

# Illustrative example with the OECD QSAR Toolbox workflow

Part 1: Introductory note



### Legal notice

This document aims to assist users in using the OECD QSAR Toolbox workflows in a systematic and reproducible manner. However, users are reminded that the text of the REACH Regulation is the only authentic legal reference and that the information in this document does not constitute legal advice. Usage of the information remains under the sole responsibility of the user. The European Chemicals Agency does not accept any liability with regard to the use that may be made of the information contained in this document.

### Illustrative examples with the OECD QSAR Toolbox workflow

#### Part 1: Introductory note

**Reference:** ECHA-14-R-06-EN  
**Cat. number:** ED-02-14-123-EN-N  
**ISBN:** 978-92-9244-432-7  
**DOI:** 10.2823/83029  
**Date:** February 2014  
**Language:** English

© European Chemicals Agency, 2014

Reproduction is authorised provided the source is fully acknowledged in the form “Source: European Chemicals Agency, <http://echa.europa.eu/>”, and provided written notification is given to the ECHA Communications Unit ([info@echa.europa.eu](mailto:info@echa.europa.eu)).

If you have questions or comments in relation to this document please send them (quoting the reference and issue date) using the information request form. The 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. BACKGROUND AND DEFINITIONS</b>	<b>5</b>
2.1 What is the Toolbox?	5
2.2 What are the key features of the Toolbox?	5
<b>3. GENERAL RECOMMENDATIONS APPLICABLE TO THE USE OF THE TOOLBOX</b>	<b>6</b>
3.1 Input	6
3.2 Profiling	6
3.3 Endpoint	6
3.4 Category definition	7
3.5 Data gap filling	7
3.6 Report	9
<b>4. ADDITIONAL CONSIDERATIONS</b>	<b>10</b>
<b>5. FURTHER INFORMATION</b>	<b>11</b>

# 1. Introduction

The objectives of REACH (Regulation (EC) No 1907/2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals) are set out in its preamble, including recitals 1 and 38, and in Article 13, which all underline the need to generate data by means other than animal tests whenever this is possible. Annex XI of REACH encloses the general rules for adaptation of the standard testing regime, allowing among others the use of weight of evidence, quantitative structure activity relationships ((Q)SARs), grouping of substances and read-across approaches in the case that these methodologies can provide reliable and scientifically valid results. The conditions/criteria that each approach has to meet are listed in the same annex.

If the requirements are satisfied, unnecessary testing can be avoided. Computational predictions from valid (Q)SARs may indicate the presence or absence of a certain dangerous property. They can also support a conclusion for a REACH endpoint as part of a weight-of-evidence approach.

This publication has two sections: Part 1, an introductory note, which provides background information on the OECD QSAR Toolbox including general considerations; and Part 2, which contains case studies for selected substances.

The document contains advice on how to obtain and report endpoint-specific predictions that may comply with the requirements for the registration of substances under REACH. The proposed methodology is a relatively standard approach which uses the same logic in the selection of analogues (same workflows). The examples show the expected level of information and include explanatory comments, which expand on the reasoning and the approach taken.

ECHA aims to support industry to improve the quality of the information provided in their registration dossiers. This applies both to updates of existing dossiers, and to those being prepared mainly for the registration deadline of 1 June 2018.

The objective of this document is to illustrate:

- What the Toolbox is and how to generate predictions with it.
- How to obtain a group of substances consistent from a chemical and mechanistic point of view, suitable for prediction.
- How to fill a data gap for qualitative and quantitative endpoints following a systematic and reproducible approach.
- How to explain a prediction made by using the Toolbox.

## 2. Background and definitions

The use of computational techniques to predict properties of chemicals is nowadays established in many scientific fields, from drug discovery to environmental sciences, thus new software are constantly released and new algorithms are developed.

The OECD QSAR Toolbox has been designed as a tool that could support the requirements set by REACH. However, it leaves the user to choose the criteria to form a chemical category, the experimental data to be considered and the method for filling the data gap. Thus, the result may be affected by the level of experience in the (eco)toxicology field and software knowledge of the user. Therefore, this section explains concepts and terminology of the QSAR Toolbox and provides general advice for its use in the context of prediction for filling data gaps under REACH.

### 2.1 WHAT IS THE TOOLBOX?

The Toolbox is a software that offers the possibility for grouping and filling gaps in physico-chemical, environmental and human health hazard data. Grouping chemicals into categories according to the chemical similarity or a common mode of action is crucial to the workflow. The Toolbox can serve as a source or link to experimental data. Different estimation techniques are possible for filling the data gaps. These include read-across, trend analysis or QSAR.

### 2.2 WHAT ARE THE KEY FEATURES OF THE TOOLBOX?

The Toolbox:

- Collects databases with results from experimental studies, providing references and details about the experimental conditions.
- Profiles the molecules according to their chemical structure or specific mode of action in order to define criteria for the category formation, providing the basis to explain the presence/absence of hazards or other properties. The Toolbox also includes simulators for the transformation of chemicals by different organisms and different organs (e.g. microbial metabolism, liver and skin metabolism), as well as simulators for abiotic transformations (e.g. hydrolysis, dissociation, auto-oxidation). Where observed transformation data are available, the tool could also retrieve these.
- Includes tools to predict missing experimental values by read-across and trend analysis from tested to untested chemicals within a category and/or by (Q)SAR models.

## 3. General recommendations applicable to the use of the Toolbox

This chapter lists a series of recommendations applicable to the use of the Toolbox. The layout follows the order of the modules of the prediction workflow included in the software.

### 3.1 INPUT

In the “Input” module, the user defines the identity of the target chemical for the prediction procedure. The identity of the chemical can be entered by name, CAS number, by drawing the molecule, by SMILES or by selecting it from a list, a database, an inventory or a file. In some cases, more than one chemical identity could be retrieved (e.g. several CAS numbers may be associated to the same structure). The user can retain only one structure for the subsequent workflow. The selected substance identities are then added to the data matrix where a colour code indicates the reliability of the chemical identifiers (green/high, yellow/moderate, red/poor). The user can then proceed to the next module.

### 3.2 PROFILING

The module “Profiling” identifies the main characteristics of the molecules in the matrix according to chemical features (e.g. organic functional groups present in the structure) or to mode of action (e.g. protein binding alerts for skin sensitisation). Indeed, not all profilers are relevant for all regulatory endpoints. While some general chemical or mechanistic profilers such as “organic functional group” or “protein binding” are relevant for many endpoints (skin sensitisation, acute toxicity, aquatic toxicity), some profilers are relevant for building categories for very specific endpoints only. It should be noted that the results of the profiling are not enough to fill a data gap. Profilers are not typically developed to predict an endpoint per se but to facilitate the grouping according to a set of rules serving as criteria to search for analogues in the Category definition step.

A general advice is to start by calculating all the profilers available in the Toolbox. In fact, a full calculation gives a broader overview of the possible reasons of concern for the chemical of interest. The colour code helps to quickly identify potential hazard properties (black font indicates no alert, red font indicates alerts; yellow indicates rules which cannot be applied for the chemical). The decision of the most suitable profilers for the target endpoint that will be used for grouping the analogues can be taken during the category definition and data gap filling steps. The use of the profiling for grouping is described in section 3.4.

Information and boundaries of the profilers are accessible at any moment by right-clicking on the name of the profiler and by selecting “Explain”.

There is no metabolism simulation selected in the Toolbox by default. The user can decide to search for observed metabolism or run simulations. The simulations can take long computational time, especially for a list of chemicals.

### 3.3 ENDPOINT

In the module “Endpoint”, the user can retrieve experimental results from the databases available in the Toolbox. Based on the different types of experimental data, the databases are organised and distributed in four basic sections:

- Physical chemical properties;
- Environmental fate and transport;
- Ecotoxicological information;
- Human health hazards.

The user can choose from which databases data should be retrieved by checking the box next to the name of the database. The user can retrieve background information for each database by right-clicking on the name of the database and by selecting "About". If gathering data from different databases, a careful check for the compatibility of data should be made. If several data points can be retrieved for a substance regarding the same endpoint, an analysis for the difference in the data points, if observed, should be made. For example, the duration of the readings could be different, or there might be interspecies differences in a similar test. Often, for modelling purposes, the experimental results are somehow aggregated and it is important to understand the rules used in this process (by default, in the successive data gap filling step the Toolbox uses the average value, can be changed in the data gap filling module/calculation options/data usage).

In the next step (Category definition), when the user searches for analogues of the target compound, the data in the selected databases are automatically retrieved and added into the data matrix.

If a reliable experimental result for a certain endpoint has been retrieved then this result could be considered for fulfilling a legal requirement. Otherwise, the user should proceed to the next step "Category definition" or directly to "Data gap filling" if he wants to use a QSAR model.

### 3.4 CATEGORY DEFINITION

In the module "Category definition", the user can group substances into chemical categories according to the profiler outcomes, e.g. specific mechanisms or modes of action or structural similarity, so that within a category data gaps can be filled by read-across or trend analysis.

The category definition is a crucial step in the prediction workflow and will affect the final result. The software leaves the user to choose how to build the category, so it is important that a consistent and reproducible approach is applied. For this purpose, it is advisable to start the categorisation with "chemical" profilers (i.e. "organic functional groups" or "US EPA new chemical categories") and then refine the result with sub-categorisation according to "mechanistic" (e.g. "protein binding") or endpoint-specific (e.g. "acute aquatic toxicity") profilers.

During the definition of the category, it is possible to select whether the analogues have to comply strictly with the functional groups/alerts of the target or if they can also contain additional functionalities/alerts. In the first instance, it is recommended to tick the "strict" option to select only close analogues of the target, excluding compounds with additional functionalities. If too few analogues are found, then the categorisation can be repeated without selecting the strict option. The experimental results of the members of the category will be displayed in a pop-up table. At this point, the user is advised to click on "Select one" and then on "OK". The sub-categorisation function is also available in the "Data gap filling" module with more functionalities and with the possibility to visualise step-by-step the chemicals removed, thus it is suggested to perform the sub-categorisation later on. Further advice about the category definitions depending on the endpoint of interest are provided further in this document (in Section 3.5 and Section 4.2) and in Part 2 of this document.

### 3.5 DATA GAP FILLING

The aim of the module "Data Gap Filling" is to predict a property of interest for the target chemical using the

experimental data of the analogues. The selection of the most appropriate endpoint in the data matrix is of crucial importance in the regulatory context. For instance, REACH indicates that the first choice test for skin sensitisation is the Local Lymph Node Assay (LLNA), while the other tests should be used only in exceptional circumstances. Thus, LLNA should be chosen with preference if enough data are available when assessing this endpoint with the Toolbox.

Three different data-gap filling tools are provided in the Toolbox:

- Read-across;
- Trend analysis;
- (Q)SAR models.

Read-across and trend analysis use the available data of the analogues to fill a data gap within a category. “(Q)SAR models” gives access to a library of external (Q)SAR models and/or predictions from them (developed outside the Toolbox and not discussed in this document), which have been integrated into the Toolbox. According to Annex XI of REACH, (Q)SAR “may indicate the presence or absence of a certain dangerous property”. However, Annex IV of the Biocidal Products Regulation (BPR) states that (Q)SAR “may indicate the presence, but not the absence of a given dangerous property”. Depending on the situation, the most relevant data gap approach should be chosen, taking into account the following considerations:

- Read-across is the appropriate data-gap filling method for endpoints expressed as “qualitative” (e.g. positive, negative, and equivocal) such as skin sensitisation (expressed as positive/negative result) or mutagenicity (mutagenicity as a positive/negative result could also be a result of aggregation).
- Trend analysis is the appropriate data-gap filling method for endpoints with “quantitative” test results (e.g., LC50 from 96-hour acute toxicity to fish).
- “(Q)SAR models” could be used to fill a data gap in weight of evidence considerations or when adequate analogues for read-across or trend analysis are not found for a target chemical.

The Toolbox offers several tools in the data gap filling module to help the user to find only data that are coherent and useful for the prediction. The user has to choose there one reference unit/scale. The data recorded with different units are consequently converted or excluded if the conversion is not possible. Another option given to the user is the possibility to pick only experimental results that satisfy certain conditions, such as standard durations or species.

For read-across and trend analysis, the Toolbox automatically plots the existing experimental data for all analogues (Y axis) according to a descriptor (X axis, Kow by default). It is advised to refine the category at this point, because the chemicals excluded in each step are highlighted in the plot and their visualisation may help to understand and better explain the final result. To do so, the user should expand the title “Select/filter data” on the right of the plot and click on “Subcategorize”. For the trend analysis approach, the data-gap filling module also provides monivariate regression equations and detailed statistics about the adequacy and uncertainty of the prediction by selecting the relevant tabs on top of the plot. When the result is considered satisfactory, the user can press the “accept prediction” button and move to the final “Report” module.

The user can export predictions to IUCLID 5. To do so, the user should click on “Return to matrix”, right-click on the predicted value and select “IUCLID 5”. A wizard will guide the user through the different steps. These steps include the selection of the predicted value to be exported, editing report information (“edit prediction info” on right click of the prediction in the data matrix once the prediction was accepted for reporting) such as information on the substance, category, prediction and adequacy, selecting an OECD harmonised template, completion of endpoint study record information such as reliability, interpretation of results, etc., and finally choosing the export format which can be through xml (\*.isz file) or directly through



WebServices. The user should also fill some fields directly in IUCLID.

### 3.6 REPORT

The Toolbox has some predefined templates for automatic report generation. These include the export of the list of analogues encoded as SMILES, data for the dependent and independent variable in a trend, information about the applicability domain, and so on. The user is advised to complete the report in the Toolbox by manually filling the editable fields in order to provide explanations of the category and to comment on the result.

## 4. Additional considerations

When forming a category, the user can refer to two wide groups of profilers:

- “chemical profilers”, which describe general chemical properties of the molecule;
- “endpoint specific” or “mechanistic” profilers, which are related to specific endpoints or modes of action.

The chemical profilers, which provide the grouping option independent of the target endpoint, are generally a good starting point. Recommended ones are:

### **US EPA new chemical categories (66 categories implemented in the Toolbox)**

The rules coded in the US-EPA New Chemical Program profiler reproduces the original categories cited in the document “TSCA New Chemicals Program (NCP)/ Chemical Categories” - an official document of U.S EPA Office of Pollution Prevention and Toxics (five of the original categories are not coded in the Toolbox, described in the software). The U.S. EPA/OPPT New Chemical Categories include classes of chemicals for which sufficient regulatory history has been accumulated so that hazard and risk concerns, and testing recommendations vary little from chemical to chemical within the new chemicals programme.

### **Organic functional groups (Nested) (489 categories)**

The Organic Functional Groups (OFG) system is designed to introduce some classification and systematisation of the various functionalities and characteristic structural fragments in organic chemicals from a large database, and identify structurally similar chemicals. Organic functional groups are specific groups of atoms and/or bonds within molecules that determine the characteristic chemical reactions of those molecules. Functional groups are an essential part of the characteristic structural fragments of organic molecules.

One functional group may change into another one, or a functional group might react with a separate molecule to build up a larger structure. The relative reactivity of a given functional group can be modified by nearby located other functional groups. Thus, functional groups are the principal “reacting units” in organic chemistry, determining the fundamental relationship between structure and chemical reactivity. The OFG profile is currently subdivided into OFG (general) and OFG (nested), which include the same functional groups. The difference is that OFG (general) displays all functional groups present in the target compounds, while OFG (nested) applies some aggregation rules.

To make the outcome repeatable and the interpretation of the result consistent, a standardised workflow is recommended. It usually starts with using a chemical profiler for a quick and efficient exclusion of the compounds not chemically related to the target. At this point, the mechanistic and endpoint specific profilers can be exploited to obtain a sub-category consistent in terms of mode of action. These sub-categorisations might not exclude any additional chemical. Nevertheless, the consistency of the category from a mechanistic point of view would be strongly supported.

## 5. Further information

Further information on the QSAR ToolBox, read-across approach and QSAR can be found in:

**OECD ToolBox official website**

<http://www.qsartoolbox.org>

**Guidance on Information Requirements and Chemical Safety Assessment, Chapter R.6: QSARs and grouping of chemicals**

[http://echa.europa.eu/documents/10162/13632/information\\_requirements\\_r6\\_en.pdf](http://echa.europa.eu/documents/10162/13632/information_requirements_r6_en.pdf)

**Practical Guide 6: How to report read-across and categories**

[http://echa.europa.eu/documents/10162/13655/pg\\_report\\_readacross\\_en.pdf](http://echa.europa.eu/documents/10162/13655/pg_report_readacross_en.pdf)

**Practical Guide 5: How to report (Q)SARs**

[http://echa.europa.eu/documents/10162/13655/pg\\_report\\_qsars\\_en.pdf](http://echa.europa.eu/documents/10162/13655/pg_report_qsars_en.pdf)

**Read-across illustrative example**

<http://echa.europa.eu/en/support/grouping-of-substances-and-read-across>

Adequate documentation of the data used to justify a proposed read-across is crucial for the evaluator when assessing the scientific validity of the read-across approach. Additional useful information can be found in:

**Practical Guide 3: How to report robust study summaries**

[http://echa.europa.eu/documents/10162/13643/pg\\_report\\_robust\\_study\\_summaries\\_en.pdf](http://echa.europa.eu/documents/10162/13643/pg_report_robust_study_summaries_en.pdf)

**Guidance for identification and naming of substances under REACH and CLP**

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

**Guidance on Information Requirements and Chemical Safety Assessment, Chapter R.7(c), Section R.7.12 Guidance on Toxicokinetics**

[http://echa.europa.eu/documents/10162/13632/information\\_requirements\\_r7c\\_en.pdf](http://echa.europa.eu/documents/10162/13632/information_requirements_r7c_en.pdf)

**Information on the Experts Workshop on Read-Across Assessment with active support from Cefic-LRI held at ECHA on 03 October 2012 is available on the ECHA website:**

[http://echa.europa.eu/en/view-article/-/journal\\_content/c6dd5b17-7079-433a-b57f-75da9bcb1de2](http://echa.europa.eu/en/view-article/-/journal_content/c6dd5b17-7079-433a-b57f-75da9bcb1de2)

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

ECHA-14-R-06-EN - ED-02-14-123-EN-N - ISBN: 978-92-9244-432-7 - DOI: 10.2823/83029



Publications Office