	12.18	12.84	0.66	-9.92	0	0.87
22	α - Pinene	80-56-8	11.24	11.35	0.11	-9.30	0	0.87
23	β-Caryophyllene	87-44-5	10.72	11.75	1.03	-9.45	0	0.87
24	1.2-Dimethylbenzene	95-47-6	15.42	15.32	-0.10	-9.29	1	0.90
25	**1.2.4-Trimethylbenzene**	95-63-6	14.74	14.71	-0.03	-9.08	1	0.89
26	3-Methylpentane	96-14-0	15.69	15.78	0.09	-11.45	0	0.84
27	**α- Phellandrene**	99-83-2	10.07	10.35	0.28	-8.85	0	0.88
28	**γ- Terpinene**	99-85-4	10.53	10.72	0.19	-9.00	0	0.88
29	**α - Terpinene**	99-86-5	9.74	9.80	0.06	-8.62	0	0.88
30	**1.4-Methylisopropylbenzene**	99-87-6	15.00	15.03	0.03	-9.26	1	0.89
31	Methoxybenzene	100-66-3	15.68	15.34	-0.34	-9.11	1	0.92
32	1.3.5-Trimethylbenzene	108-67-8	15.10	15.15	0.05	-9.27	1	0.89

33	**Ethyl propionate**	105-37-3	16.46	16.95	0.49	-11.36	0	0.90
34	1.4-Dimethylbenzene	106-42-3	15.34	15.07	-0.27	-9.18	1	0.90
35	n-Butane	106-97-8	16.18	15.42	-0.76	-11.35	0	0.83
36	**1-Butene**	106-98-9	13.91	13.17	-0.74	-10.15	0	0.86
37	1.3-Butadiene	106-99-0	13.01	12.21	-0.80	-9.47	0	0.89
38	1-Butyne	107-00-6	15.34	15.32	-0.02	-10.77	0	0.89
39	Acrolein	107-02-8	14.96	16.07	1.11	-10.69	0	0.94
40	3-Chloro-1-propene	107-05-1	15.27	15.28	0.01	-10.23	0	0.95
41	Methyl formate	107-31-3	17.52	17.20	-0.32	-11.15	0	0.94
42	**2-Methylpentane**	107-83-5	15.69	15.30	-0.39	-11.25	0	0.84
43	2.4-Dimethylpentane	108-08-7	15.84	15.34	-0.50	-11.25	0	0.84
44	**1.3-Dimethylbenzene**	108-38-3	15.63	15.37	-0.26	-9.31	1	0.90
45	**1.3.5-Trimethylbenzene**	108-67-8	15.10	15.15	0.05	-9.27	1	0.89
46	**Methylbenzene**	108-88-3	16.18	15.84	-0.34	-9.44	1	0.91
47	1-Propylacetate	109-60-4	16.30	16.68	0.38	-11.25	0	0.90
48	**n-Pentane**	109-66-0	16.09	15.37	-0.72	-11.30	0	0.83
49	Ethyl formate	109-94-4	16.70	16.89	0.19	-11.16	0	0.92
50	Pyrrole	109-97-7	10.34	11.41	1.07	-8.93	0	0.92
51	Tetrahydrofuran	109-99-9	14.31	13.90	-0.41	-10.27	0	0.88
52	Furan	110-00-9	11.84	12.89	1.05	-9.38	0	0.94
53	Thiophene	110-02-1	13.41	13.45	0.04	-9.54	0	0.95
54	n-Hexane	110-54-3	15.98	15.36	-0.62	-11.28	0	0.84
55	**Propyl formate**	110-74-7	16.27	16.65	0.38	-11.16	0	0.91
56	Cyclohexane	110-82-7	15.87	15.68	-0.19	-11.29	0	0.85
57	**Cyclohexene**	110-83-8	12.28	12.06	-0.22	-9.59	0	0.87
58	Pyridine	110-86-1	15.82	14.53	-1.29	-10.10	0	0.93
59	n-Octane	111-65-9	15.74	15.43	-0.31	-11.27	0	0.84
60	**n-Nonane**	111-84-2	15.62	15.46	-0.16	-11.28	0	0.84
61	1-Propene	115-07-1	14.13	13.15	-0.98	-10.11	0	0.86
62	Dimethyl ether	115-10-6	14.52	14.74	0.22	-10.69	0	0.87
63	2-Methyl-1-propene	115-11-7	12.50	12.34	-0.16	-9.80	0	0.86
64	Tetralin	119-64-2	14.06	15.23	1.17	-9.25	1	0.90
65	Bicyclo(2.2.1)-2.5-heptadiene	121-46-0	11.99	12.84	0.85	-9.66	0	0.90
66	N.N-Dimetyl-aniline	121-69-7	15.70	15.11	-0.59	-9.18	1	0.90
67	**Myrcene**	123-35-3	10.98	11.58	0.60	-9.32	0	0.88
68	**Crotonaldehyde**	123-73-9	14.29	15.18	0.89	-10.48	0	0.92
69	n-Decane	124-18-5	15.59	15.48	-0.11	-11.28	0	0.84
70	**β- Pinene**	127-91-3	11.63	12.32	0.69	-9.70	0	0.87
71	**Ethyl acetate**	141-78-6	16.85	16.85	0.00	-11.24	0	0.91
72	Cyclopentene	142-29-0	12.33	11.98	-0.35	-9.53	0	0.88
73	**n-Heptane**	142-82-5	15.86	15.39	-0.47	-11.27	0	0.84
74	cis-1.2-Dichloroethene	156-59-2	15.86	14.69	-1.17	-9.49	0	1.01
75	trans-1.2-Dichloroethene	156-60-5	15.97	14.75	-1.22	-9.52	0	1.01
76	Azulene	275-51-4	9.41	9.84	0.43	-8.17	0	0.93
77	Diethyl sulfide	352-93-2	11.38	10.50	-0.88	-8.86	0	0.88
78	2.2.3-Trimethylbutane	464-06-2	15.65	15.54	-0.11	-11.33	0	0.84
79	Longifolene	475-20-7	12.17	12.58	0.41	-9.82	0	0.87
80	1.4-Benzodioxan	493-09-4	15.22	15.39	0.17	-9.02	1	0.93
81	**Bicyclo(2.2.1)-2-heptene**	498-66-8	12.61	12.74	0.13	-9.78	0	0.88

82	2-Butyne	503-17-3	13.17	14.31	1.14	-10.35	0	0.89
83	2-Methyl-2-butene	513-35-9	11.02	11.31	0.29	-9.39	0	0.86
84	2.3-Dimethyl-1.3-butadiene	513-81-5	11.64	11.34	-0.30	-9.23	0	0.88
85	**1.2.3-Trimethylbenzene**	526-73-8	14.73	15.12	0.39	-9.25	1	0.89
86	2.2.4-Trimethylpentane	540-84-1	16.13	15.35	-0.78	-11.24	0	0.84
87	1.3.5-Cycloheptatriene	544-25-2	11.92	11.26	-0.66	-8.95	0	0.91
88	Methyl propionate	554-12-1	16.48	16.67	0.19	-11.17	0	0.91
89	**2-Carene**	554-61-0	10.73	11.22	0.49	-9.24	0	0.87
90	2.3-Dimethyl-2-butene	563-79-1	10.24	10.69	0.45	-9.15	0	0.86
91	**Terpinolene**	586-62-9	10.02	10.96	0.94	-9.10	0	0.88
92	**cis-2-Butene**	590-18-1	12.46	12.00	-0.46	-9.66	0	0.86
93	1-Methylcyclohexene	591-49-1	10.77	11.36	0.59	-9.33	0	0.87
94	**1.3-Cyclohexadiene**	592-57-4	10.91	10.91	0.00	-8.92	0	0.89
95	**1.3-Ethylmethylbenzene**	620-14-4	15.36	15.36	0.00	-9.35	1	0.89
96	1.4-Ethylmethylbenzene	622-96-8	15.21	15.11	-0.10	-9.25	1	0.89
97	**trans-2-Butene**	624-64-6	12.42	11.99	-0.43	-9.66	0	0.86
98	**1-Pentyne**	627-19-0	15.12	15.12	0.00	-10.76	0	0.88
99	**1.4-Cyclohexadiene**	628-41-1	12.27	11.56	-0.71	-9.19	0	0.89
100	**Cycloheptene**	628-92-2	12.32	12.30	-0.02	-9.73	0	0.87
101	1-Hexyne	693-02-7	14.80	14.80	0.00	-10.68	0	0.88
102	Bicyclo(2.2.2)-2-octene	931-64-6	12.84	12.94	0.10	-9.91	0	0.88
103	cis-1.3-Pentadiene	1574-41-0	11.85	11.34	-0.51	-9.18	0	0.88
104	**trans-1.3-Pentadiene**	2004-70-8	11.80	11.31	-0.49	-9.17	0	0.88
105	cis - Ocimene	3338-55-4	10.62	10.66	0.04	-8.94	0	0.88
106	**Sabinene**	3387-41-5	11.00	12.18	1.18	-9.64	0	0.87
107	trans - Ocimene	3779-61-1	10.98	10.91	-0.07	-9.05	0	0.88
108	**1.3-Cycloheptadiene**	4054-38-0	11.19	11.70	0.51	-9.31	0	0.89
109	trans-trans-2.4-Hexadiene	5194-51-4	10.80	10.59	-0.21	-8.92	0	0.88
110	Limonene	5989-27-5	10.95	11.48	0.53	-9.32	0	0.88
111	α-Humulene	6753-98-6	10.46	11.51	1.05	-9.33	0	0.88
112	3 - Carene	13466-78-9	11.09	11.43	0.34	-9.33	0	0.87
113	**Ocimene**	13877-91-3	10.66	10.73	0.07	-8.97	0	0.88
114	Crotonaldehyde	4170-30-3	14.29	15.18	0.89	-10.48	0	0.92

FOOTNOTES

[1] HOMO is the highest occupied molecular orbital (quantum-chemical descriptor).

[2] nBnz is the number of aromatic rings (constitutional descriptor).

[3] MATS1m is the 2D-Moran autocorrelation weighted by atomic masses (2D-autocorrelation descriptor).

APPENDIX 3 TABLE OF CHEMICALS STUDIED AND REPORTED IN PAPER [2] ALONG WITH VALUES OF THEIR DESCRIPTORS

List of studied chemicals (test set in **bold**), experimental, predicted –logk_{O3}, residuals and selected molecular descriptors values.

ID	Chemicals	CAS	logkO3exp. [9,10]	$logkO_3$ pred.	Residuals	(Homo-Lumo)gap[1]	nAB[2]	AMW[3]	NDB[4]	MATS7e[5]
1	Benzene	71-43-2	22.16	21.02	1.14	-10.15	6	6.51	0	0
2	Ethane	74-84-0	22.92	23.67	-0.75	-15.87	0	3.76	0	0
3	Ethene	74-85-1	17.48	18.05	-0.57	-11.87	0	4.68	1	0
4	Ethyne	74-86-2	19.42	20.07	-0.65	-13.15	0	6.51	0	0
5	Methylamine	74-89-5	19.67	18.18	1.49	-12.51	0	4.44	0	0
6	Propane	74-98-6	23.17	22.65	0.52	-15.22	0	4.01	0	0
7	Propyne	74-99-7	17.89	19.19	-1.3	-12.81	0	5.72	0	0
8	1-Chloroethene	75-01-4	18.61	18.02	0.59	-10.54	0	10.42	1	0
9	Vinyl fluoride	75-02-5	18.16	18.26	-0.1	-11.31	0	7.67	1	0
10	Ethylamine	75-04-7	19.56	18.4	1.16	-12.63	0	4.51	0	0
11	Acetaldehyde	75-07-0	19.47	18.06	1.41	-11.51	0	6.29	1	0
12	2-Methylpropane	75-28-5	22.7	22.66	0.04	-15.20	0	4.15	0	0
13	1.1-Dichloroethene	75-35-4	20.43	19.46	0.97	-10.07	0	16.16	1	0
14	1.1-Difluoroethane	75-37-6	24.22	23.6	0.62	-14.80	0	8.26	0	0
15	1.1-Difluoroethene	75-38-7	18.72	18.88	-0.16	-10.98	0	10.67	1	0
16	Trimethylamine	75-50-3	17.01	16.99	0.02	-11.80	0	4.55	0	0
17	**Isoprene**	78-79-5	16.9	16.52	0.38	-9.59	3	5.24	0	0
18	Methacrolein	78-85-3	17.96	17	0.96	-10.36	0	6.37	2	0
19	Methyl vinyl ketone	78-94-4	17.4	17.56	-0.16	-10.69	0	6.37	2	0
20	**α-Pinene**	80-56-8	15.48	15.6	-0.12	-10.32	0	5.24	1	0
21	**o-Xylene**	95-47-6	21.16	19.97	1.19	-9.68	6	5.9	0	0
22	o-Cresol	95-48-7	18.59	19.74	-1.15	-9.34	6	6.76	0	0
23	**1.2.4-Trimethylbenzene**	95-63-6	20.89	20.41	0.48	-9.46	6	5.72	0	0.75
24	**α -Phellandrene**	99-83-2	13.92	15.02	-1.1	-9.18	3	5.24	0	-0.675
25	**γ-Terpinene**	99-85-4	15.55	15.23	0.32	-10.05	0	5.24	2	-0.675
26	α -Terpinene	99-86-5	13.06	14.52	-1.46	-8.89	3	5.24	0	-0.675
27	1.3.5-Trimethylbenzene	108-67-8	20.66	19.93	0.73	-9.69	6	5.72	0	0
28	p-Xylene	106-42-3	21.4	20.67	0.73	-9.54	6	5.9	0	0.8
29	p-Cresol	106-44-5	18.33	19.96	-1.63	-9.28	6	6.76	0	0.284

30	n-Butane	106-97-8	23.01	22.18	0.83	-14.91	0	4.15	0	0
31	**1-Butene**	106-98-9	16.99	17.12	-0.13	-11.33	0	4.68	1	0
32	1.3-Butadiene	106-99-0	17.09	16.83	0.26	-9.73	3	5.41	0	0
33	1-Butyne	107-00-6	17.75	18.83	-1.08	-12.68	0	5.41	0	0
34	Acrolein	107-02-8	18.13	17.51	0.62	-10.51	0	7.01	2	0
35	m-Xylene	108-38-3	21.22	20.01	1.21	-9.70	6	5.9	0	0
36	**m-Cresol**	108-39-4	18.71	19.82	-1.11	-9.39	6	6.76	0	0
37	**Methylbenzene**	108-88-3	19.92	20.31	-0.39	-9.82	6	6.14	0	0
38	**1-Pentene**	109-67-1	17.28	17.1	0.18	-11.32	0	4.68	1	0
39	Ethyl nitrite	109-95-5	18.93	16.99	1.94	-10.61	0	7.51	1	0
40	Pyrrole	109-97-7	16.8	17.87	-1.07	-10.04	3	6.71	0	0
41	**Furan**	110-00-9	17.62	18.11	-0.49	-9.99	3	7.56	0	0
42	Thiophene	110-02-1	19.14	17.72	1.42	-9.35	3	9.35	0	0
43	**Cyclohexene**	110-83-8	15.98	16.33	-0.35	-10.76	0	5.14	1	0
44	1-Octene	111-66-0	17.09	17.1	-0.01	-11.32	0	4.68	1	0
45	1-Propene	115-07-1	16.84	17.04	-0.2	-11.28	0	4.68	1	0
46	2-Methyl-1-propene	115-11-7	16.94	16.43	0.51	-10.93	0	4.68	1	0
47	Tetrafluoroethene	116-14-3	19.04	19.73	-0.69	-10.11	0	16.67	1	0
48	Hexafluoro-1-propene	116-15-4	19.11	19.97	-0.86	-10.25	0	16.67	1	0
49	Myrcene	123-35-3	14.9	15.42	-0.52	-9.63	0	5.24	3	-0.648
50	Dimethylamine	124-40-3	17.58	17.54	0.04	-12.13	0	4.51	0	0
51	β-Pinene	127-91-3	16.68	16.45	0.23	-10.81	0	5.24	1	0
52	**Cyclopentene**	142-29-0	15.56	16.19	-0.63	-10.66	0	5.24	1	0
53	cis-1.2-Dichloroethene	156-59-2	19.21	18.96	0.25	-9.78	0	16.16	1	0
54	trans-1.2-Dichloroethene	156-60-5	18.75	18.95	-0.2	-9.78	0	16.16	1	0
55	Trifluoroethene	359-11-5	18.85	19.1	-0.25	-10.43	0	13.67	1	0
56	1.1.1-Trifluoroethane	420-46-2	25.27	25.83	-0.56	-15.59	0	10.51	0	0
57	1.2-Propadiene	463-49-0	18.72	18.35	0.37	-11.29	0	5.72	2	0
58	2.2-Dimethylpropane	463-82-1	23.01	23.56	-0.55	-15.69	0	4.25	0	0
59	Bicyclo[2.2.1]-2-heptene	498-66-8	14.67	16.67	-2	-10.87	0	5.54	1	0
60	2-Butyne	503-17-3	19.48	17.97	1.51	-12.17	0	5.41	0	0
61	2-Methyl-2-butene	513-35-9	15.4	15.63	-0.23	-10.46	0	4.68	1	0
62	**2.3-Dimethyl-1.3-butadiene**	513-81-5	16.58	16.36	0.22	-9.52	3	5.14	0	0
63	1.2.3-Trimethylbenzene	526-73-8	20.8	19.91	0.89	-9.68	6	5.72	0	0
64	1.3.5-Cycloheptatriene	544-25-2	16.27	17.87	-1.6	-8.99	5	6.14	0	0
65	**β-Phellandrene**	555-10-2	15.75	15.34	0.41	-9.40	3	5.24	0	-0.729
66	**3.3-Dimethyl-1-butene**	558-37-2	17.28	17.27	0.01	-11.42	0	4.68	1	0
67	**2-methyl-1-butene**	563-46-2	16.8	16.44	0.36	-10.94	0	4.68	1	0
68	**2.3-Dimethyl-1-butene**	563-78-0	16.89	16.52	0.37	-10.98	0	4.68	1	0
69	2.3-Dimethyl-2-butene	563-79-1	14.82	15.13	-0.31	-10.17	0	4.68	1	0
70	**Terpinolene**	586-62-9	14	14.96	-0.96	-10.11	0	5.24	2	-1
71	**cis-2-Butene**	590-18-1	15.9	16.19	-0.29	-10.79	0	4.68	1	0
72	**4-Methyl-1-cyclohexene**	591-47-9	16.09	16.29	-0.2	-10.76	0	5.06	1	0
73	1-Methyl-1-cyclohexene	591-49-1	15.78	15.76	0.02	-10.45	0	5.06	1	0
74	1-Hexene	592-41-6	16.9	17.69	-0.79	-11.32	0	4.68	1	0.5
75	**1.3-Cyclohexadiene**	592-57-4	14.71	16.06	-1.35	-9.21	3	5.72	0	0
76	**1-Heptene**	592-76-7	17.09	16.95	0.14	-11.31	0	4.68	1	-0.125
77	2-Heptene	592-77-8	16.1	15.93	0.17	-10.70	0	4.68	1	-0.1

78	2.3.3.-Trimethylbutene	594-56-9	17.08	16.47	0.61	-10.95	0	4.68	1	0
79	**trans-3-Methyl-2-pentene**	616-12-6	15.25	15.45	-0.2	-10.36	0	4.68	1	0
80	**trans-2-Butene**	624-64-6	15.48	16.18	-0.7	-10.78	0	4.68	1	0
81	cis-2-Pentene	627-20-3	15.7	16.27	-0.57	-10.83	0	4.68	1	0
82	2.5-Dimethyl-1.5-hexadiene	627-58-7	16.85	17.83	-0.98	-10.76	0	5.01	2	0.571
83	1.4-Cyclohexadiene	628-41-1	16.19	16.59	-0.4	-10.27	0	5.72	2	0
84	Cycloheptene	628-92-2	15.5	16.43	-0.93	-10.84	0	5.06	1	0
85	**trans-2-Pentene**	646-04-8	15.5	16.26	-0.76	-10.83	0	4.68	1	0
86	**4-Methyl-1-pentene**	691-37-2	16.98	17.02	-0.04	-11.27	0	4.68	1	0
87	**1-Methyl-1-cyclopentene**	693-89-0	15.17	15.64	-0.47	-10.36	0	5.14	1	0
88	2-Ethylbutene	760-21-4	16.89	15.98	0.91	-10.67	0	4.68	1	0
89	**2-Methyl-1-pentene**	763-29-1	16.77	16.12	0.65	-10.75	0	4.68	1	0
90	2-Methyl-1.4-pentadiene	763-30-4	16.89	16.93	-0.04	-10.60	0	5.14	2	0
91	trans-3-hexene-2.5-dione	820-69-9	17.08	17.91	-0.83	-10.01	0	7.01	3	0.32
92	1-Decene	872-05-9	16.97	17.09	-0.12	-11.31	0	4.68	1	0
93	**cis-3-Methyl-2-pentene**	922-62-3	15.34	15.71	-0.37	-10.51	0	4.68	1	0
94	2-Cyclohexen-1-one	930-68-7	17.91	18.21	-0.3	-10.31	3	6.41	0	0
95	Bicyclo[2.2.2]-2-octene	931-64-6	16.14	16.93	-0.79	-11.05	0	5.41	1	0
96	cis-Cyclooctane	931-87-3	15.43	16.53	-1.1	-10.91	0	5.01	1	0
97	**2.4-Dimethyl-1.3-butadiene**	1118-58-7	16.1	14.9	1.2	-9.41	0	5.14	2	0
98	**trans-1.2-Difluoroethene**	1630-77-9	17.68	18.48	-0.8	-10.75	0	10.67	1	0
99	**cis-1.2-Difluoroethene**	1630-78-0	18.59	18.5	0.09	-10.76	0	10.67	1	0
100	**1.2-Dimethylcyclohexene**	1674-10-8	15.68	15.24	0.44	-10.16	0	5.01	1	0
101	2.5-Dihydrofuran	1708-29-8	16.79	16.07	0.72	-10.33	0	6.37	1	0
102	2(Cl-methyl)-3-Cl-1-propene	1871-57-4	18.41	19.07	-0.66	-10.96	0	10.42	1	0
103	**trans-1.3-Pentadiene**	2004-70-8	16.37	14.94	1.43	-9.42	0	5.24	2	0
104	1.3.5-Hexatriene	03/12/2235	16.58	15.64	0.94	-8.69	0	5.72	3	0.75
105	**Dihydromyrcene**	2436-90-0	15.17	16.05	-0.88	-10.43	0	4.94	2	-0.436
106	cis-Ocimene	3338-55-4	14.7	14.62	0.08	-9.18	0	5.24	3	-0.675
107	**1.3-Cycloheptadiene**	4054-38-0	15.81	16.94	-1.13	-9.77	3	5.54	0	0
108	cis-2.trans-4-Hexadiene	5194-50-3	15.5	15.16	0.34	-9.16	0	5.14	2	0.6
109	trans-2.trans-4-Hexadiene	5194-51-4	15.43	15.18	0.25	-9.17	0	5.14	2	0.6
110	**D-Limonene**	5989-27-5	15.19	15.51	-0.32	-10.43	0	5.24	2	-1
111	Tetramethylhydrazine	09/12/6415	16.9	16.22	0.68	-11.27	0	4.9	0	0
112	Formaldehyde hydrazone	6629-91-0	16.6	16	0.6	-10.31	0	6.3	1	0
113	cis-5-Decene	7433-78-5	15.92	16.07	-0.15	-10.77	0	4.68	1	-0.077
114	**cis-3-Hexene**	03/09/1942	15.82	16.76	-0.94	-10.78	0	4.68	1	0.5
115	cis-4-Octene	7642-15-1	16.02	16.17	-0.15	-10.78	0	4.68	1	0
116	**cis-1.3-Dichloropropene**	10061-01-5	18.82	17.96	0.86	-10.07	0	12.33	1	0
117	trans-1.3-Dichloropropene	10061-02-6	18.17	17.75	0.42	-9.95	0	12.33	1	0
118	trans-3-Hexene	13269-52-8	15.77	16.59	-0.82	-10.68	0	4.68	1	0.5
119	3-Carene	13466-78-9	15.92	16.51	-0.59	-10.42	0	5.24	1	0.625
120	trans-4-Octene	14850-23-8	15.85	16.07	-0.22	-10.72	0	4.68	1	0
121	**cis-3-hexene-2.5-dione**	17559-81-8	17.75	17.47	0.28	-9.75	0	7.01	3	0.32
122	Crotonaldehyde	4170-30-3	18.13	16.87	1.26	-10.29	0	6.37	2	0
123	Octafluoro-2-butene	360-89-4	20.2	20.19	0.01	-10.38	0	16.67	1	0
124	3-Penten-2-one	625-33-2	16.67	16.81	-0.14	-10.33	0	6.01	2	0
125	Carvomenthene	5502-88-5	15.28	14.85	0.43	-10.42	0	4.94	1	-0.689

FOOTNOTES

[1] (HOMO-LUMO) Gap is the difference in energy between the highest occupied molecular orbital (HOMO) and the lowest occupied molecular orbital (LUMO) (quantum-chemical descriptor).

[2] nAB is the number of conjugated double bonds (constitutional descriptor).

[3] AMW is the average molecular weigh (constitutional descriptor).

[4] nDB is the number of isolated double bonds (constitutional descriptor).

[5] MATS7e is the 2D-Moran autocorrelation weighted by electronegativities (2D-autocorrelation descriptor).

APPENDIX 4 CHEMICALS STUDIED AND REPORTED IN PAPER [3] ALONG WITH VALUES OF THEIR MOLECULAR DESCRIPTORS

List of studied chemicals (test set in **bold**) experimental oxidation rate constants (k_{OH}, k_{O3}, k_{NO3}), PC1 from PCA (ATPINdex), PC1 calculated and selected molecular descriptors values.

ID	Chemicals	CAS	-logk(OH)	-logk(NO_3)	-logk(O_3)	PC1 from PCA	PC1 Mod.2	Homo[1]	nBnz[2]	Me[3]	DELS[4]
1	Formaldehyde	50-00-0	11.47	15.24	23.68	-2.91	-3.33	-10.63	0	1.05	2.00
2	Ethane	74-84-0	12.57	17.40	23.00	-4.33	-3.95	-11.98	0	0.96	0.00
3	Ethene	74-85-1	11.07	15.70	17.76	-1.37	-1.59	-10.64	0	0.96	0.00
4	Ethyne	74-86-2	12.09	16.54	19.32	-2.85	-2.52	-11.01	0	0.97	0.00
5	1-Propyne	74-99-7	11.23	15.64	19.74	-1.93	-1.88	-10.89	0	0.97	1.19
6	1-Chloroethene	75-01-4	11.16	15.37	18.62	-1.54	-1.26	-9.84	0	1.02	1.56
7	Acetaldehyde	75-07-0	10.80	14.56	20.22	-1.36	-2.06	-10.70	0	1.02	3.61
8	2-Methylpropane	75-28-5	11.63	16.00	22.70	-3.05	-2.91	-11.59	0	0.96	1.00
9	1.1-Dichloroethene	75-35-4	10.96	14.91	20.43	-1.65	-1.91	-9.74	0	1.07	3.11
10	2-Methyl-1.3-butadiene	78-79-5	10.00	12.17	16.85	0.76	1.22	-9.31	0	0.97	2.03
11	Camphene	79-92-5	10.27	12.18	18.05	0.24	0.79	-9.92	0	0.97	3.82
12	**α - Pinene**	80-56-8	10.27	11.21	16.08	0.95	1.64	-9.30	0	0.97	3.13
13	β-Caryophyllene	87-44-5	9.70	10.69	13.94	2.08	2.31	-9.45	0	0.97	5.73
14	**1.2-Dimethylbenzene**	95-47-6	10.86	15.42	21.16	-1.87	-1.30	-9.29	1	0.97	1.19
15	o-Cresol	95-48-7	10.38	10.70	18.59	0.44	0.26	-9.04	1	0.99	5.84
16	1.2.4-Trimethylbenzene	95-63-6	10.49	14.74	20.89	-1.29	-0.72	-9.08	1	0.97	1.79
17	**α - Phellandrene**	99-83-2	9.50	10.07	13.92	2.42	2.48	-8.85	0	0.97	3.22
18	**γ - Terpinene**	99-85-4	9.75	10.53	15.55	1.71	2.09	-9.00	0	0.97	2.90
19	**α - Terpinene**	99-86-5	9.44	9.75	13.06	2.76	2.63	-8.62	0	0.97	2.52
20	1.3.5-Trimethylbenzene	108-67-8	10.24	15.10	20.66	-1.11	-1.02	-9.27	1	0.97	1.88
21	1.4-Dimethylbenzene	106-42-3	10.85	15.34	21.40	-1.88	-1.06	-9.18	1	0.97	1.35
22	p-Cresol	106-44-5	10.33	10.70	18.33	0.54	0.29	-8.95	1	0.99	5.51
23	n-Butane	106-97-8	11.60	16.37	23.01	-3.19	-2.59	-11.35	0	0.96	0.72
24	**1-Butene**	106-98-9	10.50	13.92	16.96	-0.20	-0.33	-10.15	0	0.96	1.08
25	1.3-Butadiene	106-99-0	10.18	13.00	17.09	0.31	0.74	-9.47	0	0.97	1.44
26	1-Butyne	107-00-6	11.10	15.34	19.52	-1.69	-1.53	-10.77	0	0.97	1.56
27	Acrolein	107-02-8	10.70	14.96	18.55	-1.01	-1.50	-10.69	0	1.01	4.33
28	**1.3-Dimethylbenzene**	108-38-3	10.63	15.63	21.22	-1.73	-1.29	-9.31	1	0.97	1.32
29	**m-Cresol**	108-39-4	10.19	10.92	18.22	0.62	0.06	-9.10	1	0.99	5.63
30	**1.3.5-Trimethylbenzene**	108-67-8	10.24	15.10	20.66	-1.11	-1.02	-9.27	1	0.97	1.88

31	Methylbenzene	108-88-3	11.23	16.17	19.87	-2.10	-1.76	-9.44	1	0.97	0.69
32	Pyrrole	109-97-7	9.96	10.34	16.80	1.30	0.88	-8.93	0	0.99	0.72
33	**Furan**	110-00-9	10.39	11.84	17.62	0.32	0.06	-9.38	0	1.01	2.17
34	**Cyclohexene**	110-83-8	10.17	12.28	15.98	0.77	0.66	-9.59	0	0.96	1.07
35	1-Propene	115-07-1	10.58	14.03	16.95	-0.29	-0.38	-10.11	0	0.96	0.72
36	2-Methyl-1-propene	115-11-7	10.29	12.48	16.92	0.40	0.30	-9.80	0	0.96	1.11
37	Bicyclo(2.2.1)-2.5-heptadiene	121-46-0	9.92	12.00	14.33	1.43	0.74	-9.66	0	0.97	2.37
38	N.N-Dimetyl-aniline	121-69-7	9.83	15.70	17.04	-0.11	-1.51	-9.18	1	0.98	0.84
39	**Myrcene**	123-35-3	9.67	10.98	14.90	1.81	1.83	-9.32	0	0.97	3.79
40	Crotonaldehyde	123-73-9	10.44	14.29	18.05	-0.49	-0.73	-10.48	0	1.00	4.64
41	**β - Pinene**	127-91-3	10.10	11.60	16.68	0.86	1.19	-9.70	0	0.97	3.85
42	Cyclopentene	142-29-0	10.17	12.34	15.56	0.84	0.46	-9.53	0	0.97	0.94
43	cis-1.2-Dichloroethene	156-59-2	11.58	15.86	19.21	-2.19	-1.48	-9.49	0	1.07	3.05
44	trans-1.2-Dichloroethene	156-60-5	11.63	15.97	18.55	-2.12	-1.52	-9.52	0	1.07	3.05
45	**Bicyclo(2.2.1)-2-heptene**	498-66-8	10.31	12.61	14.67	0.85	0.21	-9.78	0	0.97	1.53
46	2-Butyne	503-17-3	10.56	13.17	19.48	-0.61	-1.08	-10.35	0	0.97	0.72
47	**2-Methyl-2-butene**	513-35-9	10.06	11.03	15.37	1.34	0.83	-9.39	0	0.96	0.57
48	**1.2.3-Trimethylbenzene**	526-73-8	10.49	14.73	20.80	-1.26	-1.08	-9.25	1	0.97	1.63
49	1.3.5-Cycloheptatriene	544-25-2	10.01	11.93	16.27	0.94	1.43	-8.95	0	0.97	0.85
50	2.3-Dimethyl-2-butene	563-79-1	9.96	10.24	14.94	1.75	1.41	-9.15	0	0.96	0.96
51	cis-2-Butene	590-18-1	10.25	12.46	15.70	0.71	0.15	-9.66	0	0.96	0.00
52	**1.3-Cyclohexadiene**	592-57-4	9.79	10.92	14.71	1.76	1.58	-8.92	0	0.97	1.07
53	**trans-2-Butene**	624-64-6	10.19	12.41	15.89	0.73	0.16	-9.66	0	0.96	0.00
54	1.4-Cyclohexadiene	628-41-1	10.00	12.18	16.19	0.90	1.23	-9.19	0	0.97	1.44
55	Cycloheptene	628-92-2	10.13	12.32	15.50	0.90	0.47	-9.73	0	0.96	1.19
56	Bicyclo(2.2.2)-2-octene	931-64-6	10.39	12.84	16.14	0.38	0.03	-9.91	0	0.97	1.66
57	**Sabinene**	3387-41-5	9.93	11.00	16.09	1.30	1.21	-9.64	0	0.97	3.65
58	trans - Ocimene	3779-61-1	9.67	10.98	14.90	1.81	1.91	-9.05	0	0.97	2.63
59	**1.3-Cycloheptadiene**	4054-38-0	9.86	11.19	15.81	1.38	0.98	-9.31	0	0.97	1.32
60	trans-trans-2.4-Hexadiene	5194-51-4	9.87	10.80	15.43	1.56	1.47	-8.92	0	0.96	0.00
61	d-Limonene	5989-27-5	9.77	10.91	15.19	1.67	1.69	-9.32	0	0.97	3.36
62	α-Humulene	6753-98-6	9.54	10.46	13.93	2.28	2.30	-9.33	0	0.97	5.12
63	Δ3 - Carene	13466-78-9	10.06	11.04	15.92	1.22	1.55	-9.33	0	0.97	3.04
64	Ocimene	13877-91-3	9.60	10.65	14.70	2.00	2.05	-8.97	0	0.97	2.63
65	**Crotonaldehyde**	4170-30-3	10.44	14.29	18.05	-0.49	-0.73	-10.48	0	1.00	4.64

FOOTNOTES

[1] HOMO is the highest occupied molecular orbital (quantum-chemical descriptor).

[2] nBnz is the number of aromatic rings (constitutional descriptor).

[3] Me is the mean atomic Sanderson electronegativity (constitutional descriptor).

[4] DELS is the molecular electrotopological variation (charge distribution) (topological descriptor).

APPENDIX 5 STATISTICAL PARAMETERS

Standard Deviation Error in Prediction (*SDEP*)

$$SDEP = \sqrt{\frac{PRESS}{n}}$$

Where PRESS is the Predictive Error Sum of Squares and it is calculated from:

$$PRESS = \sum_i (y_i - \hat{y}_{i/i})^2$$

Standard Deviation Error in Calculation (*SDEC*)

$$SDEC = \sqrt{\frac{RSS}{n}}$$

Where RSS is the Residual Sum of Squares and it is calculated from:

$$RSS = \sum_i (y_i - \hat{y}_i)^2$$

Standard Error of Estimate (*s*)

$$s = \sqrt{\frac{RSS}{n - p'}}$$

Where y_i and $\hat{y}_i$ are respectively the observed and predicted values of the dependent variable, $\hat{y}_{i/i}$ is predicted values of the dependent variable when the observation is kept out, p' is the number of model variables plus one, and n the number of the objects used to calculate the model.

ANNEX 3
QSARS FOR MUTAGENICITY AND CARCINOGENICITY

Dr Romualdo Benigni
Laboratory of Comparative Toxicology and Ecotoxicology
Istituto Superiore di Sanita'
Viale Regina Elena 299
00161 Rome
Italy

TABLE OF CONTENTS

1. INTRODUCTION

153. Dr Romualdo Benigni (Istituto Superiore di Sanita', Rome) has retrospectively applied, under the terms of a JRC contract, Setubal Principles 1-6 to the following five QSAR models relating to classes of chemical mutagens and carcinogens:

1. QSAR 1: mutagenicity of aromatic and heterocyclic amines in the Salmonella typhimurium TA98

2. QSAR 2: mutagenicity of aromatic and heterocyclic amines in the Salmonella typhimurium TA100 strain

3. QSAR 3: mutagenicity of nitro-polycyclic aromatic hydrocarbons (Salmonella typhimurium TA100 strain)

4. QSAR 4: rodent skin carcinogenicity of aromatic hydrocarbons and heterocycles

5. QSAR 5: rodent carcinogenicity of aromatic amines.

154. These QSAR models are applications of the classical Hansch approach to individual classes of chemicals, supposedly acting through a common mechanism of action in well defined biological end-points.

155. For each individual QSAR, the results of this investigation are provided in a discursive form and in tabular form (Appendices 2-6). In the discursive part, a summary of the information provided by the authors is reported, followed by the results of our re-analysis of the original data, including the application of cross-validation procedures.

2. APPLICATION OF THE SETUBAL PRINCIPLES TO QSAR 1

2.1. Defined endpoint (Principle 1)

156. This QSAR is associated with a defined toxicity endpoint (genetic mutation), addressed by an officially recognised test method (Method B.13/14 Mutagenicity – Reverse Mutation test using bacteria – Annex V to *Directive 67/548/EEC*).

2.2. Defined algorithm (Principle 2)

157. The data and QSAR models are reported in reference 1. The mutagenic potency in TA98 strain (+ S9 activation system) was modelled by:

log TA98 = 1.08(±0.26) log *P* + 1.28(±0.64)HOMO – 0.73(±0.41)LUMO + 1.46(±0.56)*IL* + 7.20(±5.4) (4.6)

$n = 88$, $r = 0.898$ ($r^2 = 0.806$), $s = 0.860$, $F_{1,83} = 12.6$

158. The mutagenic potency (log TA98) is expressed as log (revertants/nmol). The AM1 molecular orbital energies are given in eV. HOMO is the energy of the highest occupied molecular orbital, LUMO is the energy of the lowest unoccupied molecular orbital, and *IL* is an indicator variable that assumes a value of 1 for compounds with three or more fused rings.

2.3. Mechanistic basis (Principle 3)

159. Overall, the principal factor affecting the relative mutagenicity of the aminoarenes was their hydrophobicity (logP). Mutagenicity increased with increasing HOMO values; this positive correlation seems reasonable because 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 by the authors, even though similar results were obtained by other authors on different sub-sets of chemicals (2).

2.4. Domain of applicability (Principle 4)

160. The applicability domain of the model is defined explicitly by the authors, relatively to the substructures studied. The chemical set spans a large range of basic structures (aniline, biphenyl, anthracene, phenanthrene, fluorene, pyrene, fluoranthene, chrysene, quinoline, carbazole, phenazine). The ranges of the chemical descriptors are not reported explicitly, and can be derived from the original paper, as follows:

logP: 1.12 - 4.20;

HOMO: -10.018 - -7.528;

LUMO: -1.691 - 0.722;

IL: 0 - 1.

2.5. Internal performance (Principle 5)

161. The goodness-of-fit reported by the authors is the correlation coefficient (see above). The data were re-analysed by us for this work. A Multiple Linear Regression (MLR) analysis reproduced the original QSAR equation, with

$r^2 = 0.808$; Adjusted $r^2 = 0.798$; $q^2 = 0.784$.

The Cross-Validated r^2 (q^2) is = 1 – (sum of squares of the predictive residuals / sum of squares of the mean-centered response data).

162. We performed cross-validation on the data. Three leave-many-out procedures were applied, leaving out: a) 10%; b) 25%; and c) 50% of the training set. Each procedure was applied ten times. The average q^2 (with Standard Deviation) were respectively:

10%: $q^2 = 0.798$ (0.072)

25%: $q^2 = 0.786$ (0.097)

50%: $q^2 = 0.710$ (0.196)

It appears that the model is robust when subjected to cross-validation.

2.6. External validation for predictivity (Principle 6)

163. The QSAR model has not been assessed for its predictivity of the activity of external compounds.

3. APPLICATION OF THE SETUBAL PRINCIPLES TO QSAR 2

3.1. Defined endpoint (Principle 1)

164. This QSAR is associated with a defined toxicity endpoint (genetic mutation), addressed by an officially recognised test method (Method B.13/14 Mutagenicity – Reverse Mutation test using bacteria – Annex V to *Directive 67/548/EEC*).

3.2. Defined algorithm (Principle 2)

165. The data and QSAR models are reported in reference (1). The mutagenic potency in TA100 strain (+ S9 activation system) was modelled by:

$$\log TA100 = 0.92(\pm 0.23) \log P + 1.17(\pm 0.83)HOMO - 1.18(\pm 0.44)LUMO + 7.35(\pm 6.9)$$

$n = 67$, $r = 0.877$ ($r^2 = 0.769$), $s = 0.708$, $F_{1,65} = 99.23$

The mutagenic potency (log TA100) is expressed as log (revertants/nmol). The AM1 molecular orbital energies are given in eV.

3.3. Mechanistic basis (Principle 3)

166. The model is very similar to that derived for the TA98 mutagenicity. TA100 QSAR lacked the *IL* term present in the TA98 model. It was hypothesized that larger amines are more capable of inducing frameshift mutations (TA98 is specific for frame-shift mutations, whereas TA100 is specific for base-pair-substitution mutations), and that this effect is not accounted for by the increase of log *P* for increasing sizes of the molecules.

3.4. Domain of applicability (Principle 4)

167. The applicability domain of the model is defined explicitly by the authors, relatively to the substructures studied. The chemical set spans a large range of basic structures (aniline, biphenyl, anthracene, phenanthrene, fluorene, pyrene, fluoranthene, chrysene, quinoline, carbazole, phenazine). The ranges of the chemical descriptors are not reported explicitly, and can be derived from the original paper, as follows:

logP: 1.16 - 4.98

LUMO: -1.334 - 0.702

HOMO: -8.695 - -7.528

3.5. Internal performance (Principle 5)

168. The goodness-of-fit reported by the authors is the correlation coefficient (see above). The data have also been re-analysed by us. A Multiple Linear Regression (MLR) analysis reproduced the original QSAR equation, with

r^2 = 0.771; Adjusted r^2 = 0.761; q^2 = 0.740.

169. We performed cross-validation on the data. Three leave-many-out procedures were applied, leaving out: a) 10%; b) 25%; and c) 50% of the training set. Each procedure was applied ten times. The average q^2 (with Standard Deviation) were respectively:

10%: q^2 = 0.669 (0.172)

25%: q^2 = 0.675 (0.142)

50%: q^2 = 0.745 (0.068)

It appears that the model is robust when subjected to cross-validation.

3.6. External validation for predictivity (Principle 6)

170. The QSAR model has not been assessed for its predictivity of the activity of external compounds.

4. APPLICATION OF THE SETUBAL PRINCIPLES TO QSAR 3

4.1. Defined endpoint (Principle 1)

171. This QSAR is associated with a defined toxicity endpoint (genetic mutation), addressed by an officially recognised test method (Method B.13/14 Mutagenicity – Reverse Mutation test using bacteria – Annex V to *Directive 67/548/EEC*).

4.2. Defined algorithm (Principle 2)

172. The data and QSAR models were derived from the C-QSAR database and program. The dataset name is BIO_5492. The access to C-QSAR was a generous gift of Corwin Hansch; for a general presentation of C-QSAR, see reference (3). The original data are reported in Appendix 1. The mutagenic potency in TA100 strain (without S9 activation system) was modelled by:

log TA100 = -0.923(±0.362) LUMO + 1.282(±0.142)MR – 8.244(±1.150)

n = 41, r = 0.949 (r^2 = 0.901), q^2 = 0.885, s = 0.558

The mutagenic potency (log TA100) is expressed as log (revertants/nmol). The AM1 molecular orbital energies are given in eV. MR is Molar Refractivity.

4.3. Mechanistic basis (Principle 3)

173. The QSAR found is clearly related to the mechanisms of action of the aromatic nitro compounds, in particular to the chemical reduction involved in the metabolic activation step.

4.4. Domain of applicability (Principle 4)

174. Regarding the applicability domain, a list of basic substructures is not provided explicitly in the database. Both the basic substructures and the range of descriptor values can de derived from the data relative to the individual compounds (see Appendix 1). The basic structures include: benzene, pyrene, fluorene, fluorenone, fluoranthene, naphthalene, acenaphthene, toluene, carbazole, quinoline. The ranges of the chemical descriptors are as follows:

LUMO: -2.84 0 - -1.220

MR: 3.76 - 8.08

4.5. Internal performance (Principle 5)

175. The goodness-of-fit reported by the authors is the correlation coefficient (see above). Our re-analysis (MLR) of the original data reproduced the reported QSAR equation and statistical parameters.

r^2 = 0.900; Adjusted r^2 = 0.894; q^2 = 0.885.

176. We performed cross-validation on the data. Three leave-many-out procedures were applied, leaving out: a) 10%; b) 25%; and c) 50% of the training set. Each procedure was applied ten times. The average q^2 (with Standard Deviation) were respectively:

10%: q^2 = 0.804 (0.176)

25%: q^2 = 0.889 (0.072)

50%: q^2 = 0.860 (0.061)

It appears that the model is robust when subjected to cross-validation.

4.6. External validation for predictivity (Principle 6)

177. The QSAR model has not been assessed for its predictivity of the activity of external compounds.

5. APPLICATION OF THE SETUBAL PRINCIPLES TO QSAR 4

5.1. Defined endpoint (Principle 1)

178. This QSAR is associated with a defined toxicity endpoint (genetic mutation), addressed by an officially recognised test method (Method B.13/14 Mutagenicity – Reverse Mutation test using bacteria – Annex V to *Directive 67/548/EEC*).

5.2. Defined algorithm (Principle 2)

179. The data and QSAR models are in reference 4. The carcinogenicity in mice was modelled by:

$$\log Iball = 0.55(\pm 0.09) \log P - 1.17(\pm 0.14) \log (\exists\, 10^{\log P} + 1) + 0.39(\pm 0.11)\mathrm{LK} + 0.47(\pm 0.26)\mathrm{HOMO} + 1.93(\pm 2.4)$$

$n = 161$, $r = 0.845$ ($r^2 = 0.714$), $s = 0.350$, $\log \exists = -6.81$, $F_{1,155} = 12.8$

where:

Iball index = (tumor incidence) (100%) / (mean latent period in days) with

tumor incidence = (number of animal with tumors) / (number of animals alive when the first tumor appears);

LK is an indicator variable assigned a value 1 for all chemicals where a substituent is attached to a L or K region.

The electronic parameters were calculated with the AM1 procedure.

5.3. Mechanistic basis (Principle 3)

180. The QSAR model is in agreement with the theories regarding K-region (9,10 bond in phenanthrene and analogs) and L-region (region between the C14 and C17 positions of phenanthrene) activation as being responsible for carcinogenicity of these compounds (Fig. 1). A positive coefficient of LK means that a substitution in an L or K region inhibits metabolism at these points and then leads to increased potency of these congeners, other factors being equal.

5.4. Domain of applicability (Principle 4)

181. The basic substructure is in Fig. 1; details on the structural variations are in the original paper. The ranges of chemical descriptors values are not reported explicitly in the paper, and are as follows:

logP: 3.30 - 11.01

LK: 0 - 1

HOMO: -9.13 - -7.54

5.5. Internal performance (Principle 5)

182. The goodness-of-fit reported by the authors is the correlation coefficient (see above). Our re-analysis (MLR) of the original data reproduced the reported QSAR equation and statistical parameters:

r^2 = 0.713; Adjusted r^2 = 0.705; q^2 = 0.689.

183. We performed cross-validation on the data. Three leave-many-out procedures were applied, leaving out: a) 10%; b) 25%; and c) 50% of the training set. Each procedure was applied ten times. The average q^2 (with Standard Deviation) were respectively:

10%: q^2 = 0.682 (0.083)

25%: q^2 = 0.689 (0.071)

50%: q^2 = 0.684 (0.056)

It appears that the model is robust when subjected to cross-validation.

5.6. External validation for predictivity (Principle 6)

184. The QSAR model has not been assessed for its predictivity of the activity of external compounds.

6. APPLICATION OF THE SETUBAL PRINCIPLES TO QSAR 5

6.1. Defined endpoint (Principle 1)

185. This QSAR is associated with a defined toxicity endpoint (carcinogenicity), addressed by an officially recognised test method (Method B.32 Carcinogenicity test – Annex V to Directive 67/548/EEC).

6.2. Defined algorithm (Principle 2)

186. The data and QSAR models are in reference (5). The rodent carcinogenicity data (overall yes/no score from four experimental groups: rat, mouse, male, female) were analysed with Linear Discriminant Analysis. The carcinogenicity was modelled by:

$w = -2.86\ L(R) + 2.65\ B_5(R) - 1.16\ HOMO + 1.76\ LUMO + 0.40\ MR_3 + 0.58\ MR_5 + 0.54\ MR_6 - 1.55\ I(An) + 0.74\ I(NO_2) - 0.55\ I(BiBr)$

$w_{(mean,Class1)} = -1.56 \quad N1 = 13$

$w_{(mean,Class2)} = 0.38 \quad N2 = 53$

where N1 = number of non-carcinogens (Class 1) and N2 = number of carcinogens (Class 2). L(R) (length) and B_5(R) (maximal width) are Sterimol parameters. HOMO and LUMO were calculated by AM1. MR_3, MR_5, MR_6 are the MR contributions of substituents in position 3, 5, and 6 to the amino group. I(An), I(NO_2), and I(BiBr) are indicator variables that take value = 1 for anilines, for the presence of a NO_2 group, and for biphenyls with a bridge between the phenyl rings, respectively.

187. The equation correctly reclassified 87.9% of the compounds (Class1, 84.6%; Class2, 88.7%).

6.3. Mechanistic basis (Principle 3)

188. The mechanistic meaning of the parameters is discussed in detail in the original paper, mainly in terms of enhancing / inhibiting the metabolic activation of the amines (rate limiting step).

6.4. Domain of applicability (Principle 4)

189. Regarding the applicability domain of the model, the original paper lists the basic substructures to which it applies (aniline, biphenyl, naphthalene, fluorene). The ranges of chemical descriptors values are not reported explicitly, and are as follows:

L(R): 2.06 - 5.97

B5(R): 1.00 - 4.04

HOMO: -9.544 - -7.989

LUMO: -1.594 - 0.438

MR3: 0.10 - 0.80

MR5: 0.09 - 1.49

MR6: 0.09 - 0.60

I(An): 0 - 1

I(NO2): 0 - 1

I(BiBr) 0 - 1

6.5. Internal performance (Principle 5)

190. The goodness-of-fit reported by the authors consists of the accuracy, sensitivity and specificity parameters (see above). Our re-analysis of the original data reproduced the reported QSAR equation and statistical parameters.

191. We performed cross-validation on the data. Three leave-many-out procedures were applied, leaving out: a) 10%; b) 25%; and c) 50% of the training set. In addition, each procedure was applied in two different ways, by generating test sets: 1) with the same proportion Class1/Class2 present in the whole sample of chemicals; 2) without the above constraint. Each procedure was applied ten times.

192. To permit an immediate appreciation of, and easy comparison between the results of the various analyses, these are displayed in: a) tabular form (Class1 and Class2 correct classification rates) (Table 1); and b) graphical form, through a Receiver Operating Characteristics (ROC) Graph (Fig. 2). A ROC graph has the advantage of comparing simultaneously the different aspects of the performance of several systems: the axes on a ROC graph display independently the information relative to the prediction of positive and negative chemicals, and all the systems can be plotted together in the same graph. Moreover, according to the ROC curve theory the diagonal line in the plot represents random responses, whereas the top left corner is obviously the ideal performance. The most finely tuned systems are those in the left upper triangle, as close as possible to the corner. Thus, a ROC graph permits to gather visually a large amount of information (see theory in reference 6).

193. From Table 1 and Fig. 2, it appears that the model is robust when subjected to cross-validation.

6.6. External validation for predictivity (Principle 6)

194. The QSAR model has not been assessed for its predictivity of the activity of external compounds.

7. CONCLUSIONS ON THE SETUBAL PRINCIPLES AND THEIR APPLICABILITY

195. All five QSAR models considered in this work are associated with defined endpoints of regulatory importance, take the form of an unambiguous algorithm, and agree with and support known mechanisms of action. The original papers do not report explicitly the domains of applicability relatively to the ranges of values of the chemical descriptors, whereas the lists of substructures are usually given. The reported goodness-of-fit measures do not include cross-validation.

196. We applied cross-validation procedures for the purposes of this work: all five models appeared to be robust when subjected to cross-validation. The QSAR models have not been assessed for their predictivity of the activity of external compounds; however, they could be regarded as sufficiently well developed to undergo an independent, external validation process. This further check could be particularly useful especially for chemical classes (e.g. aromatic amines) of remarkable environmental and industrial importance. The Setubal principles appear to be a reliable basis for assessing the scientific / regulatory value of QSAR models.

8. REFERENCES

Benigni, R., Giuliani, A., Gruska, A. & Franke, R. (2003). QSARs for the Mutagenicity and Carcinogenicity of the Aromatic Amines. In *Quantitative Structure-Activivity Relationship (QSAR) Models of Mutagens and Carcinogens.* (Benigni, R., Ed). Chapter 4. CRC Press: Boca Raton.

Debnath, A. K., Debnath, G., Shusterman, A. J. & Hansch, C. (1992). A QSAR Investigation of the Role of Hydrophobicity in Regulating Mutagenicity in the Ames Test: 1. Mutagenicity of Aromatic and Heteroaromatic Amines in Salmonella Typhimurium TA98 and TA100. *Environ. Mol. Mutagen.* **19**, 37-52.

Franke, R., Gruska, A., Giuliani, A. & Benigni, R. (2001). Prediction of Rodent Carcinogenicity of Aromatic Amines: a Quantitative Structure-Activity Relationships Model. *Carcinogenesis* **22**, 1561-1571.

Hansch, C., Hoekman, D., Leo, A., Weininger, D., Selassie, C. D. (2002). Chem-Bioinformatics: Comparative QSAR at the Interface Between Chemistry and Biology. *Chem. Revs. 102*, 783-812.

Provost, F. & Fawcett, T. (2001). Robust Classification for Imprecise Environment. *Machine Learn. J.* **42**, 5-11.

Zhang, L., Sannes, K., Shusterman, A. J. & Hansch, C. (1992). The Structure-Activity Relationships of Skin Carcinogenicity of Aromatic Hydrocarbons and Heterocycles. *Chem. Biol. Interact.* **81**, 149-180.

Table 1 Percentage of correct classification for Class 1 (non-carcinogens) and Class 2 (carcinogens) chemicals under different procedures

	Class 1	**Class 2**
Whole set	84.6	88.7
Leave-one-out	76.9	79.3
10% [a]	70.0 (42.)	84.0 (14.7)
25% [a]	79.2 (18.9)	80.4 (12.2)
50% [a]	66.8 (26.7)	74.5 (17.3)
10% [b]	70.0 (25.8)	90.0 (17.5)
25% [b]	66.7 (22.2)	82.3 (12.6)
50% [b]	66.7 (13.6)	81.2 (6.1)

FOOTNOTES

Standard Deviations are given within brackets.

a : no constraints regarding the rate Class1 / Class2

b : rate Class1 / Class2 as in the whole data set.

Figure 1 L, K, and bay regions of polycyclic hydrocarbons.

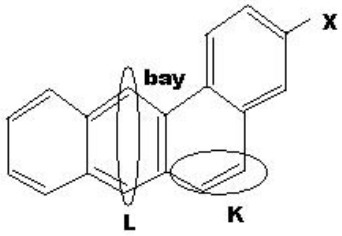
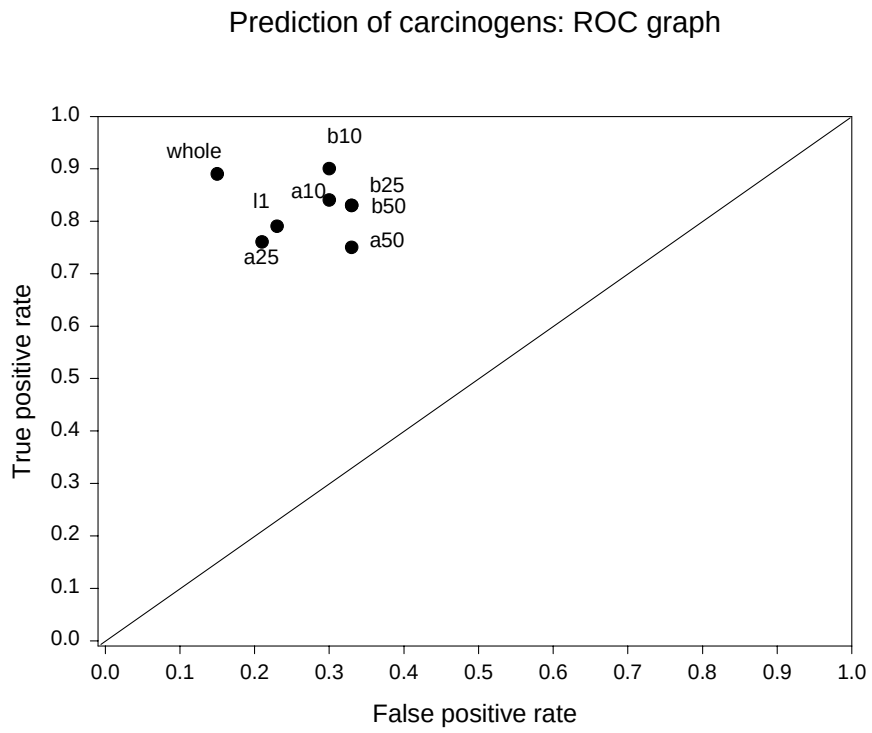


Figure 2 Performance of the QSAR model for the rodent carcinogenicity of aromatic amines, *per se* and under different cross-validation procedures: ROC Graph



True positive rate: percentage of carcinogens correctly classified

False positive rate: percentage of non-carcinogens erroneously classified as carcinogens;

Whole: whole data set

l1: leave-one-out

a10 – a50: cross-validation leaving out 10, 25, 50% of chemicals, without constraints on the Class1 /Class2 rate

b10 – b50: cross-validation leaving out 10, 25, 50% of chemicals, with the Class1/Class2 rate as in the whole data set

9. APPENDICES

APPENDIX 1 NITRO AROMATIC HYDROCARBONS: DESCRIPTORS AND ACTIVITY IN SALMONELLA TYPHIMURIUM TA100 STRAIN

		YPRED	DEV	LOGA	ELUMO	CLOGP	CMR
1	* 1,3,6-Trinitropyrene	5.56	-1.69	3.87	-2.89	4.18	8.69
2	* 2,4,7-Trintro-9-Fluorenone	3.9	-1.63	2.27	-2.83	2.5	7.43
3	1,3-Dinitropyrene	4.25	0.38	4.63	-2.32	4.44	8.08
4	1,6-Dinitropyrene	4.24	-0.15	4.09	-2.3	4.44	8.08
5	1,8-Dinitropyrene	4.23	0.51	4.74	-2.29	4.44	8.08
6	2,7-Dinitro-9-Fluorenone	2.63	0.06	2.69	-2.31	2.74	6.82
7	1-Nitrofluoranthene	2.79	0.21	3	-1.58	4.69	7.47
8	2-Nitroanthracene	1.83	1.22	3.05	-1.64	4.23	6.68
9	2,7-Di-Nitrofluorene	2.51	-1.24	1.27	-2.23	3.56	6.78
10	3-Nitrofluoranthene	2.84	0.47	3.31	-1.63	4.69	7.47
11	8-Nirtofluoranthene	2.8	-0.2	2.6	-1.59	4.69	7.47
12	1-Nitropyrene	2.87	-0.7	2.17	-1.67	4.69	7.47
13	7-Nitrofluoranthene	2.61	-0.52	2.09	-1.39	4.69	7.47
14	2-Nitronaphthalene	-0.45	0.82	0.37	-1.51	3.06	4.99
15	2-Nitrofluorene	1.08	0	1.08	-1.53	3.82	6.17
16	2-Nitrophenanthrene	1.57	0.22	1.79	-1.36	4.23	6.68
17	1-Nitronaphthalene	-0.68	0.96	0.28	-1.27	3.06	4.99
18	5-Nitroacenaphthene	0.24	0.73	0.97	-1.22	3.51	5.74
19	1,3-Di-Nitrobenzene	-1.04	0.53	-0.51	-2.07	1.62	4.13
20	1,3,5-Tri-Nitrobenzene	0.08	0.64	0.72	-2.73	1.37	4.52
21	4-Nitrotoluene	-2.26	0.16	-2.1	-1.25	2.38	3.76
22	2,3-Di-Nitrotoluene	-0.93	-0.33	-1.26	-1.84	2.05	4.38
23	2,4-Di-Nitrotoluene	-0.79	-0.5	-1.29	-2	2.05	4.38
24	2,5-Di-Nitrotoluene	-0.57	-0.06	-0.63	-2.23	2.05	4.38
25	2,6-Di-Nitrotoluene	-0.99	-0.35	-1.34	-1.79	1.97	4.38
26	3,4-Di-Nitrotoluene	-0.81	-0.49	-1.3	-1.97	2.13	4.38
27	3,5-Di-Nitrotoluene	-0.75	0.03	-0.72	-2.04	2.13	4.38
28	2,3,4-Tri-Nitrotoluene	0.46	-0.38	0.08	-2.5	1.79	4.99
29	2,3,5-Tri-Nitrotoluene	0.77	-0.31	0.46	-2.84	1.79	4.99
30	2,3,6-Tri-Nitrotoluene	0.67	-0.12	0.55	-2.73	1.71	4.99
31	2,4,5-Tri-Nitrotoluene	0.77	0.35	1.12	-2.84	1.79	4.99
32	2,4,6-Tri-Nitrotoluene	0.61	-0.45	0.16	-2.67	1.71	4.99
33	3,4,5-Tri-Nitrotoluene	0.52	0.49	1.01	-2.56	1.87	4.99
34	1-Me-2-Nitronaphthalene	-0.05	0.13	0.08	-1.3	3.48	5.45
35	3-Me-2-Nitronaphthalene	-0.08	-0.62	-0.7	-1.27	3.48	5.45

36	1,3-Di-Nitronaphthalene	0.8	0.06	0.86	-2.03	2.8	5.6
37	1,5-Di-Nitronaphthalene	0.74	0.17	0.91	-1.95	2.8	5.6
38	1,8-Di-Nitronaphthalene	0.68	0.44	1.12	-1.89	2.8	5.6
39	* 1,3,6,8-Tetra-Nitronaphthalene	3.45	-3.97	-0.52	-3.19	2.29	6.82
40	* 2,4,5,7-Tetra-Nitro-9-Fluorenone	5.08	-2.62	2.46	-3.26	2.25	8.04
41	2-Nitropyrene	2.71	0.16	2.87	-1.49	4.69	7.47
42	* 1,3,6,8-Tetranitropyrene	6.85	-3.67	3.18	-3.44	3.92	9.3
43	5-Nitroquinolene	-0.79	0.09	-0.7	-1.44	1.91	4.78
44	6-Nitroquinolene	-0.6	-0.45	-1.05	-1.64	1.91	4.78
45	2-Nitrocarbazole	0.71	-1.01	-0.3	-1.35	2.93	6.02
46	* 3-Nitrocarbazole	0.48	-1.48	-1	-1.09	2.93	6.02
47	4-Nitrocarbazole	0.67	-0.97	-0.3	-1.3	2.93	6.02
48	* 9-Nitroanthracene	1.86	-1.6	0.26	-1.67	4.23	6.68
49	* 1-Nitrobenzo(E)Pyrene	5.06	-3.41	1.65	-1.7	5.87	9.16
50	* 6-Nitrobenzo(A)Pyrene	5.39	-4.69	0.7	-2.05	5.87	9.16
51	* 6-Nitrochrysene	3.98	-1.77	2.21	-1.62	5.41	8.36

FOOTNOTES

The information reported derives from the BIO_5492 of the C-QSAR database and program (see details in the text).

YPRED: activity re-calculated by the QSAR model;

DEV: difference between the experimental and the re-calculated activity;

LOGA: mutagenic potency (logTA100 in the text);

ELUMO: energy of LUMO (LUMO in the text);

CLOGP: logP as calculated by the C-QSAR program;

CMR: MR as calculated by the C-QSAR program;

* : compounds not used for the generation of the QSAR model.

APPENDIX 2 SUMMARY REPORT OF THE APPLICATION OF THE SETUBAL PRINCIPLES TO QSAR 1 AND QSAR 2

PRINCIPLE	CONSIDERATIONS	YES / NO
1) Defined endpoint	**1.1)** Does the model have a clearly defined scientific purpose?	
	(i.e. does it make predictions of a clearly defined physicochemical, biological or environmental effect?)	YES
	1.2) Does the model have the potential to address (or partially address) a clearly defined regulatory need?	YES
	(i.e. does it make predictions of a specific endpoint associated with a specific test method or test guideline)?	
	1.3) Is information given about important experimental conditions that affect the measurement and therefore the prediction	YES
	(e.g. sex, species, temperature, exposure period)	
	1.4) Are the units of measurement of the endpoint given?	YES
2) Defined algorithm	**2.1)** In the case of a SAR, is there an explicit description of the substructure, including an explicit identification of its substituents?	
	2.2) In the case of a QSAR, is the equation explicitly defined, including definitions of all descriptors used[3]?	YES
3) Mechanistic basis	**3.1)** In the case of a SAR, is there a description of the molecular events that underlie the reactivity of the molecule?	
	(e.g. a description of how substructural features could act as nucleophiles or electrophiles, or form part or all of a receptor-binding region)	
	3.2) In the case of a QSAR, do the descriptors have a physicochemical interpretation that is consistent with a known mechanism (of biological action)?	YES
	3.3) Are any literature references cited in support of the purported mechanistic basis of the (Q)SAR?	YES

4) Domain of applicability	**4.1)** In the case of a SAR, is the substructure associated with any inclusion and/or exclusion rules on its applicability to groups of chemicals?	
	4.2) In the case of a SAR, is the substructure associated with rules regarding the modulatory effects of the substructure's molecular environment?	
	4.3) In the case of a QSAR, are the descriptor and response variables associated with inclusion and/or exclusion rules that define the variable ranges for which the QSAR is applicable (i.e. makes reliable estimates)?	The substructures and parameters values are reported.
5) Internal performance	**5.1)** Are full details of the training set given, including details of chemical names, structural formulae, CAS numbers (if available), and data for all descriptor and response variables?	YES (except CAS)
	5.2) If the data used to the develop the model were based upon the processing of raw data (e.g. the averaging of replicate values):	
	a) is there an adequate description of the data processing?	
	b) are the raw data provided?	
	5.3) Is there a specification of the statistical method(s) used to develop the QSAR (including details of any software packages used)?	YES
	5.4) Is the QSAR associated with basic statistics for its goodness-of-fit to the training set?	YES
	(e.g. r^2 values and standard error of the estimate in the case of regression models)	
	5.5)	NO
	a) Is the QSAR associated with any statistics based on cross-validation or resampling?	
	b) If yes, is the number or samples used indicated?	

6) Predictivity (External validation)	**6.1)** Does application of the appropriate statistical method(s) to the training set result in the same (Q)SAR model? **6.2)** Is there any information to indicate that the (Q)SAR has been validated previously, using a test set that is independent of the training set? **6.3)** If an external validation has been performed, is the following information available: a) the number of test structures? b) the identities of the test structures? c) the approach for selecting the test structures? d) the statistical analysis of the predictive performance of the model? (e.g. including sensitivity, specificity, and positive and negative predictivities for classification models) e) a comparison of the predictive performance of the model against previously-defined quantitative performance criteria?

APPENDIX 3 SUMMARY REPORT OF THE APPLICATION OF THE SETUBAL PRINCIPLES TO QSAR 3

PRINCIPLE	CONSIDERATIONS	YES / NO
1) Defined endpoint	**1.1)** Does the model have a clearly defined scientific purpose?	
	(i.e. does it make predictions of a clearly defined physicochemical, biological or environmental effect?)	YES
	1.2) Does the model have the potential to address (or partially address) a clearly defined regulatory need?	YES
	(i.e. does it make predictions of a specific endpoint associated with a specific test method or test guideline)?	
	1.3) Is information given about important experimental conditions that affect the measurement and therefore the prediction	Reference to original papers is given
	(e.g. sex, species, temperature, exposure period)	
	1.4) Are the units of measurement of the endpoint given?	YES
2) Defined algorithm	**2.1)** In the case of a SAR, is there an explicit description of the substructure, including an explicit identification of its substituents?	
	2.2) In the case of a QSAR, is the equation explicitly defined, including definitions of all descriptors used[3]?	YES
3) Mechanistic basis	**3.1)** In the case of a SAR, is there a description of the molecular events that underlie the reactivity of the molecule?	
	(e.g. a description of how substructural features could act as nucleophiles or electrophiles, or form part or all of a receptor-binding region)	
	3.2) In the case of a QSAR, do the descriptors have a physicochemical interpretation that is consistent with a known mechanism (of biological action)?	YES
	3.3) Are any literature references cited in support of the purported mechanistic basis of the (Q)SAR?	YES

4) Domain of applicability	**4.1)** In the case of a SAR, is the substructure associated with any inclusion and/or exclusion rules on its applicability to groups of chemicals?	
	4.2) In the case of a SAR, is the substructure associated with rules regarding the modulatory effects of the substructure's molecular environment?	
	4.3) In the case of a QSAR, are the descriptor and response variables associated with inclusion and/or exclusion rules that define the variable ranges for which the QSAR is applicable (i.e. makes reliable estimates)?	The substructures and parameters values are reported.
5) Internal performance	**5.1)** Are full details of the training set given, including details of chemical names, structural formulae, CAS numbers (if available), and data for all descriptor and response variables?	YES (except CAS)
	5.2) If the data used to the develop the model were based upon the processing of raw data (e.g. the averaging of replicate values):	
	a) is there an adequate description of the data processing?	
	b) are the raw data provided?	
	5.3) Is there a specification of the statistical method(s) used to develop the QSAR (including details of any software packages used)?	They are given in the general documentation for C-QSAR
	5.4) Is the QSAR associated with basic statistics for its goodness-of-fit to the training set?	YES
	(e.g. r^2 values and standard error of the estimate in the case of regression models)	
	5.5)	NO
	a) Is the QSAR associated with any statistics based on cross-validation or resampling?	
	b) If yes, is the number or samples used indicated?	

6) Predictivity (External validation)	**6.1)** Does application of the appropriate statistical method(s) to the training set result in the same (Q)SAR model? **6.2)** Is there any information to indicate that the (Q)SAR has been validated previously, using a test set that is independent of the training set? **6.3)** If an external validation has been performed, is the following information available: a) the number of test structures? b) the identities of the test structures? c) the approach for selecting the test structures? d) the statistical analysis of the predictive performance of the model? (e.g. including sensitivity, specificity, and positive and negative predictivities for classification models) e) a comparison of the predictive performance of the model against previously-defined quantitative performance criteria?

APPENDIX 4 SUMMARY REPORT OF THE APPLICATION OF THE SETUBAL PRINCIPLES TO QSAR 4

PRINCIPLE	CONSIDERATIONS	YES / NO
1) Defined endpoint	**1.1)** Does the model have a clearly defined scientific purpose?	YES
	(i.e. does it make predictions of a clearly defined physicochemical, biological or environmental effect?)	
	1.2) Does the model have the potential to address (or partially address) a clearly defined regulatory need?	YES
	(i.e. does it make predictions of a specific endpoint associated with a specific test method or test guideline)?	
	1.3) Is information given about important experimental conditions that affect the measurement and therefore the prediction	NO
	(e.g. sex, species, temperature, exposure period)	
	1.4) Are the units of measurement of the endpoint given?	YES
2) Defined algorithm	**2.1)** In the case of a SAR, is there an explicit description of the substructure, including an explicit identification of its substituents?	
	2.2) In the case of a QSAR, is the equation explicitly defined, including definitions of all descriptors used[3]?	YES
3) Mechanistic basis	**3.1)** In the case of a SAR, is there a description of the molecular events that underlie the reactivity of the molecule?	
	(e.g. a description of how substructural features could act as nucleophiles or electrophiles, or form part or all of a receptor-binding region)	
	3.2) In the case of a QSAR, do the descriptors have a physicochemical interpretation that is consistent with a known mechanism (of biological action)?	YES
	3.3) Are any literature references cited in support of the purported mechanistic basis of the (Q)SAR?	YES

4) Domain of applicability	**4.1)** In the case of a SAR, is the substructure associated with any inclusion and/or exclusion rules on its applicability to groups of chemicals?	
	4.2) In the case of a SAR, is the substructure associated with rules regarding the modulatory effects of the substructure's molecular environment?	
	4.3) In the case of a QSAR, are the descriptor and response variables associated with inclusion and/or exclusion rules that define the variable ranges for which the QSAR is applicable (i.e. makes reliable estimates)?	The substructures and parameters values are reported
5) Internal performance	**5.1)** Are full details of the training set given, including details of chemical names, structural formulae, CAS numbers (if available), and data for all descriptor and response variables?	YES (except CAS)
	5.2) If the data used to the develop the model were based upon the processing of raw data (e.g. the averaging of replicate values):	
	a) is there an adequate description of the data processing?	
	b) are the raw data provided?	
	5.3) Is there a specification of the statistical method(s) used to develop the QSAR (including details of any software packages used)?	YES
	5.4) Is the QSAR associated with basic statistics for its goodness-of-fit to the training set?	YES
	(e.g. r^2 values and standard error of the estimate in the case of regression models)	
	5.5)	NO
	a) Is the QSAR associated with any statistics based on cross-validation or resampling?	
	b) If yes, is the number or samples used indicated?	

6) Predictivity (External validation)	**6.1)** Does application of the appropriate statistical method(s) to the training set result in the same (Q)SAR model? **6.2)** Is there any information to indicate that the (Q)SAR has been validated previously, using a test set that is independent of the training set? **6.3)** If an external validation has been performed, is the following information available: a) the number of test structures? b) the identities of the test structures? c) the approach for selecting the test structures? d) the statistical analysis of the predictive performance of the model? (e.g. including sensitivity, specificity, and positive and negative predictivities for classification models) e) a comparison of the predictive performance of the model against previously-defined quantitative performance criteria?

APPENDIX 5 SUMMARY REPORT OF THE APPLICATION OF THE SETUBAL PRINCIPLES TO QSAR 5

PRINCIPLE	CONSIDERATIONS	YES / NO
1) Defined endpoint	**1.1)** Does the model have a clearly defined scientific purpose?	YES
	(i.e does it make predictions of a clearly defined physicochemical, biological or environmental effect?)	
	1.2) Does the model have the potential to address (or partially address) a clearly defined regulatory need?	YES
	(i.e. does it make predictions of a specific endpoint associated with a specific test method or test guideline)?	
	1.3) Is information given about important experimental conditions that affect the measurement and therefore the prediction	YES
	(e.g. sex, species, temperature, exposure period)	
	1.4) Are the units of measurement of the endpoint given?	YES
2) Defined algorithm	**2.1)** In the case of a SAR, is there an explicit description of the substructure, including an explicit identification of its substituents?	
	2.2) In the case of a QSAR, is the equation explicitly defined, including definitions of all descriptors used[3]?	YES
3) Mechanistic basis	**3.1)** In the case of a SAR, is there a description of the molecular events that underlie the reactivity of the molecule?	
	(e.g. a description of how substructural features could act as nucleophiles or electrophiles, or form part or all of a receptor-binding region)	
	3.2) In the case of a QSAR, do the descriptors have a physicochemical interpretation that is consistent with a known mechanism (of biological action)?	YES
	3.3) Are any literature references cited in support of the purported mechanistic basis of the (Q)SAR?	YES

4) Domain of applicability	**4.1)** In the case of a SAR, is the substructure associated with any inclusion and/or exclusion rules on its applicability to groups of chemicals?	
	4.2) In the case of a SAR, is the substructure associated with rules regarding the modulatory effects of the substructure's molecular environment?	
	4.3) In the case of a QSAR, are the descriptor and response variables associated with inclusion and/or exclusion rules that define the variable ranges for which the QSAR is applicable (i.e. makes reliable estimates)?	The substructures and parameters values are reported
5) Internal performance	**5.1)** Are full details of the training set given, including details of chemical names, structural formulae, CAS numbers (if available), and data for all descriptor and response variables?	YES (except CAS)
	5.2) If the data used to the develop the model were based upon the processing of raw data (e.g. the averaging of replicate values):	
	a) is there an adequate description of the data processing?	
	b) are the raw data provided?	
	5.3) Is there a specification of the statistical method(s) used to develop the QSAR (including details of any software packages used)?	YES
	5.4) Is the QSAR associated with basic statistics for its goodness-of-fit to the training set?	YES
	(e.g. r^2 values and standard error of the estimate in the case of regression models)	
	5.5)	Leave-one-out
	a) Is the QSAR associated with any statistics based on cross-validation or resampling?	
	b) If yes, is the number or samples used indicated?	

6) Predictivity (External validation)	**6.1)** Does application of the appropriate statistical method(s) to the training set result in the same (Q)SAR model? **6.2)** Is there any information to indicate that the (Q)SAR has been validated previously, using a test set that is independent of the training set? **6.3)** If an external validation has been performed, is the following information available: a) the number of test structures? b) the identities of the test structures? c) the approach for selecting the test structures? d) the statistical analysis of the predictive performance of the model? (e.g. including sensitivity, specificity, and positive and negative predictivities for classification models) e) a comparison of the predictive performance of the model against previously-defined quantitative performance criteria?

ANNEX 4
A "GLOBAL" MULTI-CASE MODEL FOR *IN VITRO* CHROMOSOMAL ABERRATIONS IN MAMMALIAN CELLS

Jay Niemelå and Eva Wedebye
Danish Environmental Protection Agency
Ministry of Environment,
Chemicals Division
Strandgade 29
1401 Copenhagen K
Denmark

TABLE OF CONTENTS

1. INTRODUCTION

197. Drs Jay Niemela and Eva Wedebye (Danish EPA, Copenhagen, Denmark) have developed a model for the prediction of chromosomal aberrations in mammalian cells, by using MULTICASE ©, an artificial intelligence QSAR system capable of making predictions for large numbers of diverse chemical structures.

198. In Section 2, the authors report the development of the model, and provides an assessment of its performance of this model, as a way of gaining experience in the task of identifying models of possible interest in future regulatory applications.

199. In Section 3, the applicability of the Setubal principles to the MULTICASE model is discussed. A summary report of the application of the principles is provided in Appendix 1.

200. The authors express their thanks to Dr Motoi Ishidate, Jr and Dr Toshio Sofuni for creating, and providing access to, the data set used to develop the MULTICASE model.

2. DEVELOPMENT AND ASSESSMENT OF THE MULTI-CASE MODEL

2.1. Description of the *in vitro* chromosome aberration test

201. The test system and its purpose is described in OECD Guideline for the Testing of chemicals, No. 473 (1).

> "The purpose of the in vitro chromosome aberration test is to identify agents that cause structural chromosome aberrations in cultured mammalian cells. Structural aberrations may be of two types, chromosome or chromatid. With the majority of chemical mutagens, induced aberrations are of the chromatid type, but chromosome-type aberrations also occur. An increase in polyploidy may indicate that a chemical has the potential to induce numerical aberrations. However, this guideline is not designed to measure numerical aberrations and is not routinely used for that purpose. Chromosome mutations and related events are the cause of many human genetic diseases and there is substantial evidence that chromosome mutations and related events causing alterations in oncogenes and tumour suppressor genes of somatic cells are involved in cancer induction in humans and experimental animals."

2.2. Description of the MULTICASE platform

202. 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. All fragments are assigned a MULTICASE activity score according to the activity of the parent structure. If the parent compound is "inactive it is assigned a score of 10, while fragments from active parents are given a score of 45. 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.

203. First, major biophore classes are defined. Then within each biophore class fragments that modify the activity of the biophores are identified. MULTICASE then computes various physicochemical properties for the training set chemicals (such as LogP octanol-water, charge densities, molecular orbital energies and two-dimensional distance descriptors).

204. When making a prediction, MULTICASE notes the presence of biophores or biophobes and the calculated physicochemical properties if applicable. The contribution of modulators (such as LogP) is also given. Information is provided if the substance is outside the domain of the model in the form of warnings for the presence of fragments not present in the training set and not covered by the model. In addition, factors contributing to uncertainty such as biophores of limited statistical power, or the presence of inactivating fragments associated with an active prediction (or the opposite) are provided. While 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.

2.3. Data sources

205. The test data used in this model were taken from a single source, the *Data Book of Chromosomal Aberration Test In Vitro* (2). This book is written in Japanese, but all tables are in English and the authors were provided with English translations for everything except the Introduction. The Introduction is identical to that used in the previous version of the book, published in English by Dr. Motoi Ishidate (3), which was also available to the authors.

206. 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. This cell line has been used almost exclusively in Japan to test hundreds of chemicals over more than two decades, as opposed to the Chinese Hamster Ovary (CHO) cell lines that are more common in Europe and the United States. Much of the test information has been published in numerous scientific articles during the years over which it has been generated. An example is provided by Ishidate *et al.* (4).

207. The reason for using this source of information was partly to model Chromosomal Aberrations in the CHL cell line, and partly because the Data Book (2) is a very detailed and well-documented source of information for preparing (Q)SAR models, whether they be "local," "global," or anything in between. Nevertheless, it must be understood that when using historical data, not all tests will have been performed in complete compliance with the newest version of the Test Guidelines.

2.4. Selection of training set data

208. Test results for a total of 901 substances are presented in the Data Book (2). The chemicals were chosen for a variety of reasons, including use in foods. A number fall into the class commonly referred to as UVCB's, or chemicals that cannot be represented by a complete structure diagram and specific

molecular formula. These were excluded for the obvious reason that it is impossible to model a chemical for which a structure is not available. However, we found that this is not always a totally unambiguous process, so we made the best judgement we could. Inorganic chemicals were also excluded, as our modelling platform cannot deal with them. A very small number of chemicals were excluded because we were unsure of the true identity (inconsistencies between chemical name, CAS number and structure/molecular weight that we were unable to resolve). A few stereo-isomers with conflicting results were also removed as they cannot be distinguished by SMILES notation (a computer code for 2D structures), which is required by our model system.

209. We made a toxicological decision to include chemicals as being positive if they were active in inducing either aberrations or polyploidy. While the current test guideline does not specify testing for a length of time, which would allow polyploidy to be assessed, much of the CHL data does and we felt the information was too valuable to lose (18 chemicals). We also decided to retain chemicals even if the test had not been performed both in the presence and absence of metabolic activation. Under current regulatory practices, metabolic activation would be a requirement for all tests.

210. Beyond this we attempted to use the judgement of the authors in their interpretation of the final test result. This included dropping 16 of 18 chemicals that the authors considered inconclusive in repeat tests (we kept two because while they were inconclusive for polyploidy, they were clearly positive for structural aberrations).

211. Seventy-eight chemicals were excluded because the authors considered them False Positive (only active at dose of more than 10 mM where effects could be due to osmotic pressure).

212. As our model system cannot handle salts (e.g. sodium salts, hydrochlorides), further interpretation was necessary. In the majority of cases there was no conflict with regard to results of testing ionised or non-ionised forms. However, in certain cases there were. We decided that for some simple organic acids that were active but where the salt was clearly inactive, to consider these as being inactive in accordance with the advice, given in the OECD Guidelines and Morita et al. (5), that particularly low pH may lead to false positive predictions. We do not know if this decision is right or wrong in relation to use of results of this *in vitro* system for predicting *in vivo* effects, but it will clearly affect the performance of the model.

213. We also made a few decisions based on additional data from the literature: vitamin B2 (Riboflavin, CAS 83-88-5) tested positive in insoluble form, but was negative in soluble form. We retained the negative results, as the mechanism for the insoluble compound appears to be physical (6). After some consideration, saccharin (CAS 81-07-2) and EDTA (CAS 60-00-4) were entered as negatives, in agreement with Ashby et al. (7), even though there was conflicting information for some of the salts.

214. Finally, about 40 chemicals having only equivocal results were excluded. This is also an arbitrary decision, but we felt that equivocal results were not likely to lead to a better training set.

215. At this stage, 513 chemicals remained. Their identities (Appendix 2) can be compared with the total list of 901 (Appendix 3), for those interested in seeing which chemicals were left out. There were 263 positive and 250 negative substances in the training set, giving the nearly 50:50 split considered ideal for our model.

216. SMILES structures were entered for all substances. The positives were assigned a numerical score of 45 and the negatives were assigned a score of 10, as required by our software (see Section 2.2). In the Data Book (2), quantitative information is available for positives in the form of D20 (concentration required to induce aberrations in 20% of the metaphases) and TR scores (relative measure of the incidence

of aberrant cells with chromatid exchanges per unit concentration). However, we were unsure how to convert this information into MULTICASE scores in a transparent manner, and therefore did not use it at this stage.

2.5 Assessment of internal performance

217. We allowed the model to predict all of the chemicals it contained and obtained the results given in Table 1 for all predictions, and the results given in Table 2 for all predictions within the domain.

218. While this information may be useful for some purposes, the very high sensitivities and specificities are an artefact of the modelling process, and give no information about model predictivity. It is essential to do either cross validation and/or external validation to assess the predictivity of MULTICASE models (as with any other complex model).

219. There are no tools built into the system to automate a cross validation process. For this reason we have not performed LOO (leave-one-out) analysis, since this would require making 513 new models, each of which requires about 10 minutes to complete.

2.5.1 Cross-validation (10% out)

220. Instead we began by randomising our training set into ten separate models from which 10% of the chemicals had been removed, and letting each of these ten models make predictions for the missing 10%. The detailed results can be seen in Table 3.

221. We pooled the results into a single file and analysed it as follows: all Inconclusive or Marginal predictions were excluded from the exercise (they are either always "right" or always "wrong."). Only chemicals predicted as Active or Inactive were retained.

222. Without taking any account of the model's ability to identify a domain, the following result was obtained:

Out of a total of 513 predictions, 470 were Active or Inactive. On the basis of these 470 predictions, the following statistics were calculated:

Sensitivity = (145/233) x 100 = 62.23%

Specificity = (192/237) x 100 = 81.01%

Concordance = (337/470) x 100 = 71.70%

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%

223. To determine the domain, a chemical was considered as being outside the domain if: a) there were warnings of unknown fragment combinations; b) there were warnings of inadequate statistical power in biophore identification; or c) there were inactivating fragments for an active prediction, or activating

fragments present for an inactive prediction. There were 335 active or inactive predictions within the model domain.

2.5.2 Cross-validation (50% out)

224. Exactly the same procedure was done by randomly removing groups of 50% and letting each new model make predictions for the other half of the chemicals. This was done for a total of 100 models.

225. Without taking account of the domain, the following result was obtained:

Out of a total of 25,649 predictions, 23022 were active or inactive without taking account of the domain. On the basis of these 23022 predictions, the following statistics were calculated:

Sensitivity = (6895/11642) x 100 = 59.22%

Specificity = (8861/11380) x 100 = 77.86%

Concordance = (15755/23022) x 100 = 68.43%

226. Taking the domain into account, we obtained, for 14619 chemicals 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%

2.5.3 Randomisation test

227. 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 all chemicals, concordance was 55.566%, or for chemicals estimated as being within the domain, 49.692%.

2.6 External validation

228. 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., 8)

229. Of a total of 98 substances (Appendix 4), two were removed: dicyclopentadiene (CAS 77-73-6), because it was already in our training set, and Pigment Green No. 7 (CAS 14832-145), a copper complex that cannot be modelled in this system. When the 96 remaining chemicals (39 active and 57 inactive) were predicted by the model, we obtained the following statistics for the 69 substances within the domain:

Sensitivity = (13/24) x 100 = 54.167%

Specificity = (37/45) x 100 = 82.222%

Concordance = (50/69) x 100 = 72.64%

230. On further examination of the data set, it was noticed that one substance (4-(1-Methylpropyl)phenol, CAS 99-71-8) was actually a false positive (only active at very high concentration, and ultimately judged inactive following an *in vitro* micronucleus test). If this substance is removed, in addition to eight chemicals where chromosomal aberrations were induced under non-physiological culture conditions (pH<6), the following results for the 62 chemicals within the domain are:

Sensitivity = (10/17) x 100 = 58.824%

Specificity = (37/45) x 100 = 82.222%

Concordance = (47/62) x 100 = 75.806%

231. As some attempt was made to avoid positives associated with low pH values while making the training set, it may be justified to consider the latter result as the most representative validation. Although we discourage over-interpretation of results from small external validation sets on their own, this at least does not deviate significantly from the cross validation results: concordance and specificity of about 75%, well in excess of sensitivity.

2.7 Assessment of predictive values

232. In order to assess predictive values, it is necessary to know the number of positive test results to be expected from the population being examined. This can be quite difficult to determine. In the total Japanese HPV data set, 40% (39/98) were positive for chromosomal aberrations. If the false positive and the low pH chemicals are removed, this is about 31% (30/98). For the entire CHL data set (901 substances) we found the percentage of structural aberrations, excluding false negatives and inconclusives to be 39% (475/779). In the near future, we intend to use our model to predict the percentage of true positives in all of our discrete EINECS chemicals (European Inventory of Existing Chemical Substances).

233. Using sensitivities and specificities from our largest cross-validation exercise (100 x leave-out 50%; sensitivity 63.9, specificity 83.4 and concordance 74%) gives the following estimate for a “chemical universe” with the same percentage of positive substances as for all the substances tested in the CHL databook.

Positive Predictive Value (PPV) 71.13%

Negative Predicted Value (NPV) 78.33%

Obviously, when dealing with specific chemicals or groups where very high or very low prevalence of the effect is expected, these PPV and NPV figures can change dramatically.

2.8 Discussion and Conclusion

234. The model found 15 statistically significant biophores and 19 biophobes (see Annex 6). These will be examined in detail in the future to see if they can shed light on the otherwise complex mechanistic understanding of chromosomal aberrations. A few of the more obvious biophores identified include nitro-aromatics, certain PAHs and anilines.

235. The predictivity of the model is not as high as we had hoped, but still provides a very useful addition to our present battery of (Q)SAR models for assessing chemicals. One can speculate as to why results are not better (for example, we obtain concordances of 80% in Ames test models). This could

possibly be related to the fact that the data set is historical and that metabolic activation was only used in 170 of our 513 training set chemicals, whereas it is a requirement according to the current OECD test guideline.

236. It has been a great pleasure to work with the Japanese CHL Chromosomal Aberration data. We will continue to expand our model, first by including data from the external validation set and later with new information as it becomes available. We hope that the training set can be further optimised to increase predictive power, and are exploring the possibility of hybrid models including 3D and quantum mechanical descriptors.

3. APPLICATION OF THE SETUBAL PRINCIPLES

3.1. Defined endpoint (Principle 1)

237. By definition, a model for an OECD Guideline endpoint fulfils this condition. The scientific purpose is to gain mechanistic knowledge to help interpret possible *in vivo* effects such as carcinogenicity, mutagenicity and, in certain cases, reproductive toxicity.

238. However, we are not aware of results from the chromosomal aberration test in mammalian cells *in vitro* alone leading directly to regulatory actions. This will also apply to predictions from the (Q)SAR. These must be used together with other information on effects (measured or estimated) to provide an overall assessment of the substance in question before administrative decisions can be made.

239. For the CHL model training set, information on experimental conditions (and where applicable) units of measure are provided in a clear and consistent manner. The information is well organised and the level of detail is higher than what is found in most scientific publications.

3.2. Defined algorithm (Principle 2)

240. MULTICASE is a statistical method that identifies activating or deactivating fragment combinations in the training set. It provides a detailed description of the substructures, including an explicit identification of the substituents.

3.3. Mechanistic basis (Principle 3)

241. MULTICASE does not have any preconceived knowledge of molecular events that explain activity of a molecule. However, many of the resulting predictions have modes of action that are obvious to persons with expert knowledge for the endpoint in question. For example, in this model, certain groups are characterised by E_{lumo} (energy of the lowest unoccupied molecular orbital) profiles, which can explain their reactivity towards DNA.

242. Other groups represent different mechanisms of action, known or unknown, which require further examination to try and elucidate their toxicological modes of action. Provided that the model is sufficiently predictive, we consider this additional ability to suggest new hypotheses based on chemical substructures to be an extremely desirable feature, rather than the opposite. While we have not yet discovered receptor-mediated mechanisms in the chromosomal aberration model, we have seen clear

indications of this in other MULTICASE models for endpoints where ligand binding and activation/deactivation is known to occur (e.g. models for estrogenicity and anti-androgenicity).

3.4. Domain of applicability (Principle 4)

243. The model is associated with inclusion and/or exclusion rules on it applicability to groups of chemicals. The substructures are also associated with rules regarding the modulatory effects of the substructures of the molecular environment (biophores and biophobes). Warnings are routinely provided as part of the prediction process in all cases where a substance is outside of the model domain. Due to the complexity of the system, there is no practical way to make a list of all these rules beforehand. However, with knowledge of the substances in the training set there is nothing to prevent human intervention to override a domain decision of the basis of expert judgement, should this be considered necessary.

3.5. Internal performance (Principle 5)

244. Full details of the training set are given, including details of chemical names, structural formulae, CAS numbers (where available) and data for all descriptor and response variables. The raw data is available and can be compared with the list of substances appearing in the final training set.

245. Details of the statistical methodology used by the software are publicly available. Basic statistics for goodness-of-fit to the training set are available as specificity, sensitivity and chi-squared.

246. The (Q)SAR is associated with statistics based of cross-validation. Cross-validations were performed for a subset of 10x10% excluded, and for 100x50% excluded. The 10x50% cross-validation was also performed on the on the model after the toxicity data had been scrambled.

3.6. External validation for predictivity (Principle 6)

247. An external validation was undertaken based on the best data we could find in the open literature. We did not apply statistical methods to assess how well the test set and training set data were related. The selection of chemicals is a simple result of which substances from the OECD HPVC program had been selected and tested for CHL chromosomal aberrations by Japan during a certain period in time. The identities of the test substances and test conditions are well described.

248. A statistical analysis for specificity, sensitivity and concordance was conducted, giving results that were broadly similar to the cross-validations.

4. CONCLUSIONS ON THE SETUBAL PRINCIPLES AND THEIR APPLICABILITY

249. We disagree with the Setubal Principles in the suggestion that cross-validation cannot provide an estimate of predictivity of a model. We believe that it can, and that this is recognised in the literature (e.g. Eriksson et. al., 9).

250. We also have doubts with regard to the usefulness of some “definitive” conclusions made on the basis of small external validation sets, using data that is sometimes not even publicly available, and where

selection biases in the test sets are difficult or impossible to assess and/or where it is unclear how the validation set represents the domain of the model.

251. While concordances were about 75% in both of the cross validations and in the external validation, because the external validation set contained only 62 chemicals within the model domain, we felt that our 100x50% cross validation gave the best estimate of predictivity and used these results to estimate the positive and negative predictive values.

252. For our purposes, predictivity is sufficient to allow this model to take its place along with many others in the Danish EPA's (Q)SAR system where it will assist us in using expert judgement to assess the possible toxic effects of chemical substances.

253. Concerning possible regulatory use by other parties, we would be happy to make the model available to anyone using the full development version of MULTICASE, or make the training set available to anyone wishing to use other modelling systems. But it is important to emphasise that this model is not "static." It will be updated by inclusion of the external validation test set chemicals and assessed once again by cross-validations. Further changes and additions will take place as new test data for this endpoint become available, or in accordance with any changes in our understanding of the model or mechanisms which lead to further improvements.

5. REFERENCES

Ashby, J. & Ishidate, M. Jr. (1986). Clastogenicity in vitro of the Na, K, Ca and Mg. Salts of Saccharin; and of magnesium chloride; consideration of significance. *Mutation Research* **163**, 63-73.

Eriksson, et al. (1999). *Introduction to Multi- and Megavariate Data Analysis using Projection Methods (PCA & PLS)*. Umetrics AB; Umeå, Sweden.

Ishidate, M. Jr., Haronois, M.C. & Sofoni, T. (1988). A Comparative analysis of data on the clastogenicity of 951 chemicals tested in mammalian cell cultures. *Mutation Research* **195**, 151-213.

Ishidate, Motoi Jr., Ed. (1988). *Data Book of Chromosomal Aberration Test In Vitro, Revised Edition*. Elsevier; Amsterdam, New York, Oxford.

Kawaguchi, Y., Hayashi, H., Sato, M. & Shindo, Y. (1997). Needle crystals of Vitamin B2 induce polyploidy in Chinese hamster lung (CHL/IU) cells. *Mutation Research* **373**, 1-7.

Kusakabe, H., Ymakage, K., Wakuri, S., Sasaki, K., Nakagawa, Y., Watanabe, M., Hayashi, M., Sufuni, T., Ono, H. & Tnanka, N. (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.

Morita, T., Nagaki, T., Fukuda, I. & Okumura, K. (1992). Clastogenicity of low pH to various cultures mammalian cells. *Mutation Research* **268**, 297-305.

OECD (1997). OECD Guidelines for the Testing of Chemicals No. 473: Genetic Toxicology: *In Vitro* Mammalian Cytogenetic Test. Organisation for Economic Cooperation and Development; Paris, France.

Sofuni, T., Ed. (1998). *Data Book of Chromosomal Aberration Test In Vitro, Revised Edition*.. Life-Science Information Center; Tokyo, Japan.

Table 1 Internal performance for all predictions

	Active	**Inactive**	**Total**	**Accuracy**
Predicted +	262	4	266	98.5%
Predicted -	1	246	247	99.6%
Total	263	250	513	
Percentage	99.6 (sensitivity)	98.4 (specificity)		

Results including 31 molecules having inconclusive (+) or (-) values

Chi square = 493.244; Phi square = 0.961

Expected Correct Predictions (ECP) = 50.05 %

Observed Correct Predictions (OCP) = 99.03 %

Table 2 Internal performance for those 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.

6. APPENDICES

APPENDIX 1 SUMMARY REPORT OF THE APPLICATION OF THE SETUBAL PRINCIPLES

PRINCIPLE	CONSIDERATIONS	YES / NO
1) Defined endpoint	**1.1)** Does the model have a clearly defined scientific purpose?	Yes
	(i.e does it make predictions of a clearly defined physicochemical, biological or environmental effect?)	
	1.2) Does the model have the potential to address (or partially address) a clearly defined regulatory need?	Yes
	(i.e. does it make predictions of a specific endpoint associated with a specific test method or test guideline)?	
	1.3) Is information given about important experimental conditions that affect the measurement and therefore the prediction	Yes
	(e.g. sex, species, temperature, exposure period)	
	1.4) Are the units of measurement of the endpoint given?	Yes
2) Defined algorithm	**2.1)** In the case of a SAR, is there an explicit description of the substructure, including an explicit identification of its substituents?	Yes
	2.2) In the case of a QSAR, is the equation explicitly defined, including definitions of all descriptors used[3]?	
3) Mechanistic basis	**3.1)** In the case of a SAR, is there a description of the molecular events that underlie the reactivity of the molecule?	Yes/No
	(e.g. a description of how substructural features could act as nucleophiles or electrophiles, or form part or all of a receptor-binding region)	

	3.2) In the case of a QSAR, do the descriptors have a physicochemical interpretation that is consistent with a known mechanism (of biological action)?	
	3.3) Are any literature references cited in support of the purported mechanistic basis of the (Q)SAR?	
4) Domain of applicability	**4.1)** In the case of a SAR, is the substructure associated with any inclusion and/or exclusion rules on its applicability to groups of chemicals?	Yes
	4.2) In the case of a SAR, is the substructure associated with rules regarding the modulatory effects of the substructure's molecular environment?	Yes
	4.3) In the case of a (Q)SAR, are the descriptor and response variables associated with inclusion and/or exclusion rules that define the variable ranges for which the QSAR is applicable (i.e. makes reliable estimates)?	
5) Internal performance	**5.1)** Are full details of the training set given, including details of chemical names, structural formulae, CAS numbers (if available), and data for all descriptor and response variables?	Yes
	5.2) If the data used to the develop the model were based upon the processing of raw data (e.g. the averaging of replicate values):	Yes
	a) is there an adequate description of the data processing?	
	b) are the raw data provided?	
	5.3) Is there a specification of the statistical method(s) used to develop the QSAR (including details of any software packages used)?	Yes
	5.4) Is the QSAR associated with basic statistics for its goodness-of-fit to the training set?	NA
	(e.g. r^2 values and standard error of the estimate in the case of regression models)	
	5.5)	NA
	a) Is the QSAR associated with any statistics based on cross-validation or resampling?	
	b) If yes, is the number or samples used indicated?	

6) Predictivity (External validation)	**6.1)** Does application of the appropriate statistical method(s) to the training set result in the same (Q)SAR model?	Yes
	6.2) Is there any information to indicate that the (Q)SAR has been validated previously, using a test set that is independent of the training set?	No
	6.3) If an external validation has been performed, is the following information available:	Yes
	a) the number of test structures?	
	b) the identities of the test structures?	
	c) the approach for selecting the test structures?	
	d) the statistical analysis of the predictive performance of the model?	
	(e.g. including sensitivity, specificity, and positive and negative predictivities for classification models)	
	e) a comparison of the predictive performance of the model against previously-defined quantitative performance criteria?	

APPENDIX 2 TRAINING SET OF 513 CHEMICALS

VERY LARGE TABLE, THEREFORE OMITTED FROM THIS REPORT. AVAILABLE ON REQUEST FROM THE AUTHORS.

APPENDIX 3 TOTAL LIST OF 911 CHEMICALS

VERY LARGE TABLE, THEREFORE OMITTED FROM THIS REPORT. AVAILABLE ON REQUEST FROM THE AUTHORS.

APPENDIX 4 LIST OF 98 CHEMICALS USED TO PERFORM AN EXTERNAL VALIDATION

Chemical class	CAS number	Chemical name	Test: CA	Test: PP	Prediction	Predictivity *	Domain flag from MC
Aniline	87-59-2	2,3-Dimethylaniline	POS	-	MAR	-	Inside
	95-64-7	3,4-Dimethylaniline	NEG	NEG	POS	FP	Inside
	103-69-5	N-Ethylaniline	POS	NEG	POS	TP	Inside
	100-61-8	N-Methylaniline	POS	NEG	POS	TP	Inside
	97-52-9	4-Nitro-o-anisidine	POS	NEG	POS	TP	Inside
	108-44-1	m-Toluidine	NEG	NEG	POS	FP	Inside
Sulfonic acid	70-55-3	4-Methylbenzenesulfonamide	NEG	NEG	NEG	TN	Inside
	121-47-1	**3-Aminobenzenesulfonic acid**	**POS**	**NEG**	**INC**	**-**	*Outside*
	88-44-8	**2-Amino-5-methylbenzene-sulfonic acid**	**POS**	**NEG**	**INC**	**-**	**Outside**
	88-53-9	**2-Amino-5-chloro-4-methylbenzenesulfonic acid**	**POS**		**NEG**	**FN**	**Inside**
Halogenated benzene	106-37-6	1,4-Dibromobenzene	POS	NEG	NEG	FN	Inside
	89-61-2	1,4-Dichloro-2-nitrobenzene	POS	NEG	NEG	FN	Inside
	3209-22-1	1,2-Dichloro-3-nitrobenzene	POS	POS	POS	TP	Inside
	611-06-3	2,4-Dichloronitrobenzene	NEG	NEG	POS	FP	Inside
	95-73-8	2,4-Dichloro-1-methylbenzene	NEG	NEG	NEG	TN	Inside
	98-08-8	Trifouromethylbenzene	NEG	NEG	NEG	TN	Inside
Phenol	123-30-8	4-Aminophenol	POS	NEG	POS	TP	Inside
	2581-34-2	3-Methyl-4-nitrophenol	POS	NEG	POS	TP	Inside
	99-71-8	4-(1-Methylpropyl)phenol	POS	NEG	-	-	Removed**
	89-83-8	Thymol	POS	NEG	NEG	FN	Inside
	98-54-4	p-tert-Butylphenol	POS	POS	POS	TP	Inside
	1879-09-0	6-tert-Butyl-2,4-xylenol	NEG	NEG	NEG	TN	Inside
	140-66-9	p-tert-Octylphenol	NEG	NEG	POS	FP	Inside
Bisphenol	119-47-1	2,2'-Methylenebis(6-tert-butyl-p-cresol)	NEG	NEG	NEG	TN	Inside
	96-69-5	4,4'-Thiobis(6-tert-butyl-m-cresol)	NEG	NEG	NEG	TN	Outside
Organic phosphate	107-66-4	Dibutyl phosphate	NEG	NEG	NEG	TN	Inside
	26444-49-5	Diphenyl cresyl phosphate	POS	NEG	INC	-	Outside
	1806-54-8	Tris(2-theylhexyl)phosphate	NEG	NEG	NEG	TN	Inside
	26967-76-0	Tris(p-cymenyl)phosphate	NEG	NEG	POS	FP	Outside

Chemical class	CAS number	Chemical name	Test: CA	Test: PP	Prediction	Predictivity *	Domain flag from MC
	78-51-3	Tris(2-butoxyethyl)phosphate	NEG	NEG	NEG	TN	Inside
	512-56-1	Thimethyl phosphate	NEG	NEG	POS	FP	Inside
PAH	83-32-9	Acenaphthalene	POS	NEG	NEG	FN	Inside
	81-16-3	**2-Amino-1-naphthalensulfonic acid**	**POS**	**NEG**	**POS**	**TP**	**Inside**
	840-65-3	Dimethyl-2,6-naphthalenecarboxylate	NEG	NEG	NEG	TN	Inside
	2216-69-5	1-Methoxynaphthalene	POS	NEG	POS	TP	Inside
	86-87-3	**1-Naphthylacetic acid**	**POS**	**NEG**	**NEG**	**FN**	**Outside**
	842-18-2	Potassium 7-hydroxy-1,3-naphthalenedisulfonate	NEG	NEG	NEG	TN	Inside
	5460-09-3	Monosodium 4-amino-5-hydroxy-2,7-naththalenedisulfonate	NEG	NEG	NEG	TN	Inside
	82-45-1	1-aminoanthraquinone	NEG	NEG	POS	FP	Inside
Pigment	5281-04-9	D & C Red No. 7	NEG	NEG	NEG	TN	Inside
	14832-14-5	Pigment green No. 7	NEG	NEG	-	-	Outside – contains copper
Heterocyclic compound	95-33-0	N-cyclohexyl-2-benzothiazolrsulfonamide	NEG	NEG	NEG	TN	Outside
	4979-32-2	N,N-Dicyclohexyl-2-benzothiazolesulfonamide	NEG	POS	NEG	FN	Outside
	95-31-8	N-tert-butyl-2-benzothiazolesulfonamide	POS	NEG	NEG	FN	Outside
	110-02-1	Thiophene	NEG	NEG	NEG	TN	Outside
	126-33-0	Tetrahydrothiophene-1,1-dioxide	NEG	NEG	NEG	TN	Inside
	583-39-1	2-mercaptobenzimidazole	POS	NEG	NEG	FN	Outside
Aldehyde	90-02-8	2-Hydroxybenzaldehyde	POS	POS	MAR	-	Outside
Alkyl benzene	1477-55-0	1,3-Bis(aminoethyl)benzene	NEG	NEG	NEG	TN	Inside
	95-63-6	1,2,4-Trimethylbenzene	NEG	NEG	NEG	TN	Inside
	105-05-5	1,4-Diethylbenzene	NEG	NEG	NEG	TN	Inside
	98-83-9	1-Methylethenylbenzene	NEG	NEG	NEG	TN	Inside
	25321-09-9	Diisopropylbenzene	NEG	NEG	NEG	TN	Inside
	1321-74-0	Divinylbenzene	NEG	NEG	NEG	TN	Inside
Alcohol or ether	4461-52-3	Methoxymethanol	POS	POS	NEG	FN	Outside
	111-41-1	N-(Aminoethyl)ethanolamine	NEG	POS	NEG	FN	Inside
	123-42-2	Diacetone alcohol	NEG	NEG	NEG	TN	Inside
	110-63-4	1,4-Butnediol	NEG	NEG	NEG	TN	Inside
	4457-71-0	3-Methyl-1,5-pentanediol	NEG	NEG	NEG	TN	Inside
	584-03-2	1,2-Butanediol	NEG	NEG	NEG	TN	Inside
	126-30-7	2,2-Dimethyl-1,3-propanediol	NEG	NEG	NEG	TN	Inside
	77-99-6	2-Etheyl-2-hydroxymethyl-1,3-propanediol	NEG	NEG	NEG	TN	Inside
	108-65-6	Propylene glycol monomethyl ether acetate	NEG	NEG	NEG	TN	Inside
	115-77-5	Pentaerythritol	NEG	NEG	NEG	TN	Inside
	6846-50-0	2,2,4-Trimethyl-1,3-pentanediol diisobutyrate	NEG	NEG	NEG	TN	Outside

Chemical class	CAS number	Chemical name	Test: CA	Test: PP	Prediction	Predictivity *	Domain flag from MC
	24800-44-0	Tripropylene glycol	NEG	NEG	NEG	TN	Inside
	3452-97-9	3,5,5-Trimethylhexan-1-ol	NEG	NEG	NEG	TN	Inside
	102-76-1	**Gycerol triacetate**	**POS**	**NEG**	**NEG**	**FN**	**Inside**
	105-45-3	**Methyl acetoacetate**	**POS**	**POS**	**NEG**	**FN**	**Inside**
Carboxylic acid or ester	526-78-3	2,3-Dibromosuccinic acid	NEG	NEG	POS	FP	Inside
	99-96-7	**4-Hydroxybenzoic acid**	**POS**	**NEG**	**POS**	**TP**	**Inside**
	105-99-7	Dibutyl adipate	POS	NEG	NEG	FN	Inside
	623-91-6	Diethyl fumarate	POS	POS	POS	TP	Outside
	2439-35-2	2-(Dimethylamino)ethyl acrylate	POS	POS	POS	TP	Inside
	868-77-9	2-Hydroxyethyl methacrylate	POS	POS	NEG	FN	Outside
	106-91-2	2,3-Epoxypropyl methacrylate	POS	POS	POS	TP	Outside
	105-16-8	2-(Dimethylamino)ethyl methacrylate	POS	POS	NEG	FN	Outside
	923-26-2	2-Hydroxypropyl methacrylate	POS	POS	NEG	FN	Outside
	111-82-0	Methyl Dodecanoate	NEG	NEG	NEG	TN	Inside
	3319-31-1	1,2,4-Tris(2-ethylhexyl) 1,2,4-benzenetricarboxylate	NEG	NEG	MAR	-	Outside
Cyanide	78-97-7	2-Hydroxypropanenitrile	POS	POS	NEG	FN	Outside
	78-67-1	2,2-Azobis(2-methylpropionitrile	NEG	NEG	NEG	TN	Outside
	626-17-5	1,3-Dicyacobenzene	NEG	NEG	NEG	TN	Outside
	623-26-7	1,4-Dicyanobenzene	NEG	NEG	NEG	TN	Outside
	108-80-5	Isocyanuric acid	NEG	NEG	NEG	TN	Inside
Non-cyclic alkenes	7756-94-7	Triisobutylene	NEG	NEG	NEG	TN	Inside
	760-23-6	3,4-Dichloro-1-butene	POS	POS	POS	TP	Inside
Non-cyclic alkanes	629-62-9	n-Pentadecane	NEG	NEG	NEG	TN	Inside
	544-76-3	n-Hexadecane	NEG	NEG	NEG	TN	Inside
	109-69-3	1-Chrolobutane	NEG	NEG	POS	FP	Inside
	1120-21-4	Undecane	NEG	NEG	NEG	TN	Inside
	4390-04-9	2,2,4,4,6,8,8-Heptamethylnonane	NEG	NEG	NEG	TN	Inside
Others	538-75-0	Dicyclohexylcarbodiimide	NEG	NEG	NEG	TN	Outside
	11070-44-3	Tetrahydromethyl-1,3-isobenzofuranedione	NEG	POS	NEG	FN	Outside
	3048-65-5	3a,4,7,7a-Tetrahydro-1H-indene	POS	NEG	NEG	FN	Inside
	77-73-6	Dicyclopentadiene	NEG	NEG	-	-	Removed - in training set
	16219-75-3	5-Etheylidene-2-norborene	NEG	NEG	NEG	TN	Inside
	96-29-7	Etheyl methyl ketoxime	NEG	NEG	NEG	TN	Outside

FOOTNOTES

* TP: True Positive (predicted positive and positive test)

FP: False Positive (predicted positive and negative test)

TN: True Negative (predicted negative and negative test)

FN: False Negative (predicted negative and positive test)

** According to Kusakabe et. al (9), the induction of CA for CAS 99-71-8 was in the end judged as a false positive, because it is only active at very high concentrations. This was confirmed by a negative result in an *in vitro* micronucleus test using CHL/IU cells..

Chemicals in "**bold format**" are the eight chemicals where Chromosomal Aberrations were induced under non-physiological culture conditions (pH<6).

ANNEX 5
QSARS FOR PREDICTING THE NO OBSERVED EFFECT LEVEL (NOEL) IN HUMANS

Dr Edwin Matthews and Dr Joseph F. Contrera
US Food & Drug Administration
Center for Drug Evaluation and Research
5600 Fishers Lane
20857 Rockville
Maryland
United States

TABLE OF CONTENTS

1. INTRODUCTION

254. Dr Edwin Matthews (US Food & Drug Administration) has retrospectively applied Setubal Principles 1-6 to a MC4PC (MCASE) model for predicting the threshold for toxic effects of chemicals in humans.

255. In addition, Dr Joseph F. Contrera (US Food & Drug Administration) has retrospectively applied Setubal Principles 1-6 to a MDL QSAR model for predicting the threshold for toxic effects of chemicals in humans.

256. A summary (in tabular form, Appendices 1 and 2) of the extent to which the Setubal principles have been met by these models is attached.

257. A more-detailed explanation (when necessary) following the items in Appendices 1 and 2 is reported below.

2. APPLICATION OF THE SETUBAL PRINCIPLES TO MODEL 1

2.1. Defined endpoint (Principle 1)

258. The MC4PC (sometimes referred to as MCASE) model was developed by the Center for Drug Evaluation and Research (CDER) of the US FDA (1). The model makes predictions of the No Effect Level (NOEL) in humans, using data derived from pharmaceutical clinical trials (2).

259. It is questionable whether the NOEL is a clearly defined endpoint, since the actual tissue/organ in which the threshold effect occurs is not necessarily defined.

260. The units of the prediction are given as mg/kg-bw/day.

261. The model is used by FDA to provide decision support information for regulatory and research issues. For example, the data can be used to: 1) evaluate the toxicities of contaminants in FDA regulated substances; 2) estimated the start dose for pharmaceuticals in human clinical trials; and 3) prioritise substances for review in relation to the FDA Modernization Act of 1997 (FDAMA), which established a pre-market notification system for Food Contact Substances (FCS). This pre-market notification process places the burden on FDA to object to a notification within 120 days or an FCS may be legally marketed on the 121st day (3).

2.2. Defined algorithm (Principle 2)

262. The model is based on 134 defined structural alerts, which were correlated with toxicity in humans by using a training set of 1309 chemicals.

263. Full details of this training set are given for 1233 non-proprietary pharmaceuticals (2, 4) in the US FDA Maximum Recommended Therapeutic Dose (MRTD) Database. The database was compiled from

data in *Martindale: The Extra Pharmacopoeia* (1973, 1983, 1993) and in *The Physicians' Desk Reference* (1995, 1999). MRTD values for the 1233 chemicals in this dataset range from 0.00001 to 1000 mg/kg-bw/day.

264. Most of the MRTD values in the database were determined from pharmaceutical clinical trials that employed an oral route of exposure. In contrast, roughly 5% of the pharmaceuticals in the MRTD database (antineoplastics and anesthetics) were administered intravenously and/or intramuscularly. When separate MRTDs were reported for different routes of exposure, only the oral MRTD was included in the database. In addition, some pharmaceuticals have different MRTD values for male and female adults, children, or elderly patients. In this situation only MRTD values for the average adult patient were used.

265. Pharmaceuticals that are administered orally usually have MRTDs reported as mg/day. The mg/day unit was converted to mg/kg-body weight (bw)/day based upon an average adult weighing 60 kg. In contrast, the dose unit for most antineoplastic drug MRTDs is reported as mg/m^2 which was converted to mg/kg-bw/day using the formula mg/kg-bw/day = $mg/m^2/37$ for an average adult. Additionally, a few drugs had MRTDs reported in parts per million (ppm), which were converted to mg/kg-bw/day on the basis that 1000 ppm equals 25 mg/kg-bw/day for an average 60 kg adult.

266. To develop the MC4PC model, a database of the maximum recommended therapeutic dose (MRTD) of marketed pharmaceuticals was compiled. The MRTD is sometimes referred to as the maximum recommended daily dose (MRDD) of the pharmaceutical. Chemicals with low MRTDs were classified as high-toxicity compounds, whereas chemicals with high MRTDs were classified as low-toxicity compounds. Two separate training data sets were constructed to identify specific structural alerts associated with high and low toxicity chemicals.

267. The human MRTD and NOEL of a pharmaceutical are directly related to one another. Based upon our analyses of the therapeutic dose ranges for pharmacologic effects of drugs in our database, the overwhelming majority of drugs demonstrate efficacy over a small range of treatment doses. An analysis of the MRTD database revealed that most drugs do not demonstrate efficacy or adverse effects at a dose approximately 1/10 the MRTD (data not presented). Based upon this observation, NOEL is defined as MRTD/10 in this study. For a few noteworthy pharmaceutical categories (e.g. some chemotherapeutics and immunosuppressants), the clinically effective dose may be a dose that is accompanied by substantial adverse effects. In such cases, the true NOEL value may be less than 1/10 the MRTD. On the other hand, for chemicals that are not pharmaceuticals there is no MRTD and the NOEL can be considered a dose above which any compound related effect is likely to be considered an adverse effect and a manifestation of toxicity.

2.3. Mechanistic basis (Principle 3)

268. Many of the structural alerts are associated with specific clinical indications of drugs, but the alerts are not necessarily associated with mechanistic hypotheses or explanations.

2.4. Domain of applicability (Principle 4)

269. In addition to identifying structural alerts, MCASE identifies modulators of activity, and calculations are made of the statistical significance of the structural alerts and the modulators.

270. A domain of applicability is defined for this model in terms of molecular coverage of the test molecule relative to the molecules in the training data set. If the model identifies two or more 2-3 atom molecular fragments in the test molecule that are not present in the training data molecular library, the test molecule is evaluated as not covered and thereby outside of the domain of the model. In addition, the model has an adjunct expert system software that has human expert inclusion and/or exclusion rules which

automatically evaluate MC4PC molecular descriptor information. Furthermore, the model provides calculations of the statistical significance and provides a measure of certainty in the structural alerts and modulators.

271. The following types of chemicals were excluded during the development of the MCASE model, due to their unsuitability for QSAR modelling: inorganic chemicals, high molecular weight polymers (>5000 Daltons), fibers, salts, mixtures of organic chemicals and small molecules (<100 Daltons).

272. The particular MCASE model described here exhibited good coverage (89.9-93.6%) for three classes of chemicals: pharmaceuticals, direct food additives, and food contact substances (2). It should be noted that food contact substances are typically industrial chemicals used in the manufacture of products that come in contact with food.

2.5. Internal performance (Principle 5)

273. An internal validation experiment (2) showed that predictions for high-toxicity and low-toxicity chemicals were good (positive predictivity 92.5%; specificity 95.2%; false positives 4.8%; sensitivity 74.0%; false negatives 26.0%), and differences between experimental and predicted MRTDs were small (0.27 - 0.70 log-fold).

2.6. External validation for predictivity (Principle 6)

274. The model has not been externally validated, by using a test set that is independent of the training set. However, the FDA has compiled a new external validation set of 160 chemicals that will be used for this purpose. The chemicals were obtained using the same sources cited above.

3. APPLICATION OF THE SETUBAL PRINCIPLES TO MODEL 2

3.1. Defined endpoint (Principle 1)

275. The MDL-QSAR model was developed by the Center for Drug Evaluation and Research (CDER) of the US FDA. The model makes predictions of the No Effect Level (NOEL) in humans, using data derived from pharmaceutical clinical trials (2, 5).

276. It is questionable whether the NOEL is a clearly defined endpoint, since the actual tissue/organ in which the threshold effect occurs is note necessarily defined.

277. The units of the prediction are given as mg/kg-bw/day.

278. The model is used by FDA to provide decision support information for regulatory and research issues. For example, the data can be used to: 1) evaluate the toxicities of contaminants in FDA regulated substances; 2) estimated the start dose for pharmaceuticals in human clinical trials; and 3) prioritise substances for review in relation to the FDA Modernization Act of 1997 (FDAMA), which established a pre-market notification system for Food Contact Substances (FCS). This pre-market notification process places the burden on FDA to object to a notification within 120 days or an FCS may be legally marketed

on the 121st day (3). The model may also be used to estimate the safe starting dose for human subjects in phase I clinical trials.

3.2. Defined algorithm (Principle 2)

279. The model is based on approximately 77 defined electrotopological molecular E-state and connectivity descriptors, which were correlated with toxicity in humans by using a training set of 1309 chemicals.

280. To develop the MDL QSAR model, a database of the maximum recommended therapeutic dose (MRTD) of marketed pharmaceuticals was compiled. The MRTD is sometimes referred to as the maximum recommended daily dose (MRDD) of the pharmaceutical. Chemicals with low MRTDs were classified as high-toxicity compounds, whereas chemicals with high MRTDs were classified as low-toxicity compounds. Two separate training data sets were constructed to identify specific molecular descriptors associated with high and low toxicity chemicals.

281. The human MRTD and NOEL of a pharmaceutical are directly related to one another. Based upon our analyses of the therapeutic dose ranges for pharmacologic effects of drugs in our database, the overwhelming majority of drugs demonstrate efficacy over a small range of treatment doses. An analysis of the MRTD database revealed that most drugs do not demonstrate efficacy or adverse Effects at a dose approximately 1/10 the MRTD (data not presented). Based upon this observation, NOEL is defined as MRTD/10 in this study. For a few noteworthy pharmaceutical categories (e.g. some chemotherapeutics and immunosuppressants), the clinically effective dose may be a dose that is accompanied by substantial adverse effects. In such cases, the true NOEL value may be less than 1/10 the MRTD. On the other hand, for chemicals that are not pharmaceuticals there is no MRTD and the NOEL can be considered a dose above which any compound related effect is likely to be considered an adverse effect and a manifestation of toxicity.

3.3. Mechanistic basis (Principle 3)

282. Many of the structural alerts are associated with specific clinical indications of drugs, but the alerts are not necessarily associated with mechanistic hypotheses or explanations.

3.4. Domain of applicability (Principle 4)

283. The MDL QSAR identifies relevant structural electrotopological descriptors and calculations are made of the correlation and statistical significance of descriptors to biological endpoints of interest.

284. A domain of applicability is not explicitly defined for this model. However, the calculations of the statistical significance provide a measure of certainty in the structural descriptors.

285. The following types of chemicals were excluded during the development of the MCASE model, due to their unsuitability for QSAR modeling: inorganic chemicals, high molecular weight polymers (>5000 Daltons), fibers, salts, mixtures of organic chemicals and small molecules (<100 Daltons).

286. The particular MDL QSAR model described here exhibited good coverage (>71%) for chemicals that were mainly pharmaceuticals and industrial chemicals.

3.5. Internal performance (Principle 5)

287. The training set is described above (Section 2.5).

288. An internal validation experiment with 120 compounds, and an external validation study with 160 compounds, showed that predictions for high-toxicity and low-toxicity chemicals were good with 74-78% of predicted MRTD values falling within 0.1-10 times the actual MRTD.

3.6. External validation for predictivity (Principle 6)

289. The model has been externally validated by using a test set of 160 compounds that is independent of the training set. The chemicals were obtained using the same sources cited above.

4. CONCLUSIONS ON THE SETUBAL PRINCIPLES AND THEIR APPLICABILITY

290. The principles and the associated checklist are useful for identifying key pieces of information about (Q)SAR models, but not all of the items in the checklist are applicable to all types of model. For example, structual alerts that have been identified by clinical studies, or defined on the basis of expert knowledge, cannot be assessed for statistical significance by using a training set, since the structural alerts were not obtained by the application of a statistical algorithm to a training set.

5. REFERENCES

http://www.epa.gov/nheerl/dsstox/sdf_dbpcan.html

http://www.fda.gov/cder/Offices/OPS_IO/ICSAS.htm#ComToxCFSAN

http://www.fda.gov/cder/Offices/OPS_IO/MRTD.htm

http://www.multicase.com/products/prod01.htm

Matthews, E. J., Kruhlak, N. L., Benz, R. D. & Contrera, J. F. (2004). Assessment of the Health Effects of Chemicals in Humans: I. QSAR Estimation of the Maximum Recommended Therapeutic Dose (MRTD) and No Effect Level (NOEL) of Organic Chemicals Based on Clinical Trial Data. *Current Drug Discovery Technologies* **1**(1).

6. APPENDICES

APPENDIX 1 SUMMARY REPORT OF THE APPLICATION OF THE SETUBAL PRINCIPLES TO MODEL 1

PRINCIPLE	CONSIDERATIONS	YES / NO
1) Defined endpoint	**1.1)** Does the model have a clearly defined scientific purpose? (i.e does it make predictions of a clearly defined physicochemical, biological or environmental effect?)	Yes
	1.2) Does the model have the potential 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)?	Yes
	1.3) Is information given about important experimental conditions that affect the measurement and therefore the prediction (e.g. sex, species, temperature, exposure period)	Yes
	1.4) Are the units of measurement of the endpoint given?	Yes
2) Defined algorithm	**2.1)** In the case of a SAR, is there an explicit description of the substructure, including an explicit identification of its substituents?	Yes, structural alerts
	2.2) In the case of a QSAR, is the equation explicitly defined, including definitions of all descriptors used[3]?	No
3) Mechanistic basis	**3.1)** In the case of a SAR, is there a description of the molecular events that underlie the reactivity of the molecule? (e.g. a description of how substructural features could act as nucleophiles or electrophiles, or form part or all of a receptor-binding region)	Sometimes
	3.2) In the case of a QSAR, do the descriptors have a physicochemical interpretation that is consistent with a known mechanism (of biological action)?	Sometimes
	3.3) Are any literature references cited in support of the purported mechanistic basis of the (Q)SAR?	Yes
4) Domain of	**4.1)** In the case of a SAR, is the substructure associated with	Yes

applicability	any inclusion and/or exclusion rules on its applicability to groups of chemicals?	
	4.2) In the case of a SAR, is the substructure associated with rules regarding the modulatory effects of the substructure's molecular environment?	Yes
	4.3) In the case of a QSAR, are the descriptor and response variables associated with inclusion and/or exclusion rules that define the variable ranges for which the QSAR is applicable (i.e. makes reliable estimates)?	Yes, rules in adjunct expert system
5) Internal performance	**5.1)** Are full details of the training set given, including details of chemical names, structural formulae, CAS numbers (if available), and data for all descriptor and response variables?	Yes, for non-proprietary chemicals
	5.2) If the data used to the develop the model were based upon the processing of raw data (e.g. the averaging of replicate values): a) is there an adequate description of the data processing? b) are the raw data provided?	Yes
	5.3) Is there a specification of the statistical method(s) used to develop the QSAR (including details of any software packages used)?	Yes, pair-wise t-tests
	5.4) Is the QSAR associated with basic statistics for its goodness-of-fit to the training set? (e.g. r^2 values and standard error of the estimate in the case of regression models)	No
	5.5) a) Is the QSAR associated with any statistics based on cross-validation or resampling? b) If yes, is the number or samples used indicated?	No
6) Predictivity (External validation)	**6.1)** Does application of the appropriate statistical method(s) to the training set result in the same (Q)SAR model?	Not done
	6.2) Is there any information to indicate that the (Q)SAR has been validated previously, using a test set that is independent of the training set?	In progress (8/4/04)
	6.3) If an external validation has been performed, is the following information available: a) the number of test structures?	160 chemicals / pharmaceuticals
	b) the identities of the test structures? c) the approach for selecting the test structures? d) the statistical analysis of the predictive performance of the model? (e.g. including sensitivity, specificity, and positive and negative predictivities for classification models) e) a comparison of the predictive performance of the model against previously-defined quantitative performance criteria?	In progress (8/4/04)

APPENDIX 2 SUMMARY REPORT OF THE APPLICATION OF THE SETUBAL PRINCIPLES TO MODEL 2

PRINCIPLE	CONSIDERATIONS	YES / NO
1) Defined endpoint	**1.1)** Does the model have a clearly defined scientific purpose? (i.e. does it make predictions of a clearly defined physicochemical, biological or environmental effect?)	Yes
	1.2) Does the model have the potential 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)?	Yes
	1.3) Is information given about important experimental conditions that affect the measurement and therefore the prediction (e.g. sex, species, temperature, exposure period)	Yes
	1.4) Are the units of measurement of the endpoint given?	Yes
2) Defined algorithm	**2.1)** In the case of a SAR, is there an explicit description of the substructure, including an explicit identification of its substituents?	Yes, molecular descriptors
	2.2) In the case of a QSAR, is the equation explicitly defined, including definitions of all descriptors used[3]?	Sometimes
3) Mechanistic basis	**3.1)** In the case of a SAR, is there a description of the molecular events that underlie the reactivity of the molecule? (e.g. a description of how substructural features could act as nucleophiles or electrophiles, or form part or all of a receptor-binding region)	Sometimes
	3.2) In the case of a QSAR, do the descriptors have a physicochemical interpretation that is consistent with a known mechanism (of biological action)?	Sometimes
	3.3) Are any literature references cited in support of the purported mechanistic basis of the (Q)SAR?	Yes
4) Domain of applicability	**4.1)** In the case of a SAR, is the substructure associated with any inclusion and/or exclusion rules on its applicability to groups of chemicals?	No
	4.2) In the case of a SAR, is the substructure associated with rules regarding the modulatory effects of the substructure's molecular environment?	Yes
	4.3) In the case of a QSAR, are the descriptor and response	Yes

	variables associated with inclusion and/or exclusion rules that define the variable ranges for which the QSAR is applicable (i.e. makes reliable estimates)?	
5) Internal performance	**5.1)** Are full details of the training set given, including details of chemical names, structural formulae, CAS numbers (if available), and data for all descriptor and response variables?	Yes, for non-proprietary chemicals
	5.2) If the data used to the develop the model were based upon the processing of raw data (e.g. the averaging of replicate values): a) is there an adequate description of the data processing? b) are the raw data provided?	Yes
	5.3) Is there a specification of the statistical method(s) used to develop the QSAR (including details of any software packages used)?	Yes, linear regression / discriminant analysis
	5.4) Is the QSAR associated with basic statistics for its goodness-of-fit to the training set? (e.g. r^2 values and standard error of the estimate in the case of regression models)	Yes
	5.5) a) Is the QSAR associated with any statistics based on cross-validation or resampling? b) If yes, is the number or samples used indicated?	Yes
6) Predictivity (External validation)	**6.1)** Does application of the appropriate statistical method(s) to the training set result in the same (Q)SAR model?	Not done
	6.2) Is there any information to indicate that the (Q)SAR has been validated previously, using a test set that is independent of the training set?	In progress (8/4/04)
	6.3) If an external validation has been performed, is the following information available: a) the number of test structures? b) the identities of the test structures? c) the approach for selecting the test structures? d) the statistical analysis of the predictive performance of the model? (e.g. including sensitivity, specificity, and positive and negative predictivities for classification models) e) a comparison of the predictive performance of the model against previously-defined quantitative performance criteria?	Yes 160 chemicals / pharmaceuticals Manuscript submitted for publication (8/4/04)

ANNEX 6
ECOSAR

Etje Hulzebos
National Institute of Public Health and Environment (RIVM)
Utrecht
Netherlands

TABLE OF CONTENTS

1. INTRODUCTION

291. Etje Hulzebos (National Institute of Public Health and Environment, NL) has applied Setubal principles 1-6 to ECOSAR. The work was carried out in consultation with Ruth Poshumus.

292. A summary (in tabular form, Appendix 1) of the extent to which the Setubal principles have been met in ECOSAR are attached.

293. A more detailed explanation following the items in the Appendix 1 is reported below.

2. DESCRIPTION OF ECOSAR

294. ECOSAR is developed and maintained by the US EPA as a tool for making quantitative effect assessments for aquatic organisms, to identify possible concerns for New Chemicals submitted to the EPA under the premanufacture notification (PMN) process.

295. ECOSAR predicts the toxicity of chemicals to aquatic organisms such as fish, daphnids and algae by using QSARs that are based on log Kow as the sole descriptor. The program estimates acute, and in some cases chronic, toxicity. ECOSAR allows access to over 100 QSARs developed for 46 chemical classes according to the ECOSAR manual (1). In a more recent user's guide for ECOSAR (version 0.99d), more than 50 classes are stated to be available (2). The latest version of ECOSAR is 0.99g.

296. The QSARs are mostly empirically derived. ECOSAR categorises the chemical in one of the available classes. It shows the acute and chronic aquatic toxicity prediction for which a QSAR is available. If the program cannot identify a specific class, ECOSAR categorises the chemical as a neutral organic and gives the predictions based on this. ECOSAR categorises a chemical in more than one chemical class when more than one active chemical group is identified. Then ECOSAR provides predictions for all classes.

297. ECOSAR is a user-friendly program and can be run with a CAS number or with a description of the chemical structure using the Simplified Molecular Input Line Entry System (SMILES). The user either enters the Log Kow of the chemical or ECOSAR generates it. After giving a prediction, ECOSAR warns when the estimated L(E)C50 or chronic value for a chemical is above the water solubility of the substance. ECOSAR generates water solubility from the log Kow. In addition, the upper limits are given for the log Kow above which values the predictions are not longer reliable.

298. ECOSAR is not suitable for inorganics, dyes, polymers and for substances for which the SMILES notation is difficult to derive. To run ECOSAR with charged ions, user modifications are necessary (2). For surfactants, QSARs are available but they are not based on Log Kow as a descriptor.

3. APPLICATION OF THE SETUBAL PRINCIPLES

3.1. Defined endpoint (Principle 1)

299. ECOSAR predicts defined endpoints as required by the US EPA regulatory framework, such as acute L(E)C50 and long-term NOECs for fish, daphnids and algae.

3.2. Defined algorithm (Principle 2)

300. The QSAR equations are based on linear regression analysis, using log Kow as the sole descriptor for predicting the L(E)C50 values (except for the class of surfactants). There is no explicit description of the chemical classes and in or exclusion rules.

3.3. Mechanistic basis (Principle 3)

301. The QSAR 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. The chemical classes are empirically derived. Nevertheless, a mechanistic basis is often apparent, e.g. reactive chemicals will be more toxic than neutral organics.

3.4. Domain of applicability (Principle 4)

302. ECOSAR will provide a prediction for all chemicals for which a SMILES notation can be made. Inorganics, dyes and polymers cannot be run.

303. The training set and the log Kow values of these chemicals are given in the manual (1). The prediction also gives the maximum limit of the log Kow value that should be used. The prediction also shows which predicted L(E)C50 values are expected to be higher than the water solubility.

3.5. Internal performance (Principle 5)

304. The ECOSAR manual gives the linear regression equations, r^2 of these regressions, and the numbers of chemicals used to derive the regression lines.

305. The main criteria used by Hulzebos and Posthumus (3) for good QSARs were:

a. number of chemicals in the training set > 4 and

b. r^2 should at least be 0.7.

306. This resulted in 96/112 (78%) unreliable and 27/123 (22%) reliable QSARs for 123 QSARs. These reliable QSARs and classes are given in Table 1.

3.6. External validation for predictivity (Principle 6)

307. An external validation was done in 1993 comparing the outcome of New Chemical notifications in the EU with the predictions of the experts of US EPA, in which the use of ECOSAR was included (4).

308. For acute fish toxicity, 82% of the predictions were within a factor of 10 of the experimental values, whereas for acute Daphnia toxicity, this percentage was 71%. Hulzebos and Posthumus (3) verified ECOSAR predictions for 70 chemicals from the IPCS database and some EU New notified chemicals. For three classes, neutral organics, esters and phenols, a sufficient number of chemicals ($\geq$ 25) were available to validate these classes. At least 63% of the chemicals falling into these three classes were predicted within a factor of 10 of the experimental values.

4. CONCLUSIONS ON THE SETUBAL PRINCIPLES AND THEIR APPLICABILITY

4.1. Conclusions about ECOSAR

309. ECOSAR predicts aquatic toxicity for the standard species and endpoints needed for regulatory purposes. The program can assess organic chemicals for which a SMILES notation can be derived. The computerised program and the manual together are very transparent considering the training set and the model statistics. For every chemical class, individual training sets and model equations are given. Each QSAR in ECOSAR needs to be checked for good modelling practice, despite the fact that all models have the same linear regression bases.

310. The 27 reliable QSARs can be validated. The other 96 QSARs are well worth developing, because so far it seems that the predictions made by the non-reliable QSARs can be as good as the predictions made by the reliable QSARs (Hulzebos and Posthumus, 3).

5. REFERENCES

ECOSAR (1996). Technical reference manual. See website:
http://www.epa.gov/oppt/newchems/sarman.pdf

ECOSAR User manual (1998). See website:

http://www.epa.gov/oppt/newchems/manual.pdf

Hulzebos EM & Posthumus, R. (2003). (Q)SARs: Gatekeepers against risk on chemicals?, *SAR and QSAR in Environmental Research* **14**, 285-316.

OECD (1994). US EPA/EC Joint project on the evaluation of (quantitative) structure activity relationships. OECD Technical Report No 88. Paris, France.

Table 1 Reliable QSARs and classes in ECOSAR, according to the criteria of Hulzebos and Posthumus

Chemical Class	**QSAR**
Methacrylates:	96h LC50 fish
Aliphatic amines	96h LC50 fish 48h EC50 Daphnids 96h algae
Aromatic amines	96h LC50 fish 14d LC50 fish
Epoxides, mono	14d LC50 fish
Esters	96h LC50 fish
Hydrazines	96h LC50 fish
Hydrazines, semicarbazide, aryl, ortho/meta/para substituted	6h EC50 algae
Ketones, aliphatic, di	96h LC50 fish 48h EC50 Daphnids
Neutral organics	96h LC50 fish (FW+SW) 14d LC50 fish ChV fish 28d BCF fish 48h LC50 Daphnids 96h LC50 shrimp ChV Daphnids 96h algae ChV algae
Phenols	96h LC50 fish ChV Fish ChV Daphnids
Dinitrophenols	48h EC50 Daphnids
Urea, substituted	4h EC50 algae

6. APPENDICES

APPENDIX 1 SUMMARY REPORT OF THE APPLICATION OF THE SETUBAL PRINCIPLES

PRINCIPLE	CONSIDERATIONS	YES / NO
1) Defined endpoint	**1.1)** Does the model have a clearly defined scientific purpose? (i.e does it make predictions of a clearly defined physicochemical, biological or environmental effect?)	Yes
	1.2) Does the model have the potential 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)?	Yes
	1.3) Is information given about important experimental conditions that affect the measurement and therefore the prediction (e.g. sex, species, temperature, exposure period)	Yes
	1.4) Are the units of measurement of the endpoint given?	Yes
2) Defined algorithm	**2.1)** In the case of a SAR, is there an explicit description of the substructure, including an explicit identification of its substituents?	No
	2.2) In the case of a QSAR, is the equation explicitly defined, including definitions of all descriptors used[3]?	Yes
3) Mechanistic basis	**3.1)** In the case of a SAR, is there a description of the molecular events that underlie the reactivity of the molecule? (e.g. a description of how substructural features could act as nucleophiles or electrophiles, or form part or all of a receptor-binding region)	No
	3.2) In the case of a QSAR, do the descriptors have a physicochemical interpretation that is consistent with a known mechanism (of biological action)?	Yes
	3.3) Are any literature references cited in support of the purported mechanistic basis of the (Q)SAR?	Yes
4) Domain of applicability	**4.1)** In the case of a SAR, is the substructure associated with any inclusion and/or exclusion rules on its applicability to	Unknown

	groups of chemicals?	
	4.2) In the case of a SAR, is the substructure associated with rules regarding the modulatory effects of the substructure's molecular environment?	Unknown
	4.3) In the case of a QSAR, are the descriptor and response variables associated with inclusion and/or exclusion rules that define the variable ranges for which the QSAR is applicable (i.e. makes reliable estimates)?	Yes
5) Internal performance	**5.1)** Are full details of the training set given, including details of chemical names, structural formulae, CAS numbers (if available), and data for all descriptor and response variables?	Yes for chemical names. No for structural formula and Cas no.
	5.2) If the data used to the develop the model were based upon the processing of raw data (e.g. the averaging of replicate values): a) is there an adequate description of the data processing? b) are the raw data provided?	a) Yes b) Mostly, yes
	5.3) Is there a specification of the statistical method(s) used to develop the QSAR (including details of any software packages used)?	Yes
	5.4) Is the QSAR associated with basic statistics for its goodness-of-fit to the training set? (e.g. r^2 values and standard error of the estimate in the case of regression models)	R^2 is given, but not the standard error of the estimate
	5.5) a) Is the QSAR associated with any statistics based on cross-validation or resampling? b) If yes, is the number or samples used indicated?	No
6) Predictivity (External validation)	**6.1)** Does application of the appropriate statistical method(s) to the training set result in the same (Q)SAR model?	Yes
	6.2) Is there any information to indicate that the (Q)SAR has been validated previously, using a test set that is independent of the training set?	Yes
	6.3) If an external validation has been performed, is the following information available: a) the number of test structures? b) the identities of the test structures? c) the approach for selecting the test structures? d) the statistical analysis of the predictive performance of the model? (e.g. including sensitivity, specificity, and positive and negative predictivities for classification models) e) a comparison of the predictive performance of the model against previously-defined quantitative performance criteria?	a) Yes b) Yes c) Yes d) Yes e) No

ANNEX 7
BIOWIN

Theo Traas
National Institute of Public Health and Environment (RIVM)
Utrecht
Netherlands

TABLE OF CONTENTS

1. INTRODUCTION

311. Theo Traas (National Institute of Public Health and Environment, NL) has applied Setubal Principles 1-6 to BIOWIN. The work was carried out in consultation with Etje Hulzebos and Ruth Poshumus.

312. A summary (in tabular form, Appendix 1) of the extent to which the Setubal principles have been met in BIOWIN are attached.

313. A more detailed explanation following the items in the Appendix 1 is reported below.

2. DESCRIPTION OF BIOWIN

314. The Biodegradation Probability Program (BIOWIN) estimates the probability for the rapid aerobic biodegradation of an organic chemical in the presence of mixed populations of environmental micro-organisms. It also gives a first classification of the timeframe for biodegradation. The program gives results from six different models, as explained later. The results are listed individually, and the classification based on these results is given for each individual model. For example, a summary of BIOWIN output for benzene is given in Figure 1.

315. Estimates are based upon fragment constants that were developed using multiple linear and non-linear regression analyses. A discussion of the method used to derive the linear and non-linear fragment constants is presented in a journal article by Howard et al. (1). Experimental biodegradation data for the multiple linear and non-linear regressions were obtained from Syracuse Research Corporation's (SRC) data base of evaluated biodegradation data (2).

316. BIOWIN version 3 added estimates for the time required to achieve primary and ultimate biodegradation. Some modifications are made to the linear and non-linear estimates. A journal article by Boethling et al. (3) gives a complete description of the ultimate/primary methodology.

317. BIOWIN version 4 added two new predictive models for assessing a chemical’s biodegradability in the Japanese MITI (Ministry of International Trade and Industry) biodegradation test. The new models use an approach similar to the linear/non-linear regression models noted above. A journal article giving a complete description of the MITI Biodegradation models was published recently (Tunkel et al., 4).

3. APPLICATION OF THE SETUBAL PRINCIPLES

318. The check list of considerations for the application of the Setubal principles was applied to the BIOWIN program (see Appendix 1) and is elaborated in this section. Each category is discussed briefly, after which the questions are answered and discussed in some detail. The BIOWIN program has been applied to a small number of test compounds to view the output of the program. The cited supporting papers together with the help file for BIOWIN were used to answer each question in the check list. Many principles could only be addressed by going back to the original research papers.

3.1. Defined endpoint (Principle 1)

319. A problem with judging the applicability of the Setubal principles to the BIOWIN program is that the definition of 'fast biodegradation' or 'readily biodegradable' of the BIODEG and MITI models is not explained in the help file. This can only be found when going back to the supporting papers, as suggested for the BIODEG model in the electronic manual (2). For the MITI model, the manual states that the MITI tests are among the few approved as ready biodegradability test guidelines of the OECD (Organisation for Economic Cooperation and Development). What exactly constitutes 'fast biodegradation' or 'readily biodegradable' is not defined.

320. Therefore, the program delivers a precise probability for a qualitative judgement on biodegradability. One might expect a prediction of the microbial degradation rate (in d-1) with a confidence interval. Each legislative framework could then have its own scale of judging degradability as 'acceptable' or not. However, microbial degradation rates depend on far more than just the chemical structure. Additional factors include the amount of micro-organisms, taxonomic representation, viability of the biomass in the substrate, nutrient status etc.

321. *Does the model have a clearly defined scientific purpose? YES*

The objective of the model is clearly stated in the BIOWIN help text: 'estimates the probability for the rapid aerobic biodegradation of an organic chemical in the presence of mixed populations of environmental micro-organisms'.

This question can be difficult to answer without additional information on the QSARs or the program implementing it. In this specific case, it is an interesting problem that the probability predicted is *precise* while the endpoint *rapid biodegradation* is not defined clearly in the program or elsewhere.

The model is a full or partial replacement for biodegradation tests, as needed for regulatory acceptance of substance use or admission on the market. This objective is mentioned in all supporting papers in the BIOWIN help files.

3.2. Defined algorithm (Principle 2)

322. The BIOWIN models are described in detail in the supporting papers, including the fragment contribution to the prediction of the model. Problems are expected when judging the validity of the models for a specific compound (see: domain of applicability).

323. *In the case of a SAR, is there an explicit description of the substructure, including an explicit identification of its substituents? YES / NO*

The BIODEG and BIOWIN models show output where the substructures are explicitly listed with associated coefficients. The primary/ultimate biodegradation model does not do this, although a similar reasoning seems to be behind it (Boethling et al., 3).

3.3. Mechanistic basis (Principle 3)

324. One may question the requirement to use this principle. If the prediction of the model is based on empirical relationships instead of knowledge of metabolic pathways, it may still predict well. For BIOWIN where degradation is considered to be performed by a community of micro-organisms, it is perhaps not realistic to expect such a mechanistic basis.

325. *In the case of a SAR, is there a description of the molecular events that underlie the reactivity of the molecule? (e.g. a description of how substructural features could act as nucleophiles or electrophiles)? NO*

The models described in BIOWIN are not (semi)mechanistic but purely empirical, without reference to microbial metabolism. Given the fact that degradation is considered to be performed by a *community* of micro-organisms, it is perhaps not realistic to expect this.

3.4. Domain of applicability (Principle 4)

326. When using the BIOWIN model, no indication is given for which chemical structures the program is validated, or for which structures unreliable predictions may be expected. The supporting papers list the structures that were taken into account in the development of the BIOWIN models, but a warning could be given for missing structural fragments. The only warning in the electronic manual is about sulphonic acid salts. With the exception of sodium, potassium and lithium salts, BIOWIN does not detect sulphonic acid salts. When more and more complex new chemicals are introduced, the chances are that a chemical contains fragments not included in the regression analyses. Moreover, interactions between structural fragments of such chemicals are also not taken into account.

327. *In the case of a SAR, is the substructure associated with any inclusion and/or exclusion rules on its applicability to groups of chemicals? YES / NO*

For some structural features in the primary/ultimate biodegradation model, exclusion rules are mentioned (Boethling & Sabljic, 5). For the BIODEG and MITI models, this does not seem to be the case, judging from the original papers.

328. In the case of a SAR, is the substructure associated with rules regarding the modulatory effects of the substructure's molecular environment? NO

Not that I can notice in the supporting papers.

329. *In the case of a QSAR, are the descriptor and response variables associated with inclusion and/or exclusion rules that define the variable ranges for which the QSAR is applicable?*

Not applicable (?)

3.5. Internal performance (Principle 5)

330. The supporting papers report on the validity of the regression models for the training set and report the percentage of correct prediction for the training set. In the case of the MITI model, a thorough cross-validation based on a resampling strategy was followed. The supporting papers do not list the raw

data used, but these are available as tables (BIOWIN help files) or databases allowing a re-evaluation. The uncertainty in the model estimates (output) is not always reported in the supporting papers. A full specification of parameter uncertainty is necessary for improved uncertainty estimates in probabilistic risk assessment.

331. *Are full details of the training set given, including details of chemical names, structural formulae, CAS numbers (if available), and data for all descriptor and response variables? NO*

For all BIOWIN models, no specific details are given as defined above, but (partial) information is available from the authors, or is referred to in the references of the supporting papers.

332. *If the data used to develop the model were based upon the processing of raw data (e.g. the averaging of replicate values):*

a. is there an adequate description of the data processing? YES

The supporting papers adequately document the data processing.

b. are the raw data provided? YES

The raw data are listed are listed as tables in the BIOWIN help files

c. is there a specification of the statistical method(s) used to develop the QSAR (including details of any software packages used)? YES

The methods used to develop the QSARs are documented in the supporting papers, and consist of linear and non-linear models.

333. *Is the QSAR associated with basic statistics for its goodness-of-fit to the training set? (e.g. r^2 values and standard error of the estimate in the case of regression models) YES*

Basic goodness-of-fit statistics were reported for the BIODEG model (including SE of the estimate), the biodegradation model (Boethling & Sabljic, 1989) For the MITI model, the paper reports that they have been calculated but they are not reported (4).

334. *Is the QSAR associated with any statistics based on cross-validation or resampling? YES*

Internal validation is reported as percentage correctly classified for the training set and a discussion of poorly predicted chemicals for the BIODEG model. The MITI model has been subjected to a resampling strategy to optimise the predictions.

3.6. External validation for predictivity (Principle 6)

335. External validation has been performed and judged ‘adequate’. Adequate must be defined. It has two meanings: adequate with respect to the *precision* of the estimate (how far is it off the true value) but also in terms of *accuracy* (are groups consistently over- or underestimated). Although general performance is given in terms of % correct classification, the uncertainty in the model estimate requires a far more rigorous definition than is currently the case. Since we may want to use (Q)SAR estimates as input to other risk models, the model output should allow the risk assessor to carry over the uncertainty of the QSAR estimate to the receiving risk model.

336. *Does application of the appropriate statistical method(s) to the training set result in the same (Q)SAR model? YES*

For all models, a percentage correct classification is given for the training set, which is generally high (90% or more).

337. Is there any information to indicate that the (Q)SAR has been validated previously, using a test set that is independent of the training set? YES

The BIODEG and MITI models are applied to an independent validation set and report the percentage correct classification. The primary/ultimate biodegradation papers do not report an independent validation.

338. *If an external validation has been performed, did the validation exercise include:*

a. an adequate number of test structures?

This principle is hard to apply since 'adequate number' may depend on personal views. If adequate means say at least 10% of the number in the training set, then all validations for BIODEG and MITI are 'adequate'.

b. an appropriate selection of test structures?

This principle needs expert opinion and also depends on the size of a). Especially in the procedure described for the MITI database, with a large training set, this is expected. For the BIODEG models, this is probably less convincing.

c. an adequate statistical analysis of the predictive performance of the model?

Again, the meaning of 'adequate' must be defined. It has two meanings: adequate with respect to the precision of the estimate (how far is it off the true value) but also in terms of accuracy (are groups consistently over- or underestimated). Although general performance is given in terms of % correct classification, the uncertainty in the estimate requires a far more rigorous definition than is currently the case.

d. a comparison of the predictive performance of the model with previously defined acceptability criteria?

The acceptability of prediction accuracy is discussed for the MITI model by Tunkel et al. (4). Here, prediction accuracy is judged acceptable if > 80%. For the BIODEG model, a quantitative predictive capability is not defined *a priori.*

4. CONCLUSIONS ON THE SETUBAL PRINCIPLES AND THEIR APPLICABILITY

339. In most cases, the criteria were found helpful to determine if a QSAR is sufficiently well developed and validated to be considered scientifically valid, and therefore ready to be considered for regulatory use. Some definitions need to be refined, such as 'adequate' when judging validation. Ít is

advised to put more emphasis on statistical aspects (internal validation) to allow the full use of QSAR estimates in modern (probabilistic) risk assessment methods.

5. REFERENCES

Boethling, R.S. & Sabljic, A. (1989) Screening-level model for aerobic biodegradability based on a survey of expert knowledge. *Environmental Science and Technoogy* **23**, 672-679.

Boethling, R.S., Howard, P.H., Meylan, W., Stiteler, W., Beaumann, J. & Tirado N. (1994). Group contribution method for predicting probability and rate of aerobic biodegradation. *Environ. Sci. Technol*. **28**, 459-65.

Howard, P.H., Boethling, R.S., Stiteler, W.M., Meylan, W.M., Hueber, A.E., Beauman, J.A. & Larosche, M.E. (1992). Predictive model for aerobic biodegradability developed from a file of evaluated biodegradation data. *Environmental Toxicology and Chemistry* **11**, 593-603.

Howard, P.H., Hueber, A.E. & Boethling, R.S. (1987). Biodegradation data evaluation for structure/biodegradability relations. *Environmental Toxicology and Chemistry* **6**, 1-10.

Tunkel, J., Howard, P.H., Boethling, R.S., Stiteler, W. & Loonen, H. (2004). Predicting Ready Biodegradability in the MITI Test. *Environmental Toxicology and Chemistry*. Accepted for publication.

Figure 1 BIOWIN results for benzene

```
---------- BIOWIN v4.00 Results ----------
Benzene
SMILES : c1ccccc1
MOL FOR: C6 H6
MOL WT : 78.11
------------------------------------------------------------------------

A Probability ≥ 0.5 indicates --> Biodegrades Fast
A Probability Less Than 0.5 indicates --> Does NOT Biodegrade Fast
   Linear Model Prediction    :  Biodegrades Fast   (1.0296)
   Non-Linear Model Prediction:  Biodegrades Fast      (0.9999)

Result Classification: 5.00 -> hours  4.00 -> days    3.00 -> weeks
(Primary & Ultimate)    2.00 -> months    1.00 -> longer
   Ultimate Biodegradation Timeframe:  Weeks-Months     (2.4406)
   Primary  Biodegradation Timeframe:  Days-Weeks   (3.3922)

A Probability ≥ 0.5 indicates --> Readily Degradable
A Probability Less Than 0.5 indicates --> NOT Readily Degradable
   MITI Linear Model Prediction    :  Readily Degradable  (0.5291)
   MITI Non-Linear Model Prediction:  Readily Degradable  (0.7294)
```

6. APPENDICES

APPENDIX 1 SUMMARY REPORT OF THE APPLICATION OF THE SETUBAL PRINCIPLES

PRINCIPLE	CONSIDERATIONS	YES / NO
1) Defined endpoint	**1.1)** Does the model have a clearly defined scientific purpose? (i.e does it make predictions of a clearly defined physicochemical, biological or environmental effect?)	Yes
	1.2) Does the model have the potential 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)?	Yes
	1.3) Is information given about important experimental conditions that affect the measurement and therefore the prediction (e.g. sex, species, temperature, exposure period)	Yes, MITI test Others no
	1.4) Are the units of measurement of the endpoint given?	Yes
2) Defined algorithm	**2.1)** In the case of a SAR, is there an explicit description of the substructure, including an explicit identification of its substituents?	Yes (see text)
	2.2) In the case of a QSAR, is the equation explicitly defined, including definitions of all descriptors used[3]?	Yes
3) Mechanistic basis	**3.1)** In the case of a SAR, is there a description of the molecular events that underlie the reactivity of the molecule? (e.g. a description of how substructural features could act as nucleophiles or electrophiles, or form part or all of a receptor-binding region)	No
	3.2) In the case of a QSAR, do the descriptors have a physicochemical interpretation that is consistent with a known mechanism (of biological action)?	No
	3.3) Are any literature references cited in support of the purported mechanistic basis of the (Q)SAR?	Yes

4) Domain of applicability	**4.1)** In the case of a SAR, is the substructure associated with any inclusion and/or exclusion rules on its applicability to groups of chemicals?	*Yes/No* (see text)
	4.2) In the case of a SAR, is the substructure associated with rules regarding the modulatory effects of the substructure's molecular environment?	No
	4.3) In the case of a QSAR, are the descriptor and response variables associated with inclusion and/or exclusion rules that define the variable ranges for which the QSAR is applicable (i.e. makes reliable estimates)?	Yes
5) Internal performance	**5.1)** Are full details of the training set given, including details of chemical names, structural formulae, CAS numbers (if available), and data for all descriptor and response variables?	No
	5.2) If the data used to the develop the model were based upon the processing of raw data (e.g. the averaging of replicate values): a) is there an adequate description of the data processing? b) are the raw data provided?	a) Yes b) Yes
	5.3) Is there a specification of the statistical method(s) used to develop the QSAR (including details of any software packages used)?	Yes
	5.4) Is the QSAR associated with basic statistics for its goodness-of-fit to the training set? (e.g. r^2 values and standard error of the estimate in the case of regression models)	Yes
	5.5) a) Is the QSAR associated with any statistics based on cross-validation or resampling? b) If yes, is the number or samples used indicated?	Yes
6) Predictivity (External validation)	**6.1)** Does application of the appropriate statistical method(s) to the training set result in the same (Q)SAR model?	Yes
	6.2) Is there any information to indicate that the (Q)SAR has been validated previously, using a test set that is independent of the training set?	Yes
	6.3) If an external validation has been performed, is the following information available: a) the number of test structures? b) the identities of the test structures? c) the approach for selecting the test structures? d) the statistical analysis of the predictive performance of the model? (e.g. including sensitivity, specificity, and positive and negative predictivities for classification models) e) a comparison of the predictive performance of the model against previously-defined quantitative performance criteria?	*a) Yes* *b) Yes* *c) See text* *d) No* e) Yes

ANNEX 8
DEREK FOR WINDOWS

Etje Hulzebos
National Institute of Public Health and Environment (RIVM)
Utrecht
Netherlands

TABLE OF CONTENTS

1. INTRODUCTION

340. Etje Hulzebos (National Institute of Public Health and Environment, NL) has applied Setubal principles 1-6 to DEREK for Windows. The work was carried out in consultation with Lidka Maslankiewicz.

341. A summary (in tabular form, Appendix 1) of the extent to which the Setubal principles have been met in DEREK is attached.

342. A more detailed explanation following the items in the Appendix 1 is reported below.

2. DESCRIPTION OF DEREK FOR WINDOWS 6.0

343. DEREKfW (Deductive Estimation of Risk from Existing Knowledge for Windows) is a rule-based expert system. The expert system is based on knowledge, collected by experts, on toxicological effects of certain types of chemicals for specific endpoints. DEREKfW predicts the toxicological properties of chemicals based on their molecular structure. The system is biased towards structures showing effects.

344. DEREKfW for Windows contains about 303 structural fragments, called toxicophores, and rules that identify mostly adverse effects (1-3). In general, effects are predicted for both humans and mammals if differences can be distinguished.

345. DEREKfW can be searched with ISIS draw or MOL files (2-dimensional structures). Only molecules up to 64 atoms can be screened and the program is not suitable for polymers, but can assess some metals (2, 4).

346. DEREKfW is a user-friendly program. The program alerts the user to the presence of structural alerts, gives references for the predictions and, if available, chemicals that show the effect are used as positive examples. It also gives the domains of the alerts. Kinetic parameters, e.g. for (un)expected dermal absorption, are included as well. Negative predictions are based on kinetic parameters that give evidence for low availability. For example, when low dermal absorption is predicted, skin irritation and sensitisation are expected to be ‘improbable’. No negative structural alerts are available for assessing the human toxicological endpoints. Structural alerts embedded in structures outside the domain of the alert are predicted as ‘no alert available’, and are predicted negative. The user can judge the prediction based on the data given, so the prediction can be considered transparent. However, the way in which the reasoning rules are presented is not always transparent. DEREKfW gives qualitative outcomes, presented in the form of eight levels of likelihood: improbable, doubtful, equivocal, plausible, probable and certain for 36 human toxicological endpoints.

347. DEREK is a commercial program. The information in the program that is visible to the user depends on the licence agreement. The predictions are shown as described above. In addition to the prediction part, the program also has an editor part. This editor part shows all the alerts of DEREK for all endpoints. In this part, tools are provided for users to add their own in-house alerts.

3. APPLICATION OF THE SETUBAL PRINCIPLES

3.1. Defined endpoint (Principle 1)

348. DEREKfW contains 36 human health endpoints (see Table 1). The definition of these endpoints needs to be retrieved from the references, which are shown when the chemical is predicted positive for that endpoint. DEREKfW does not define endpoints.

3.2. Defined algorithm (Principle 2)

349. DEREKfW predicts effects by using a number of structural alerts, varying from one to 77, depending on the toxicological endpoint (see Table 1).

350. The program gives the domain of the alert, such as adjacent groups that do not interfere with the activity of the structural alert. Other adjacent groups than given will give the prediction 'no alert shown'.

3.3. Mechanistic basis (Principle 3)

351. DEREKfW is mechanistically and empirically based. References are available for all predictions directly associated with a structural alert. The predictions that are based on kinetic properties are usually stand-alone.

3.4. Domain of applicability (Principle 4)

352. Molecules up to 64 atoms can be screened, and some metals can be assessed, but the program is not suitable for polymers. No further inclusion or exclusion rules on chemical types are given.

3.5. Internal performance (Principle 5)

353. The structural alerts in DEREKfW are proposed by an Expert Group. No further details are known.

3.6. External validation for predictivity (Principle 6)

354. For the endpoints irritation, sensitisation, mutagenicity, carcinogenicity and teratogenicity, external validations have been performed (see Table 2).

355. Hulzebos and Posthumus (5) ran DEREK with 70 chemicals. Barratt and Langowski (6) used the German BgVV list of contact allergens for validating and updating DEREK. Zinke et al (7) used the same database to validate DEREK for the sensitisation endpoint and gave comments on which structural alerts work well, which need to be adapted and which might need to be left out. Cariello et al. (4) validated DEREKFW for mutagenicity. Pearl et al. (8) validated DEREKfW with experimental results from Ames, carcinogenicity and teratogenicity tests applied to 123 drugs and 516 non-drugs. Benigni (9) and Benigni

and Zito (10) compared the results of DEREK predictions with the outcome of the NTP carcinogenicity tests, among other predictions. The results of the DEREK validations are given in Table 2.

4. CONCLUSIONS ON THE SETUBAL PRINCIPLES AND THEIR APPLICABILITY

4.1. Conclusions about DEREK

356. DEREKfW focuses on positive structural alerts, which have been identified by experts. DEREKfW is limited in that it identifies only 'activating' fragments, meaning that negative predictions are mostly based on the lack of structural alerts (Pearl et al. 2001), except for some predictions based on kinetic properties. Only qualitative outcomes are provided. The endpoints sensitisation, mutagenicity and carcinogenicity seem to be best developed as a considerable number of structural alerts associated with the effect are available. For sensitisation in particular, validation of and commenting on each of the structural alerts has already been carried out (7). For mutagenicity, and possibly for carcinogenicity, this type of work could also be performed.

5. REFERENCES

Barratt MD and Langowski JJ (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.Sciences* **39**, 294-298.

Benigni R. (1997). The first US National Toxicology Program exercise on the prediction of rodent carcinogenicity: definitive results. *Mutation Research* **387**, 35-45.

Benigni R. and Zito R. (2004). The second National Toxicology Program comparative exercise on the prediction of rodent carcinogenicity: definitive results. *Mutation Research* **566**, 49-63.

Cariello, N.D., Wilson, J.D., Britt, B.H., Wedd, D.J., Burlinson, B., Gombar, V. (2002). Comparison of the computer programs DEREKFW and TOPKAT to predict bacterial mutagenicity, *Mutagenesis* **17**, 321-329.

ECETOC (2003). (Q)SARs: Evaluation of the commercially available software for human health and environmental endpoints with respect to chemical management application. Technical report No 89. Brussels, Belgium.

Greene, H., Judson, H.N., Langowski, J.J., and Marchant, C.A. (1999). Knowledge-based expert systems for toxicity and metabolism prediction: DEREKfW, StAR and METEOR, *SAR and QSAR in Environmental Research* **10**, 299-314.

Hulzebos EM and Posthumus R. (2003). (Q)SARS: Gatekeepers against risk on chemicals? *SAR and QSAR in Environmental Research* **14**, 285-316.

LHASA (2001). DEREKfW TM for windows version 5.0. LHASA Limited, School of chemistry, University of Leeds, UK.

Pearl, G.M., Livingstone-Carr & Durham, S.K. (2001). Integration of Computational Analysis as a sentinental tool in toxicological assessments. *Current Topics in Medicinal Chemistry* **1**, 247-255.

Sanderson, D.M. and Earnshaw, C.G. (1991). Computer prediction of possible toxic action from chemical structure; The DEREKfW system. *Human & Experimental Toxicology* **10**, 261-273.

Zinke, S., Gerner, I. & Schlede, E. (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.

Table 1 Endpoints and structural alerts in DEREKfW

Toxicological endpoint	Number of structural alerts
Alpha-mu-globulin nephropathy	3
Anaphylasis	1
Anticholinoesterase activity	2
Bladder urothelial hyperplasia	1
Carcinogenicity	46
Cerabral oedema	1
Chloracne	3
Cumulative effects on white cell count type effect	1
Cyanide type effect	1
Developmental toxicity	3
Genotoxicity	1
Hepatotoxicity	2
High acute toxicity	4
Irritation of the eye and respiratory track	1
Irritation of eye	3
Irritation of gastrointestinal track	2
Irritation of respiratory track	2
Irritation of skin and eye	9
Irritation of skin, eye and respiratory track	16
Lachrymation	1
Methaemoglobinaemia	1
Mutagenicity	77
Neurotoxicity	6
Occupational asthma	1
Oestrogenicity	4
Peroxisome proliferation	7
Photoallergenicity	6
Pulmonar toxicity	1
Respiratory sensitisation	13
Skin sensitisation	61
Teratogenicity	5
Testicular toxicity	1
Thyroid toxicity	14
Uncoupler of oxidative phosphorylation	1
Total	303

Table 2 Results of external validation exercises on DEREK

	False negatives		**Specificity**		**Sensitivity**		**False positives**		**Predictivity**		**N**	**Ref**
		%		%		%		%		%		
Irritation	23/24	96	26/27	96	1/24	4	0/27	0	27/51	53	51	1
Sensitisation	5/8	63	17/18	94	3/8	37	1/18	6	20/26	77	26	1
	12/83	14			71/83	85					83	2
	253/403	63	541/636	85	150/403	37	95/636	15	691/1039	67	1039	3
Genotoxicity	1/10	10	25/34		9/10	90	9/34	26	34/44	77	44	1
	44/82	54	226/327	69	38/82	46	101/327	31	264/409	65	409	4
		6						24		70	516	5
		8						31		61	123	5
Carcinogenicity	2/16	13	8/13		14/16	87	5/13	38	22/29	76	29	1
		13		40		75		30		57	142	5
										58.8	44	6
										43	30	7
Teratogenicity		28						0		72	34	5

FOOTNOTES

False negatives = ratio of predicted negative to total number of experimental positives

False positives = ratio of predicted positives to total number of experimental negative

Specificity = ratio of correctly predicted negatives to total number of experimental negatives

Sensitivity = ratio of correctly predicted positives to total number of experimental positives

Predictivity = ratio of correctly predicted positives and negatives to the total number of predicted compounds

The above definitions are based on ECETOC Technical Report 89 (11)

References used in the table:

Hulzebos and Posthumus, 2003
Barratt and Langowski, 1999
Zinke et al. (2002)
Cariello et al. (2002)
Pearl et al. (2001)
Benigni (1997)
Benigni and Zito (2004)

6. APPENDICES

APPENDIX 1 SUMMARY REPORT OF THE APPLICATION OF THE SETUBAL PRINCIPLES

PRINCIPLE	CONSIDERATIONS	YES / NO
1) Defined endpoint	**1.1)** Does the model have a clearly defined scientific purpose? (i.e does it make predictions of a clearly defined physicochemical, biological or environmental effect?)	Yes
	1.2) Does the model have the potential 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)?	Yes, for classification and labelling
	1.3) Is information given about important experimental conditions that affect the measurement and therefore the prediction (e.g. sex, species, temperature, exposure period)	Partly, references are often provided, which describe the test methods used
	1.4) Are the units of measurement of the endpoint given?	No, the predictions are qualitative
2) Defined algorithm	**2.1)** In the case of a SAR, is there an explicit description of the substructure, including an explicit identification of its substituents?	Yes
	2.2) In the case of a QSAR, is the equation explicitly defined, including definitions of all descriptors used[3]?	Not applicable
3) Mechanistic basis	**3.1)** In the case of a SAR, is there a description of the molecular events that underlie the reactivity of the molecule? (e.g. a description of how substructural features could act as nucleophiles or electrophiles, or form part or all of a receptor-binding region)	Yes, in the reasoning rules and in the references
	3.2) In the case of a QSAR, do the descriptors have a physicochemical interpretation that is consistent with a known mechanism (of biological action)?	
	3.3) Are any literature references cited in support of the purported mechanistic basis of the (Q)SAR?	Yes
4) Domain of applicability	**4.1)** In the case of a SAR, is the substructure associated with any inclusion and/or exclusion rules on its applicability to	Yes

	groups of chemicals?	
	4.2) In the case of a SAR, is the substructure associated with rules regarding the modulatory effects of the substructure's molecular environment?	Yes, domains for SAR itself. No other active groups are taken into account.
	4.3) In the case of a QSAR, are the descriptor and response variables associated with inclusion and/or exclusion rules that define the variable ranges for which the QSAR is applicable (i.e. makes reliable estimates)?	
5) Internal performance	**5.1)** Are full details of the training set given, including details of chemical names, structural formulae, CAS numbers (if available), and data for all descriptor and response variables?	Partly, for non-confidential data
	5.2) If the data used to develop the model were based upon the processing of raw data (e.g. the averaging of replicate values): a) is there an adequate description of the data processing? b) are the raw data provided?	It is several experts judgement a) in the reasoning rules b) partly
	5.3) Is there a specification of the statistical method(s) used to develop the QSAR (including details of any software packages used)?	Not applicable
	5.4) Is the QSAR associated with basic statistics for its goodness-of-fit to the training set? (e.g. r^2 values and standard error of the estimate in the case of regression models)	No, expert judgement
	5.5) a) Is the QSAR associated with any statistics based on cross-validation or resampling? b) If yes, is the number or samples used indicated?	Not applicable
6) Predictivity (External validation)	**6.1)** Does application of the appropriate statistical method(s) to the training set result in the same (Q)SAR model?	Not applicable, but some of the expert judgement used is generally accepted
	6.2) Is there any information to indicate that the (Q)SAR has been validated previously, using a test set that is independent of the training set?	Yes, see Table 2.
	6.3) If an external validation has been performed, is the following information available: a) the number of test structures? b) the identities of the test structures? c) the approach for selecting the test structures? d) the statistical analysis of the predictive performance of the model? (e.g. including sensitivity, specificity, and positive and	See Table 1 a) yes b) sometimes c) sometimes

	negative predictivities for classification models) e) a comparison of the predictive performance of the model against previously-defined quantitative performance criteria?	d) not applicable e) no

ANNEX 9
(Q)SAR MODELS FOR SKIN SENSITISATION IN DEREKFW VERSION 7.00

Ms Grace Patlewicz
Safety and Environmental Assurance Centre (SEAC)
Unilever Colworth
Colworth House
Sharnbrook
Bedford MK44 1LQ
UK

TABLE OF CONTENTS

1. INTRODUCTION

357. Grace Patlewicz (SEAC, Unilever Colworth, UK) has retrospectively applied, under the terms of a JRC contract, Principles 1-6 to the (Q)SAR models for skin sensitization in DEREK for Windows (1), as published in the following references and elsewhere:

1. Greene, N., Judson, N.P., Langowski, J.J., Marchant, C.A. (1999). Knowledge-based expert systems for toxicity and metabolism prediction: DEREKfW, StAR and METEOR. *SAR and QSAR in Environmental Research* **10**, 299-314.

2. Sanderson, D.M., Earnshaw, C.G. (1991). Computer prediction of possible toxic action from chemical structure; The DEREK system. *Human & Experimental Toxicology* **10**, 261-273.

3. Zinke, S., Gerner, I. & Schlede, E. (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.

358. A summary of the extent to which the Setubal principles have been met in all the QSAR models of the above papers is given in tabular form (Appendix 1)

359. A more-detailed explanation (when necessary) following the items in Appendix 1 is reported below.

2. EXPLANATION OF THE DEREK SYSTEM

360. DEREKfW 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, but its main strengths lie in the areas of mutagenicity, carcinogenicity and skin sensitisation.

361. The expert knowledge incorporated into the DEREKfW system originated from Sanderson and Earnshaw (2). These workers identified a series of 'structural alerts' associated with certain types of toxic activity.

362. The DEREK knowledge base was written, developed and continues to be enhanced by LHASA (Logic and Heuristics Applied to Synthetic Analysis) 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 DEREKfW system and the evolution of its toxicity knowledge base. The development of rules is carried out in collaboration with these organizations, using the software through a committee called the DEREK collaborative group. This consists of toxicologists who represent LHASA Ltd and customers who meet at regular intervals to give advice and guidance on the rule development work and predictions made by the program. The rule

development process is a continuous cycle of literature reviews, coding, validation and refinement. The performance of rules is explored using published datasets. In addition, comments and suggestions from members of the Collaborative group are used to refine the rules in the systems. DEREK rules describe generalised structure-activity relationships and do not record internally the specific chemical structures on which they are based. It is therefore 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. Hence DEREK has benefited from experience and knowledge that is not always in the public domain.

363. All the rules in DEREK are based either on hypotheses relating to mechanisms of action of a chemical class or on observed empirical relationships, the ideas for which come from a variety of sources. Information used in the development of rules for DEREK includes published data and ideas along with suggestions from toxicological experts in industry, regulatory bodies and academia.

364. The rules in DEREK fail to generate metabolites explicitly (the sister product METEOR is required for this), but they do account for some of the various metabolic pathways that chemicals can undergo *in vivo* and *in vitro*. This is limited to those classes of chemicals where mechanisms of interaction with biological systems are adequately understood and where the metabolite is easily identified (3,4).

365. The program has a graphical interface for the input of structures and the display of results. Structures can be entered using an ISIS/Draw™ sketchpad or imported as mol files using other chemical drawing packages, such as ChemDraw™. Batch processing of many structures is also possible through the use of sd files.

366. The toxicity predictions made by DEREK are the result of two processes. The program checks whether any alerts in the knowledge base match toxicophores in the query structure. The reasoning engine then assesses the likelihood of a structure being toxic. There are 9 levels of confidence: certain, probable, plausible, equivocal, doubted, improbably, impossible, open, contradicted. The reasoning model considers the following information:

a. The toxicological endpoint

b. The alerts that match toxicophores in the query structure

c. The physicochemical property values calculated for the query structure

d. The presence of an exact match between the query structure and a supporting example within the knowledge base

367. For skin sensitisation and photoallergenicity, DEREK uses a calculation of skin permeability, which is estimated by Log Kp derived from the Log P (octanol/water partition coefficient) value and Molecular weight. DEREK uses an estimated calculation of the Log P developed by Moriguchi (5). A Clog P (BioByte Corp, USA) plug in can be used to override the Moriguchi calculation of Log P. Human log Kp values are calculated from the molecular weight and log P values of a chemical by using the Potts and Guy equation (6). This equation is derived from a data set of ninety three chemicals with a molecular weight range of 18 to >750, and a log P range of -3 to +6.

368. The rules for skin sensitisation are given in Appendix 2.

369. DEREKfW displays the alerts that match a query structure as a hierarchy called the prediction tree. The prediction tree will include the toxicity endpoint, the species and reasoning outcome for that

endpoint, the number and name of the alerts, 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. This will be only a subset of the references used rather than an exhaustive list, and the example compounds in these references are not an exhaustive list. DEREK is supplied with a small number of examples provided to support all the skin sensitisation and photoallergenicity alerts and a small number of alerts relating to other endpoints. The database does not contain all the examples available in the public domain nor the total number of compounds used as a basis for the alert.

370. DEREKfW is effectively an archive of current knowledge of structure-toxicity relationships, containing some 312 alerts in version 7.00. Only molecules of up to 64 atoms can be screened, and the program is not suitable for polymers.

371. DEREKFW can be used where a qualitative outcome is sufficient for the prediction needed. DEREKFW is not an appropriate tool where a measure of potency is required, as is the case in a risk assessment. DEREKFW identifies structural alerts for a potential effect including adjacent atoms but more remote fragments are not taken into account. The training set for the structural alerts is not provided. Only a handful of example compounds with positive results are given. This depends on the structural alert and may vary from 0-7. A subset of the reasoning and literature references for the structural alert are provided. DEREKfW structural alerts are based on expert judgements underlined with scientific data. Therefore every structural alert would need to be assessed against the Setubal principles. This would require sourcing of the original data (in some cases proprietary) used to derive the alert. No negative structural alerts are available for assessing skin sensitisation.

3. APPLICATION OF THE SETUBAL PRINCIPLES

3.1. Defined endpoint (Principle 1)

372. DEREKfW is able to make a qualitative prediction of skin sensitisation. It is unable to provide any information on the potency of a skin sensitiser. The skin sensitisation knowledge encoded within DEREK includes both public and proprietary data generated through a number of different test methods. Older alerts are based on guinea pig data, whereas more recent alerts have been based on data generated in the Local Lymph Node Assay. Information about the experimental conditions is only given in the references associated with a given alert. Since only a subset of these are fully referenced, the quality of the data used in the derivation of an alert cannot be fully verified. Version 7.00 of DEREKfW contains 61 alerts for skin sensitisation (Table 1).

3.2. Defined algorithm (Principle 2)

373. DEREKfW 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.

374. 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.

3.3. Mechanistic basis (Principle 3)

375. 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. 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.

3.4. Domain of applicability (Principle 4)

376. DEREKfW 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. Physical properties (Log P and MW) are used to limit the domain for skin sensitization, by accounting for skin permeability (where dermal absorption is relevant)

377. DEREKfW has limited means of flagging which chemistries are covered in the rulebase and which are not. It is known that up to 64 atom chemicals can be screened. The program is not suitable for polymers. DEREKfW identifies structural alerts including adjacent atoms but more remote fragments are not taken into account. There are no negative alerts for skin sensitisation.

3.5. Internal performance (Principle 5)

378. DEREKfW does not give full details of the training data used to develop an alert. Only a subset of the references used to develop a specific alert are provided. A handful of example chemicals supporting the alerts for sensitisation are provided. For the end user, there is no means of determining what specific data were used for the development of an alert.

3.6. External validation for predictivity (Principle 6)

3.6.1. Validation by Zinke et al. (7)

379. An external validation for skin sensitization, using the BGVV database, was performed by Zinke *et al*. (7). The BGVV database includes 1039 chemicals that have reliable data for the assessment of skin sensitising potential. The results (Table 2) 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*. (7) gave comments on which structural alerts worked well, which needed to be adapted, and which might need to be left out.

3.6.2. Validation by Seaman et al (8)

380. A total of 78 chemicals which underwent testing using the LLNA to identify moderate and severe skin sensitisers were also evaluated by DEREK by Seaman *et al* (8). They obtained a concordance of 59%, a sensitivity of 79% and a specificity of 47% (Table 3). 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. The combined DEREK and LLNA data is summarised in Table 3. 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.

3.6.3. Validation using 89 chemicals from Henkel

381. A total of 89 compounds (mostly aromatic amines) taken from Henkel were evaluated using DEREK v 3.6.0. 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 (9).

3.6.4. Validation using 80 chemicals from IUCLID

382. 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. The results (Table 4) indicated a concordance of 62.5%, a sensitivity of 62.5%, and a specificity of 62.5% (10).

4. KEY POINTS ABOUT DEREK

1. DEREKfW is essentially a knowledge archive of structure-toxicity relationships.

2. DEREKfW is limited in that it identifies only ‘activating’ fragments, meaning the negative prediction is based solely on the lack of structural alerts.

3. Only qualitative outcomes are provided, no measure of potency is provided.

4. Training sets of chemicals containing these structural alerts are not provided.

5. DEREKfW 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.

6. 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.

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

8. DEREKfW 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.

9. Development of DEREKfW is incremental, focussing on each chemical class in turn.

10. DEREKfW 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.

5. CONCLUSIONS ON THE SETUBAL PRINCIPLES AND THEIR APPLICABILITY

383. In general, the principles were quite helpful to assess whether a model was sufficiently robust and validated to be considered scientifically valid and useful in regulatory programs. However despite the pricniples, judgements are subjective and open to interpretation, e.g. how much information is sufficient to satisfy the domain of applicability principle. The principles are useful starting points but additional information on the algorithms and training sets are needed to verify statistics and conduct/verify existing external validations. This would help to ensure that the Setubal principles were being applied consistently.

6. REFERENCES

Delbanco, E.H. (2002). Use of the prediction software DEREK in the Hazard assessment of raw materials. *Naunyn Schmiedeberg's Arch Pharmacol* Suppl **365**, R 639.

ECETOC Technical Report No. 89. (Q)SARs: Evaluation of the commercially available software for human health and environmental endpoints with respect to chemical management applications.

Greene, N. (2002). Computer systems for the prediction of toxicity: an update. *Advanced Drug Delivery Reviews*, **54**, 417-431.

Greene, N., Judson, H.P., Langowski, J.J., Marchant, C.A. (1999). Knowledge-based expert systems for toxicity and metabolism prediction: DEREKfW, StAR and METEOR. *SAR and QSAR in Environmental Research*, **10**, 299-314.

LHASA (2003). DEREKfW TM for windows version 7.0. LHASA Limited, School of chemistry, University of Leeds, UK.

Moriguchi, I., Hirono, S., Liu, Q., Nakagome, I., Matsushita, Y. (1992). Simple method of calculating octanol/water partition coefficient. *Chem. Pharm. Bull.* **40(1)**, 127-130.

Potts, R.O., Guy, R.H. (1992). Predicting skin permeability. *Pharmaceutical Research*, **9**, 663-669.

Sanderson, D.M., Earnshaw, C.G. (1991). Computer prediction of possible toxic action from chemical structure; The DEREKfW system. *Human & Experimental Toxicology*, **10**, 261-273.

Seaman, C.W., Guerriero, F.J., Sprague, G.L. (2001). The use of DEREK (a structure/toxicity prediction program) in the identification of skin sensitizers. *Toxicologist*, **60**, 1452.

Zinke, S., Gerner, I., Schlede, E. (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.

Table 1 Structural alerts for skin sensitization in DEREKfW version 7.0

No	*Alert Name*	Patterns	References	Examples
1	Carboxylic acid halide	1	4	4
2	Acid azide	1	4	4
3	Sulphonyl halide	1	2	1
4	Sulphonyl azide	1	1	1
5	Acid anhydride or analogue	1	6	3
6	Diacyl peroxide	1	1	2
7	Phenyl ester	1	4	4
8	Phenyl carbonate	1	1	3
9	Isocyanate	1	4	3
10	Isothiocyanate	1	1	1
11	Ring-strained amide, ester, thioamide or thioester	2	6	2
12	Thioester	1	5	3
13	Haloalkane	1	5	4
14	Alkyl sulphate or sulphonate	1	7	4
15	Activated benzene	7	13	7
16	Quinone	2	5	1
17	Hydroquinone or precursor	4	4	1
18	Catechol or precursor	1	4	3
19	Aldehyde	6	5	4
20	1,3-Diketone	1	1	3
21	alpha,beta-Unsaturated aldehyde, amide, ester, ketone, nitrile or nitro compound	13	13	5
22	Precursor of alpha,beta-unsaturated aldehyde, amide, ester or ketone	7	3	2
23	Precursor of alpha,beta-unsaturated aldehyde, amide, ester, ketone, nitrile or nitro compound	8	3	3
24	Aldehyde precursor	5	2	3
25	Enol ether	1	0	0
26	Formaldehyde donor	2	6	2
27	Aromatic primary or secondary amine	2	4	3
28	Aromatic azo compound	2	5	3
29	N-Haloimide	1	1	1
30	N-Chlorosulphonamide	1	2	1
31	Thiol or thiol exchange agent	7	22	7

32	Epoxide	5	2	2
33	Isothiazolinone	2	4	3
34	Diamine	5	14	6
35	Quaternary ammonium salt	1	2	1
36	Activated N-heterocycle	3	2	3
37	Activated pyridine, quinoline or isoquinoline	14	5	2
38	Phenol or precursor	3	13	3
39	Resorcinol or precursor	1	4	1
40	Gallate or precursor	2	2	4
41	Thiuram mono- or di-sulphide	2	6	4
42	Metal salt	8	11	4
43	Imine or alpha,beta-unsaturated imine	1	1	1
44	Thiosulphate or thiosulphonate	1	5	3
45	alpha,beta-Unsaturated sulphone	1	1	1
46	N-Nitro or N-nitroso compound	2	8	4
47	Hydrazine or precursor	1	3	2
48	Alk-2-ynyl halide, alcohol or alcohol precursor	1	0	0
49	Cyanate, cyanamide or cyanogen halide	1	3	1
50	Hydroxylamine or precursor	2	3	2
51	Allyl hydroperoxide	1	5	2
52	Squarate	3	2	2
53	Carbodiimide	1	2	2
54	Glycidyl ether, amine, ester or amide	5	5	2
55	Persulphate	2	6	2
56	Alkyl ester of phosphoric or phosphonic acid	1	4	3
57	Tin or tin compound	1	5	3
58	1,2-Dicarbonyl compound or precursor	4	7	4
59	Diazonium salt	1	4	1
60	Bay-region polycyclic aromatic hydrocarbon	1	12	2
61	Thiocyanate	1	13	3

Footnote to Table 1

Patterns is the number of different substructure fragments used in the alert description.
References is the number of references in the alert description.
Examples is the number of example compounds for the alert description.

Table 2 External validation of the DEREK skin sensitisation rulebase by Zinke et al (7)

	Predicted as positive	**Predicted as negative**	**Total**
Positive in experiment	150	253	403
Negative in experiment	85	541	636
Total	235	794	1039

Table 3 External validation of the DEREK skin sensitisation rules by Seaman et al (8)

	Predicted as Positive	**Predicted as Negative**	**Total**
LLNA positive	23	6	29
LLNA negative	26	23	49
Total	49	29	78
	LLNA negative/DEREK positive	**LLNA negative/DEREK negative**	
GPMT positive	10	5	15
GPMT negative	11	13	24
Total	21	18	39

Table 4 External validation of DEREK skin sensitisation rules using 80 chemicals from IUCLID (10)

	Predicted as positive	**Predicted as negative**	**Total**
Positive in GPMT	20	12	32
Negative in GPMT	18	30	48
Total	38	42	80

7. APPENDICES

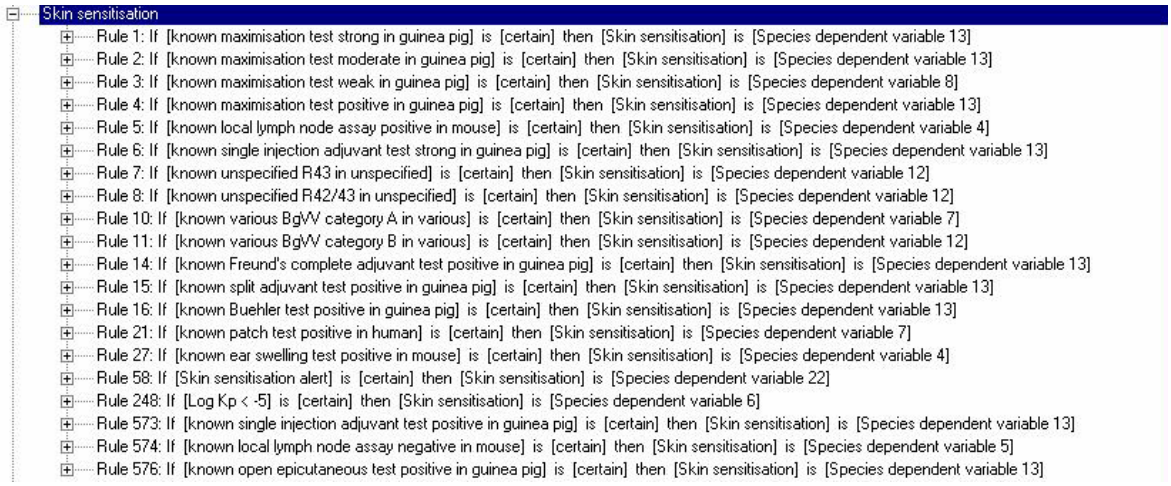
APPENDIX 1 SUMMARY REPORT OF THE APPLICATION OF THE SETUBAL PRINCIPLES

PRINCIPLE	CONSIDERATIONS	YES / NO
1) Defined endpoint	**1.1)** Does the model have a clearly defined scientific purpose? (i.e does it make predictions of a clearly defined physicochemical, biological or environmental effect?)	Yes
	1.2) Does the model have the potential 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)?	Yes, but predictions are not for a specific test method, since rules are based on all types of data including GPMT, Buehler and LLNA data.
	1.3) Is information given about important experimental conditions that affect the measurement and therefore the prediction (e.g. sex, species, temperature, exposure period)	No, but some information should be available in the references cited for a given rule.
	1.4) Are the units of measurement of the endpoint given?	No, qualitative predictions only
2) Defined algorithm	**2.1)** In the case of a SAR, is there an explicit description of the substructure, including an explicit identification of its substituents?	Yes, this is given in the alert description
	2.2) In the case of a QSAR, is the equation explicitly defined, including definitions of all descriptors used?	Not applicable
3) Mechanistic basis	**3.1)** In the case of a SAR, is there a description of the molecular events that underlie the reactivity of the molecule? (e.g. a description of how substructural features could act as	Yes, in certain cases, a reasoning is

	nucleophiles or electrophiles, or form part or all of a receptor-binding region)	given for the prediction as well as the Kp (skin permeability) value
	3.2) In the case of a QSAR, do the descriptors have a physicochemical interpretation that is consistent with a known mechanism (of biological action)?	Not applicable
	3.3) Are any literature references cited in support of the purported mechanistic basis of the (Q)SAR?	Yes, bibliographic references used in the development of the rules are given, but only a subset of those used to develop the rule
4) Domain of applicability	**4.1)** In the case of a SAR, is the substructure associated with any inclusion and/or exclusion rules on its applicability to groups of chemicals?	Yes, these are given in the comments for a given alert and in the reasoning rules, e.g. Kp
	4.2) In the case of a SAR, is the substructure associated with rules regarding the modulatory effects of the substructure's molecular environment?	Yes for the environment that is directly attached to the structural alert, but not always for more remote parts of the structure
	4.3) In the case of a (Q)SAR, are the descriptor and response variables associated with inclusion and/or exclusion rules that define the variable ranges for which the QSAR is applicable (i.e. makes reliable estimates)?	Not applicable
5) Internal performance	**5.1)** Are full details of the training set given, including details of chemical names, structural formulae, CAS numbers (if available), and data for all descriptor and response variables?	References are given but these are not exhaustive. Sometimes these can be very brief and if the alert is based on confidential data, access is not available to

		question that particular rule. No audit trail to training data for the development of a specific rule.
	5.2) If the data used to the develop the model were based upon the processing of raw data (e.g. the averaging of replicate values): a) is there an adequate description of the data processing? b) are the raw data provided?	Not applicable
	5.3) Is there a specification of the statistical method(s) used to develop the QSAR (including details of any software packages used)?	Only applicable to the algorithm for Kp prediction
	5.4) Is the QSAR associated with basic statistics for its goodness-of-fit to the training set? (e.g. r^2 values and standard error of the estimate in the case of regression models)	Not generally applicable, except to the Kp prediction
	5.5) a) Is the QSAR associated with any statistics based on cross-validation or resampling? b) If yes, is the number or samples used indicated?	Not applicable
6) Predictivity (External validation)	**6.1)** Does application of the appropriate statistical method(s) to the training set result in the same (Q)SAR model?	Not applicable
	6.2) Is there any information to indicate that the (Q)SAR has been validated previously, using a test set that is independent of the training set?	Yes, some limited external validation has been reported in the literature, and documented as part of the ECETOC Technical Report 89
	6.3) If an external validation has been performed, is the following information available: a) the number of test structures? b) the identities of the test structures? c) the approach for selecting the test structures? d) the statistical analysis of the predictive performance of the model? (e.g. including sensitivity, specificity, and positive and negative predictivities for classification models) e) a comparison of the predictive performance of the model against previously-defined quantitative performance criteria?	a) yes b) in some cases (not BgVV validation) c) yes, limited information is provided in the examples cited d) yes e) No

APPENDIX 2 SKIN SENSITISATION RULES IN DEREK



ANNEX 10
CERI BIODEGRADATION PREDICTION SYSTEM

Y. Sakuratani, K. Kasai, J. Yamada and Y. Noguchi
Chemical Management Center,
National Institute of Technology and Evaluation (NITE)
Tokyo
Japan

TABLE OF CONTENTS

1. INTRODUCTION

384. Dr Yuki Sakuratani and his colleagues at the Chemical Management Center of the National Institute of Technology and Evaluation have performed a validation exercise on the biodegradation prediction system (Expert System) developed by the Chemicals Evaluation and Research Institute (CERI).

385. Section 1 of this report describes the outcome of this validation exercise, and Section 2 describes the application of the Setubal principles to the expert system.

2. VALIDATION OF THE BIODEGRADABILITY PREDICTION SYSTEM

2.1. Description of the CERI Biodegradability Expert System

386. The biodegradation prediction system (Expert System) has been developed by the Japanese Chemicals Evaluation and Research Institute (CERI) on the basis of biodegradation data for the existing chemicals under the Japanese Chemical Substance Control Law (CSCL; 1,2). This is the prediction system for the biodegradation of chemical substances based on the OECD 301C test data and can be accessed free of charge through the CERI web-site (1).

387. The CERI Expert System predicts the biodegradability of chemicals which are classified by their molecular structures. First, the chemical substances with a molecular weight greater than 500 are separated from those below 500. Then, those with a molecular weight below 500 are categorized into 9 subgroups by their molecular structure. The chemical substances, separated into a total of 10 groups, are further classified by the type and number of subgroups to achieve degradability prediction. Finally the results of the prediction are indicated as follows:

1. "Degradable": High probability of 60% or more biodegradation by the 28th day of the MITI degradation test.

2. "Difficult to degrade": Very low possibility of 60% or more biodegradation by the 28th day of the MITI degradation test.

3. Difficult to predict

388. The prediction method of the system is based on two separate systems: expert flow and SAR equation. If the predictions by both systems differed, priority was given to the results from the expert system over the other.

389. The validation of the system was performed by using new chemicals of CSCL (3, 4). This report describes evaluated results of the expert system.

2.2. Validation set

390. The authors selected 965 chemicals from the new chemicals notified by 2001 under CSCL, which have specified structures, and no stable biodegradation intermediates. Among those chemicals, 359 chemicals are found to be biodegradable with a BOD value of greater than 60% in the MITI test, and 606 chemicals are "difficult to degrade" with a BOD below 60%.

2.3. Results and Discussion

391. The validation results are shown in Table 1. 84 chemicals were classified to be difficult to predict, 154 chemicals were predicted to be degradable, and 142 chemicals out of those (92%) were considered biodegradable in the MITI test. 727 chemicals were predicted to be difficult to degrade and 565 chemicals out of those (78%) were actually difficult to degrade in the MITI test. The number of chemicals predicted to be degradable or "difficult to degrade" was 881 chemicals in total, and 707 chemicals out of those were properly classified (80%). In the case of the prediction being degradable, the hitting rate was more than 90% for all groups. On the contrary, in the prediction of "difficult to degrade" group, the hitting rate was particularly low at 47% for aliphatic chain compounds and 54% for mono heterocyclic, aliphatic monocyclic compounds respectively.

392. The trends of the prediction for each group are as follows:

1. Chemicals with molecular weights greater than 500

   120 chemicals were predicted to be "difficult to degrade" and 116 chemicals were "difficult to degrade" in the actual test (the hitting rate: 97%). Substances failed in the predictions were mostly found to have had the structure containing a sugar chain or ester group. No chemicals in this group were predicted to be degradable.

2. Aliphatic Chain Compounds

   105 chemicals were predicted to be degradable with the hitting rate of 90% (95 chemicals were properly predicted). 73 substances were categorized to be "difficult to degrade" and only 34 substances were properly predicted (the hitting rate: 47%). Failed predictions were mostly for chemicals containing isocyanate, ester and tertiary amine group.

3. Polycyclic, Condensed Cyclic, and Bridged Cyclic Compounds

   151 chemicals were predicted to be "difficult to degrade" and 137 out of these were correctly predicted (the hitting rate: 91%). Failed predictions were mainly for chemicals with a steroid structure (3 chemicals) or an ester structure (8 chemicals). There were no chemicals predicted to be degradable.

4. Mono Heterocyclic and Aliphatic Monocyclic Compounds

   18 chemicals were predicted to be degradable and 17 out of these were correctly predicted (the hitting rate: 94%). 111 substances were classified into the category of "difficult to degrade" and 60 were predicted correctly (the hitting rate: 54%). Failed cases were mainly observed for chemicals with a cyclic ester ring (13 chemicals) or a hetero 5 membered ring.

5. Pyridine and Pyrimidene compounds

2 substances were predicted to be degradable with 100% hits. 22 chemicals were predicted to be "difficult to degrade" resulting in 20 hits out of them (the hitting rate: 91%). And the two cases of failed predictions were for the compounds of aldehyde substituted for pyridine.

6. Aromatic 6 Membered Ring + Non-aromatic Ring

   One substance was predicted to be degradable with a right prediction. 60 substances were predicted to be "difficult to degrade" with 45 of them being correctly predicted (the hitting rate: 75%). Failures were mainly found for those with nitrogen containing 5 membered ring as non-aromatic ring compounds (9 chemicals).

7. Monocyclic Benzene Derivatives

   28 substances were predicted to be degradable, and 27 of them were correctly predicted (the hitting rate: 96%). 71 substances were predicted to be "difficult to degrade" resulting in the hit of 58 (the hitting rate: 82%). No trend was identified in the failed cases.

8. Naphthalene and Quinone Compounds

   5 substances were predicted to be "difficult to degrade" with 100% accurate prediction. No substance in this group was predicted to be degradable.

9. 2 Benzene Rings

   84 substances were predicted to be "difficult to degrade" with 69 out of these correctly predicted (the hitting rate: 82%). Failed predictions were for ester containing compounds (7 chemicals). There was no compound predicted to be degradable in this group.

10. 2 Heterocyclic Rings

    30 substances were predicted to be "difficult to degrade" with 21 out of these correctly predicted (the hitting rate: 70%). Failed predictions were mainly for epoxy containing compounds.

393. Table 2 shows the hitting rate of substances validated by various degradation prediction systems. This biodegradation prediction system (expert system) was found particularly superior for predicting degradable chemicals compared to other systems.

2.4. Conclusions

394. The biodegradation prediction system (Expert System) was validated using new chemicals of CSCL in Japan. The hitting rate of the system for the prediction of degradable chemicals was as high as 90%, which is greater than that of any other system.

395. The predictions for "difficult to degrade" groups had a relatively low hitting rate of 78%. In case of the prediction of the chemicals with aliphatic chain, mono heterocyclic and aliphatic monocyclic group, the hitting rate was low. While partial structures of ester group were seen in the chemicals which were predicted to be "difficult to degrade", they were actually degradable in the test. This might be due to the effects caused by hydrolysis in the water. It may be difficult to make exact predictions for such chemicals based on the flow classification of 2D chemical structures.

3. APPLICATION OF THE SETUBAL PRINCIPLES

3.1. Defined endpoint (Principle 1)

396. The model predicts biodegradation of chemicals in the mixture of aerobic micro-organism in the environment. The defined endpoint is the 28th day degradation of the 301C test of OECD Test Guideline.

397. This model predicts whether the biodegradation rate of chemical substances is more than 60% or not on the 28th day in the 302C test of the OECD Test Guideline

398. Chemical substances for prediction are classified into the following three categories:

1. "Degradable": High probability of 60% or more biodegradation by the 28th day of the MITI degradation test.

2. "Difficult to degrade": Very low probability of 60% or more biodegradation by the 28th day of the MITI degradation test.

3. Difficult to predict

Information about the testing conditions is adequately given. Therefore, the principle is satisfied.

3.2. Defined algorithm (Principle 2)

399. No. The model predicts degradability by flow classification of the chemical structure. The flow is developed based on the empirical knowledge of experts (1, 2). This is not a clear and readily applicable algorithm. Therefore, principle 2 is not satisfied.

3.3. Mechanistic basis (Principle 3)

400. No. The model predicts degradability by flow classification of chemical structure and the flow is developed based on the empirical knowledge of experts (1, 2). It has no clear mechanistic basis. For this reason, principle 3 is not satisfied.

3.4. Domain of applicability (Principle 4)

401. No. The CERI web site explains that the system failed to predict some natural chemicals with specific degradation patterns. Also, it outputs "unpredictable" in case of difficult prediction due to the lack of data (1, 2). In order to satisfy principle 4, it is necessary to provide a much clearer applicability domain.

3.5. Internal performance (Principle 5)

402. This principle is not applicable for this system, because of the prediction made by expert flow.

3.6. External validation for predictivity (Principle 6)

403. Yes. The model has been developed by CERI on the basis of biodegradation data for the existing chemicals under the Japanese Chemical Substance Control Law (CSCL). An external validation for this model has been performed by NITE, using about 1000 new chemicals under the CSCL (3, 4). The results showed an 80% accuracy rate.

404. The identities of the test structures are available. However, the structures of these chemicals cannot be disclosed, because they are new chemicals. The hitting rate of skeleton structures and predictability for sub-structures are shown in the references (3, 4).

405. The approach for selecting the test structures is described. Organic chemicals with verified structures were selected from new chemicals. These seem to reflect the trends of chemical development today (3,4).

406. The statistical assessment of the predictive performance of the model is difficult because it is an expert model. The hitting rate is shown for skeleton structures respectively (3, 4).

407. Principle 6 is therefore satisfied.

4. CONCLUSIONS ON THE SETUBAL PRINCIPLES AND THEIR APPLICABILITY

408. The model has been developed on the basis of biodegradation test data (OECD 301C) for the existing chemicals under the CSCL, with the intention of predicting the results of the biodegradation test (OECD301C). The model has been externally validated by the biodegradability data of new chemicals under the CSCL. Consequently, Setubal principles 1 and 6 are satisfied. Principles 2 through 4 could not be satisfied, since this system is a black box model. It might have been necessary to show the process of prediction to secure its transparency.

409. In conclusion, the Setubal Principles were found to be useful to clarify the features and the points to be improved for the model. The Setubal Principles are appropriate on the whole as a measure of the regulatory application of QSARs. However, it may be difficult to apply principles 2 through 4 to such an expert model. Also, the applicability domain in principle 4, and the accuracy of the model in principles 5 and 6, should be improved.

5. REFERENCES

Hiromatsu K, Yakabe Y, Katagiri K & Nishihara T. (2000). Prediction for biodegradability of chemicals by an empirical flowchart. *Chemosphere* **41** 1749-1754.

http://qsar.cerij.or.jp/cgi-bin/QSAR/index.cgi?e

http://www.nedo.go.jp/

NEDO Project Report “Acceleraten of Safety Inspection for Existing Chemical Substances (2001)”, Chap.7, NEDO (2002). [in Japanese].

NEDO Project Report ”Acceleraten of Safety Inspection for Existing Chemical Substances (2002)” Chap.7, NEDO (2003). [in Japanese].

Table 1 Validation of Biodegradation Prediction System (Expert Prediction)

Substance Group	Number of Chemicals used for Validation		Degradable Prediction			Non-Dagradable Prediction			Difficult to Predict
	Degradable in the Test	Non-Degradable in the Test	Number of Prediction	Number of Hittings	Hitting Rate	Number of Prediction	Number of Hittings	Hitting Rate	
Molecular Weight of more than 500	5	116	0	0		120	116	97	1
Aliphatic Chain	155	54	105	95	90	73	34	47	31
Polycyclic, Condensed Cyclic and Bridged Cyclic	14	137	0	0		151	137	91	0
Mono Hetero Cyclic and Aliphatic Mono Cyclic	68	61	18	17	94	111	60	54	0
Pyridine and Pyrimidine	5	20	2	2	100	22	20	91	1
Aromatic 6 Membered Ring + Non-Aromatic Ring	17	45	1	1	100	60	45	75	1
Monocyclic Benzene Derivatives	67	76	28	27	96	71	58	82	44
Naphthalene and Quinone	1	7	0	0		5	5	100	3
2 Benzene Rings	18	69	0	0		84	69	82	3
2 Heterocyclics	9	21	0	0		30	21	70	0
Total	359	606	154	142	92	727	565	78	84

Table 2 Comparison of Hitting Rate of Biodegradation Prediction Systems

Prediction Model	Prediction	Number of Chemicals Predicted		Number of Hitting		Hitting Rate (%)	
Biodegradability Estimate System (Expert prediction)	Degradable	154	881	142	707	92	80
	Difficult to Degrade	727		565		78	
Biodegradability Estimate System (SAR Prediction)	Degradable	371	507	147	248	40	49
	Difficult to Degrade	136		101		74	
TOPKAT	Degradable	379	871	222	615	59	71
	Difficult to Degrade	492		393		80	
BIOWIN (MITI Linear Model Prediction)	Degradable	376	944	294	745	78	79
	Difficult to Degrade	568		451		79	
BIOWIN (MITI Non-Linear Model Prediction)	Degradable	288	899	225	695	78	77
	Difficult to Degrade	611		470		77	

6. APPENDICES

APPENDIX 1 SUMMARY REPORT OF THE APPLICATION OF THE SETUBAL PRINCIPLES

PRINCIPLE	CONSIDERATIONS	YES / NO
1) Defined endpoint	**1.1)** Does the model have a clearly defined scientific purpose?	Yes
	(i.e. does it make predictions of a clearly defined physicochemical, biological or environmental effect?)	
	1.2) Does the model have the potential to address (or partially address) a clearly defined regulatory need?	Yes
	(i.e. does it make predictions of a specific endpoint associated with a specific test method or test guideline)?	
	1.3) Is information given about important experimental conditions that affect the measurement and therefore the prediction	Yes
	(e.g. sex, species, temperature, exposure period)	
	1.4) Are the units of measurement of the endpoint given?	Yes
2) Defined algorithm	**2.1)** In the case of a SAR, is there an explicit description of the substructure, including an explicit identification of its substituents?	No
	2.2) In the case of a QSAR, is the equation explicitly defined, including definitions of all descriptors used[3]?	
3) Mechanistic basis	**3.1)** In the case of a SAR, is there a description of the molecular events that underlie the reactivity of the molecule?	No
	(e.g. a description of how substructural features could act as nucleophiles or electrophiles, or form part or all of a receptor-binding region)	
	3.2) In the case of a QSAR, do the descriptors have a physicochemical interpretation that is consistent with a known mechanism (of biological action)?	
	3.3) Are any literature references cited in support of the purported mechanistic basis of the (Q)SAR?	No

4) Domain of applicability	**4.1)** In the case of a SAR, is the substructure associated with any inclusion and/or exclusion rules on its applicability to groups of chemicals?	Yes
	4.2) In the case of a SAR, is the substructure associated with rules regarding the modulatory effects of the substructure's molecular environment?	No
	4.3) In the case of a (Q)SAR, are the descriptor and response variables associated with inclusion and/or exclusion rules that define the variable ranges for which the QSAR is applicable (i.e. makes reliable estimates)?	
5) Internal performance	**5.1)** Are full details of the training set given, including details of chemical names, structural formulae, CAS numbers (if available), and data for all descriptor and response variables?	
	5.2) If the data used to the develop the model were based upon the processing of raw data (e.g. the averaging of replicate values):	
	a) is there an adequate description of the data processing?	
	b) are the raw data provided?	
	5.3) Is there a specification of the statistical method(s) used to develop the QSAR (including details of any software packages used)?	
	5.4) Is the QSAR associated with basic statistics for its goodness-of-fit to the training set?	
	(e.g. r^2 values and standard error of the estimate in the case of regression models)	
	5.5)	
	a) Is the QSAR associated with any statistics based on cross-validation or resampling?	
	b) If yes, is the number or samples used indicated?	
6) Predictivity (External validation)	**6.1)** Does application of the appropriate statistical method(s) to the training set result in the same (Q)SAR model?	Not applicable
	6.2) Is there any information to indicate that the (Q)SAR has been validated previously, using a test set that is independent of the training set?	Yes
	6.3) If an external validation has been performed, is the following information available:	a) Yes
	a) the number of test structures?	b) Yes
	b) the identities of the test structures?	
	c) the approach for selecting the test structures?	c) Yes
	d) the statistical analysis of the predictive performance of the	

	model? (e.g. including sensitivity, specificity, and positive and negative predictivities for classification models) e) a comparison of the predictive performance of the model against previously-defined quantitative performance criteria?	d) Yes

ANNEX 11
RAT ORAL CHRONIC TOXICITY MODELS IN TOPKAT

Roger Breton
31 De Brignoles,
Gatineau, Quebec
J8T 8E3
Canada

TABLE OF CONTENTS

1. INTRODUCTION

410. Dr Roger Breton (Ottawa, Canada) has retrospectively applied, under the terms of a JRC contract, Setubal Principles 1-6 to the rat oral chronic toxicity (LOAEL) model in TOPKAT. The work was carried out in consultation with Chandrika Moudgal and Raghuraman Venkatapathy (National Center for Environmental Assessment, Cincinnati, Ohio), and Charles Pittinger (The Cadmus Group, Cincinnati, Ohio).

411. A summary (in tabular form, Appendix 1) of the extent to which the Setubal principles have been met in all the QSAR models of the above papers is attached.

412. A more-detailed explanation following the items in the Appendix 1 is reported below. In particular, there is a discussion on the availability of the model statistics, including cross-validated statistics. Furthermore, this report discusses whether the model is sufficiently validated for regulatory use.

1.1. Description of TOPKAT

413. TOPKAT, The Open Practical Knowledge Acquisition Toolkit is a commercial computational toxicology package that uses chemical structural information (2D descriptors of structural fragments) and QSAR models to estimate a range of human health toxicological and non-human ecological endpoints. Predictions are made for untested chemicals by computing the relevant descriptors of structural fragments that contribute towards a given endpoint.

414. TOPKAT has the capability to predict chronic rat LOAELs for a wide range of chemicals, and it is currently being used by a number of research and regulatory institutions worldwide. The U.S. EPA, Office of Research and Development (ORD) scientists at the National Center for Environmental Assessment (NCEA) in Cincinnati, Ohio, apply TOPKAT in a research program to aid the risk assessment process, and have compiled a substantial historical database of LOAEL predictions for a diverse group of chemicals. Health Canada is currently considering the use of TOPKAT for categorizing substances on the Designated Substances List under the Canadian Environmental Protection Act (CEPA).

415. TOPKAT is a commercial program licensed by the Accelrys Company (accelrys.com/products/topkat), which purchased the licensing rights from Health Designs, Inc. (HDI), the company that developed the software. Though the models training sets are available to the licensed user, the algorithms are considered to be proprietary assets and are not disclosed even to licensed users. For this reason, essential features (algorithms) of TOPKAT needed to validate the methods used to predict output are inaccessible. Without knowledge of the the underlying algorithms in TOPKAT, transparency of the model is somewhat limited.

2. APPLICATION OF THE SETUBAL PRINCIPLES

2.1. Defined endpoint (Principle 1)

416. The chronic oral LOAEL (Lowest Observed Adverse Effect Level) was selected as a measure of potential health effects that might result from long-term consumption of a drinking water contaminant. The LOAEL is defined (1) as "the lowest concentration or amount of a substance, found by experiment or observation, which causes an (adverse) alteration of morphology, functional capacity, growth, development, or life span of a target organism distinguishable from normal (control) organisms of the same species and strain under defined conditions of exposure."

2.2. Defined algorithm (Principle 2)

417. The current version of the rat chronic (LOAEL) module in the TOPKAT package comprises five statistically significant and cross-validated quantitative structure-activity relationship (QSAR) models (e.g., single benzenes, multiple benzenes, heteroaromatics, alicyclics, acyclics).

418. The original model was derived using 44 variables including molecular fragments, shape and connectivity descriptors. The total number of descriptors for the revised model is not reported. Predictions of toxicity are obtained by using a Simplified Molecular Input Line Entry System (SMILES) notation.

2.3 Mechanistic basis (Principle 3)

419. The models are constructed using various statistical methods and have little or no mechanistic basis. However, certain features in the models (moiety value, 2D descriptors) will enable users to draw conclusions of possible modes of action in conjunction with other studies published in the literature that relate the descriptors used in a QSAR model to specific health endpoints. However, the TOPKAT models are comprised of a wide range of chemical classes as opposed to congeneric series. Hence, it is almost impossible to form a hypothesis regarding the actual mechanism.

2.4. Domain of applicability (Principle 4)

420. To minimize inappropriate chemical predictions (beyond the "optimum predictive space" of the model) and inaccurate results, an automated module in the LOAEL model assesses whether the query structure is within the range of the octanol-water partition coefficients of the compounds from which the LOAEL was developed, as well as the estimated and/or experimental rat oral LD50. The latter step is taken to check whether the LOAEL is more toxic than the LD50. This automated module is designed for operation with the TOPKAT interface, which (i) automatically determines whether the submitted structure belongs in the Optimum Prediction Space (OPS) of the model (i.e., multivariate descriptor space), and (ii) allows the user to compute a QSTR similarity distance for chemicals in the model database in order to evaluate the reliability of the QSTR-based assessment. With this information, one can determine whether the query structure lies in an information-rich region of the model data space, and if similar compounds are well predicted by the model.

421. TOPKAT contains a sophisticated, automated series of 10 unique cautionary statements to guide users in appropriate application of the model, and inappropriate use or interpretation of the output. This is a valuable asset of the program. Examples of such cautionary statements include:

1. Predicted chronic LOAEL is more that the computed/experimental acute LD50

2. Fragment Unrecognized.

3. SMILES not recognized by TOPKAT (either a nonsense SMILES notation, or one beyond TOPKAT's ability to recognize.

4. Outside Prediction Space (OPS), the domain of descriptor space capable of making a prediction. Certain results may be valid even if TOPKAT says a chemical is outside the OPS. At the beginning of the report, TOPKAT indicates "Query Outside Optimum Prediction Space (OPS), and one of the following messages: (a) Distance from OPS Exceeds Permissible Limit" or (b) Distance from OPS within Permissible Range". In the former case, the prediction is invalid while in the latter, the prediction is considered valid.

5. "TOPKAT Error", indicating a missing element from the training set.

2.5. Internal performance (Principle 5)

422. The training set for the "original" model was based on 234 substances and has been published by Mumtaz et al. (2). The original model only had one equation for all substances. However, since that period, the model has been revised and now includes 393 substances in the training set (TOPKAT User Guide, Version 6.1). Uniform experimental LOAEL data (393 values) selected after critical review of the open literature (TOPKAT User Guide, Version 6.1; Mumtaz et al. [2], National Cancer Institute/National Toxicology Program technical reports, and the U.S. EPA databases were used to develop these models). All data were for oral rat chronic studies of at least 1 year's duration. The U.S. EPA data consisted of peer-reviewed LOAEL values. Lowest doses at which an adverse effect was reported were used to develop the models. Each LOAEL QSAR model reports the computed chronic LOAEL value in the rat in chemical weight/body weight units, along with 95% confidence limits.

423. The 159 additional substances in the training set of the revised model (393 substances) has not been published in the literature so far and is considered proprietary information by Accelrys, and cannot be distributed to the public due to license terms and conditions. The statistics for the original model are available in Mumtaz et al. (2). Statistics are also available from Venkatapathy et al. (3).

424. Mumtaz et al. (2) used a "leave one out" procedure of the original training set, and achieved higher correlations of predicted versus empirical values. These might be considered "goodness-of-fit" validations as opposed to the model performance validations conducted by NCEA and the Cadmus study. Logically, using the training sets upon which the algorithms were developed resulted in more accurate predictions (e.g., >90 percent of predictions were within a factor of 5 of empirical data). The Accelrys correlations are greater, as (presumably) these represent comparisons for each of the five sub-modules and the exact training sets (not available) used in model development.

2.6. External validation for predictivity (Principle 6)

425. A number of independent validation studies have been performed with TOPKAT (Table 1). The results of the Cadmus study (4) were consistent with preliminary model performance data collected by NCEA (Venkatapathy et al., 5, 6), despite the distinct differences in study design criteria (NCEA used two data sets: a) chronic data from bioassays 2 years in duration, and b) subchronic and chronic data. Both data sets considered studies on rats only, and excluded chemicals yielding a TOPKAT cautionary statement.).

426. Approximately 33 percent of NCEA's actual and predicted LOAELs were within a factor of 2, versus 20 percent in the Cadmus study, 60 percent were within a factor of 5 (vs. 53 percent in the Cadmus study), 72 percent were within a factor of 10 (vs. 68 percent in the Cadmus study), and 98 percent were within a factor of 100 (vs. 95 percent in the Cadmus study). Thus, performance observed in the Cadmus

study was slightly lower than that observed by NCEA, but the results can be considered remarkably similar given the significant differences in data selection required by the Cadmus study.

3. CONCLUSIONS ON THE SETUBAL PRINCIPLES AND THEIR APPLICABILITY

427. In general, the criteria were helpful to determine if a model is sufficiently robust and validated to be considered scientifically valid and useful in regulatory programs. Additional information on the algorithms and training sets are needed to verify statistics and conduct/verify existing external validations.

4. REFERENCES

Glossary for Chemists of Terms Used in Toxicology (1993). *Pure and Applied Chemistry*, Vol. 65, Number 9, pp. 2003-2122.

Mumtaz, M.M., L.A. Knauf, D.J. Reisman, W.B. Peirano, C.T. DeRosa, V.K. Gombar, K. Enslein, J.R. Carter, B.W. Blake, K.I. Huque and V.M.S. Ramanujam. (1995). Assessment of effect levels of chemicals from quantitative structure-activity relationship (QSAR) models. I. Chronic lowest-observed-adverse-effect level (LOAEL).

The Cadmus Group, Inc. (2003). Evaluation of the use of QSAR models to generate data for use in screening the CCL universe to the PCCL. Discussion draft for NDWAC CCL workshop.

Venkatapathy, R.; Moudgal, C.; Swartout, J. and Bruce, R. M. (2003). "An assessment of the performance of the rat chronic LOAEL model for a wide variety of chemicals by TOPKAT, a commercial QSAR software". Toxicology and Risk Assessment Conference, Fairborn, OH, April, 2003.

Venkatapathy, R.; Moudgal, C.; Swartout, J. and Bruce, R. M. (2003). "An assessment of the performance of the rat chronic LOAEL model in TOPKAT, a commercial QSAR software". Third Indo-US workshop on Mathematical Chemistry, Duluth, MN, August, 2003.

Venkatapathy, R.; Moudgal, C.; Swartout, J.; Bruce, R. M. (2004). "Assessment of the rat chronic LOAEL model in TOPKAT, a QSAR software for toxicity prediction". in prep.

Table 1 Cross-validation and validation studies performed with the TOPKAT model

Validation	Percent Predictions within Factors of Empirical Estimates				
	Factors:	2	5	10	100
Accelrys Reported Goodness-of-Fit; Training Set	All five models (averaged)	76	99		
Mumtaz et al. 1995 Training Set	All Models	55	94	100	
NCEA Preliminary Data	All Models, chronic duration, rat only, no error-coded data	33	60	72	98
The Cadmus Group, Inc. Separate test set	All Models, ≥28 exposure duration, rat and mouse data, no error-coded data	20	53	68	95

5. APPENDICES

APPENDIX 1 SUMMARY REPORT OF THE APPLICATION OF THE SETUBAL PRINCIPLES

PRINCIPLE	CONSIDERATIONS	YES / NO
1) Defined endpoint	**1.1)** Does the model have a clearly defined scientific purpose? (i.e. does it make predictions of a clearly defined physicochemical, biological or environmental effect?)	Yes, within the definition of the rat oral chronic LOAEL endpoint, with multiple no effects considered
	1.2) Does the model have the potential 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)?	Yes, to categorize chemicals according to predicting chronic, non-cancer effect potency
	1.3) Is information given about important experimental conditions that affect the measurement and therefore the prediction (e.g. sex, species, temperature, exposure period)	Partly. References provided to licensed users for database chemicals.
	1.4) Are the units of measurement of the endpoint given?	Yes
2) Defined algorithm	**2.1)** In the case of a SAR, is there an explicit description of the substructure, including an explicit identification of its substituents?	Yes, fragments within the five sub-modules included in the model.
	2.2) In the case of a QSAR, is the equation explicitly defined, including definitions of all descriptors used[3]?	Partly. Considered proprietary information, except as defined in Mumtaz et al. (1995) where definitions of

		descriptors are provided.
3) Mechanistic basis	**3.1)** In the case of a SAR, is there a description of the molecular events that underlie the reactivity of the molecule? (e.g. a description of how substructural features could act as nucleophiles or electrophiles, or form part or all of a receptor-binding region)	No
	3.2) In the case of a QSAR, do the descriptors have a physicochemical interpretation that is consistent with a known mechanism (of biological action)?	Partly. Must look through the literature
	3.3) Are any literature references cited in support of the purported mechanistic basis of the (Q)SAR?	No
4) Domain of applicability	**4.1)** In the case of a SAR, is the substructure associated with any inclusion and/or exclusion rules on its applicability to groups of chemicals?	Yes, explicit cautionary statements are included
	4.2) In the case of a SAR, is the substructure associated with rules regarding the modulatory effects of the substructure's molecular environment?	Yes, moiety effects provided
	4.3) In the case of a (Q)SAR, are the descriptor and response variables associated with inclusion and/or exclusion rules that define the variable ranges for which the QSAR is applicable (i.e. makes reliable estimates)?	Yes, TOPKAT issues messages if predictions are outside the model domain
5) Internal performance	**5.1)** Are full details of the training set given, including details of chemical names, structural formulae, CAS numbers (if available), and data for all descriptor and response variables?	Yes for the "original" model published by Mumtaz et al. Yes for current model but only to licensed users.
	5.2) If the data used to ~~the~~ develop the model were based upon the processing of raw data (e.g. the averaging of replicate values): a) is there an adequate description of the data processing? b) are the raw data provided?	a) Yes b) Yes, licensed users only
	5.3) Is there a specification of the statistical method(s) used to develop the QSAR (including details of any software packages used)?	Yes, in the User's Guide.
	5.4) Is the QSAR associated with basic statistics for its goodness-of-fit to the training set?	Yes for the original model published by Mumtaz et al., as well

	(e.g. r^2 values and standard error of the estimate in the case of regression models)	as the correlations on the Accelrys web-site.
	5.5) a) Is the QSAR associated with any statistics based on cross-validation or resampling? b) If yes, is the number or samples used indicated?	Yes for original model, Yes for revised model (User Manual)
6) Predictivity (External validation)	**6.1)** Does application of the appropriate statistical method(s) to the training set result in the same (Q)SAR model?	Need model equation to compare, but likely yes
	6.2) Is there any information to indicate that the (Q)SAR has been validated previously, using a test set that is independent of the training set?	Yes, personal communication by NCEA.
	6.3) If an external validation has been performed, is the following information available: a) the number of test structures? b) the identities of the test structures? c) the approach for selecting the test structures? d) the statistical analysis of the predictive performance of the model? (e.g. including sensitivity, specificity, and positive and negative predictivities for classification models) e) a comparison of the predictive performance of the model against previously-defined quantitative performance criteria?	NCEA has conducted preliminary analyses for validation. a) yes b) yes c) yes d) some e) yes