

## OECD (Q)SAR Toolbox v.4.4.1

Step-by-step example on how to use the (Q)SAR editor:  
to create a (Q)SAR model based on a single linear  
regression equation or  
to upload an external (Q)SAR model via web services

# Outlook

- **Background**
- Objectives
- (Q)SAR models
- The exercise

## Background

- This is a step-by-step presentation designed to provide guidance to the Toolbox users on how to use the (Q)SAR Editor to create their own (Q)SAR models and to disseminate them to other users.
- The created (Q)SAR could be used for predicting purposes.
- Also the user will become familiar with functionalities of the (Q)SAR managing tool.
- Two examples will be illustrated:
  - Building (Q)SAR using a single linear regression
  - Building (Q)SAR using web service link.

**Note:** Please note that building of custom items (such as profilers, (Q)SAR models as well as importing of custom databases) is only enabled in single user mode. So, if your Toolbox is installed in multiuser mode, you will be not able to follow this tutorial.

# Outlook

- Background
- **Objectives**
- (Q)SAR models
- The exercise

## Objectives

- **This presentation demonstrates how to build and use a new (Q)SAR module including:**
  - naming of the (Q)SAR model;
  - define target endpoint and units as well as input of QMRF info;
  - define equation by a mathematical expression or by a web service link;
  - importing training set (test set) of the model;
  - define applicability domain of the model;
  - add statistics for the model;
  - save the model;
  - use the model for predicting a list of chemicals;
  - disseminate the models to users.

# Outlook

- Background
- Objectives
- **(Q)SAR models**
- The exercise

# (Q)SAR models

## Definition

“Structure-activity relationship (SAR) and quantitative structure-activity relationship (QSAR) models - collectively referred to as (Q)SARs - are mathematical models that can be used to predict the physicochemical, biological and environmental fate properties of compounds from the knowledge of their chemical structure.”<sup>1</sup>

---

<sup>1</sup>ECHA/Support/QSAR models: <https://echa.europa.eu/support/registration/how-to-avoid-unnecessary-testing-on-animals/qsar-models>

# (Q)SAR models

## Overview

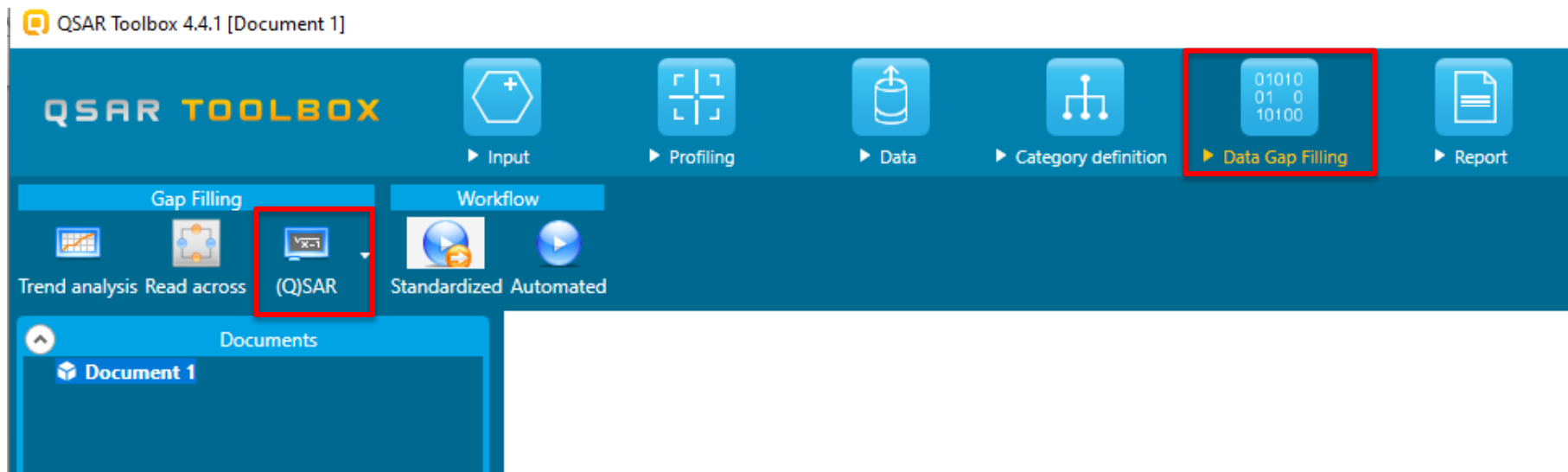
- Toolbox contains many predefined (Q)SAR models and implementation of new (Q)SAR models is also feasible.
- Single or multiple regression models (based on linear regression) and models based on read-across approach can be implemented.
- Building a custom (Q)SAR is possible via two ways:
  - Once you are in the stage of Data gap filling (read-across or trend analysis approach)\*
  - Independently from a read-across/trend analysis
- The purpose of this tutorial is to exemplify how a (Q)SAR based on a single linear regression can be created independently from a read-across/trend analysis.

\*Creating a custom (Q)SAR in the Gap Filling module is illustrated in tutorial: Step-by-step example for building a (Q)SAR model



# (Q)SAR models

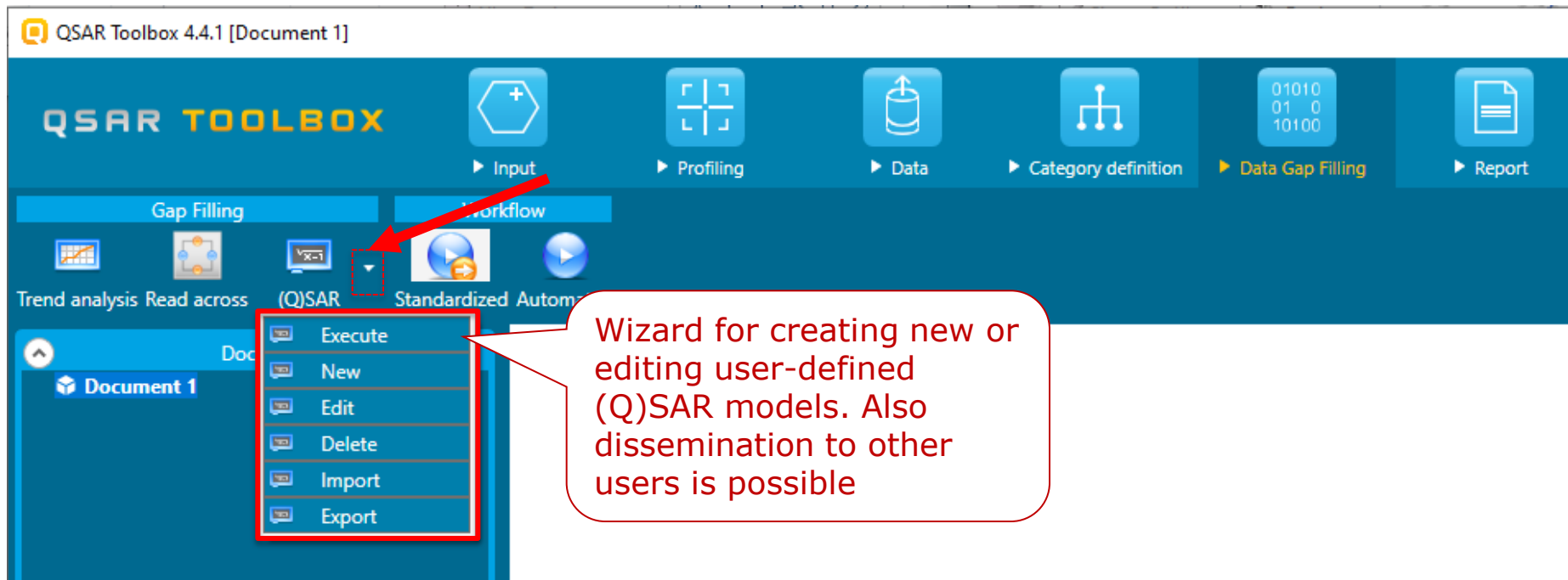
## Implementation in Toolbox



Button (Q)SAR is located under Data Gap Filling module

# (Q)SAR models

## Implementation in Toolbox



# Outlook

- Background
- Objectives
- (Q)SAR models
- **The exercise**

# The Exercise

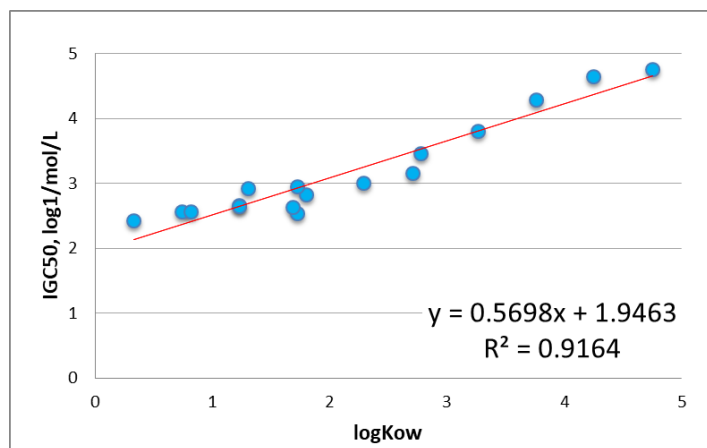
## Example 1

- In the first example we will use a (Q)SAR editor for building a (Q)SAR with the following conditions:

### Endpoint

<b>Effect</b>	Growth
<b>Endpoint</b>	IGC50
<b>Endpoint unit</b>	log 1/mol/L
<b>Species</b>	Tetrahymena pyriformis
<b>Duration</b>	48 h

### Algorithm



### Applicability domain

Parametric requirements:

$$0.1 \leq \log Kow \leq 5$$

**AND**

Structural requirements:

*Aldehydes*

### Training set and statistics

17 training set chemicals with experimental IGC50 data, [mol/L]  
 Coefficient of determination,  $R^2 = 0.92$   
 Coefficient of determination – leave one out,  $Q^2 = 0.894$   
 Sum of squared residuals,  $SSR = 0.77$   
 Fisher function,  $F = 157$

### Validation set

10 validating set chemicals with IGC50 data, [mol/L]  
 Coefficient of determination,  $R^2 = 0.89$   
 Coefficient of determination – leave one out,  $Q^2 = 0.828$   
 Sum of squared residuals,  $SSR = 0.22$

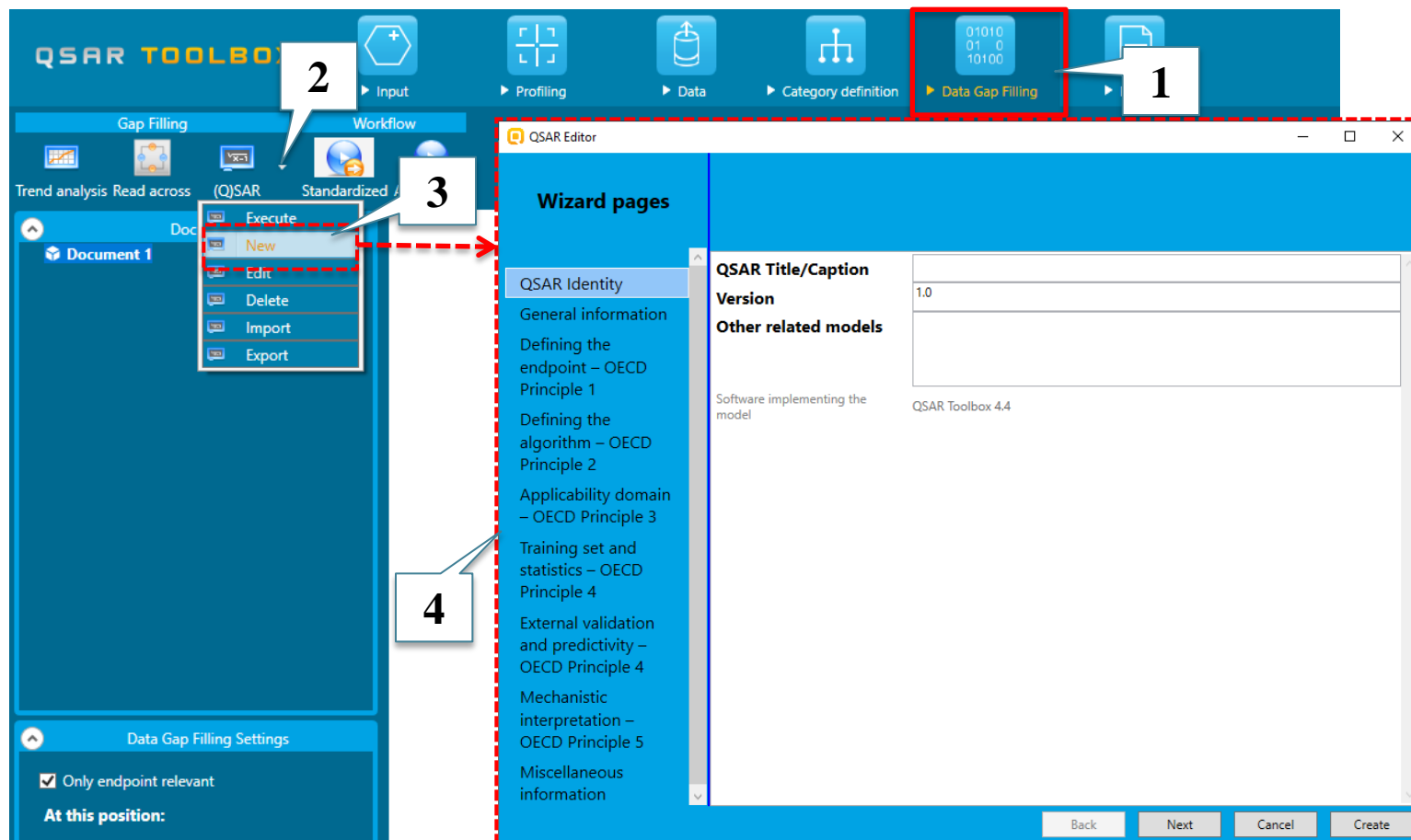
## The Exercise

### Start building a new (Q)SAR

We are going to create a new (Q)SAR module:

- Open the Toolbox.
- Move to the Data Gap Filling module  
(see next screen shot).

# Building a new (Q)SAR



1. Go to the **Data Gap Filling** module;
2. Click on the drop-down menu;
3. Select **New**;
4. **(Q)SAR Editor** wizard appears.

## Building a new (Q)SAR

Once the (Q)SAR Editor is opened, there are two types of sections that should be filled:

- Important sections - mandatory for correct work of the (Q)SAR
  - The (Q)SAR title;
  - The endpoint (**IGC50**) and its unit (**1/mol/L**);
  - The mathematical equation ( **$y=0.57*\log Kow+1.94$** ).
- Additional sections – not mandatory for correct work of the (Q)SAR but recommended according to the five OECD principles\*
  - Applicability domain could be defined (**parametric ( $0.1 \leq \log Kow \leq 5$ ) and structural boundaries (aldehyde)**);
  - Training set/test set could be imported along with statistical information (**list with 17 aldehydes with observed IGC50 data (training set) and 10 aldehydes (validation set)**)
  - Additional QMRF information could be added, too (such as author, dependent variables, description of the algorithm etc.)
- **The forthcoming slides illustrate the consecutive filling of the fields of the (Q)SAR Editor with the above information**

\*OECD principles: <https://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf>; guidance: [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)

# Building a new (Q)SAR

## *Name of the model*

QSAR Editor

**Wizard pages**

- QSAR Identity
- General information**
- Defining the endpoint – OECD Principle 1
- Defining the algorithm – OECD Principle 2
- Applicability domain – OECD Principle 3
- Training set and statistics – OECD Principle 4
- External validation and predictivity – OECD Principle 4
- Mechanistic interpretation – OECD Principle 5
- Miscellaneous information

**General information**

**QSAR Title/Caption** Acute aquatic toxicity (IGC50) of simple aldehydes (LMC)

**Version** 1.0

**Other related models**

Software implementing the model QSAR Toolbox 4.4

First we need to add a name of the custom model. This field is mandatory for building a model

1. Add the name of the (Q)SAR model in the **(Q)SAR Title/Caption** field. In our case it is **"Acute aquatic toxicity (IGC50) of simple aldehydes (LMC)";**
2. Fields **"Version"** and **"Software implementing the model"** are automatically populated. You could add information in the empty **"Other related models"** panel, if there are other models related to this one;
3. Move to section **General information;**



# Building a new (Q)SAR

## General information

**Wizard pages**

- 1 QSAR Identity
- General information
- Defining the endpoint – OECD Principle 1
- 4 Defining the algorithm – OECD Principle 2
- Applicability domain – OECD Principle 3
- Training set and statistics – OECD Principle 4
- External validation and predictivity – OECD Principle 4
- Mechanistic interpretation – OECD Principle 5
- Miscellaneous information

**Form fields:**

- Date:** Friday, September 27, 2019
- Author(s):** Laboratory of Mathematical Chemistry (LMC)
- Model updates:**
- Date of model updates:**
- Model developer(s):** Laboratory of Mathematical Chemistry (LMC)
- Date of development and/or publication:**
- Reference(s) to main scientific papers:**

1. Section **General information** is selected;
2. Field **"Date"** is automatically filled;
3. In our case **"Author(s)"** and **"Model developer(s)"** fields are populated (e.g. Laboratory of Mathematical Chemistry (LMC)). You could add additional information to these and other fields. As already mentioned these fields are not mandatory;
4. Move to section **"Defining the endpoint – OECD Principle 1"**.

# Building a new (Q)SAR

## Define the endpoint

In this section we will add the endpoint of the model (IGC50). This is a mandatory step in order for the model to work properly.

1. Section **"Defining the endpoint – OECD Principle 1"** is selected;
2. Click on **Define** button to define the target endpoint;
3. Select **Ecotoxicological Information, Aquatic Toxicity**;
4. Consecutively add **IGC50** endpoint and metadata as shown above (Effect: **Growth**; Species: **Tetrahymena pyriformis**; Duration: **48 h**);
5. Click **Finish**; Move to the **Unit** – see next slide

# Building a new (Q)SAR

## Define unit of the endpoint

**Wizard pages**

- 1 QSAR Identification
- General information
- Defining the endpoint – OECD Principle 1
- Defining the algorithm – OECD Principle 2
- Applicability domain – OECD Principle 3
- Training set and statistics – OECD Principle 4
- External validation

**Endpoint to predict**

Tree position: Ecotoxicological Information#Aquatic Toxicity

Data filters: Effect=Growth; Test organisms (species)=Tetrahymena pyriformis; Endpoint=

**Comment on the endpoint**

**Endpoint units** log(1/mol/L)

**Endpoint units** Unknown **Set** **Unset**

**Dependent variable**

**Experimental protocol**

**Data quality and reliability**

**Expressions**

- ☐ As it is
- ☐ 1/Endpoint, 1/mol/L
- ☐ log(Endpoint), log(mol/L)
- ☒ log(1/Endpoint), log(1/mol/L)

**Origin**

scale: unit:

**Destination**

Molar concentration

**unit**

- ☐  $\mu\text{mol/L}$
- ☐  $\text{mmol/L}$
- ☐  $\text{mmol/m}^3$
- ☒  $\text{mol/L}$
- ☐  $\text{mol/m}$
- ☐  $\text{nmol/L}$
- ☐  $\text{pmol/L}$

**OK** **Cancel**

Here we need to add unit of endpoint (mol/L) of the model. This is a mandatory field.

1. Keep section **"Defining the endpoint – OECD Principle 1"** selected;
2. Click on **Set** button to define the unit (in our case it is log (1/mol/L));
3. From the appeared window select **Molar concentration** and
4. Choose **mol/L** unit;
5. Select **"log(1/Endpoint), log (1mol/L)"** in order to use the correct mathematical expression for building the regression;
6. Click **OK**;
7. Now the unit is recognized by the system ("**Unknown**" is changed to **"log (1/mol/L)"**).

# Building a new (Q)SAR

## Add supporting information of the endpoint

**QSAR Editor**

**Wizard pages**

- QSAR Identity
- General information
- Defining the endpoint – OECD Principle 1
- Defining the algorithm – OECD Principle 2
- Applicability domain – OECD Principle 3
- Training set and statistics – OECD Principle 4
- External validation and predictivity – OECD Principle 4
- Mechanistic interpretation –

**Endpoint to predict**

Tree position: Ecotoxicological Information#Aquatic Toxicity

Data filters: Effect=Growth; Test organisms (species)=Tetrahymena pyriformis; Endpoint=IGC50; Duration=48 h;

**Comment on the endpoint**

**Endpoint units** log(1/mol) **1**

**Dependent variable** IGC50

**Experimental protocol** Acute aquatic toxicity test using ciliate Tetrahymena for assessing toxicity of chemicals to aquatic organisms **2**

**Data quality and variability** curated data based on expert analysis **3**

**4**

Additional information could be added to other fields within this section. They are not mandatory for correct working of the (Q)SAR. However we recommend filling all the fields in order for the (Q)SAR to meet all the requirements related to OECD Principle 1.

1. Add **"IGC50"** in the field **Dependant variable**;
2. Additional information is added to section **"Experimental protocol"** and
3. **"Data quality and variability"**;
4. Move to section **"Defining the algorithm – OECD Principle 2"**.

# Building a new (Q)SAR

## Define algorithm

**Wizard pages**

- QSAR Identity
- General information
- Defining the endpoint – Principle 1
- Defining the algorithm – OECD Principle 2**
- Applicability domain – OECD Principle 3
- Training set and statistics – OECD Principle 4
- External validation and predictivity – OECD Principle 4
- Mechanistic interpretation – OECD Principle 5

**Type of model**

**Algorithm description**

☒ Equation ☐ Web service link

**Define variables**

Number of variables:  [Define](#)

**Equation**

y = [Example:] 0 + 1\*D1

D1:  [Name: log Kow](#) [Unit: Change unit](#)

**Describe variable**

- log BCF max
- Log Koa (Air-water partition coefficient model)
- Log Koa (Henry's law constant model)
- log Kow**
- LOMO Energy
- Maximum distance
- Maximum donor delocalizability
- Mean Melting Point
- Melting Point (Adapted Joback Method)
- Melting point (Gold and Ogle method)

The section "Defining the algorithm" is one of the most important fields for correct configuration of the (Q)SAR model. In this section we will define the number of used variables and the equation itself.

1. Section "Defining the algorithm – OECD Principle 2" is selected;
2. Select "Equation" radio button;
3. In section "Define variables" you should specify the number of variables used in your custom model. In our case it is 1;
4. From the drop-down menu for the variables select those which will be used in your equation. In our case this is **log Kow**.
5. Click **Change unit** to specify the unit of the variable used. In our case this is not needed (because of the logarithmic unit). Definition of the equation continues on the next slide.

# Building a new (Q)SAR

## Define algorithm

1. Enter your model equation in section **"Equation"**. In our case write "1.94+0.57\*D1" in the empty field; The user is able to derive the equation (i.e. to build a model) by using the functionality "Save model" inside the Data Gap Filling stage.\*
2. Click **Check** button in order system to check for correctness of the defined equation;
3. A window appears informing that the equation is **"valid"**;
4. You are able to fill in the other empty fields related to **OECD Principle 2** (e.g. **Type of model Algorithm description**, etc.);
5. Move to the next section related to **"Applicability domain - OECD Principle 3"**.

\* Creating a custom (Q)SAR in the Gap Filling module is illustrated in tutorial: *Step-by-step example for building a (Q)SAR model*

# Building a new (Q)SAR

## Define applicability domain

**Wizard pages**

- QSAR Identity
- General information
- Defining the endpoint – OECD Principle 1
- Defining the algorithm – OECD Principle 2
- Applicability domain – OECD Principle 3**
- Training set and statistics – OECD Principle 4
- External validation and predictivity – OECD Principle 4
- Mechanistic interpretation – OECD Principle 5
- Miscellaneous

**Domain**

No domain defined

**Define** **Undefine**

**Description of the applicability**

**Method used to assess the applicability domain**

**Software name and version for applicability domain assessment**

**Limits of applicability**

**Domain scheme (Custom) - Profiling Scheme Browser**

Save Scheme Export Scheme Save Tests Show Tests Run All Tests

Categories

Filter:

Domain scheme In Domain

Definition Properties Training Set Literature MetaInfo Table Custom Captions Scheme

[1] In Domain

Category tree

ADD DEL AND

Query details

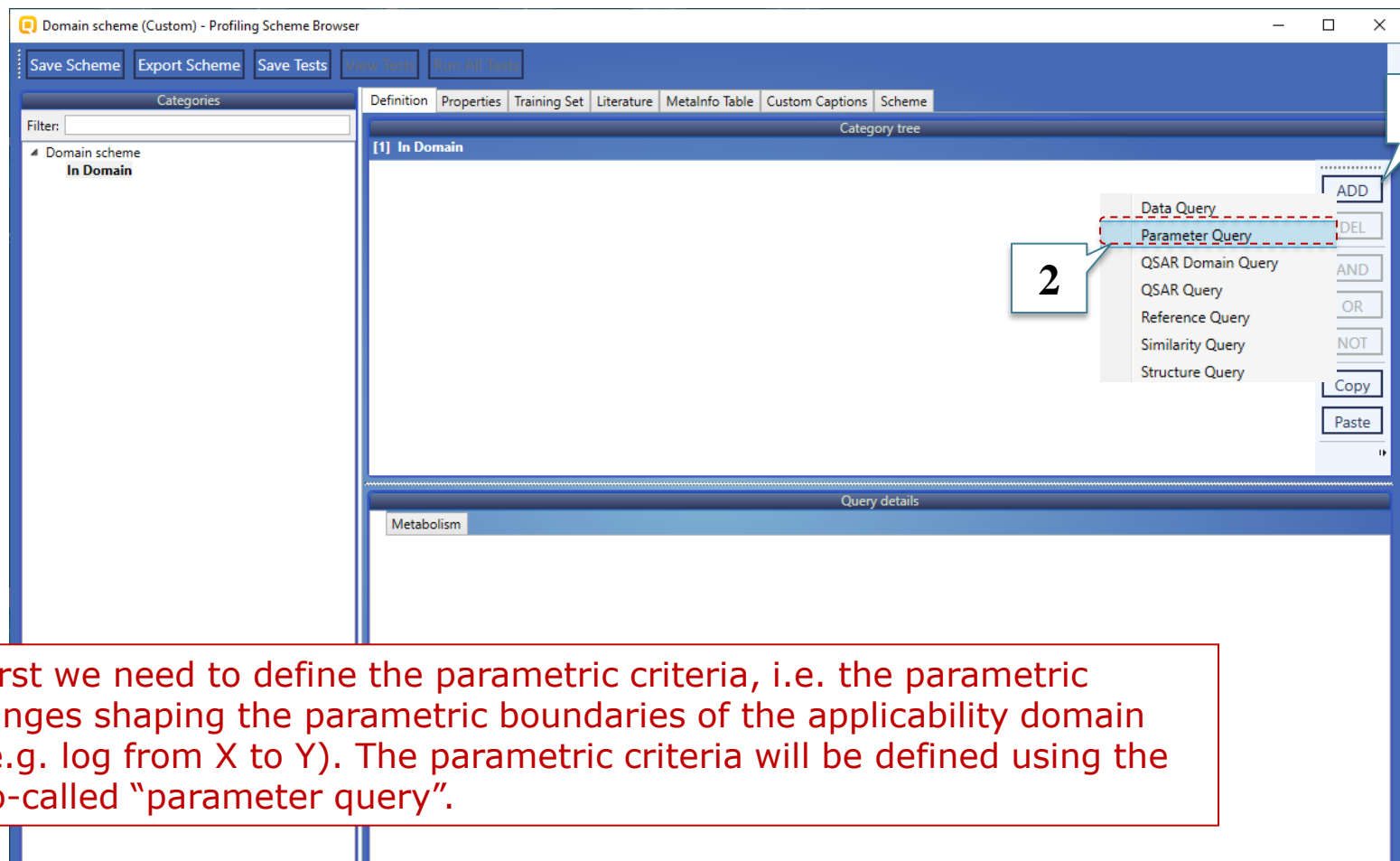
Metabolism

In section "Applicability domain - OECD Principle 3" we will define the layers of the model domain. Here we will define the parametric and structural boundaries of our (Q)SAR model.

1. Section **"Applicability domain - OECD Principle 3"** is selected;
2. There are two possibilities: to define and to undefined the already defined domain. Click the **"Define"** button;
3. A new window appears. It is explained in detail in the next slide.

# Building a new (Q)SAR

## *Define applicability domain – parametric boundary*



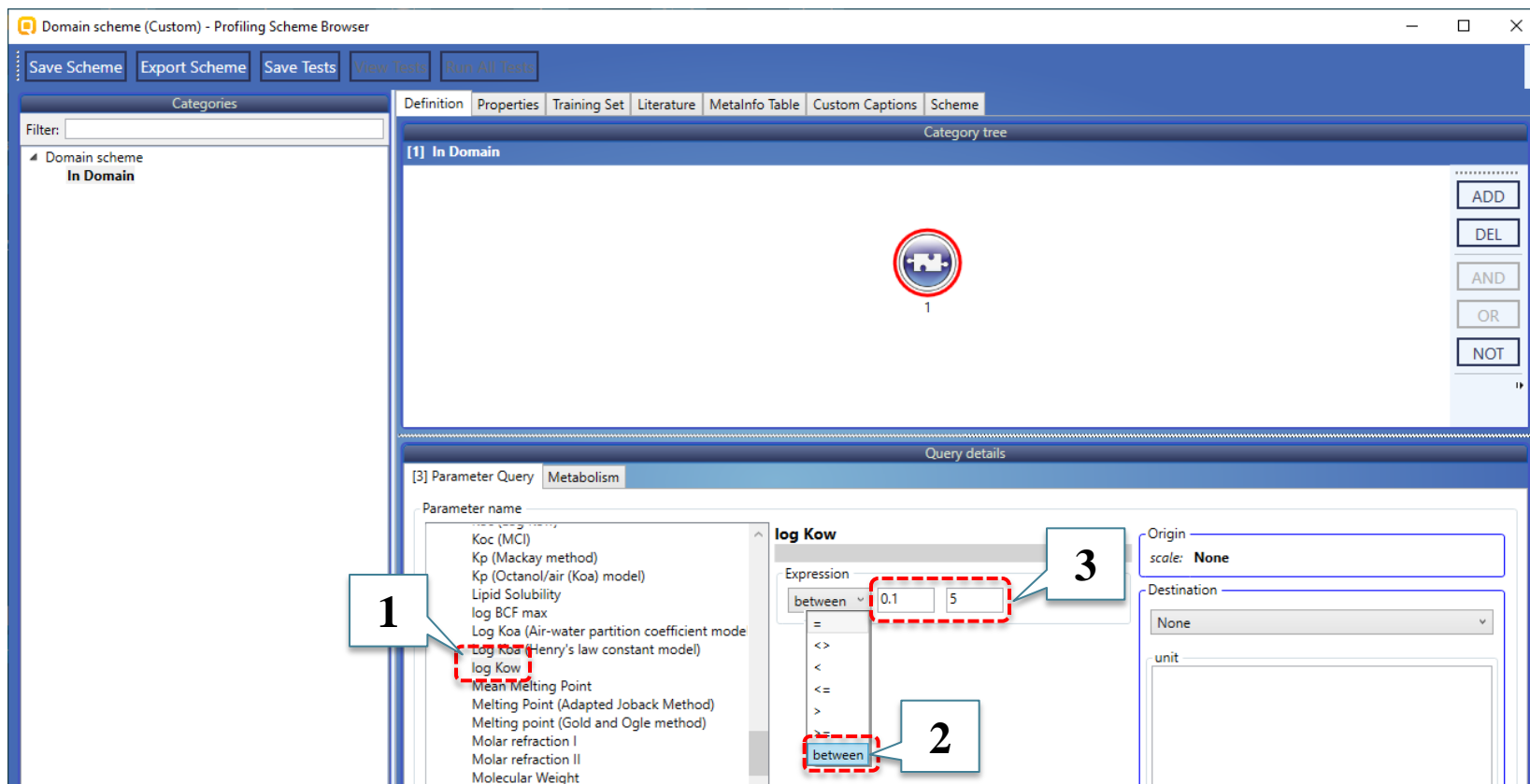
First we need to define the parametric criteria, i.e. the parametric ranges shaping the parametric boundaries of the applicability domain (e.g. log from X to Y). The parametric criteria will be defined using the so-called "parameter query".

1. Click on the **ADD** button; 2. Select **Parameter Query**.



# Building a new (Q)SAR

## Define applicability domain – parametric boundary



1. Select **log Kow** from the list with 2D parameters;
2. Select **between** from drop-down menu;
3. Specify **0.1** for the lower and **5** for the upper border of the parametric range. **The parametric boundary has been defined.**

# Building a new (Q)SAR

## Define applicability domain – structural boundary

The screenshot shows the 'Domain scheme (Custom) - Profiling Scheme Browser' window. The interface includes a top menu bar with buttons like 'Save Scheme', 'Export Scheme', 'Save Tests', 'View Tests', and 'Run All Tests'. Below this is a 'Categories' panel on the left with a filter box and a tree view showing 'Domain scheme' and 'In Domain'. The main area is divided into two panels: 'Category tree' and 'Query details'. The 'Category tree' panel shows a tree structure with 'In Domain' at the top and two icons labeled '1' and '2'. The 'Query details' panel shows a 'Reference Query' for 'Metabolism' with a list of 'Profiling schemes' on the left and 'Selected categories' and 'Available categories' on the right. The 'Selected categories' list contains 'Aldehyde', and the 'Available categories' list contains 'Acyl halide', 'Acylal', 'Acylol', 'Alcohol', 'Aldehyde', and 'Aldimine'. Red dashed boxes and arrows highlight the steps: 1. Clicking the 'ADD' button in the 'Category tree' panel. 2. Selecting 'Reference Query' from the dropdown menu. 3. Selecting 'Organic functional groups' from the 'Profiling schemes' list. 4. Finding 'Aldehyde' in the 'Available categories' list. 5. Moving the 'Aldehyde' category from the 'Available categories' list to the 'Selected categories' list using the arrow button. 6. The final state where 'Aldehyde' is in the 'Selected categories' list, defining the structural boundary.

1. Click **Add** button;
2. Select **Reference Query**;
3. Select **Organic functional group** from the list with empiric profilers;
4. Find "**Aldehyde**" category and click on it;
5. Move the selected category in the upper panel (6) using the arrow button (5); **The structural boundary has been defined.**

# Building a new (Q)SAR

## Define applicability domain – structural boundary

1. Select both queries holding the **Ctrl** button (they should become red circled);

2. Click **AND** in order to combine them logically;

3. A new node named **"AND"** appears;

4. Finally click **"Save Scheme"**;

5. Close the message;

6. Close the profiling scheme window.

# Building a new (Q)SAR

## *Define applicability domain*

The screenshot shows the 'QSAR Editor' window with the 'Define applicability domain' wizard page. The sidebar on the left lists the following pages: QSAR Identity, General information, Defining the endpoint – OECD Principle 1, Defining the algorithm – OECD Principle 2, **Applicability domain – OECD Principle 3** (highlighted with a red dashed box and callout 3), Training set and statistics – OECD Principle 4, External validation and predictivity – OECD Principle 4, Mechanistic interpretation – OECD Principle 5, and Miscellaneous.

The main content area is titled 'Domain' and contains the following sections:

- Domain**: A status bar at the top shows 'Domain defined.' with a 'Define' button and an 'Undefine' button. Callout 1 points to this status bar.
- Description of the applicability**: An empty text input field.
- Method used to assess the applicability domain**: An empty text input field.
- Software name and version for applicability domain assessment**: An empty text input field.
- Limits of applicability**: An empty text input field.

Callout 2 points to the empty input fields, indicating where the user can fill in information.

1. An indication appears that the domain has been defined;
2. The user is able to fill in the empty sections;
3. Move to the next section **"Training set and statistics – OECD Principle 4"**.

# Building a new (Q)SAR

## Import training set

The screenshot displays the QSAR Editor software interface. On the left, a 'Wizard pages' sidebar lists various steps, with 'Training set and statistics – OECD Principle 4' highlighted. The main window shows the 'Import training set chemicals and data' section, where the 'New' button is highlighted. A red dashed box outlines the 'Importing wizard' dialog box, which includes fields for 'File name', 'Number of data', 'Used separators', 'Import as inventory', 'Import to', 'Import title', and a 'Preview of file' area. Callout numbers 1, 2, and 3 point to the selected sidebar item, the 'New' button, and the wizard dialog box, respectively.

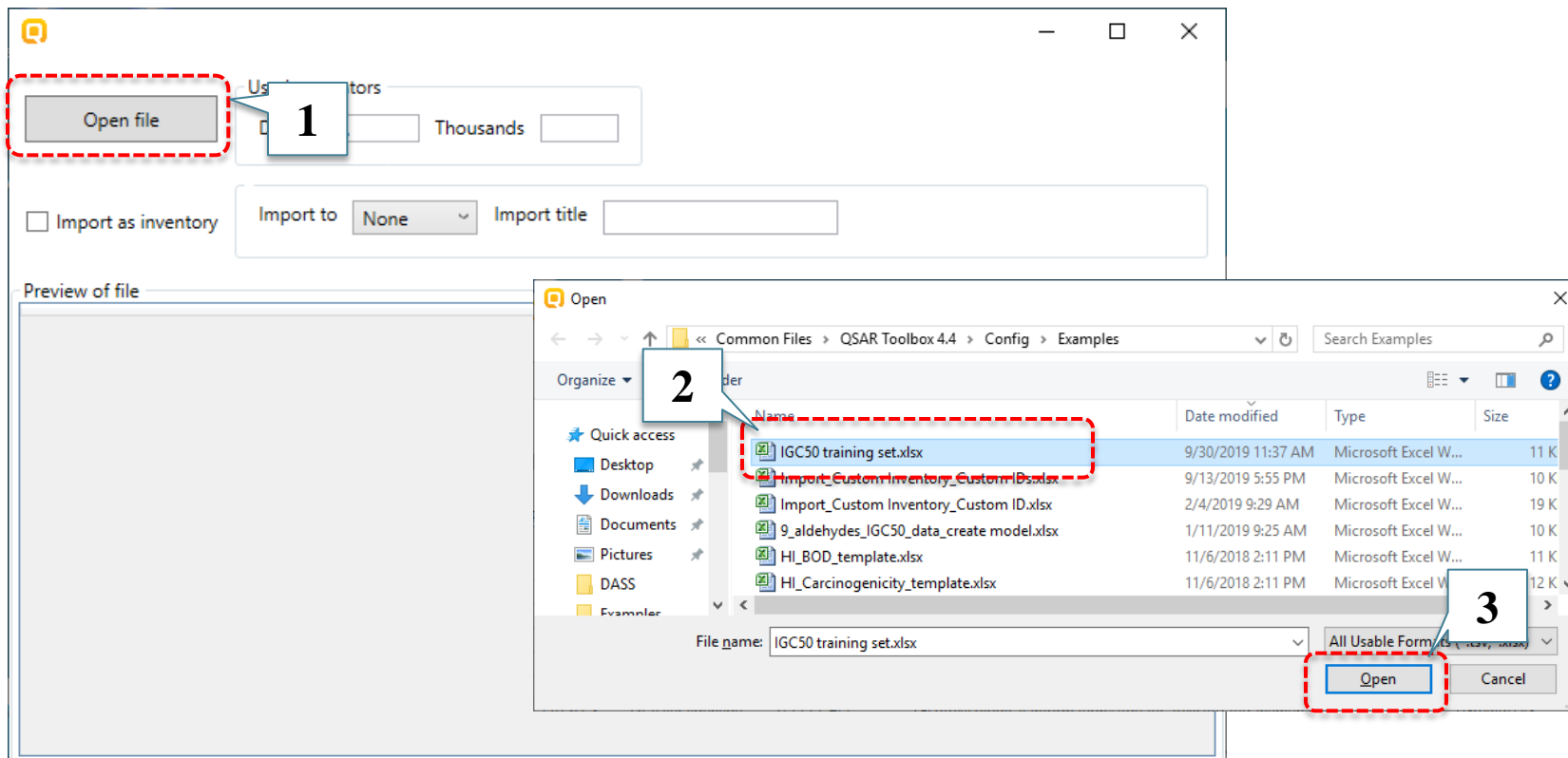
In section "Training set and statistics" we will import training set chemical with their observed data.

1. Section "Training set and statistics – OECD Principle 4" is selected;
2. Click **New** button to evoke the **Importing wizard (3)**. Import of the training set uses the same wizard used for importing the database\*. More details regarding the import of the training set is illustrated on the next few slides.

\*Further details can be found in tutorial: *Tutorial of how to Import/Export a custom database and Import/Export database via IUCLID*

# Building a new (Q)SAR

## *Import training set*



1. Click **Open file**;
2. From the Toolbox example folder (by default located here: C:\Program Files (x86)\Common Files\QSAR Toolbox 4.4\Config\Examples) select **IGC50 training set.xlsx** file;
3. Click **Open** button.

# Building a new (Q)SAR

## Import training set

Importing to IGC50 training set\_1

Open file

Used separators  
Decimal  Thousands

☐ Import as inventory

Import to  Import title

Preview of file

CAS Number	Chemical names	SMILES	EndpointPath	Test organisms species	Endpoint	Effect	Duration	MeanValue	Duration Unit
66-25-1	Hexanal	CCCCC=O	Ecotoxicological Information#Aquatic Toxicity	Tetrahymena pyriformis	IGC50	Growth	48		h
97-96-1	2-ethylbutanal	CCC(CC)C=O	Ecotoxicological Information#Aquatic Toxicity	Tetrahymena pyriformis	IGC50	Growth	48		h
112-44-7	Hendecanaldehyde	CCCCCCCCC=O	Ecotoxicological Information#Aquatic Toxicity	Tetrahymena pyriformis	IGC50	Growth	48		h
112-31-2	1-Decanal	CCCCCCCCC=O	Ecotoxicological Information#Aquatic Toxicity	Tetrahymena pyriformis	IGC50	Growth	48		h
123-38-6	Propanal	CCC=O	Ecotoxicological Information#Aquatic Toxicity	Tetrahymena pyriformis	IGC50	Growth	48		h
123-15-9	2-Methylpentanal	CCCC(C)C=O	Ecotoxicological Information#Aquatic Toxicity	Tetrahymena pyriformis	IGC50	Growth	48		h
110-62-3	n-valeraldehyde	CCCCC=O	Ecotoxicological Information#Aquatic Toxicity	Tetrahymena pyriformis	IGC50	Growth	48		h
2987-16-8	3,3-dimethylbutanal	CC(C)(C)CC=O	Ecotoxicological Information#Aquatic Toxicity	Tetrahymena pyriformis	IGC50	Growth	48		h
590-86-3	3-Methylbutanal	CC(C)CC=O	Ecotoxicological Information#Aquatic Toxicity	Tetrahymena pyriformis	IGC50	Growth	48		h
78-84-2	2-methyl-1-propanal	CC(C)C=O	Ecotoxicological Information#Aquatic Toxicity	Tetrahymena pyriformis	IGC50	Growth	48		h
124-13-0	1-octanal	CCCCCCC=O	Ecotoxicological Information#Aquatic Toxicity	Tetrahymena pyriformis	IGC50	Growth	48		h
111-71-7	Heptanal	CCCCC=O	Ecotoxicological Information#Aquatic Toxicity	Tetrahymena pyriformis	IGC50	Growth	48		h
123-72-8	Butanal	CCCC=O	Ecotoxicological Information#Aquatic Toxicity	Tetrahymena pyriformis	IGC50	Growth	48		h
112-54-9	Dodecanal	CCCCCCCCCCC=O	Ecotoxicological Information#Aquatic Toxicity	Tetrahymena pyriformis	IGC50	Growth	48		h
123-05-7	2-Ethylhexanal	CCCCC(CC)C=O	Ecotoxicological Information#Aquatic Toxicity	Tetrahymena pyriformis	IGC50	Growth	48		h
124-19-6	Nonanal	CCCCCCCCC=O	Ecotoxicological Information#Aquatic Toxicity	Tetrahymena pyriformis	IGC50	Growth	48		h
96-17-3	2-Methylbutanal	CCC(C)C=O	Ecotoxicological Information#Aquatic Toxicity	Tetrahymena pyriformis	IGC50	Growth	48		h

QSAR Editor

Wizard pages

3

Import training set chemicals and data

File name: IGC50 training set\_1

Number of data: 17

New Clear

1

2

3

Back Next Import

Back Next Import

1. Click **Next** button;
2. Click **Import** (if something is not correctly imported a message highlighted red will appear at the top of the window);
3. Details about the status of the imported file appear in the main window (such as the name of the file and the number of imported data).

# Building a new (Q)SAR

## Import training set

QSAR Editor

**Wizard pages**

- QSAR Identity
- General information
- Defining the endpoint – OECD Principle 1
- Defining the algorithm – OECD Principle 2
- Applicability domain – OECD Principle 3
- Training set and statistics – OECD Principle 4**
- External validation and predictivity – OECD Principle 4
- Mechanistic interpretation – OECD Principle 5
- Miscellaneous information

**Import training set chemicals and data**

File name: IGC50 training set\_2\_1  
Number of data: 17

New Clear

<b>Availability of the training set</b>	It is available, attached to the model	1.1
<b>Available information for the training set</b>		
<b>Descriptors values for the training set</b>	The descriptor values (log K <sub>ow</sub> ) are calculated and attached to the training set chemicals	1.2
<b>Response data for the training set</b>	The response data is attached	1.3
<b>Other information about the training set</b>		
<b>Pre-processing of data before modelling</b>		
<b>Statistics for goodness-of-fit</b>	Number of chemicals = 17 Coefficient of determination, R <sup>2</sup> = 0.92 Sum of squared residuals, SSR = 0.77 Fisher function = 157	1.4
<b>Statistics obtained by leave-one-out cross-validation</b>	Coefficient of determination – leave one out, Q <sub>2</sub> = 0.894	1.5

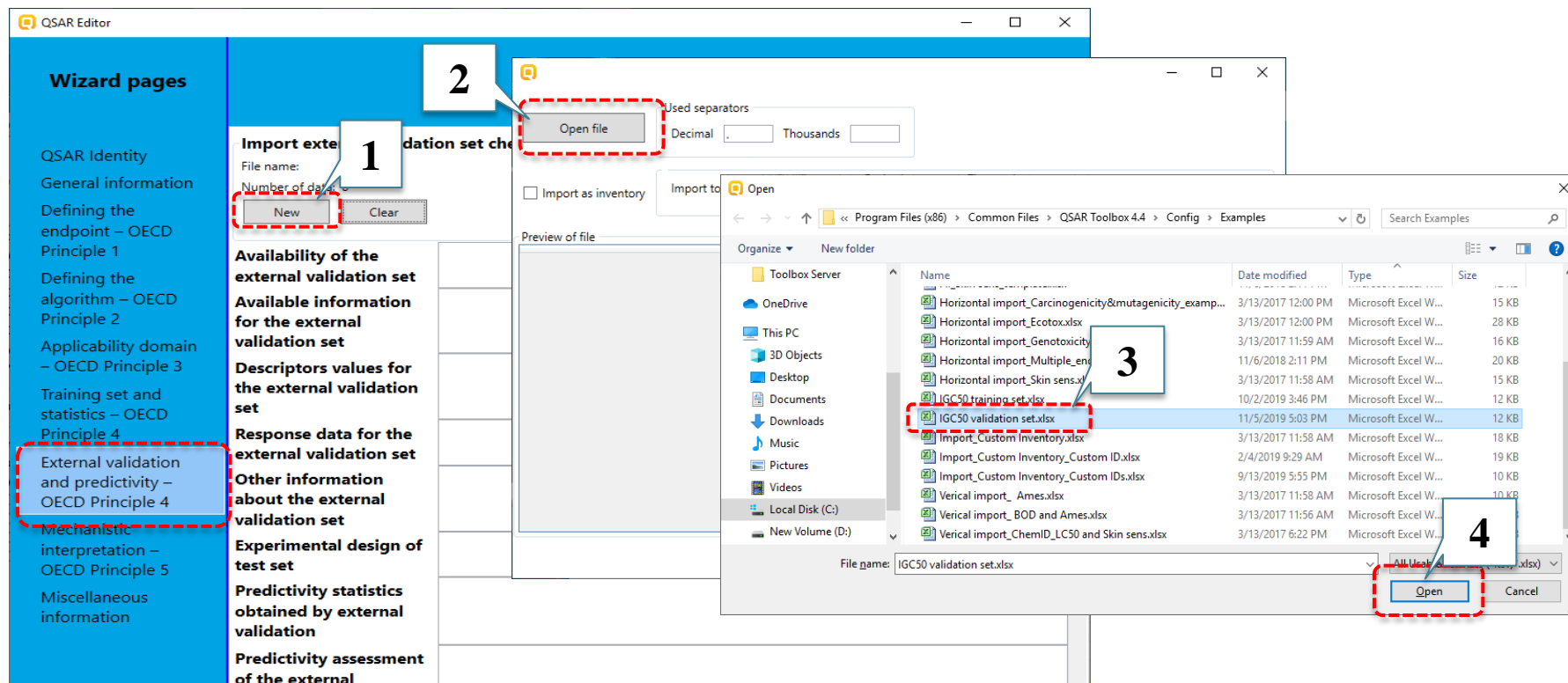
2

- Some additional information can be added to the empty fields; Please add the following text:
  - Field "Availability of the training set" – add "It is available, attached to the model"
  - Field "Descriptors values for the training set" – add "The descriptor values (log K<sub>ow</sub>) are calculated and attached to the training set"
  - Field "Response data for the training set" – add "The response data is attached"
  - Field "Statistics for goodness-of-fit" – add R<sup>2</sup>=0.92; SSR=0.77; Fisher function= 157
  - Field "Statistic obtained by leave-one-out cross validation" – add Q<sub>2</sub> = 0.894
- Move to the next section **"External validation and predictivity – OECD Principle 4"**.



# Building a new (Q)SAR

## Import validation set



Once section “**External validation and predictivity - OECD Principle 4**” is opened follow the steps:

1. Click **New** button
2. Click **Open file**;
3. From the Toolbox example folder (by default located here: C:\Program Files (x86)\Common Files\QSAR Toolbox 4.4\Config\Examples) select **IGC50 validation set.xlsx** file;
4. Click **Open** button. Then follow the steps illustrated on slide 31

# Building a new (Q)SAR

## Import validation set

**QSAR Editor**

**Wizard pages**

- QSAR Identity
- General information
- Defining the endpoint – OECD Principle 1
- Defining the algorithm – OECD Principle 2
- Applicability domain – OECD Principle 3
- Training set and statistics – OECD Principle 4
- External validation and predictivity – OECD Principle 4**
- Mechanistic interpretation – OECD Principle 5
- Miscellaneous

**Import external validation set chemicals and data**

File name: IGC50 validation set  
Number of data: 10  
[New] [Clear]

**Availability of the external validation set**  
It is available attached to the model

**Available information for the external validation set**  
Chemical names, CAS numbers and SMILES are available for the test chemicals. Experimental (IGC50) data for the validation set chemicals is available, along with calculated descriptor values (log *K<sub>ow</sub>*)

**Descriptors values for the external validation set**  
It is available attached to the model

**Response data for the external validation set**  
The response data is attached

**Other information about the external validation set**  
Statistical metrics related to validation set:  
Number of validation test chemicals = 10  
Coefficient of determination, R<sup>2</sup> = 0.89  
Coefficient of determination – leave one out, Q<sup>2</sup> = 0.828  
Sum of squared residuals, SSR = 0.22  
Randomly selected aldehydes with experimental IGC50 data

**Experimental design of test set**

**Predictivity statistics**

2.1

2.2

2.3

2.4

2.5

2.6

1. A list with chemicals from validation set appears.
2. Some additional information can be added to the empty fields; Please add the following text:
  - 2.1. Field "Availability of the external validation set" – add "It is available, attached to the model";
  - 2.2. Field "Available information for the external validation set" – additional text is added;
  - 2.3. Field "Descriptors values for the external validation set" – add "It is available attached to the model";
  - 2.4. Field "Response data for the external validation set" – add "The response data is attached";
  - 2.5. Field "Other information about the external validation set" – additional text with statistical metrics are added;
  - 2.6. Field "Experimental design of test set" – text is added.

# Building a new (Q)SAR

**QSAR Editor**

**Wizard pages**

- QSAR Identity
- General information
- Defining the endpoint – OECD Principle 1
- Defining the algorithm – OECD Principle 2
- Applicability domain – OECD Principle 3
- Training set and statistics – OECD Principle 4
- External validation and predictivity – OECD Principle 4
- Mechanistic interpretation – OECD Principle 5**

**Mechanistic basis of the model**

**A priori or a posteriori mechanistic interpretation**

**Other information about the mechanistic interpretation**

**Wizard pages**

- QSAR Identity
- General information
- Defining the endpoint – OECD Principle 1
- Defining the algorithm – OECD Principle 2
- Applicability domain – OECD Principle 3
- Training set and statistics – OECD Principle 4
- External validation and predictivity – OECD Principle 4
- Mechanistic interpretation – OECD Principle 5
- Miscellaneous information**

**Comments**

**Bibliography**

**Supporting information**

Back Next Cancel Create

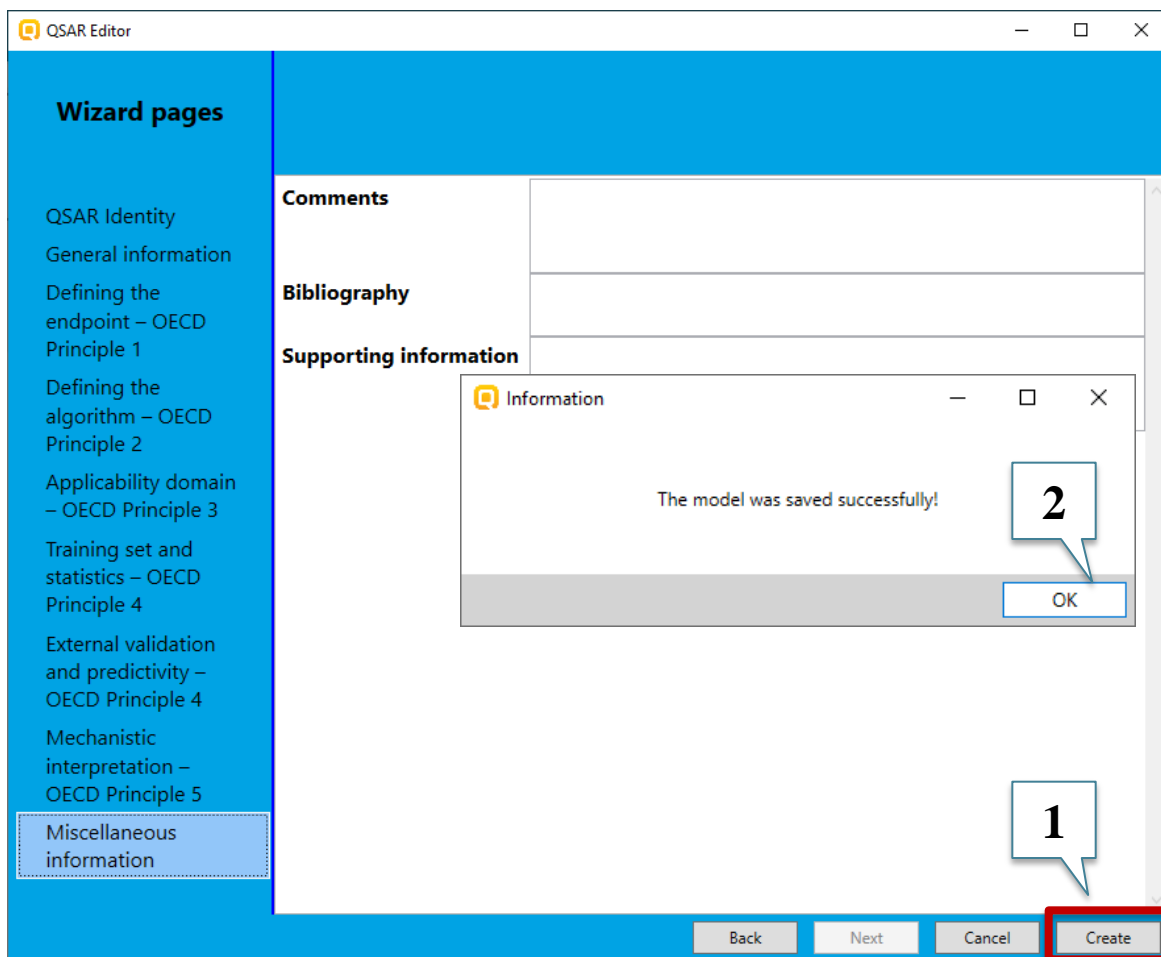
Additional information could be added to the next two sections:

- 1) "Mechanistic interpretation – OECD Principle 5"
- 2) "Miscellaneous information".

In our case these sections are left empty. As already mentioned these sections are not mandatory for correct working of the (Q)SAR model, but we recommend to fill them when you create your custom model in order for your model to follow the criteria of the five OECD Principles.

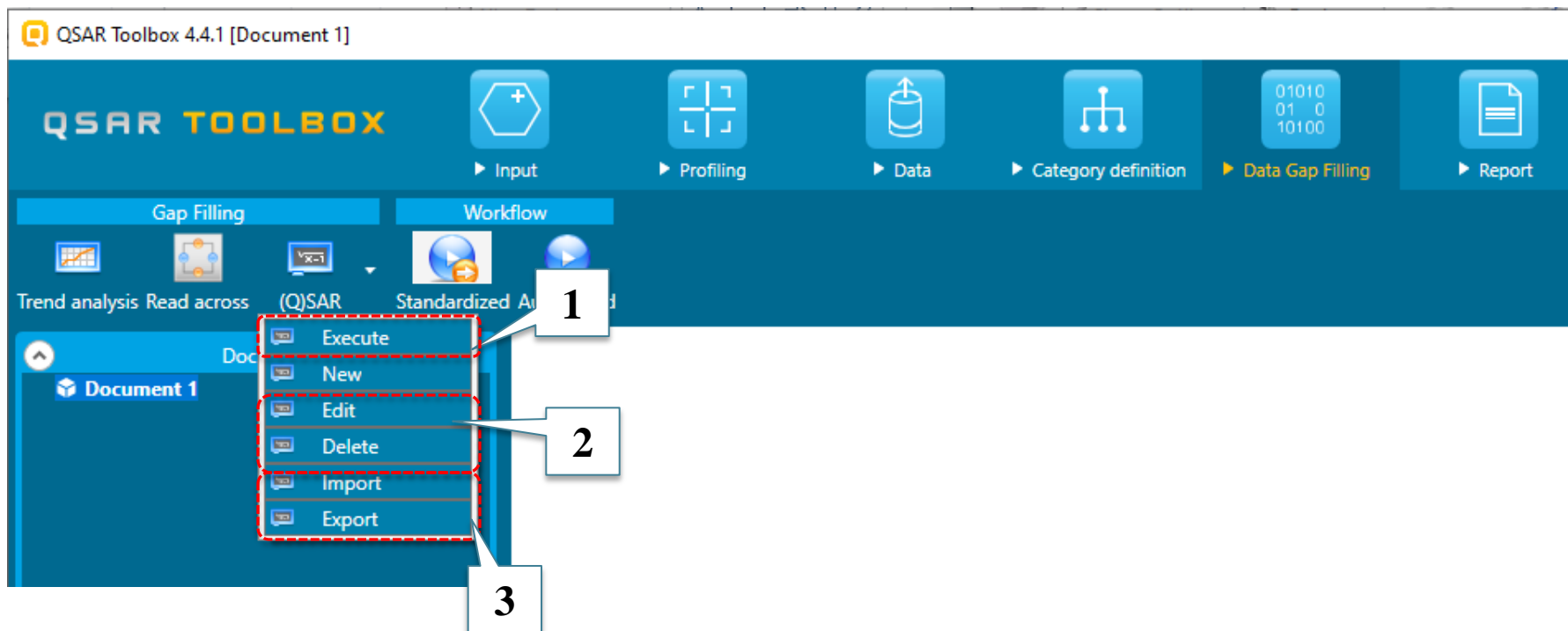
# Building a new (Q)SAR

## Create (Q)SAR model



1. Click **Create** button;
2. A message about the successful creation of the (Q)SAR appears, click **OK**

# Application of the (Q)SAR



Already created (Q)SAR models could be:

- 1) Applied to single or lists of chemicals (illustrated on the next slides);
- 2) Deleted or edited;
- 3) Imported or exported for use by other users (illustrated on the next slides).

# Application of the (Q)SAR to a list of chemicals

QSAR Toolbox 4.4.1 [Document 1]

1. Go to **Input** module;

2. Click **List** and select **From example folder**;

3. Select the file **Aldehyde analogues.smi**;

4. Click **Open**;

5. Click **No** on the appeared message

6. The chemicals appear in the data matrix.

# Application of the (Q)SAR to a list of chemicals

The screenshot illustrates the application of (Q)SAR to a list of chemicals in the QSAR Toolbox. The interface is divided into several sections:

- Top Toolbar:** Contains icons for Input, Profiling, Data, Category definition, and Data Gap Filling (highlighted with a red dashed box and labeled 1).
- Left Panel:** Shows a tree view of endpoints. The 'Ecotoxicological Information' section is expanded, and 'Aquatic Toxicity' is selected (labeled 2). A 'Data Gap Filling Settings' panel is also visible.
- Right Panel:** Displays a table of QSAR models. The first row is highlighted (labeled 4):
 

QSAR name	#	Predicted	Domain	Class	Databa	Duration	Effect	Endpoint
Acute aquatic toxicity (IGC50) of simple aldehydes (LMC) (1.0)	1	108 mg/L	In domain			48 h	Growth	IGC50
Daphnia magna 48h EC50 - Danish QSAR DB battery model (1.0)	2	219 mg/L	In domain	Br (b				EC50
Daphnia magna 48h EC50 - Danish QSAR DB Leadscope model (1.0)	3	377 mg/L	In domain	Br (b				EC50
Daphnia magna 48h EC50 - Danish QSAR DB SciQSAR model (1.0)	4	60.0 mg/L	In domain					EC50
ECOSAR: DAPHNID 48 h LC50 Mortality Aldehydes (Mono) (1.0)	5	13.0 mg/L	No domain available					LC50
ECOSAR: DAPHNID ChV Aldehydes (Mono) (1.0)	6	1.82 mg/L	No domain available	B				ChV
ECOSAR: Fish (SW) 96 h LC50 Mortality Aldehydes (Mono) (1.0)	7	10.3 mg/L	No domain available			96 h	Mortality	LC50
ECOSAR: Fish (SW) ChV Aldehydes (Mono) (1.0)	8	1.21 mg/L	No domain available					ChV
- Dialog Box:** A 'Select QSAR method' dialog box is open, showing options to 'Enter Gap filling', 'Predict selected chemical', 'Predict all chemicals' (selected), or 'Predict chemicals in domain'. The 'OK' button is highlighted (labeled 6).
- Buttons:** The 'Run' button is highlighted (labeled 5) at the bottom right of the dialog box.
- Menu:** The 'Execute' option is highlighted in the 'Document 1' menu (labeled 3).

1. Go to **Data Gap Filling** section
2. Open the **Ecotoxicological Information** part of the endpoint tree and select the **Aquatic Toxicity** level;
3. Select **Execute** from the pop-up menu; which has now opened;
4. The custom (Q)SAR appears on the top of the window, **click on it**;
5. Click **Run**;
6. Select "**Predict all chemicals**" and click **OK**

# Application of the (Q)SAR to a list of chemicals

**QSAR TOOLBOX**

Input Profiling Data Category definition Data Gap Filling Report

Gap Filling Workflow

Trend analysis Read across (Q)SAR Standardized Automated

**Documents**

Document 1

[C: 3;Md: 0;P: 3] Aldehyde analogues.sm

**Data Gap Filling Settings**

☒ Only endpoint relevant

**At this position:**

QSARs 709

Automated workflows 1

Standardized workflows 1

**Filter endpoint tree...**

Structure

Structure info

Parameters

Physical Chemical Properties

Environmental Fate and Transport

Ecotoxicological Information

Aquatic Toxicity

Growth

IGC50

48 h

Protozoa

Ciliophora

Ciliatea

Tetrahymina p... 3/3

Sediment Toxicity

Terrestrial Toxicity

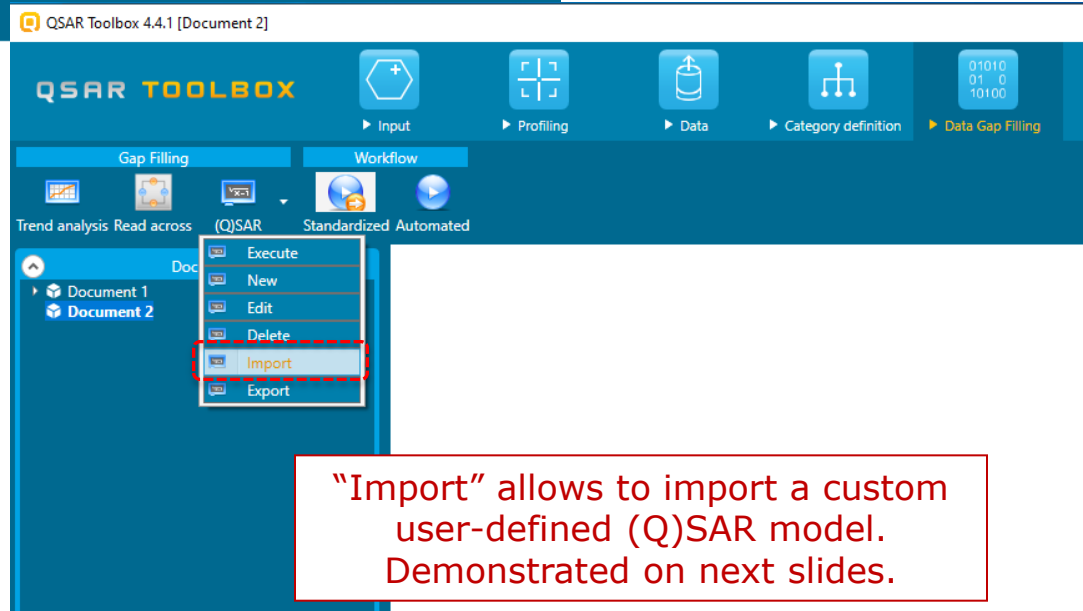
Human Health Hazards

	1	2	3
Structure	<chem>CCCCCCCC=O</chem>	<chem>CCCCCCCC=O</chem>	<chem>CCCC(C)C=O</chem>
Structure info			
Parameters			
Physical Chemical Properties			
Environmental Fate and Transport			
Ecotoxicological Information			
Aquatic Toxicity			
Growth			
IGC50			
48 h			
Protozoa			
Ciliophora			
Ciliatea			
Tetrahymina p... 3/3	Q: 108 mg/L	Q: 64.9 mg/L	Q: 71.2 mg/L
Sediment Toxicity			
Terrestrial Toxicity			
Human Health Hazards			

The predictions from the custom (Q)SAR model appear in the data matrix (1)

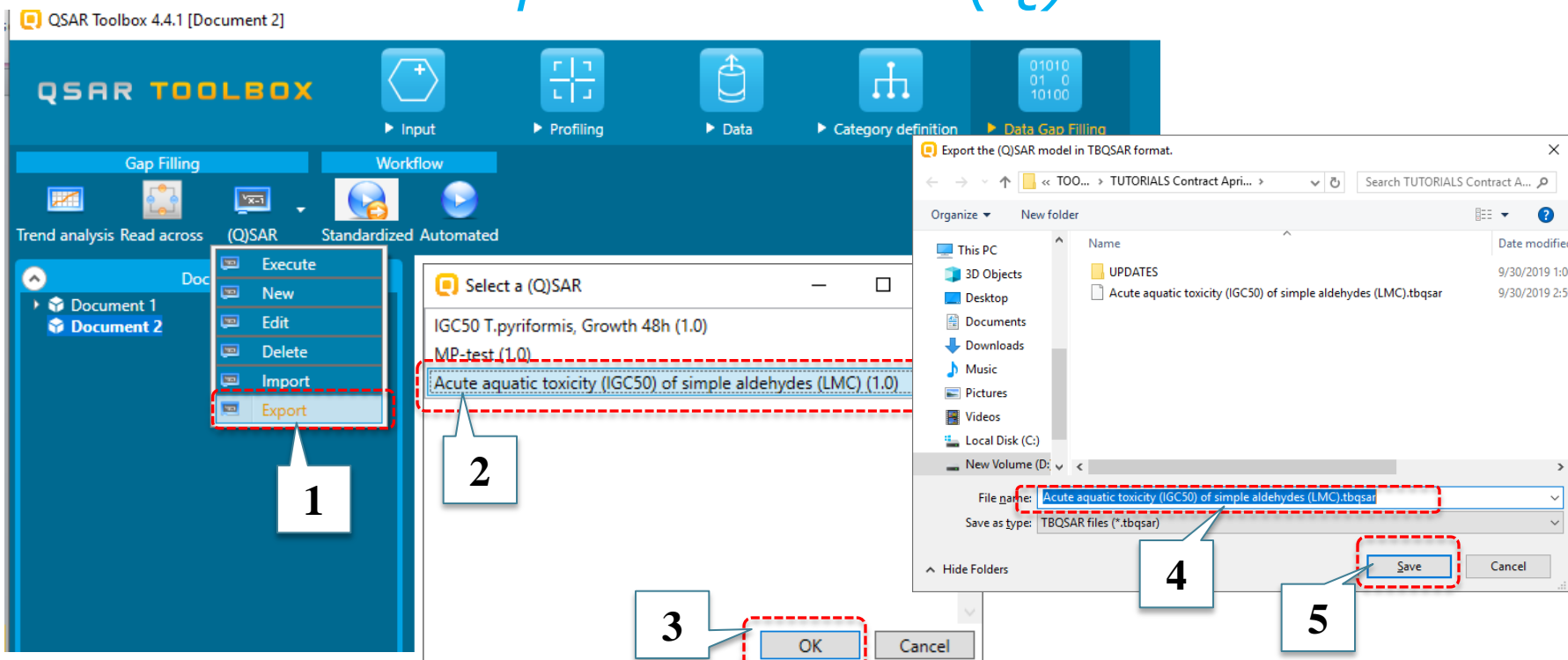


# Import/Export of the custom (Q)SAR



# Import/Export of the custom (Q)SAR

## *Export a custom (Q)SAR*

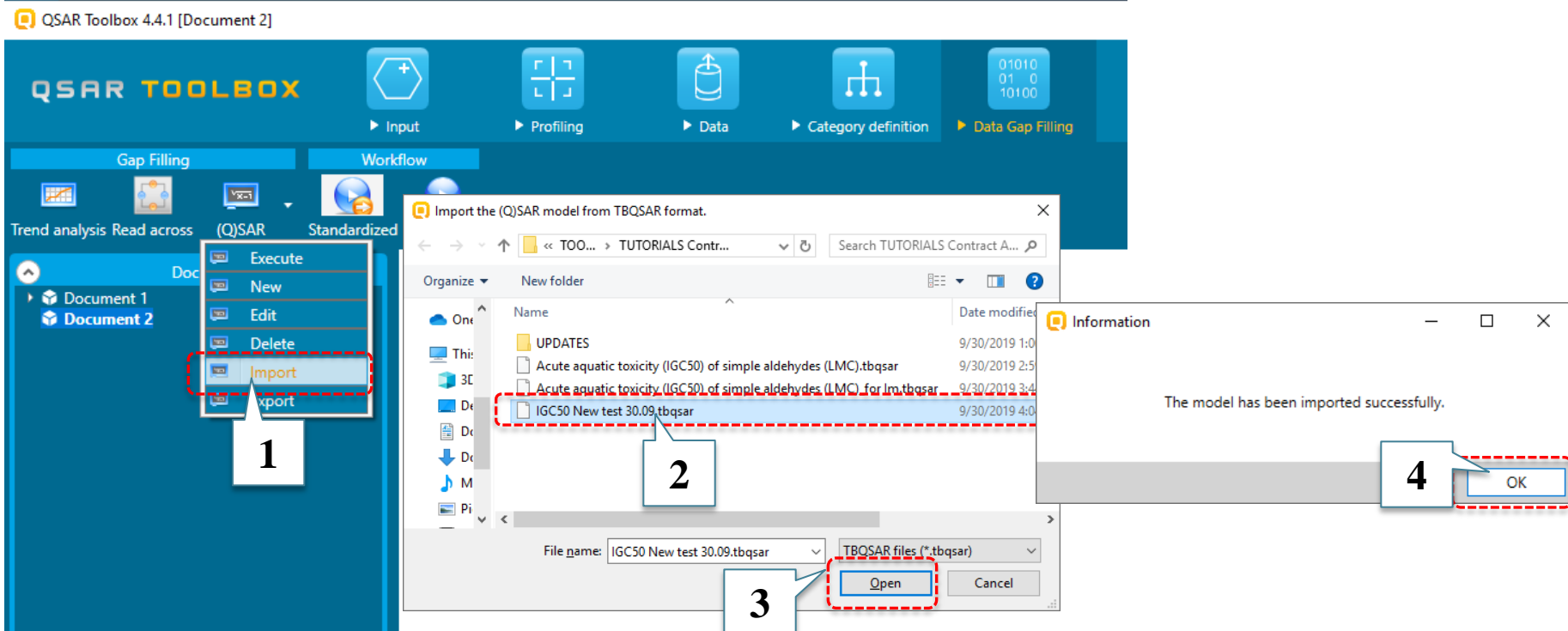


The "Export" operation packs all files associated with (Q)SAR in one file with the extension "\*.tb(Q)SAR". The file can be disseminated to other users and imported by them via Import (see next slides).

1. Click **Export** button;
2. From the list of user-defined (Q)SARs select the desired model "**Acute aquatic toxicity (IGC50) of simple aldehydes (LMC) (1.0)**";
3. Click **OK**;
4. Browse to a folder on your computer and **give** the name of the file (or keep it as it is); The (Q)SAR and its supporting information such as applicability domain, equation and training set are packed in one file with extension "\*.tb(Q)SAR";
5. Click **Save**

# Import/Export of the custom (Q)SAR

## *Import a custom (Q)SAR*



“Import” of custom (Q)SAR is allowed only for (Q)SARs built in Toolbox environment (the (Q)SAR generated by Toolbox tools).

1. Click **Import** button;
2. Browse and find your custom (Q)SAR (with extension “\*.tb(Q)SAR”); Select it;
3. Click **Open**;
4. A message appears notifying the user that the (Q)SAR is imported successfully.

## The Exercise

### Example 2

- In the second example we will build a (Q)SAR using a web service link.
- This link must be provided by (Q)SAR model developers related to a specific (Q)SAR model and endpoint (in our case we will use a link provided by <http://qsardb.org/about/citing>).
- The link will provide a predicted value for a given SMILES. The value will be returned to the Toolbox using API services.
- In the current example we will demonstrate building a (Q)SAR model for predicting melting point (MP). Details about the model are given below:

#### Endpoint

Endpoint	MP
Endpoint unit	°C (degree Celsius)

#### Web service link:

<http://qsardb.org/repository/service/predictor/10967/104/models/rf?<smi>>

**Reference:** <http://qsardb.org/repository/handle/10967/104>

## The Exercise

### Start building a new (Q)SAR

We are going to create a new (Q)SAR model:

- Open the Toolbox.
- Move to the Data Gap Filling module
- Create New (Q)SAR (already showed on slide 14)
- Specify the name of the new (Q)SAR (see next screenshot).

# Building a new (Q)SAR

The screenshot displays the QSAR Toolbox 4.4.1 interface. At the top, the 'QSAR TOOLBOX' logo is visible. Below it, a navigation bar contains several icons and labels: 'Input', 'Profiling', 'Data', 'Category definition', 'Data Gap Filling' (highlighted with a red box and callout 1), and 'Report'. A secondary bar shows 'Gap Filling', 'Workflow', 'Trend analysis', 'Read across', '(Q)SAR', 'Standardized', and 'Automated'. On the left, a 'Document 1' tab is active, with a context menu open showing options: 'Execute', 'New' (highlighted with a red dashed box and callout 2), 'Edit', 'Delete', 'Import', and 'Export' (callout 3). The main workspace is divided into two panes. The left pane, titled 'Wizard pages', lists steps: 'QSAR Identity', 'General information', 'Defining the endpoint - OECD Principle 1', 'Defining the algorithm - OECD Principle 2', 'Applicability domain - OECD Principle 3', 'Training set and statistics - OECD Principle 4', 'External validation and predictivity - OECD Principle 4', and 'Mechanistic' (callout 4). The right pane, titled 'QSAR Editor', contains a table with the following data:

QSAR Title/Caption	Version	Other related models
QSAR for predicting MP based on webservice link	1.0	

Below the table, the text 'Software implementing the model' is followed by 'QSAR Toolbox 4.4 (beta)' (callout 5).

1. Go to the **Data Gap Filling** module;
2. Click on the drop-down menu;
3. Select **New**;
4. (Q)SAR Editor wizard appears;
5. Add name of the (Q)SAR model. In this case **“(Q)SAR for predicting MP based on web service link”** is added.

## Building a new (Q)SAR

Once the (Q)SAR Editor is opened, the following two types of sections should be filled:

- Important sections mandatory for correct work of the (Q)SAR:
  - the (Q)SAR title (**already defined**);
  - the endpoint (**in our case MP**) and its unit (**in our case degree Celsius (°C)**)
  - The web service link – Requirements for the constructing the link itself are available in the Editor wizard.
- Additional sections - not mandatory for correct work of the (Q)SAR but recommended according to the five OECD principles:
  - Applicability domain could be defined;
  - Training set/test set could be imported along with statistical information
  - Additional QMRF information could be added, too (such as author, dependent variables, description of the algorithm etc.)

In this example these additional fields are not filled.

# Building a new (Q)SAR

## Define the endpoint

**Wizard pages**

- QSAR Identity
- General information
- Defining the endpoint – OECD Principle 1**
- Defining the algorithm – OECD Principle 2
- Applicability domain – OECD Principle 3
- Training set and statistics – OECD Principle 4
- External validation and predictivity – OECD Principle 4
- Mechanistic interpretation

**Endpoint to predict**  
No endpoint defined. Please define an endpoint to predict.

**Define**

**Comment on the endpoint**

**Endpoint units**  
Unknown Set Unset

**Dependent variable**

**Experimental protocol**

**Data quality and variability**

**Select endpoint**

- Physical Chemical Properties
  - Melting / freezing point

**Endpoint**  
Melting point

**Selection of additional metadata fields:**

Add Up Down Clear Remove

**Finish**

First we need to define an endpoint. In our example this is "Melting point"

1. Go to the section **Defining the endpoint – OECD Principle 1**;
2. Click **Define**;
3. Select **Melting / freezing point** from **Physical Chemical Properties**;
4. Select **Melting Point** from **Endpoint** field;
5. Click **Finish**.



# Building a new (Q)SAR

## Define the unit of the endpoint

**Wizard pages**

- QSAR Identity
- General information
- Defining the endpoint – OECD Principle 1
- Defining the algorithm – OECD Principle 2
- Applicability domain – OECD Principle 3
- Training set and statistics – OECD Principle 4
- External validation and predictivity – OECD Principle 4
- Mechanistic interpretation – OECD Principle 5

**Endpoint to predict**

Tree position: Physical Chemical Properties#Melting / freezing point

Data filters: Endpoint=Melting point;

Define

**Comment on the endpoint**

**Endpoint units** Unknown **Set** Unset

**Dependent variable**

**Experimental protocol**

**Data quality and variability**

**Origin**

scale: unit:

**Destination**

Temperature

unit

☒ °C

☐ °F

☐ K

Expressions

OK Cancel

Cancel Create

Second the unit of the defined endpoint needs to be added. In this case this is "°C (degree Celsius)"

1. Click **Set** button;
2. Select **Temperature** from the drop-down menu;
3. Chose "°C";
4. Click **OK**;

# Building a new (Q)SAR

## Define the web services link

**Wizard pages**

- QSAR Identity
- General information
- Defining the algorithm – OECD Principle 2
- Applicability domain – OECD Principle 3
- Training set and statistics – OECD Principle 4
- External validation and predictivity – OECD Principle 4
- Mechanistic interpretation – OECD Principle 5
- Miscellaneous

**Type of model**

**Algorithm description**

☐ Equation ☒ **Web service link**

Link:

**Web service link requirements:**

- 1) The web service link must contain <smi>. That way when <smi> is replaced with a SMILES value and called the web server should respond with a predicted value for that chemical.  
An example format is `http://host/service?<smi>`
- 2) The response value must be a single string in the format: "endpoint = value".  
Where "endpoint" is the name of the predicted endpoint and the value is an integer or a floating point number with a full stop (.) as a decimal separator.

Reference link:

**Descriptor selection**

**Algorithm and descriptor generation**

**Software name and**

Third step is to add the web services link, which will return the predicted values to the Toolbox based on a given SMILES. Detailed instructions on how to construct the link are available.

1. Click on section **"Defining the algorithm – OECD Principle 2"**;
2. Click **Web service link** radio button;
3. The requirements for building the link are available. Continue on next slide.

# Building a new (Q)SAR

## Define the web services link

**Wizard pages**

- QSAR Identity
- General information
- Defining the endpoint – OECD Principle 1
- Defining the algorithm – OECD Principle 2**
- Applicability domain – OECD Principle 3
- Training set and statistics – OECD Principle 4
- External validation and predictivity – OECD Principle 4
- Mechanistic interpretation – OECD Principle 5
- Miscellaneous

**Type of model**

**Algorithm description**

Equation ☒ Web service link

Link:  Check

**Web service link requirements:**

- 1) The web service link must contain <smi>. That way when <smi> is replaced with the server should respond with a predicted value for that chemical.  
An example format is `http://host/service?<smi>`
- 2) The response value must be a single string in the format: "endpoint = value".  
Where "endpoint" is the name of the predicted endpoint and the value is an integer, full stop (.) as a decimal separator.

Reference link:

**Descriptor selection**

**Algorithm and descriptor generation**

**Software name and**

**Success**

The Web service link is valid.

OK

**A web service link related to the (Q)SAR model needs to be pasted in the appropriate field. In our case the link is (also provided on slide 44):**  
<http://qsardb.org/repository/service/predictor/10967/104/models/rf?<smi>>

1. Paste the **web service link** in the field for **Link**;
2. Click **Check** to check for correctness of the link;
3. A message appears, click **OK**.

# Building a new (Q)SAR

## Define the reference link

QSAR Editor

**Wizard pages**

- QSAR Identity
- General information
- Defining the endpoint – OECD Principle 1
- Defining the algorithm – OECD Principle 2**
- Applicability domain – OECD Principle 3
- Training set and statistics – OECD Principle 4
- External validation and predictivity – OECD Principle 4
- Mechanistic interpretation – OECD Principle 5
- Miscellaneous information

**Type of model**

**Algorithm description**

☐ Equation ☒ Web service

Link:

**Information**

The model was saved successfully!

OK

Web service link requirements:  
 1) The web service link must contain the name of the predicted endpoint and the value is an integer or a floating point number.  
 An example format is http://ho...  
 2) The response value must be a...  
 Where "endpoint" is the name of the predicted endpoint and the value is an integer or a floating point number.

Reference link:

**Descriptor selection**

**Algorithm and descriptor generation**

**Software name and version for descriptor generation**

Back Next Cancel **Create**

Also a reference link associated with the (Q)SAR could be added:  
[http://\(Q\)SARdb.org/repository/handle/10967/104](http://(Q)SARdb.org/repository/handle/10967/104)

1. Paste the **reference link** in the appropriate place;
2. Additional fields could be filled, too;
3. Finally click **Create**;
4. A message opens, click **OK**.

# Application of the (Q)SAR to a list of chemicals

QSAR Toolbox 4.4.1 [Document 1]

Details for 6 (Q)SAR models

QSAR name	#	Predicted	Domain	Endpoint
(Q)SAR for predicting MP based on web service link" (1.0)	1	171 °C	No domain available	Melting point
Mean Melting Point (EPISUITE) (1.0)	2	265 °C	No domain available	Melting point
Melting Point (Adapted Joback Method) (EPISUITE) (1.0)	3	350 °C	No domain available	Melting point
Melting Point (Gold and Ogle Method) (EPISUITE) (1.0)	4	180 °C	No domain available	Melting point
Selected Melting Point (EPISUITE) (1.0)	5	214 °C	No domain available	Melting point

Select QSAR method

☐ Enter Gap filling  
☐ Predict selected chemical  
☒ Predict all chemicals

Run

Here we will apply the created custom (Q)SAR model to a list of chemicals. For this purpose an example file with 13 structures will be loaded.

1. Load "**structures\_quantitative\_metabolic\_data.smi**" from the **Example folder** (see slide 36 for more details);
2. Position mouse on the level of **Physical Chemical Properties#Melting / freezing point**;
3. Click **(Q)SAR** and **Execute**;
4. Select the user-defined (Q)SAR model;
5. Click **Run**;
6. Select **Predict all chemicals** and click **OK**.

# Application of the (Q)SAR to a list of chemicals

QSAR Toolbox 4.4.1 [Document 1]

The screenshot displays the QSAR Toolbox 4.4.1 interface. The top menu bar includes options like Input, Profiling, Data, Category definition, Data Gap Filling, and Report. The left sidebar shows the 'Documents' panel with 'Document 1' and '[C: 13;Md: 0;P: 13] structures\_quantitative'. Below it, the 'Data Gap Filling Settings' panel is visible, showing 'Only endpoint relevant' checked and 'At this position:' with 'QSARs' set to 6. The main area features a 'Filter endpoint tree...' on the left and a data matrix on the right. The matrix has 13 columns (numbered 1-13) and multiple rows of endpoints. The 'Melting / freezing point' row is highlighted with a red dashed box, showing values: 13/13, Q: 171 °C, Q: 51.1 °C, Q: 99.7 °C, Q: 86.5 °C, Q: 56.1 °C, Q: 105 °C, Q: 56.9 °C, Q: 29.5 °C, and Q: 165 °C. A callout bubble with the number '1' points to the value 'Q: 56.1 °C' in column 5.

Endpoint	1	2	3	4	5	6	7	8	9
Structure									
Melting / freezing point	13/13	Q: 171 °C	Q: 51.1 °C	Q: 99.7 °C	Q: 86.5 °C	Q: 56.1 °C	Q: 105 °C	Q: 56.9 °C	Q: 29.5 °C

Predictions from the custom (Q)SAR model appear in the data matrix (1)

# Congratulations!

- Now you know how to:
  - create a custom (Q)SAR model via two ways:
    - Mathematical equation;
    - Web service link;
  - apply the created (Q)SAR model to a list of chemicals;
  - import/export the already created custom (Q)SAR for use by other users.
- Continual use of the Toolbox will increase your skills.