

# Practical Guide 5

How to use and report (Q)SARs

# ABC

| Version     | Changes  | Date          |
|-------------|--|---------------|
| Version 1.0 | First version  | March 2010    |
| Version 2.0 | Revision of the Practical guide addressing structure and content. The update includes the following: <ul style="list-style-type: none"> <li>• Correction of broken hyperlinks throughout the document.</li> <li>• Changes in IUCLID screenshots with associated text improved for more clarity.</li> <li>• Section 2 on 'How to get started with (Q)SARs' addressing approach to tools developed in research and development projects.</li> </ul>  | December 2012 |
| Version 3.0 | Revision of the Practical guide addressing structure and content. The update includes the following: <ul style="list-style-type: none"> <li>• Update of information in sections 2 and 3</li> <li>• Addition of a section (section 4) giving practical examples on how to assess the reliability of QSAR predictions with some QSAR programs.</li> <li>• Addition of an appendix (Appendix 1) giving examples (a non-exhaustive list) of (Q)SAR programs available for each of the REACH required endpoints.</li> </ul> | March 2016    |

## Practical Guide 5 – How to use and report (Q)SARs

**Reference:** ECHA-16-B-09-EN  
**Cat. Number:** ED-AE-10-005-EN  
**ISBN:** 978-92-9247-809-4  
**ISSN:** 1831-6727  
**DOI:** 10.2823/81818  
**Publ.date:** March 2016  
**Language:** EN

© European Chemicals Agency, 2016  
Cover page © European Chemicals Agency

If you have questions or comments in relation to this document please send them (quote the reference and issue date) using the information request form. The information request form can be accessed via the Contact ECHA page at:  
<http://echa.europa.eu/contact>

### European Chemicals Agency

Mailing address: P.O. Box 400, FI-00121 Helsinki, Finland  
Visiting address: Annankatu 18, Helsinki, Finland

## Table of Contents

|  |           |
|--|-----------|
| <b>1. INTRODUCTION</b>   | <b>4</b>  |
| <b>2. HOW TO GET STARTED WITH (Q)SARS</b>  | <b>5</b>  |
| 2.1 Definitions .....  | 5         |
| 2.2 Substance characterisation .....   | 5         |
| 2.3 Experimental results .....   | 5         |
| 2.4 Conditions for using (Q)SAR results .....  | 6         |
| 2.5 Strategy for using (Q)SAR results .....  | 6         |
| <b>3. HOW TO CHECK A QSAR PREDICTION</b>   | <b>7</b>  |
| 3.1 Is the (Q)SAR model valid? .....   | 7         |
| 3.2 Does the substance fall within the applicability domain of the (Q)SAR model? .....                       | 8         |
| 3.3 Is the prediction adequate for the purpose of classification and labelling and/or risk assessment? ..... | 9         |
| 3.4 How to report a (Q)SAR prediction in IUCLID 5? .....   | 10        |
| <b>4. PRACTICAL EXAMPLES</b>   | <b>14</b> |
| 4.1 Log Kow (EPI Suite) .....  | 14        |
| 4.2 Ready biodegradability (VEGA) .....  | 17        |
| 4.3 Short-term toxicity to fish (ECOSAR) .....   | 20        |
| 4.4 Acute toxicity to rat (T.E.S.T.) .....   | 25        |
| <b>APPENDIX 1. QSAR MODELS RELATED TO REACH ENDPOINTS</b>  | <b>28</b> |
| <b>APPENDIX 2. FURTHER GUIDANCE DOCUMENTS AND LINKS</b>  | <b>34</b> |

## 1. Introduction

REACH foresees in Annex XI that the standard testing regime can be adapted by the use of non-test methods, such as (Quantitative) Structure-Activity Relationships [(Q)SARs], if certain conditions are fulfilled.

This practical guide provides an overview of important aspects to consider when predicting properties of substances using (Q)SAR models as defined in the REACH Regulation, aspects which ECHA also takes into account to evaluate (Q)SAR results. This practical guide also gives useful examples for good prediction practices based on widely used and freely available (Q)SAR software programs.

Section 2 of this document gives general information about (Q)SARs and how to use them.

Section 3 explains the conditions that need to be fulfilled to use (Q)SAR predictions under REACH. Registrants are advised to explicitly include these considerations in their registration dossiers.

Section 4 gives practical examples based on freely available and commonly used (Q)SAR programs.

Appendix 1 gives examples (a non-exhaustive list) of (Q)SAR programs available for each of the REACH required endpoints.

Appendix 2 provides links to other guidance documents and tools that give further insights on the use of QSARs.

## 2. How to get started with (Q)SARs

### 2.1 Definitions

Structure-Activity Relationship (SAR) and Quantitative Structure-Activity Relationship (QSAR) models – collectively referred to as (Q)SARs – are theoretical models that can be used to predict in a quantitative or qualitative manner the physicochemical, biological (e.g. an (eco)toxicological endpoint) and environmental fate properties of compounds from the knowledge of their chemical structure.

A SAR is a qualitative relationship that relates a (sub)structure to the presence or absence of a property or activity of interest.

A QSAR is a mathematical model relating one or more quantitative parameters, which are derived from the chemical structure, to a quantitative measure of a property or activity.

In this document, the chemical for which an endpoint is being estimated by a (Q)SAR model is referred to as the target chemical. In other sources, this target chemical is sometimes called a query compound or input structure.

### 2.2 Substance characterisation

The chemical structure needs to be well defined, following the Guidance on identification and naming of substances under REACH<sup>1</sup>. All individual constituents of multi-constituent substances should be addressed. The composition of the well-defined substances has to also include known impurities (and additives, if any).

For UVCBs, expert judgement is needed to decide whether representative structures for the substance can be identified. Stable transformation products should also be identified. A suitable structural representation for the chemical (SMILES, mol file, etc.) is usually required.

### 2.3 Experimental results

In general, if reliable and adequate experimental (measured) results are available they should prevail over estimated values for the risk assessment and the classification and labelling of the substance.

Therefore, before using (Q)SAR models to predict a specific property of a substance, a critical first step is to assemble all of the available information on the substance. There are many information sources available for this purpose and those are further explained in another guidance document<sup>2</sup>.

Among those sources, it should be noted that the OECD QSAR Toolbox includes one of the largest collections of publicly available data. Detailed information on how to use the QSAR Toolbox is provided under the following link:

<http://echa.europa.eu/support/oecd-qsar-toolbox>

In addition, most of the (Q)SAR software programs will indicate if their training set (dataset used to construct the (Q)SAR model) contains experimental results for the target chemical. In this case, the user should give priority to this existing experimental data over the predicted

---

<sup>1</sup> [http://echa.europa.eu/documents/10162/13643/substance\\_id\\_en.pdf](http://echa.europa.eu/documents/10162/13643/substance_id_en.pdf)

<sup>2</sup> See "Guidance on information requirements and chemical safety assessment - Chapter R.3: Information gathering" available at: [http://echa.europa.eu/documents/10162/13643/information\\_requirements\\_r3\\_en.pdf](http://echa.europa.eu/documents/10162/13643/information_requirements_r3_en.pdf)

data, if there is sufficient indication that the experimental data is of good quality.

## 2.4 Conditions for using (Q)SAR results

Several (Q)SAR models have been integrated in software programs that are straightforward to use. However, experience and a thorough understanding of (Q)SARs is needed to verify their reliability and adequacy.

Results of (Q)SARs may be used instead of testing when the conditions set in REACH Annex XI (1.3) are met:

- (i) a (Q)SAR model should be used whose scientific validity has been established,
- (ii) the substance should fall within the applicability domain of the (Q)SAR model,
- (iii) the prediction should be fit for the regulatory purpose, and
- (iv) the information should be well documented.

An assessment of the first three points above is expected to be included in the registration dossier if substance properties are predicted using (Q)SAR models.

Section 3 of this practical guide provides detailed information on how to do this assessment.

## 2.5 Strategy for using (Q)SAR results

In general, **it is recommended to use (Q)SAR results as part of a weight of evidence (WoE) approach** or as supporting information. For instance, (Q)SAR predictions can support results from tests that have not been performed according to good laboratory practice (GLP) or according to accepted guidelines, if those predictions concur with the experimental results. A compilation of several predictions with unassignable quality cannot provide an adaptation on its own.

When using (Q)SARs **it is recommended to run all the available (Q)SAR models** for the endpoint to be fulfilled, especially when models are independent from each other (e.g. the algorithms are based on different descriptors, structural alerts or training sets). Agreement among predictions generated from independent and scientifically-valid (Q)SAR models increases the confidence in relying on the predictions.

Predictions that fulfil only some conditions specified in REACH Annex XI (1.3) should be disregarded or the reason for providing these predictions should be explained if it is considered that there are some benefits to provide these predictions. If the remaining (valid and adequate) predictions show small quantitative differences, the most conservative result should be chosen for further consideration. If those remaining predictions show significant quantitative differences, it is up to the registrant to decide if these differences could affect the risk assessment (for demonstrating safe use) and/or classification and labelling.

If the (Q)SAR prediction outcome is a quantitative result, it should be kept in mind that **the closer to a regulatory threshold the predicted result is, the more accurate the prediction needs to be**. For instance, if a (Q)SAR model predicts a LC50 (for fish at 96 hrs) of 1.2 mg/L then this predicted value needs to be fully reliable to be sure that the actual LC50 of the substance is not below the CLP regulatory threshold of 1 mg/L. In contrast, if all (Q)SAR results (and even the worst case/over-conservative ones) do not go beyond the regulatory threshold of interest, this can support the waiving of the experimental study.

## 3. How to check a QSAR prediction

### 3.1 Is the (Q)SAR model valid?

As indicated in REACH Annex XI (1.3), the validity of the (Q)SAR model is the first condition to be fulfilled to use a QSAR result. To check this, ECHA follows the OECD principles for (Q)SAR models validation<sup>3</sup>. These are five principles saying that a (Q)SAR model should be associated with:

1. **A defined endpoint:** the model must predict the same endpoint that would be measured to fulfil the requirements listed in REACH Annexes VII to X. For instance, predictions from a model generically predicting "mutagenicity" cannot be accepted as such. The model should predict the outcome of a specific test such as "positive", "negative" or "ambiguous" in a bacterial reverse mutation assay (i.e. Ames test required in REACH Annex VII, 8.4.1). Another example of an endpoint being too broad is a global prediction of a 'repeated dose toxicity lowest observed adverse effect level (LOAEL)' from a training set of LOAEL data based on a variety of mode of actions, target organs, species or test protocols. This principle links with the adequacy of the predictions described later in the document.
2. **An unambiguous algorithm:** the algorithm underlying the model must be available to ensure transparency and reproducibility of the calculation. Predictions from a model whose algorithm is not available (to ECHA) to verify its functioning and to reproduce the predictions can hardly be accepted. In particular, special precautions are needed where non-transparent and difficult to reproduce methods have been used to build the (Q)SAR model (e.g. artificial neural networks using many structural descriptors).
3. **A defined domain of applicability:** the applicability domain (AD) and the limitations of the model have to be described to allow the assessment of the AD for the specific prediction (see Section 3.2 of this document). The most common methods for describing the AD are to consider the ranges of the individual descriptors and the presence of the structural fragments in the training set. Predictions from a model without information on the AD cannot be accepted.
4. **Appropriate measures of goodness-of-fit, robustness and predictivity:** this principle expresses the need for statistical validation of the model. Statistics on internal validation (goodness-of-fit and robustness) and external validation (predictivity) must be available. For instance, for regression models, the statistics of the regression model could be reported through the correlation coefficient ( $R^2$ ), cross-validated (e.g. from leave-one-out procedure) correlation coefficient ( $Q^2$ ) and the standard error of the model ( $s$ ). It can be noted that an  $R^2$  below 0.7, a  $Q^2$  below 0.5 or an  $s$  above 0.3 should warn the (Q)SAR user of a potential low performance of the (Q)SAR model. Regarding the external validation, it should have been done by predicting compounds from an external set, i.e. not used for the model development. Statistics on the external validation are useful to estimate the uncertainty associated with the predictions.
5. **A mechanistic interpretation, if possible:** reasoning on the causal link between the descriptors used in the model and the predicted endpoint adds confidence in the reliability of the predictions e.g. a SAR model predicting skin sensitisation can be based on structural alerts. If reasoning is provided on how the structural alerts are associated with skin sensitisation (for example, they enclose electrophilic groups able to bind to

---

<sup>3</sup> [http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?doclanguage=en&cote=env/jm/mono\(2007\)2](http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?doclanguage=en&cote=env/jm/mono(2007)2)

proteins), then confidence in the prediction would increase due to the possibility to interpret the outcome of the model.

In general, a reference to a well-documented model or a (Q)SAR Model Reporting Format (QMRF) attached to the registration dossier is recommended. See the Guidance on information requirements and chemical safety assessment, Chapter R.6: (Q)SARs and grouping of chemicals (R.6.1.9.1) for more details on the QMRF. The JRC (Q)SAR Model Database (QMRF Inventory) is intended to provide information on (Q)SAR models submitted to the JRC for peer review: [http://qsardb.jrc.it/qmrf/search\\_catalogs.jsp](http://qsardb.jrc.it/qmrf/search_catalogs.jsp)

There is no formal adoption process existing or foreseen for (Q)SAR models under REACH. The validity, applicability and adequacy of (Q)SAR models is assessed individually with the prediction generated for the target chemical.

NOTE: A valid (Q)SAR model does not necessarily produce a valid prediction. It is necessary to assess whether the substance falls within the applicability domain of the (Q)SAR model, that the results are adequate for the purpose of classification and labelling and/or risk assessment, and that adequate and reliable documentation of the applied method is provided.

### 3.2 Does the substance fall within the applicability domain of the (Q)SAR model?

It is important to verify that the target substance falls within the applicability domain (AD) of the model. The concept of AD was introduced to assess the probability of a chemical being covered by the (Q)SAR training set. Predictions outside the AD are normally not reliable and their use is hard to justify. A practical approach to check if a substance falls into the AD is to check the following elements:

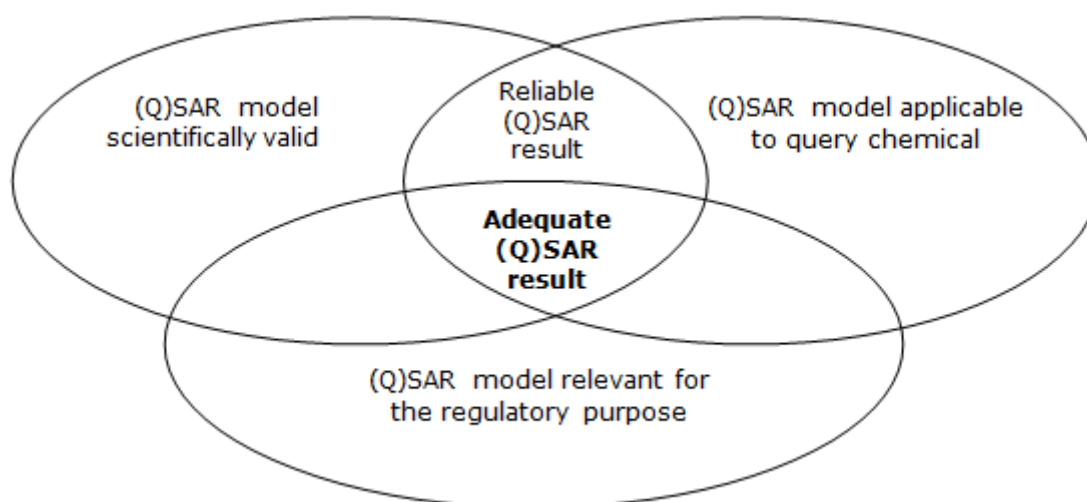
1. **Descriptor domain:** As indicated in the previous section, the AD of a model can be based on the range of the descriptors of the chemicals that are in the training sets (e.g. molecular weight, log Kow...). Therefore, if those ranges have been described, users of (Q)SAR models should check whether the target chemical falls into these ranges. It is expected that the descriptors follow normal distributions and if there are significant deviations, these should be explained. Clusters or other anomalies in the distribution of data (concerning one or both the dependent and the independent variables) may invalidate the model, and the prediction thereof.
2. **Structural fragment domain:** Users of (Q)SAR models should check whether the (sub)structures (i.e. functional groups or meaningful fragments) of their target chemical are represented in the training set. In addition, the (Q)SAR user should also check whether their target chemical has more instances of a given fragment than the maximum for all training set compounds (e.g. if the target chemical has three hydroxyl groups and any compounds in the training do not have more than two hydroxyl groups, then the target chemical may be considered outside the AD).
3. **Mechanistic and metabolic domains, if possible:** To check these points is complex but still very valuable to support the reliability of the prediction. One approach could be to use the OECD QSAR Toolbox. Within this tool, the "profiling methods" can indicate whether the chemical shows relevant mechanisms of action that are not covered by the (Q)SAR model (i.e. not covered by its algorithm/mechanistic domain) while the "Metabolism/Transformations" – also available in the module "Profiling" of the QSAR Toolbox – can indicate metabolites/degradation products that should be considered. Significant potential differences in absorption, distribution, metabolism and elimination between the target chemical and the chemicals in the training set might also invalidate the prediction from an otherwise valid model. These considerations are not explicitly addressed by the tools and might need to be considered separately from the model.



- Analogues in the training set:** Having close structural analogues in the training set of the model increases the reliability of the prediction. Therefore, if the training set is available and the software does not do it automatically, the user should search for close analogues either manually (for small sets) or with the support of IT tools that provide structural search functionalities.
- Accuracy of model predictions for analogues:** If there are substances similar to the target substance, whose experimental results for the endpoint of interest are known (e.g. analogues in the training sets, in the validation sets or from any other source), the model can be run to predict these analogues and to compare these predictions with the experimental results (to see how accurate the model is for these similar substances).
- Considerations for specific substances:** Firstly, and as mentioned in Section 2.2, special considerations should be given to UVCBs, multiconstituents, additives, impurities, metabolites and degradation products. Secondly, most of the (Q)SAR models are developed for organic chemicals and do not address the specificity of some types of chemicals such as ionisable substances (e.g. salts, weak acids and bases), large molecular weight substances (e.g. polymers), potentially hydrolysable substances (e.g. esters, carbamates), surfactants (e.g. hydrocarbon chain with hydrophilic head) and isomers (e.g. stereoisomers, tautomers).

### 3.3 Is the prediction adequate for the purpose of classification and labelling and/or risk assessment?

For a (Q)SAR prediction to be adequate, it should be not only reliable (i.e. derived from a valid QSAR model and within its applicability domain), but also relevant for regulatory decision. The adequacy of the prediction for the purpose of classification and labelling (C&L) and/or risk assessment is very much endpoint-dependent. Additional information might be needed to assess the adequacy of the prediction in the context of a regulatory decision. Therefore, the validity (are the five OECD principles on scientific validity of a model fulfilled?), applicability (can reliable predictions be expected if the model is applied to the target substance?) and relevance (is the information which is needed for the risk assessment and/or C&L generated?) need to be assessed for each individual prediction.



C&L and risk assessment are based on well-defined requirements in terms of tests (and endpoints), thresholds and uncertainty analysis. Therefore, results from (Q)SAR models should be equivalent to results obtained from the required experimental test.

Some examples of inadequacy are listed here:

- (Q)SAR models able to fully cover the complexity of higher-tier endpoints do not exist yet (e.g. repeated dose toxicity or reproductive toxicity). So far, the use of (Q)SARs as stand-alone information for these endpoints cannot be accepted. For instance, repeated dose toxicity tests provide many data points for effects in specific tissues (specific target organ toxicity) and it is not only the no observed adverse effect level (NOAEL) that matters. Indeed, effect results are needed to trigger other tests such as reproductive toxicity or for specific target organ toxicity single exposure/repeated exposure (STOT SE/RE) classification.
- If a quantitative outcome is needed (e.g. to derive a derived no-effect level (DNEL) or for classification) and the model only gives qualitative predictions (e.g. negative or positive result), then the model is probably not adequate for the purpose.
- Uncertainty associated with predictions close to regulatory thresholds needs to be examined cautiously. For instance, if the predicted oral rat LD50 does not go beyond the threshold for classification but that the standard error of the model and/or the error of the estimate is larger than this gap, then the prediction is probably not adequate.
- As required for an experimental bacterial reverse mutation assay (Ames test), the training set of the (Q)SAR model should include experimental results that cover the five bacterial strains in presence and absence of metabolic activation (S9). This information has to be included in the documentation of the model and ideally also in the prediction report.
- (Q)SAR models for fish toxicity whose experimental results for the chemicals in the training set have been performed according to OECD test guideline 204 (14-day studies) cannot be used to predict long-term toxicity to fish because the test duration is too short.
- (Q)SAR models predicting the biodegradation half-life of a compound cannot be used as a stand-alone replacement of a simulation test as they do not cover the need to identify the degradation products (REACH Annex IX, 9.2.3 requirements).

### 3.4 How to report a (Q)SAR prediction in IUCLID 5?

The information must be reported in the IUCLID 5 endpoint study record as follows.

#### **Block “Administrative data”**

- The field “Purpose flag” states whether the estimate is used as a key study, as a supporting study or in a weight of evidence approach.
- The field “Study result type” to state “(Q)SAR”.
- The field “Reliability” states the reliability score, bearing in mind that for (Q)SAR predictions it should normally be a maximum of 2.

**Administrative Data**

Purpose flag: weight of evidence  robust study summary  used for classification  used for MSDS

Data waiving:

Justification for data waiving:

Study result type: (Q)SAR Study period:

Reliability: 2 (reliable with restrictions)

### Block "Data source"

- The field "Year" is used to include the year when the software program was released or when the (Q)SAR model was published. Additionally, the "Title" field to state the name and version of the program and/or the title of the publication, and "Bibliographic source" to provide information on the (Q)SAR model.
- The field "Data access" provides information on the accessibility of the model.

**Data source**

| Reference type | Author | Year | Title                  | Bibliographic source              | Testing labor... | Report no. | Owner comp... | Company study ... | Report date |
|----------------|--------|------|------------------------|-----------------------------------|------------------|------------|---------------|-------------------|-------------|
|                |        | 2012 | EPI Suite Version 4.11 | KOWWIN - Meylan and Howard (1995) |                  |            |               |                   |             |

Add... Edit... Delete Move up Move down Select Insert

Data access: data published

### Block "Materials and methods"

Either the field "Guideline" (in table "Test guideline") or the field "Principles of method if other than guideline" should be filled in.

- In the field "Guideline", the user can select "other guideline" and provide text in the adjacent field. This text could, for instance, refer to the REACH Guidance on QSARs R.6 or to the test guidelines used to generate the data for the training set.
- Otherwise in the field "Principles of method other than guideline", the user could provide further details/references on the (Q)SAR model.

**Materials and methods**

Test guideline

| Qualifier | Guideline   | Deviations |
|-----------|---|------------|
|           | other guideline: REACH Guidance on QSARs R.6, May/July 2008 |            |

Add... Edit... Delete Move up Move down

Principles of method if other than guideline

Meylan, W.M. and P.H. Howard. 1995. Atom/fragment contribution method for estimating octanol-water partition coefficients. J. Pharm. Sci. 84: 83-92.

### Block "Test materials"

- The table "Test material identity" should include information on the substance for which the prediction was made.
- The SMILES notation should be reported in the table 'Test material identity' or in the field "Details on test material".

Test material identity

| Identifier    |                   |
|---------------|-------------------|
| IUPAC name    | 4-Methyl-2-hexene |
| EC number     | 222-281-0         |
| CAS number    | 3404-55-5         |
| other: SMILES | CCC(C)C=CC        |

Test material form

Details on test material

SMILES: CCC(C)C=CC

SMILES to be reported in one of these fields

NOTE: the registered substance may contain more than one constituent and/or impurities. In such cases, it may be useful to prepare an individual endpoint study record and a (Q)SAR prediction reporting format (QPRF) for each constituent/impurity to be able to address each chemical separately (recommended if constituents have different properties and thus different models, assessments, etc. have to be applied).

### Block “Results and discussion”

- The (Q)SAR predicted result should be reported in the structured result fields. This would allow the user to transfer information automatically from these result fields to the chemical safety report (CSR) when the IUCLID 5 CSR plugin is used. The list of fields to be filled in the “Results and discussions” block will vary depending on the endpoint.

Therefore, we recommend consulting Data Submission Manual 5 “How to complete a technical dossier for registrations and PPORD notifications” for instructions on how to fill in the results.

Results and discussions

Partition coefficient

| Type    | Partition coefficient | Temp. | pH | Remarks              |
|---------|-----------------------|-------|----|----------------------|
| log Pow | 3.49                  |       |    | QSAR predicted value |

Any other information on results incl. tables

KOWWIN predicted that 4-Methyl-2-hexene has a log Kow = 3.49

- If it is not possible to fill in all structured result fields required to pass the technical completeness check then the fields “Remarks” (at the right end of the table) or the field “Any other information on results incl. tables” could be used instead.

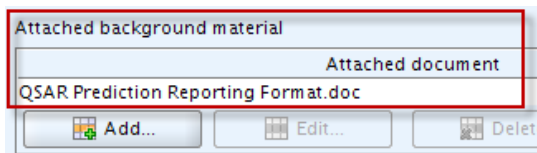
### QMRF and QPRF

As discussed previously in sections 2.4, 3.1, 3.2 and 3.3, the following information has to be reported in IUCLID 5:

- Information on the validity of the (Q)SAR model;
- Verification that the substance falls within the applicability domain of the (Q)SAR model; and

- Assessment of the adequacy of the results for the purpose of classification and labelling and/or risk assessment.

These three pieces of information should be compiled according to the (Q)SAR prediction reporting format (QPRF). The QPRF – as well as the (Q)SAR model reporting format (QMRF) if available – should be attached under “Attached background material”.



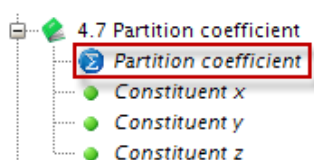
These formats are respectively available in chapters R.6.1.10.1 and R.6.1.9.1 of the Guidance on information requirements and chemical safety assessment available at: [http://echa.europa.eu/documents/10162/13632/information\\_requirements\\_r6\\_en.pdf](http://echa.europa.eu/documents/10162/13632/information_requirements_r6_en.pdf)

The QMRF is normally provided by the model developer while the QPRF is prediction-specific and should be prepared by the registrant using the information in the software report and manual.

Alternatively, these three pieces of information could be provided in the field “Any other information on results incl. tables”, in the field “Details on results” or in the field “Remarks on results including tables and figures”. However, it should be noted that the submission of the QPRF (and QMRF if available) is the preferred option.

### Endpoint study summary

In addition, it is advisable to create an endpoint study summary when more than one endpoint study record is available and to provide the overall assessment on the particular endpoint. This would enable the automatic transfer of this information to the CSR when the IUCLID 5 CSR plugin is used.



## 4. Practical examples

This section describes how to assess the reliability of QSAR predictions. The assessment depends on the software and on the target endpoint. The examples used in this section are based on computer programs that are widely used and freely available. The fact that these programs are used in these examples does not represent their endorsement by ECHA. Usually the use of QSARs is limited to experts. With these practical examples the aim is to allow less experienced people to use and interpret QSARs at least for some endpoints (like in the following examples).

The programs used in the examples can predict several endpoints. However, only one endpoint per program (corresponding to one REACH requirement) has been used for each example. In most cases, predictions for different endpoints from the same program are reported (and can be assessed) in a similar way.

The four endpoints predicted in the following examples are log Kow, ready biodegradability, short-term toxicity to fish and acute mammalian toxicity. Those endpoints have been selected as representatives of REACH Annexes VII or VIII requirements for physico-chemical properties, environmental fate, ecotoxicological and toxicological information.

### 4.1 Log Kow (EPI Suite)

#### a) Introduction

Partition coefficient n-octanol/water is a REACH requirement for all substances produced or imported above one tonne/year (REACH Annex VII). It is commonly expressed as a logarithmic value called log Kow or log P.

Many QSAR models are available to predict log Kow. KOWWIN – which is part of EPI Suite – is one of the most commonly used programs. KOWWIN uses a "fragment constant" method to predict Kow. Fragment constant methods divide the chemical structure into smaller structural fragments (atoms or larger functional groups). Each fragment is associated with a pre-assigned coefficient value called fragment constants. The predicted log Kow value is obtained by summing all the fragment constants appearing in the chemical structure.

At the time of writing this manual, the current version of EPI Suite™ was version 4.11, which has been used to prepare this example.

Link to the (Q)SAR program: <http://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface>

## b) How to check the reliability of the prediction

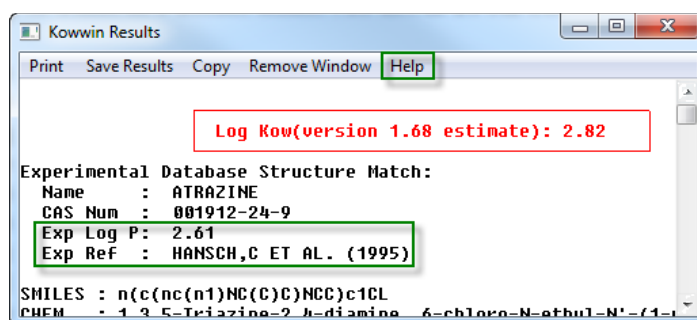
| Information on the QSAR model  | Where to find this information   | How to use this information   |
|--|--|---|
| Description of the algorithm   | Available in the KOWWIN Help <sup>4</sup> , in section "Methodology".  | See Section 3.1 of this practical guide.  |
| Statistics (goodness-of-fit and robustness)                                      | Available in the KOWWIN Help, in section "Accuracy & Domain".  | See Section 3.1 of this practical guide.  |
| Information on the applicability domain  | Where to find this information   | How to use this information   |
| General applicability domain (chemical classes covered/not covered by the model) | Available on the front page of EPI Suite (" <i>The intended application domain is organic chemicals. Inorganic and organometallic chemicals generally are outside the domain.</i> ") and in the KOWWIN Help, in sections "Ionization" and "Zwitterion Considerations". | As indicated in Section 3.2 point 6 of this practical guide, special attention should be paid for some types of chemicals. KOWWIN includes some "corrections" for ionisable and zwitterionic substances to refine the low predictivity for these substances.  |
| Descriptor domain  | Available in the KOWWIN Help, in section "Accuracy & Domain".  | The user should check that the target chemical is in the molecular weight range of the compounds in the training set (i.e. between 18 and 720).   |
| Structural fragment domain   | <p>The KOWWIN results window lists the fragments (and their numbers) found in the target chemical.</p> <p>Appendix D of the KOWWIN Help gives the maximum number of fragments that occur in any individual compound of the training set.</p>                           | <p>The user should check whether the number of each fragment found in the target chemical (column "NUM" in the KOWWIN results window) does not exceed the maximum number of this fragment that occurs in any individual compound of the training set (column "Training set / Max" of Appendix D of the KOWWIN Help).</p> <p>Notes on specific substructures:</p> <ul style="list-style-type: none"> <li>- For some substructures, KOWWIN reports correction factors. In this case, the user should do the same verification described above for the numbers of fragments.</li> <li>- For some substructures, the coefficient has been estimated (if this is the case, it will be reported as a note in the KOWWIN results window). It should be kept in mind that this estimation brings additional uncertainty to the overall prediction.</li> </ul> |

<sup>4</sup> KOWWIN Help can be accessed by clicking on the tab "Help" at the top of the KOWWIN window.

| Training set and validation set | Where to find this information   | How to use this information   |
|---------------------------------|--|---|
|                                 | <p>The training and validation sets can be downloaded through links given at the bottom of the section "Accuracy &amp; Domain" of the KOWWIN Help.</p> <p>The first link provides an Excel file with chemical names, experimental and estimated values of log Kow:<br/> <a href="http://esc.syrres.com/interkow/KowwinData.htm">http://esc.syrres.com/interkow/KowwinData.htm</a>.</p> <p>The second link provides an SDF file with the same information than the Excel file plus structural information<sup>5</sup>:<br/> <a href="http://esc.syrres.com/interkow/EpiSuiteData_ISIS_SDF.htm">http://esc.syrres.com/interkow/EpiSuiteData_ISIS_SDF.htm</a></p> | <p>The user should check whether there are close analogues in the training and validation sets (either manually in the Excel file or automatically with the SDF file). If there are close analogues, the user should check how well they have been predicted by KOWWIN.</p> |

### c) Additional information on EPI Suite

- The EPI Suite Help files do not have exactly the same table of contents between the various estimation programs (e.g. between KOWWIN, BIOWIN...). Therefore, the user should go through each specific Help file to identify the information needed to check the reliability of the prediction (as done for KOWWIN in the table above).
- Several programs of EPI Suite display the experimental result of the target chemical if it is part of the training or validation sets. For instance, KOWWIN contains a database of approximately 13 500 experimental log Kow and if an experimental log Kow value is available for the target chemical it will be displayed in the "Results" window (see green rectangle in the figure below). In this case, it is recommended to report this experimental data and the EPI Suite estimate in two different endpoint study records.



<sup>5</sup> Many tools are available to read SDF files. These tools allow users to visualise the chemicals, to search for closest structural analogues or to search for specific substructures. Some of these tools are freely available (e.g. Knime or the OECD QSAR Toolbox).



## 4.2 Ready biodegradability (VEGA)

### a) Introduction

Ready biodegradability is a REACH requirement for all substances produced or imported above one tonne/year (REACH Annex VII). The main outcome of a ready biodegradability test is the classification of the chemical either as “readily biodegradable” or as “not readily biodegradable”.

The VEGA platform contains several QSAR models for various endpoints. One of these models predicts ready biodegradability (model developed by the Istituto di Ricerche Farmacologiche Mario Negri). This model is based on structural alerts.


Four sets of substructures (i.e. fragments) are included in this model and those sets are classified as “non readily biodegradable”, “possible non readily biodegradable”, “readily biodegradable” and “possible readily biodegradable”. A target chemical is always considered non biodegradable if at least one fragment related to “non readily biodegradability” is found.

At the time of writing this manual, the current version of VegaNIC is 1.1.0, which has been used to prepare this example.

Link to the (Q)SAR program: <http://www.vega-qsar.eu/>

### b) How to check the reliability of the prediction

| Information on the (Q)SAR model             | Where to find this information  | How to use this information              |
|---|---|--|
| Description of the algorithm                | Available in the “Guide to Ready Biodegradability Model” <sup>6</sup> (in Sections 1.2, 1.4 and 1.5). | See Section 3.1 of this practical guide. |
| Statistics (goodness-of-fit and Robustness) | Available in the “Guide to Ready Biodegradability Model” (in Section 1.6).                            | See Section 3.1 of this practical guide. |

<sup>6</sup> This guide can be downloaded from the VEGA program by clicking on the tab “SELECT”, then on the tab “Environ”, then on the question mark icon  adjacent to “Ready Biodegradability model (IRFMN)”.


| Information on the applicability domain  | Where to find this information   | How to use this information   |
|--|--|---|
| General applicability domain (chemical classes covered/not covered by the model) | Partly available in the Vega prediction report.  | <p>If less than 3 “golden stars” are displayed in section 1 of the Vega report, this indicates that at least one issue has been detected for the prediction and that therefore the prediction might not be reliable. In this case, the user should investigate the issue(s) thoroughly. Note that the issue(s) are further detailed in Section 3.2 of the VEGA report.</p> <p>In addition, as indicated in Section 3.2 point 6 of this practical guide, special attention should be paid for some types of chemicals.</p>   |
| Descriptor domain  | Not provided.  | <p>The training set of the model is based on tests performed according to the OECD 301C guideline. Some substances led to unreliable results using this guideline (e.g. low water soluble, volatile or absorptive substances). Therefore, if the target substance has low water solubility, a high vapour pressure or a high log K<sub>oc</sub>, then the user should keep in mind that the prediction might be erroneous.</p> <p>In addition, if the target substance has a large molecular weight or a multi-branched alkyl structure, then the user should verify if these characteristics also appear among the compounds in the training set.</p>                |
| Structural fragment domain   | Partly available in the “Guide to Ready Biodegradability Model” and in the Vega prediction report. | <p>If the target chemical does not contain any of the fragments listed in Sections 1.4 and 1.5 of the model’s guide then no prediction is given by the tool.</p> <p>In addition, in Section 3.2 of the Vega Prediction report, if the “Atom Centered Fragments similarity check” gives an ACF index &lt;1, this would indicate that there is at least one atom centered fragment of the target chemical which has not been found in the compounds of the training set (or which is rarely present). In this case, the user should determine if these missing/rare fragments (listed in Section 4.1 of the report, if any) could have an impact on biodegradation.</p> |

| Training set and validation set | Where to find this information  | How to use this information  |
|---------------------------------|---|--|
|                                 | Available in the file called "Training set (plain text with SMILES)" <sup>7</sup> . | This file contains the SMILES of the training set compounds and of the test set compounds.<br><br>The VEGA report displays the most similar compounds found in the training set and in the test set in Section 3.1. The user should check in this section whether these compounds are closely similar to the target chemical and if their experimental outcome is in agreement with the predicted one. |

### c) Additional information on VEGA

If there is an experimental result for the target chemical in the training set or in the test set, this data will be displayed in the VEGA report (in Section 1 of the report). In this case, the user should look for further information about this test – e.g. by searching this experimental test within the QSAR Toolbox – and report the details of this test in an endpoint study record only dedicated to this experimental study.

---

<sup>7</sup> This file can be downloaded from the VEGA program by clicking on the tab "SELECT", then on the tab "Environ", then on the question mark icon  adjacent to "Ready Biodegradability model (IRFMN)".

### 4.3 Short-term toxicity to fish (ECOSAR)

#### a) Introduction

Short-term toxicity testing on fish is a REACH requirement for all substances produced or imported above 10 tonnes/year (REACH Annex VIII). The endpoint to be derived is the LC50, which is the concentration lethal to 50% of the fish.

The Ecological Structure Activity Relationships (ECOSAR) Class Program is a collection of QSAR models estimating aquatic toxicity, including short-term toxicity to fish. Most of the ECOSAR models are based on the relationships between log Kow and toxicity (LC50 or EC50) and take into account different structural classes.

NOTE: Regarding EPI Suite and ECOSAR, ECOSAR is developed and maintained as a stand-alone program. Even if the latest version of ECOSAR (v.1.11) has been included in the latest version of EPI Suite (v.4.11), it is still recommended to use the ECOSAR stand-alone program because it will inform the user if an experimental value is available for the target chemical while the ECOSAR integrated in EPI Suite does not have this functionality.

Link to the (Q)SAR program: <http://www.epa.gov/tsca-screening-tools/ecological-structure-activity-relationships-ecosar-predictive-model>

#### b) How to check the reliability of the prediction

| Information on the (Q)SAR model             | Where to find this information  | How to use this information   |
|---|---|---|
| Description of the algorithm                | Available in the ECOSAR Help <sup>8</sup> , in section "Interpreting QSAR Class Ref Docs" and in each "QSAR Equation Document" of the various chemical classes. | See Section 3.1 of this practical guide.  |
| Statistics (goodness-of-fit and Robustness) | Available in the ECOSAR Help, in each "QSAR Equation Document" of the various chemical classes.   | See Section 3.1 of this practical guide. For instance, the user should check that: <ul style="list-style-type: none"> <li>- <math>R^2</math> (i.e. correlation or determination coefficient) is not too low (e.g. not below 0.7),</li> <li>- The data points are not too few. This is especially important as several ECOSAR classes contain only one or few data points which can lead to unreliable predictions.</li> <li>- The linear regression is not built on scattered data points.</li> </ul> |

<sup>8</sup> ECOSAR Help can be accessed by clicking on the tab "Help" at the top of the ECOSAR window.

| Information on the applicability domain  | Where to find this information  | How to use this information?  |
|--|---|---|
| General applicability domain (chemical classes covered/not covered by the model) | Available in the ECOSAR Help, in sections "Chemicals That Should Not Be Profiled", "Surfactants" and "Special ECOSAR Classes / Dyes". | <p>The user should check whether its chemical is one of those "That Should Not Be Profiled" with ECOSAR or if it should be predicted using a special ECOSAR class.</p> <p>In addition, it should be kept in mind that most of the ECOSAR models are based on the relationships between log Kow and toxicity (LC50 or EC50) addressing the uptake of chemicals through the aqueous phase. For very hydrophobic or very sorptive substances, the uptake from food can also be an important exposure pathway. Moreover, other properties of the substance can trigger specific modes of action (e.g.: the substance is likely to be more reactive if it is protein binding). Finally, it should also be kept in mind that some chemical classes expressing excess toxicity (compared to the baseline toxicity model) have not yet been included in the ECOSAR program.</p> |
| Descriptor domain  | Available in the ECOSAR Help, in each "QSAR Equation Document" of the various chemical classes and in the ECOSAR results window.      | <p>The user should check that the molecular weight (MW) of the target chemical does not exceed 1 000. The MW of the target chemical is indicated in the ECOSAR results window.</p> <p>If the log Kow of the compound exceeds the maximum log Kow of the class-specific model (e.g. for Fish 96h LC50, the maximum log Kow varies between class models from 2.6 to 8.2, often being 5) or if the predicted effect (e.g. Fish 96h LC50) is lower than the water solubility of the compound, then the prediction might be unreliable and long-term toxicity data might be more appropriate. Detailed instructions on how to check these points is given in section c) below.</p>   |

| Structural fragment domain      | Can be determined from each "QSAR Equation Document" of the various chemical classes (except for the "dyes" and "surfactants" special classes).   | Each "QSAR Equation Document" gives the training set of the specific class model. In these training sets, the CAS numbers and the chemical names are provided (if not claimed confidential (CBI)). This requires manual work since the training sets are not in a structurally searchable format. Software programs <sup>9</sup> are available to help the user derive the chemical structures if needed. In addition, the whole structural domain of the model cannot be determined if there are CBI compounds in the training set. |
|---------------------------------|---|--|
| Training set and validation set | Where to find this information  | How to use this information  |
|                                 | <p>The training sets of all models (except the special classes for dyes and surfactants) are available from the "QSAR equation document" of each chemical class. However, some of the compounds in the training set have been kept confidential (CBI).</p> <p>Validation (test) sets do not exist in ECOSAR.<sup>10</sup></p> | The user should check whether there are close analogues in the training set of the specific class model (based on the CAS numbers or on the chemical names provided). If there are close analogues, the user should compare their predicted and experimental values.   |

### c) Additional information on ECOSAR

- In ECOSAR, all predictions (except for surfactants and dyes) are based on log Kow. By default, ECOSAR estimates Kow using KOWWIN. However, if users have a reliable measured log Kow they should insert it in the data entry screen (see screenshot below). This value will be taken into account by the model and will reduce the uncertainty of the prediction.

<sup>9</sup> Several software programs allow users to derive the chemical structure from the CAS number or from the chemical name. Some of these programs are freely available (e.g. Chemspider or the OECD QSAR Toolbox).

<sup>10</sup> However, there are several peer-reviewed publications assessing the external performance of ECOSAR.

- If there is an experimental result for the target chemical in the training set then this data will be displayed in the ECOSAR results window (see “Available Measured Data from ECOSAR Training Set” in the figure below). In this case, it is recommended to report this experimental data and the ECOSAR estimate in two different endpoint study records.

| CAS No      | Organism | Duration | End Pt | Measured mg/L (ppm) | Ecosar Class     | Reference |
|-------------|----------|----------|--------|---------------------|------------------|-----------|
| 000050-00-0 | Fish     | 96-hr    | LC50   | 24.1                | Aldehydes (Mono) | DUL       |

- If the compound has been allocated to a specific class (e.g. “Aldehydes, mono” class) then the effect level of this class should be taken into account and not only the one from the “Neutral organics” class (i.e. baseline toxicity potential). In general, if the program identifies several classes, it is recommended to use the most conservative effect level from any of these classes (and to consider the potential synergistic toxicity effect of these various classes).
- Each of the “QSAR Equation Document” is class-specific and some contain more information than others under the titles “APPLICATION” and “LIMITATIONS” (see examples below). Therefore, users should carefully read the “QSAR Equation Document” of the specific class/endpoint for which they want to perform a prediction.

**APPLICATION:**

This SAR may be used to estimate the toxicity of aldehydes (mono) with log Kow values of less than 5.0 and molecular weights less than 1000. Acrolein is about 1400 times more toxic than predicted by this SAR.

**LIMITATIONS:**

Aliphatic polyamines with greater than 3 aliphatic amines and/or an amine-nitrogen composition of  $\geq 25\%$  exhibit excess toxicity based on available CBI data. Insufficient data were available to construct a QSAR, but fish toxicity test data of compounds with 27% amine-nitrogen resulted in LC<sub>50</sub> values that were ~100x more toxic than estimations predicted from the aliphatic amine class.

- As indicated in the table from section b) (see Descriptor domain), if the log Kow of the compound exceeds the maximum log Kow of the class-specific model or if the predicted effect value is lower than the water solubility of the compound, then the prediction might be unreliable (see ECOSAR results window below).

The screenshot shows the 'Ecosar Results' window with the following content:

**Values used to Generate ECOSAR Profile**

Log Kow: 5.252 (EPISuite Kowwin v1.68 Estimate)  
 Wat Sol: 0.052 (mg/L, PhysProp DB exp value)

**ECOSAR v1.1 Class-specific Estimations**

**Neutral Organics**

| ECOSAR Class     | Organism | Duration | End Pt | Predicted mg/L (ppm) |
|------------------|----------|----------|--------|----------------------|
| Neutral Organics | : Fish   | 96-hr    | LC50   | 0.140 *              |
| Neutral Organics | : Fish   |          | ChV    | 0.020                |

**Note:** \* = asterisk designates: Chemical may not be soluble enough to measure this predicted effect. If the effect level exceeds the water solubility by 10X, typically no effects at saturation (NES) are reported.

**Class Specific LogKow Cut-Offs**

If the log Kow of the chemical is greater than the endpoint specific cut-offs presented below, then no effects at saturation are expected for those endpoints.

**Neutral Organics:**

Maximum LogKow: 5.0 (Fish 96-hr LC50; Daphnid LC50, Mysid LC50)

- Ecotoxicity chronic values (ChV) can be predicted with ECOSAR. However, users should pay attention to the following points:
  - The (Q)SAR models available for predicting these chronic values are often built on small to very small training sets.
  - The chronic value (ChV) is defined as the geometric mean of the no observed effect concentration (NOEC) and the lowest observed effect concentration (LOEC). However, under REACH, NOECs are the effect concentrations used to assess long-term toxicity data to aquatic organisms. A proxy to derive the NOEC is to divide the ChV by  $\sqrt{2}$ .
  - Acute-to-chronic ratios (ACRs) are used by ECOSAR when measured data are lacking within a class. Such predictions are flagged with an exclamation mark (!) in the ECOSAR results window (see screenshot below) and should be considered with caution.

The screenshot shows the 'Ecosar Results' window with the following content:

**Phenol Amines** : Fish ChV 0.565 !

**NOTE:** ! = exclamation designates: The toxicity value was estimated through application of acute-to-chronic ratios per methods outlined in the ECOSAR Methodology Document provided in the ECOSAR Help Menu.



## 4.4 Acute toxicity to rat (T.E.S.T.)

### a) Introduction

Acute toxicity by oral route is a REACH requirement for all substances produced or imported above one tonne/year (REACH Annex VII). The preferred test species according to the OECD test guidelines is the rat, and the endpoint to be derived is the LD50 (generally expressed in mg/kg body weight).

Most of the software programs that predict acute oral toxicity are commercial except T.E.S.T. (Toxicity Estimation Software Tool), which is made freely available by US EPA. This tool offers four different methods to predict acute oral toxicity to rats: hierarchical, FDA, nearest neighbor and consensus methods.

The consensus method predicts the toxicity simply by taking an average of the predicted toxicities from the three other methods (hierarchical, FDA and nearest neighbor methods). This consensus method should be the preferred one to use as it achieved the best results for prediction accuracy and leverage compared to the three other methods. To check the reliability of the predictions coming from this consensus method, the user should assess the reliability of the three other methods. Therefore, the information given in the following table addresses all of these four methods.

At the time of writing this manual, the current version of T.E.S.T. is 4.1, which has been used to prepare this example.

Link to the (Q)SAR program: <http://www2.epa.gov/chemical-research/toxicity-estimation-software-tool-test>

### b) How to check the reliability of the prediction

| Information on the (Q)SAR model | Where to find this information  | How to use this information   |
|---------------------------------|---|---|
| Description of the algorithm    | Available in the T.E.S.T. User's Guide <sup>11</sup> , in sections 1.2 and 2.2 called "QSAR Methodologies". | <p>The four methods are transparently described in Section 2.2 of the User's Guide. It should be noted that none of them involve mechanistic interpretations.</p> <p>Concerning the hierarchical and FDA methods, they are based on clustering and genetic algorithms that lead to equations (and descriptors) that vary depending on the target chemical.</p> <p>Concerning the nearest neighbor method, the predicted toxicity is the average of the toxicities of the three most similar chemicals (structural analogues) in the training set.</p> |

<sup>11</sup> This User's Guide can be accessed by clicking on the tab "Help" at the top right of the T.E.S.T. window.

| <p>Statistics (goodness-of-fit and robustness)</p>                                      | <p>Available in the T.E.S.T. User's Guide, in Sections 2.2.1, 2.2.2, 2.3.1 and 4.4.1.</p>     | <p>In Section 4.4.1 of the User's Guide, it is written that <math>R^2</math> is less than 0.6 for the hierarchical, FDA and nearest neighbour methods and that <math>\frac{R^2 - R_0^2}{R^2}</math> is more than 0.1 for all of the four methods. Therefore, those methods do not satisfy the conditions for an acceptable predictive power as indicated in Section 2.3.1 of the T.E.S.T. User's Guide.</p> <p>In addition it is stated in Section 4.4.1 of the User's Guide that <i>"The prediction statistics for this endpoint were not as good as those for the other endpoints. This is not surprising since this endpoint has a higher degree of experimental uncertainty and has been shown to be more difficult to model than other endpoints"</i>.</p>            |
|---|---|--|
| <p>Information on the applicability domain</p>  | <p>Where to find this information</p>   | <p>How to use this information</p>   |
| <p>General applicability domain (chemical classes covered/not covered by the model)</p> | <p>Available in the T.E.S.T. User's Guide, in Section 3.4.</p>                                | <p>For instance, in this Section 3.4 of the User's Guide, it says that <i>"salts, undefined isomeric mixtures, polymers, or mixtures were removed [from the training set]"</i>. Therefore, those types of substances should not be predicted with T.E.S.T.</p>   |
| <p>Descriptor domain</p>  | <p>Information can be found in the T.E.S.T. User's Guide (Sections 2.1 and 2.2.1, 2.2.2).</p> | <p>For the hierarchical and FDA methods, 797 descriptors can be used in the model equation depending on the target chemical. These methods give a prediction only if the target chemical is within the ranges of descriptor values for the chemicals in the cluster (for the descriptors appearing in the cluster model).</p> <p>Concerning the nearest neighbor method, the descriptor differences between the target chemical and its structural analogues are not accounted for by T.E.S.T.</p> <p>To increase the confidence in the predictions given by these three methods, the user could potentially check that the target chemical is within the ranges of log Kow and molecular weight of the compounds in the cluster (or of the three nearest neighbours).</p> |

|  |  |  |
|--|--|--|
| Structural fragment domain             | Information related to that can be found in the T.E.S.T. User's Guide (Sections 2.2.1 and 2.2.2).  | The hierarchical and FDA methods give a prediction only if the compounds in the cluster have at least one example of each of the fragments contained in the target chemical.<br><br>Concerning the nearest neighbor method, the structural differences between the target chemical and its structural analogues are not accounted for by T.E.S.T. Therefore, users should check this visually from the T.E.S.T. report.  |
| <b>Training set and validation set</b> | <b>Where to find this information</b>  | <b>How to use this information</b>   |
|  | The results report of each of the four methods display the predictions for the most similar chemicals in the validation set (i.e. prediction set or external test set) followed by the predictions for the most similar chemicals in the training set.<br><br>In addition, the training and validation sets can be downloaded as SDF files through the link given at the bottom of this web page:<br><a href="http://www2.epa.gov/chemical-research/toxicity-estimation-software-tool-test">http://www2.epa.gov/chemical-research/toxicity-estimation-software-tool-test</a> | In the results report of each method, for the predictions performed on the analogues in the validation and training sets, the user should check: <ul style="list-style-type: none"> <li>• whether these predicted values are close to the experimental values and,</li> <li>• whether the mean absolute error (MAE) for these analogues is not more than the MAE of the entire set (this would be displayed as a red cell instead of a green cell).</li> </ul> |

### c) Additional information on T.E.S.T.

- If there is an experimental result for the target chemical in the training set or in the validation set (i.e. prediction set or external test set) then this data will be displayed in the T.E.S.T. results report (see "aNote" in the screenshot below). The user can check this experimental value by clicking on the name of the source (in the example below by clicking on "ChemidPlus"). In this case, it is recommended to report this experimental data and the T.E.S.T. estimate in two different endpoint study records.

Prediction results

| Endpoint                                 | Experimental value (CAS= 28785-06-0)<br>Source: <a href="#">ChemidPlus</a> | Predicted value <sup>a</sup> |
|--|--|------------------------------|
| Oral rat LD <sub>50</sub> -Log10(mol/kg) | 1.97   | 1.92                         |
| Oral rat LD <sub>50</sub> mg/kg          | 1599.21  | 1762.18                      |

<sup>a</sup>Note: the test chemical was present in the external test set.

- Some methods (hierarchical, FDA and group contribution methods) give the "Prediction interval" (i.e. 90% confidence interval) in the results report. When using these, the user should check that this interval does not overlap with a regulatory threshold (e.g. CLP, persistent, bioaccumulative and toxic (PBT) or risk assessment thresholds).

## Appendix 1. QSAR models related to REACH endpoints

The QSAR computer programs listed in this appendix are widely known and are given to inform REACH registrants on the QSAR models availability for each of the REACH endpoints. However, it constitutes neither an exhaustive list of available programs nor a list of regulatory validated QSAR models. So far, most of the toxicological and ecotoxicological information required under REACH can rarely be fulfilled with QSAR predictions alone.

### a. Physicochemical endpoints

| Endpoint                        | Software tool                      | Models/Modules                               | Free or Commercial |
|---------------------------------|------------------------------------|--|--------------------|
| Melting/freezing point          | EPI Suite (US EPA)                 | MPBPVP                                       | Free               |
| Boiling point                   | EPI Suite (US EPA)                 | MPBPVP                                       | Free               |
|                                 | T.E.S.T. (US EPA)                  | Normal boiling point                         | Free               |
|                                 | ACD/Percepta (ACD/Labs)            | Boiling Point/Vapor Pressure Module          | Commercial         |
| Relative density                | T.E.S.T. (US EPA)                  | Density                                      | Free               |
| Vapour pressure                 | EPI Suite (US EPA)                 | MPBPVP                                       | Free               |
|                                 | T.E.S.T. (US EPA)                  | Vapor pressure at 25°C                       | Free               |
|                                 | ACD/Percepta (ACD/Labs)            | Boiling Point/Vapor Pressure Module          | Commercial         |
| Surface tension                 | T.E.S.T. (US EPA)                  | Surface tension at 25°C                      | Free               |
| Water solubility                | EPI Suite (US EPA)                 | WSKOW and WATERNT                            | Free               |
|                                 | T.E.S.T. (US EPA)                  | Water solubility at 25°C                     | Free               |
|                                 | ACD/Percepta (ACD/Labs)            | Aqueous Solubility Module                    | Commercial         |
|                                 | ADMET Predictor (Simulations Plus) | Physicochemical and Biopharmaceutical Module | Commercial         |
|                                 | Discovery Studio (Accelrys)        | ADMET Descriptors                            | Commercial         |
| Partition coefficient (log Kow) | EPI Suite (US EPA)                 | KOWWIN                                       | Free               |
|                                 | VEGA (IRFMN)                       | LogP Models                                  | Free               |
|                                 | ACD/Percepta (ACD/Labs)            | LogP Module                                  | Commercial         |
|                                 | ADMET Predictor (Simulations Plus) | Physicochemical and Biopharmaceutical Module | Commercial         |
|                                 | JChem (ChemAxon)                   | LogP/logD predictor                          | Commercial         |
| Flash point                     | T.E.S.T. (US EPA)                  | Flash point                                  | Free               |
| Dissociation constant           | Danish QSAR Database (DTU)         | pKa from ACD/Labs                            | Free               |
|                                 | ACD/Percepta (ACD/Labs)            | pKa Module                                   | Commercial         |
|                                 | ADMET Predictor (Simulations Plus) | Physicochemical and Biopharmaceutical Module | Commercial         |
|                                 | JChem (ChemAxon)                   | pKa predictor                                | Commercial         |
| Viscosity                       | T.E.S.T. (US EPA)                  | Viscosity at 25°C                            | Free               |

## b. Environmental fate and pathways endpoints

| Endpoint <sup>12</sup>             | Software tool               | Models/Modules                                | Free or Commercial |
|------------------------------------|-----------------------------|---|--------------------|
| Hydrolysis                         | EPI Suite (US EPA)          | HYDROWIN                                      | Free               |
| Ready biodegradability             | Danish QSAR Database (DTU)  | Not Ready Biodegradability model from DTU     | Free               |
|                                    | EPI Suite (US EPA)          | BIOWIN and BioHCwin                           | Free               |
|                                    | VEGA (IRFMN)                | IRFMN model                                   | Free               |
|                                    | CATALOGIC (LMC)             | Several OECD 301 models                       | Commercial         |
|                                    | Discovery Studio (Accelrys) | Aerobic Biodegradability model                | Commercial         |
|                                    | Meta-PC (MultiCASE)         | Aerobic Microbial Biodegradation expert rules | Commercial         |
| Bioaccumulation in aquatic species | EPI Suite (US EPA)          | BCFBAF  | Free               |
|                                    | T.E.S.T. (US EPA)           | Bioaccumulation factor                        | Free               |
|                                    | VEGA (IRFMN)                | CAESAR, Meylan and KNN/Read-Across models     | Free               |
|                                    | CASE Ultra (MultiCASE)      | EcoTox model bundle                           | Commercial         |
|                                    | CATALOGIC (LMC)             | Two BCF base-line models                      | Commercial         |
| Adsorption/desorption screening    | EPI Suite (US EPA)          | KOCWIN  | Free               |

---

<sup>12</sup> The REACH requirement “Simulation testing in water, soil or sediment and identification of degradation products” is not listed in this table because – to our knowledge – there are no QSAR tools/models available for this endpoint.

### c. Ecotoxicological endpoints

| Endpoint <sup>13</sup>                                 | Software tool                      | Models/Modules                                       | Free or Commercial |
|--|------------------------------------|--|--------------------|
| Short-term toxicity to fish                            | Danish QSAR Database (DTU)         | Fathead minnow 96h LC50 from DTU                     | Free               |
|  | ECOSAR (US EPA)                    | Fish, 96-hr, LC50                                    | Free               |
|  | T.E.S.T. (US EPA)                  | Fathead minnow LC50 (96 hr)                          | Free               |
|  | VEGA (IRFMN)                       | SarPy/IRFMN classification and KNN/Read-Across model | Free               |
|  | ADMET Predictor (Simulations Plus) | Toxicity module                                      | Commercial         |
|  | CASE Ultra (MultiCASE)             | EcoTox model bundle                                  | Commercial         |
|  | Discovery Studio (Accelrys)        | Fathead Minnow LC50                                  | Commercial         |
| Long-term toxicity to fish                             | ECOSAR (US EPA)                    | Fish, ChV <sup>14</sup>                              | Free               |
| Short-term toxicity to aquatic invertebrates (daphnia) | Danish QSAR Database (DTU)         | Daphnia magna 48h EC50 from DTU                      | Free               |
|  | ECOSAR (US EPA)                    | Daphnid, 48-hr, LC50                                 | Free               |
|  | T.E.S.T. (US EPA)                  | Daphnia magna LC50 (48 hr)                           | Free               |
|  | ADMET Predictor (Simulations Plus) | Toxicity module                                      | Commercial         |
|  | Discovery Studio (Accelrys)        | Daphnia EC50   | Commercial         |
| Long-term toxicity to aquatic invertebrates (daphnia)  | ECOSAR (US EPA)                    | Daphnid, ChV <sup>11</sup>                           | Free               |
| Toxicity to aquatic plants (algae)                     | Danish QSAR Database (DTU)         | Pseudokirchneriella s. 72h EC50 from DTU             | Free               |
|  | ECOSAR (US EPA)                    | Green Algae, 96-hr, EC50                             | Free               |
| Short-term toxicity to terrestrial invertebrates       | ECOSAR (US EPA)                    | Earthworm, 14-day, LC50                              | Free               |

<sup>13</sup> The following REACH requirements are not listed in this table because – to our knowledge – there are no QSAR tools/models available for these endpoints:

- Toxicity to aquatic micro-organisms (activated sludge respiration inhibition testing),
- Long-term toxicity to sediment organisms,
- Long-term toxicity to terrestrial invertebrates,
- Short-term toxicity to terrestrial plants,
- Long-term toxicity to terrestrial plants,
- Toxicity to terrestrial micro-organisms, and
- Long-term toxicity to birds.

<sup>14</sup> See section 4.3 c) of this practical guide for further information on these chronic values.

#### d. Toxicological endpoints

| Endpoint                          | Software tool                      | Models/Modules   | Free or Commercial |
|-----------------------------------|------------------------------------|--|--------------------|
| Acute toxicity                    | Danish QSAR Database (DTU)         | Models for acute toxicity in rodents from ACD/Labs               | Free               |
|                                   | T.E.S.T. (US EPA)                  | Oral rat LD50  | Free               |
|                                   | ACD/Percepta (ACD/Labs)            | Acute Toxicity Module  | Commercial         |
|                                   | ADMET Predictor (Simulations Plus) | Toxicity module  | Commercial         |
|                                   | CASE Ultra (MultiCASE)             | AcuteTox model bundle  | Commercial         |
|                                   | Discovery Studio (Accelrys)        | Rat oral LD50 and rat inhalation toxicity LC50                   | Commercial         |
| Skin irritation or skin corrosion | Danish QSAR Database (DTU)         | Skin irritation model  | Free               |
|                                   | OECD QSAR Toolbox                  | Skin irritation/corrosion Inclusion (and Exclusion) rules by BfR | Free               |
|                                   | ToxTree (JRC)                      | Skin irritation / skin corrosion                                 | Free               |
|                                   | ACD/Percepta (ACD/Labs)            | Irritation Module  | Commercial         |
|                                   | CASE Ultra (MultiCASE)             | SkinEye Toxicity model bundle                                    | Commercial         |
|                                   | Derek (Lhasa)                      | Irritation (of the skin) alerts                                  | Commercial         |
|                                   | Discovery Studio (Accelrys)        | Skin irritancy   | Commercial         |
| Eye irritation                    | OECD QSAR Toolbox                  | Eye irritation/corrosion Inclusion (and Exclusion) rules by BfR  | Free               |
|                                   | ToxTree (JRC)                      | Eye irritation and corrosion                                     | Free               |
|                                   | ACD/Percepta (ACD/Labs)            | Irritation Module  | Commercial         |
|                                   | CASE Ultra (MultiCASE)             | SkinEye Toxicity model bundle                                    | Commercial         |
|                                   | Derek (Lhasa)                      | Irritation (of the eye) alerts                                   | Commercial         |
|                                   | Discovery Studio (Accelrys)        | Ocular irritancy   | Commercial         |
| Skin sensitisation                | Danish QSAR Database (DTU)         | Allergic Contact Dermatitis model                                | Free               |
|                                   | OECD QSAR Toolbox                  | Protein binding alerts for skin sensitisation by OASIS           | Free               |
|                                   | ToxTree (JRC)                      | Skin sensitisation reactivity domains                            | Free               |
|                                   | VEGA (IRFMN)                       | CAESAR model   | Free               |
|                                   | ACD/Percepta (ACD/Labs)            | Irritation Module  | Commercial         |
|                                   | CASE Ultra (MultiCASE)             | SkinEye Toxicity model bundle                                    | Commercial         |
|                                   | Derek (Lhasa)                      | Skin sensitisation   | Commercial         |
|                                   | Discovery Studio (Accelrys)        | Skin sensitization   | Commercial         |
| Repeated dose toxicity            | ADMET Predictor (Simulations Plus) | Toxicity module  | Commercial         |
|                                   | CASE Ultra (MultiCASE)             | Several model bundles associated with repeated dose toxicity     | Commercial         |
|                                   | Derek (Lhasa)                      | Several endpoints associated with repeated dose toxicity         | Commercial         |

| Endpoint  | Software tool                      | Models/Modules  | Free or Commercial |
|---|------------------------------------|---|--------------------|
|   | Discovery Studio (Accelrys)        | Rat Chronic (Oral) LOAEL  | Commercial         |
|   | Leadscope                          | Several models associated with repeated dose toxicity                         | Commercial         |
| <i>In vitro</i> gene mutation in bacteria (Ames test)                         | Danish QSAR Database (DTU)         | Models for Ames test  | Free               |
|   | OECD QSAR Toolbox                  | Several profilers (alerts) associated with this endpoint                      | Free               |
|   | T.E.S.T. (US EPA)                  | Mutagenicity  | Free               |
|   | ToxTree (JRC)                      | <i>In vitro</i> mutagenicity (Ames test) alerts by ISS                        | Free               |
|   | VEGA (IRFMN)                       | CAESAR, SarPy/IRFMN, ISS and KNN/Read-Across models                           | Free               |
|   | ACD/Percepta (ACD/Labs)            | Genotoxicity Module   | Commercial         |
|   | CASE Ultra (MultiCASE)             | Bacterial mutagenicity model bundle   | Commercial         |
|   | Derek (Lhasa)                      | Mutagenicity <i>in vitro</i>  | Commercial         |
|   | Discovery Studio (Accelrys)        | Ames Mutagenicity   | Commercial         |
|   | Leadscope                          | Genetox Expert Alerts Suite and Non-human Genetic Toxicity Suite              | Commercial         |
|   | TIMES (LMC)                        | Ames mutagenicity   | Commercial         |
| Mutagenicity (other endpoints than <i>in vitro</i> gene mutation in bacteria) | Danish QSAR Database (DTU)         | Models for genotoxicity endpoints   | Free               |
|   | OECD QSAR Toolbox                  | Several profilers (alerts) associated with mutagenicity                       | Free               |
|   | ToxTree (JRC)                      | Several decision trees associated with mutagenicity                           | Free               |
|   | CASE Ultra (MultiCASE)             | EcoTox model bundle   | Commercial         |
|   | Derek (Lhasa)                      | Chromosome damage <i>in vitro</i>   | Commercial         |
|   | Leadscope                          | Non-human Genetic Toxicity Suite  | Commercial         |
|   | TIMES (LMC)                        | Several models associated with mutagenicity                                   | Commercial         |
| Reproductive toxicity   | Danish QSAR Database (DTU)         | Models for Endocrine endpoints and model for Teratogenic Potential in Humans  | Free               |
|   | VEGA (IRFMN)                       | CAESAR and PG models  | Free               |
|   | ADMET Predictor (Simulations Plus) | Toxicity module   | Commercial         |
|   | CASE Ultra (MultiCASE)             | Several model bundles associated with reproductive and developmental toxicity | Commercial         |
|   | Derek (Lhasa)                      | Several endpoints associated with reproductive toxicity                       | Commercial         |
|   | Discovery Studio (Accelrys)        | Developmental Toxicity Potential  | Commercial         |
|   | Leadscope                          | Several models associated with reproductive and developmental toxicity        | Commercial         |
|   | TIMES (LMC)                        | Androgen, AHR and Estrogen (receptor) binding affinity models                 | Commercial         |



## e. Information on the Danish (Q)SAR Database

A new version of the Danish (Q)SAR database has been released on November 2015 and is publicly available at the following link: <http://qsar.food.dtu.dk/>. This database contains (Q)SAR predictions for physicochemical properties, ecotoxicity, environmental fate, ADME and toxicity of more than 600 000 chemical structures.

When possible, models from the Technical University of Denmark and some commercial models have been modelled in the three software systems Leadscope, CASE Ultra (MultiCASE) and SciQSAR. Some model predictions from ACD/Labs and US EPA (EPI Suite and ECOSAR) have also been integrated in the database.

However, it should be noted that the database does not provide the possibility to refine the predictions as some of the source software programs do (e.g. ECOSAR). In addition, the database does not provide as much details on the results as the source software programs and it is not updated on a regular basis. Therefore, whenever possible, the predictions given by the database should be compared to the results obtained from the source software programs themselves.

A REACH registrant who would like to report in its IUCLID registration dossier a prediction coming from the Danish (Q)SAR database should also check that the (Q)SAR model is valid – by comparing the points given in Section 3.1 of this practical guide with the information given in the model's QMRF – and should attach a QPRF (for each prediction) to the IUCLID endpoint study record.

## Appendix 2. Further guidance documents and links

### a. Guidance documents providing information about (Q)SARs

A summary on how to use non-testing data obtained by applying (Q)SARs is available in the Guidance on information requirements and chemical safety assessment in Chapter R.4.3.2.1 (Q)SAR data:

[Chapter R.4: Evaluation of available information](#)

A dedicated part on computational methodologies is available in the Guidance on information requirements and chemical safety assessment in Chapter R.6.1 Guidance on (Q)SARs:

[Chapter R.6: \(Q\)SARs and grouping of chemicals](#)

Relevant tools and approaches for the endpoints of interest are offered by each endpoint specific guidance document included in the Guidance on information requirements and chemical safety assessment in:

[Chapter R.7: Endpoint specific Guidance](#)

For human health, the available (Q)SARs may be suitable mostly for hazard identification, in particular in a weight of evidence approach as described in the Guidance on information requirements and chemical safety assessment in:

[Chapter R.8: Characterisation of dose \[concentration\]-response for human health](#)

Several (Q)SAR tools which may be used to determine the Predicted No-Effect Concentrations (PNECs) are listed in the Guidance on information requirements and chemical safety assessment in Chapter R.10.2.2.2 (Q)SAR and grouping approaches:

[Chapter R.10: Characterisation of dose \[concentration\]-response for environment](#)

Information on the use of non-testing degradation and bioaccumulation data for persistent, bioaccumulative and toxic (PBT) chemicals is accessible from the Guidance on information requirements and chemical safety assessment in:

[Chapter R.11: PBT Assessment](#)

### b. Other useful links

OECD (Q)SAR Toolbox:

<http://www.qsartoolbox.org/>

OECD Global Portal (eChemPortal):

[http://www.echemportal.org/echemportal/index?pageID=0&request\\_locale=en](http://www.echemportal.org/echemportal/index?pageID=0&request_locale=en)

Data Submission Manual 5 'How to complete a technical dossier for registrations and PPORD notifications':

[http://echa.europa.eu/documents/10162/13653/dsm5\\_tech\\_dossier\\_en.pdf](http://echa.europa.eu/documents/10162/13653/dsm5_tech_dossier_en.pdf)

EUROPEAN CHEMICALS AGENCY  
ANNANKATU 18, P.O. BOX 400,  
FI-00121 HELSINKI, FINLAND  
ECHA.EUROPA.EU