QSAR TOOLBOX

The OECD QSAR Toolbox for Grouping Chemicals into Categories

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."[1]

[1]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.

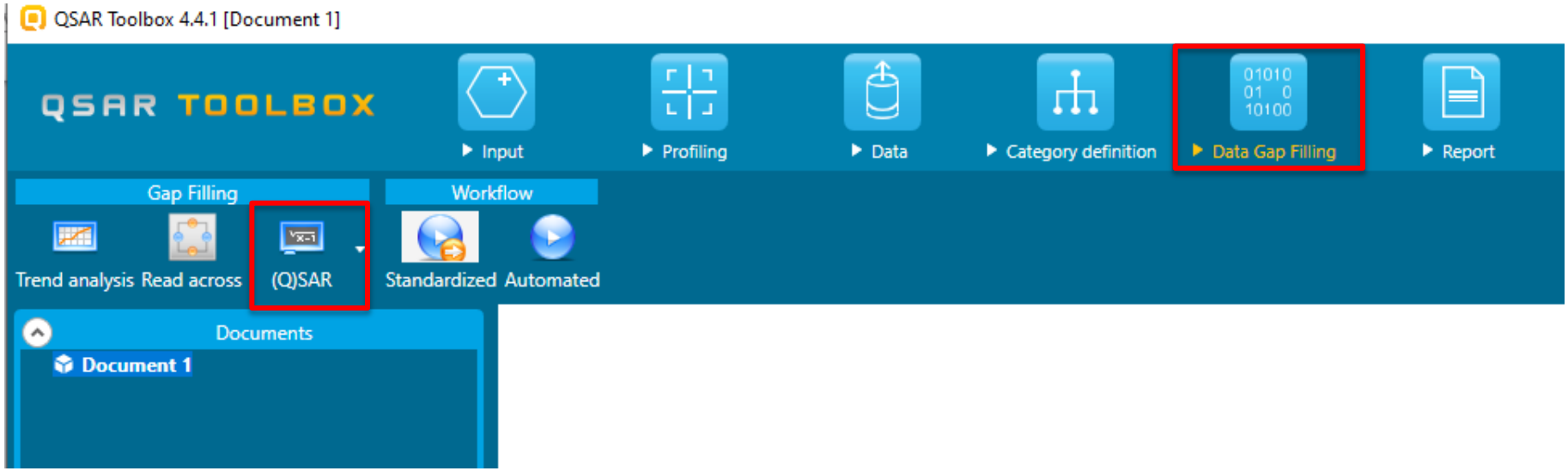
- Single or multiple regression models (based on linear regression) and models based on read-across approach can be implemented.

- 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.

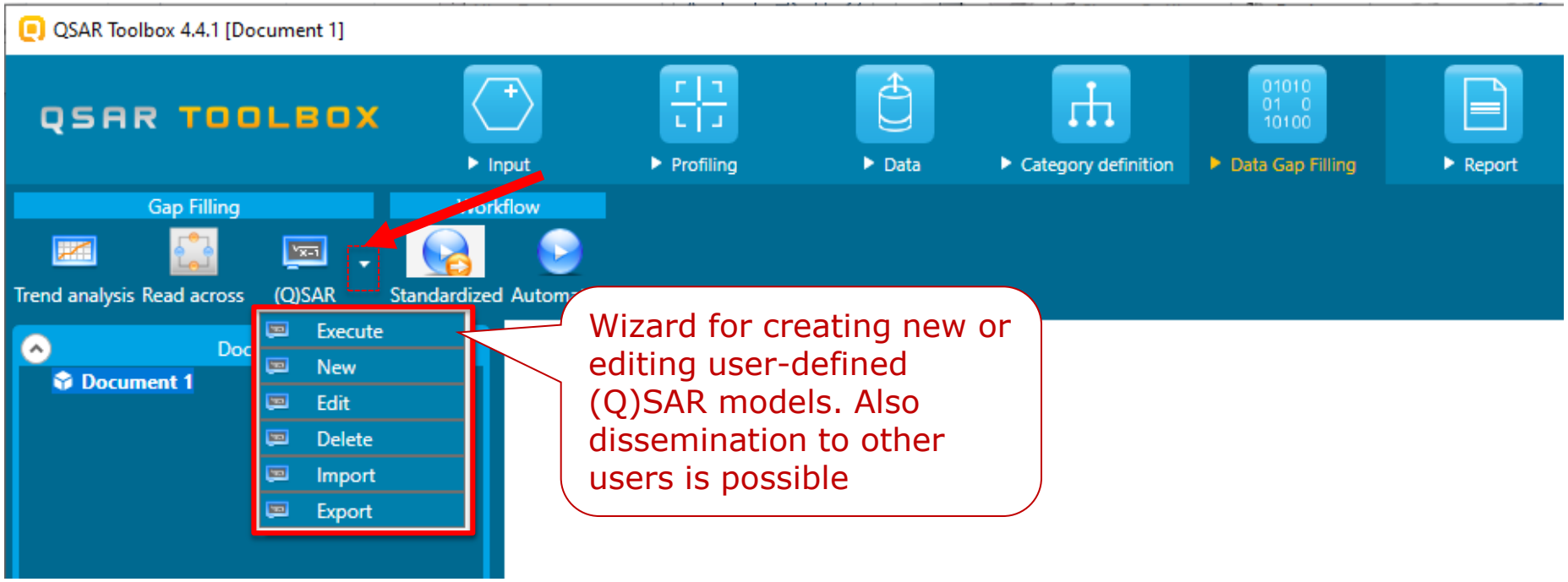
*Creating a custom (Q)SAR in the Gap Filling module is illustrated in tutorial: Step-by-step example for building a (Q)SAR model

(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**

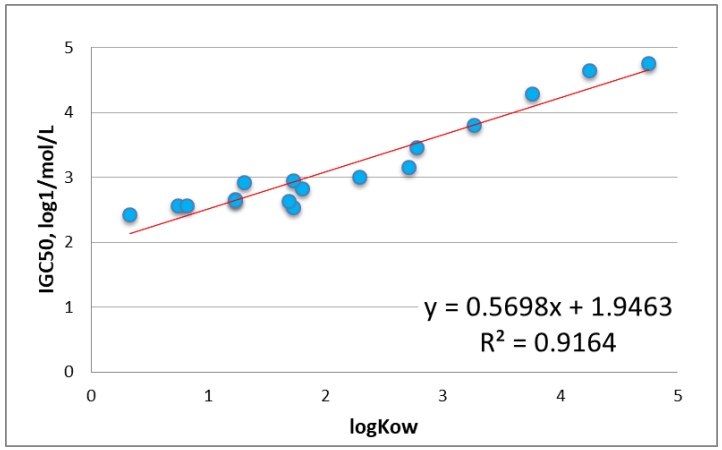
The Exercise
Example 1

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

Endpoint

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

Algorithm



Applicability domain

Parametric requirements:

0.1 ≤ log Kow ≤ 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, Q2 = 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, Q2 = 0.828
Sum of squared residuals, SSR = 0.22

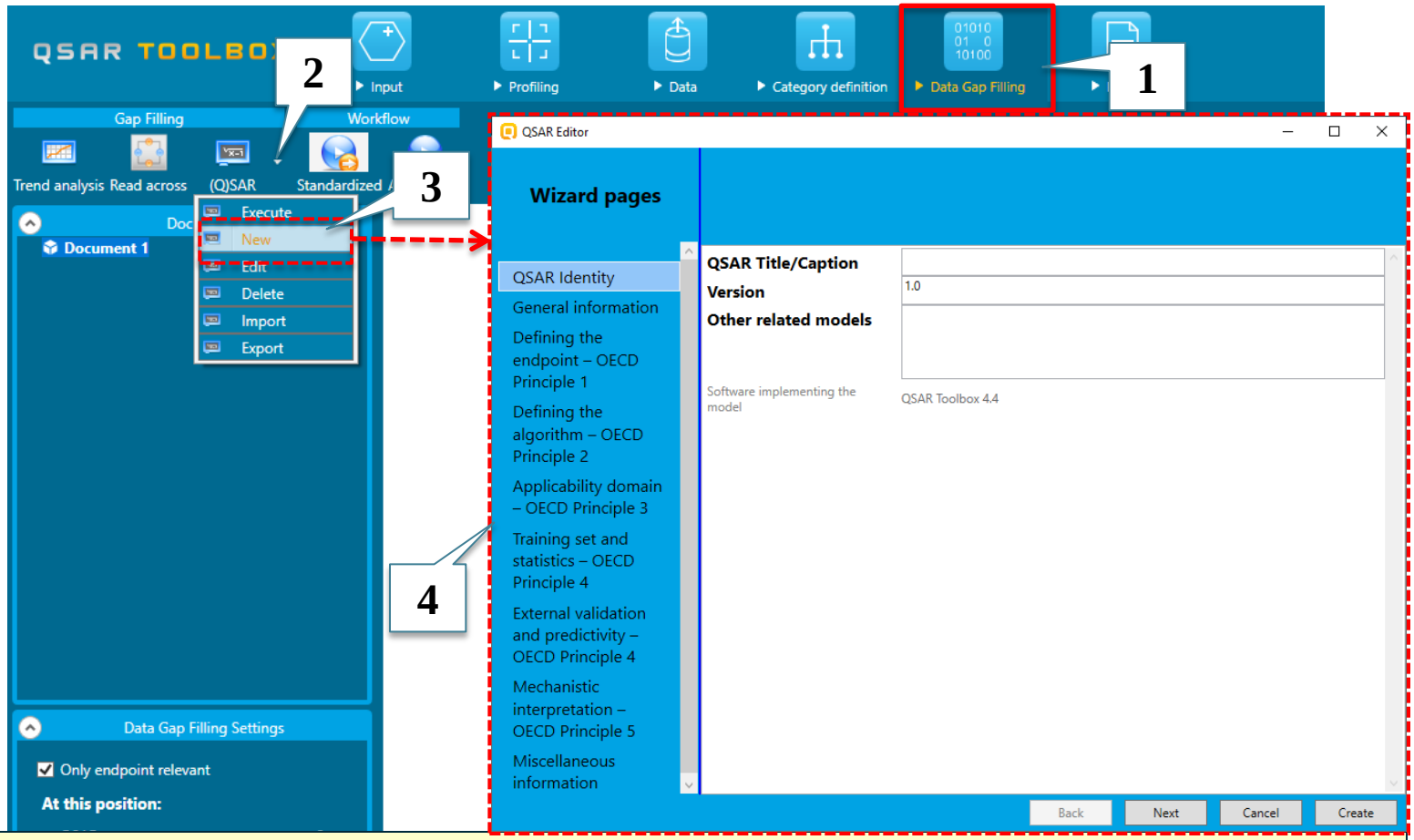
The Exercise
Start building a new (Q)SAR

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*logKow+1.94)**.

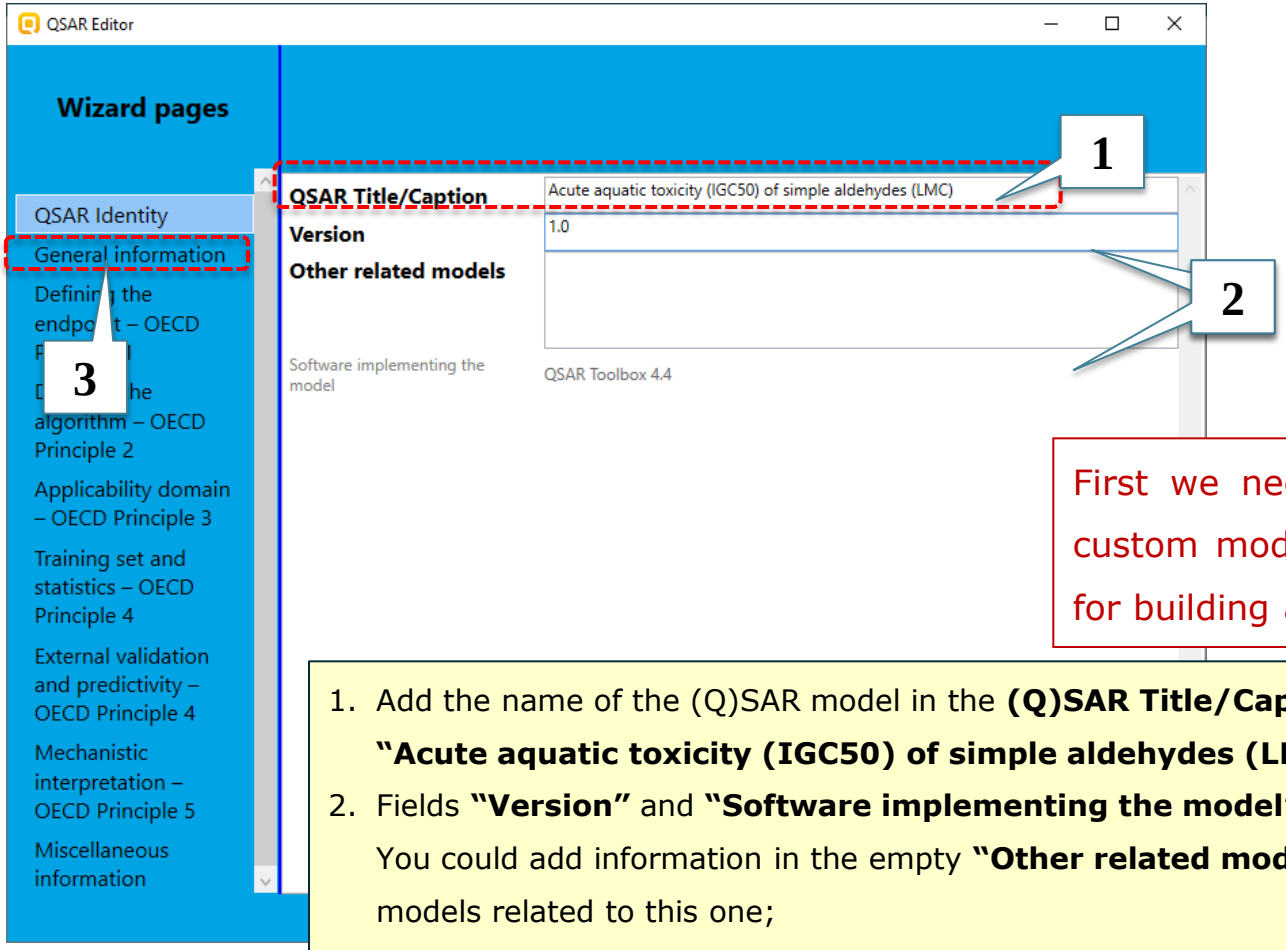
- Additional sections – not mandatory for correct work of the (Q)SAR but recommended according to the five OECD principles*
  - Applicability domain could be defined **(parametric (0.1 ≤ log Kow ≤ 5)** and **structural boundaries (aldehyde))**;
  - Training set/test set could be imported along with statistical 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

Building a new (Q)SAR

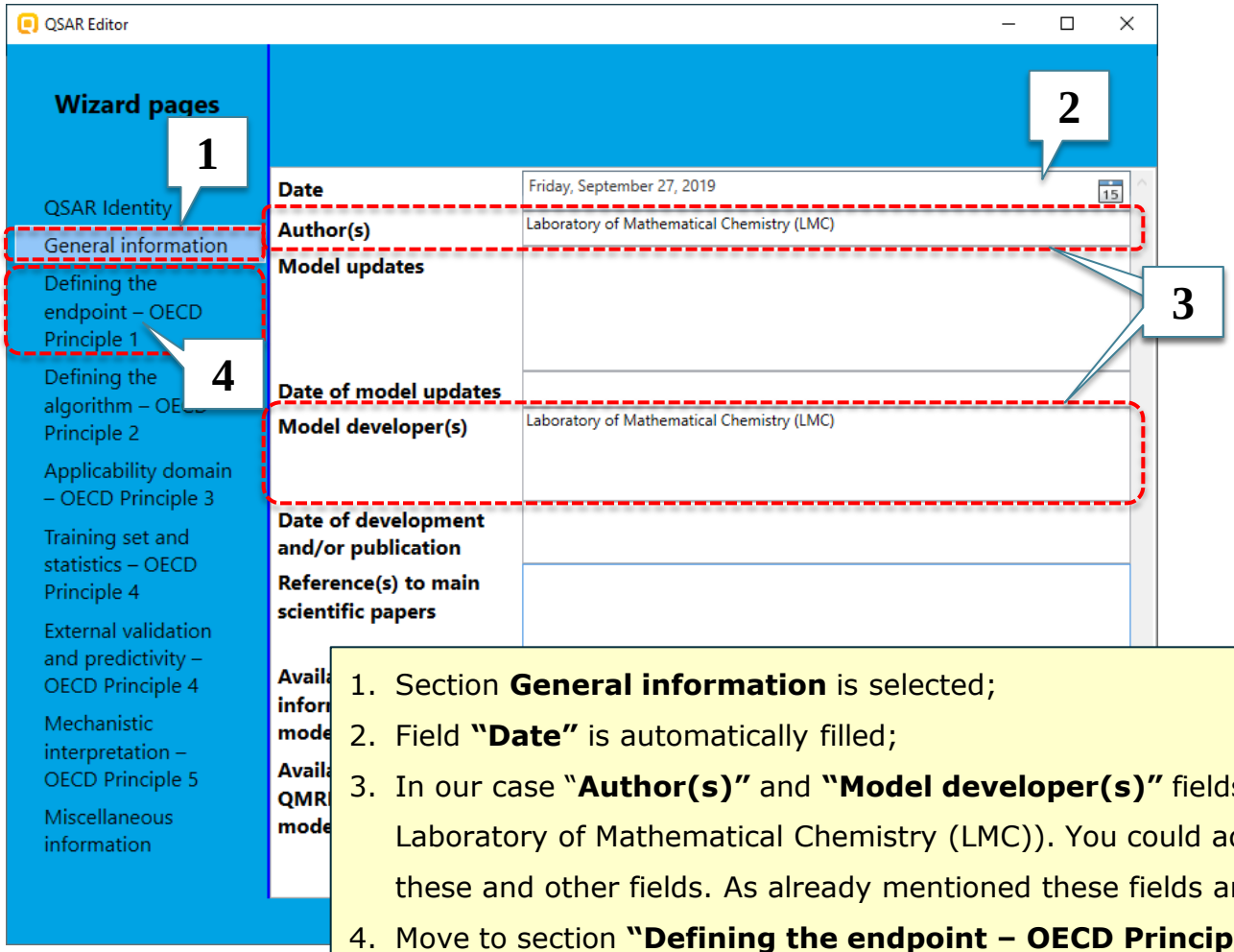
Name of the model



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

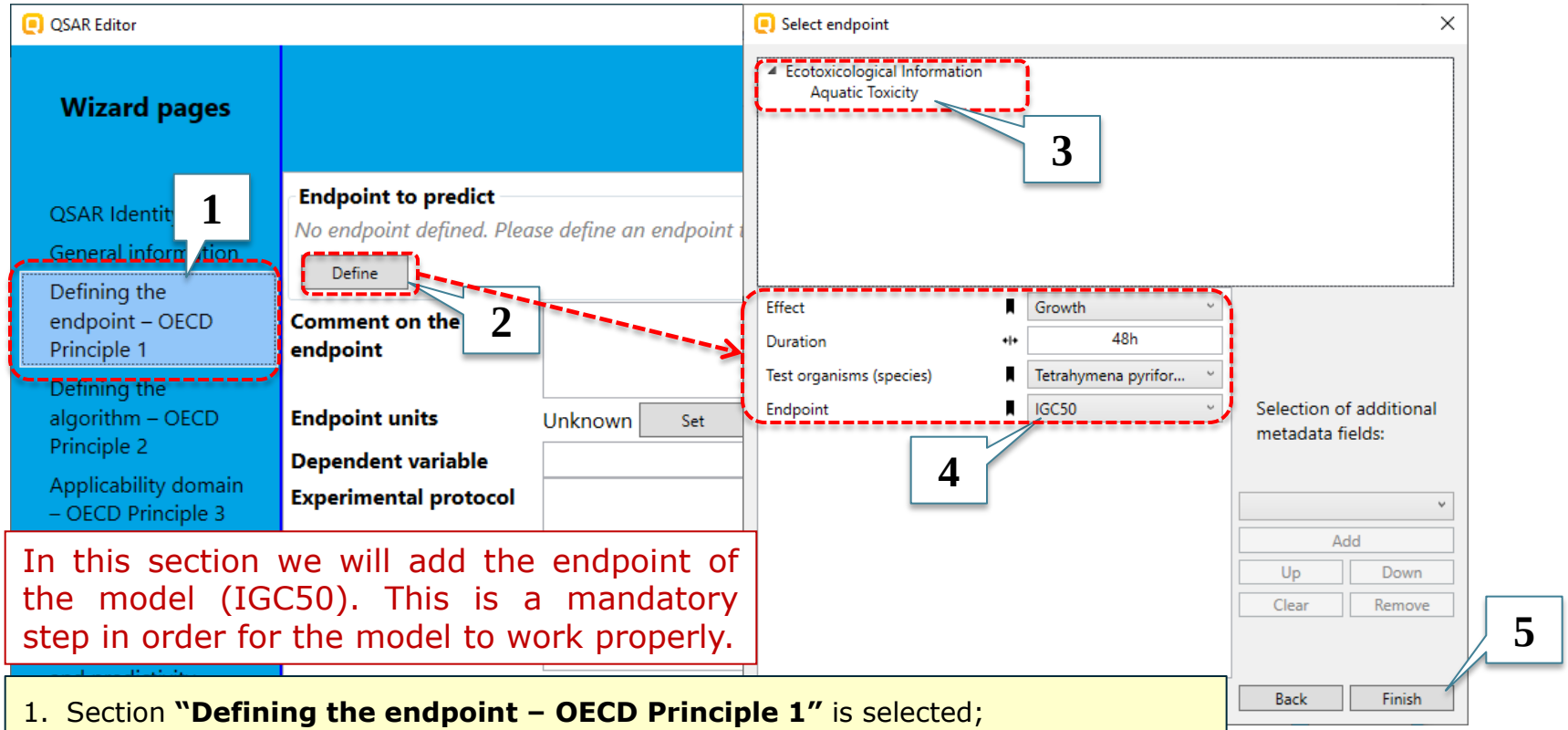
1. Add the name of the (Q)SAR model in the **(Q)SAR Title/Caption** field. In our case it is **"Acute aquatic toxicity (IGC50) of simple aldehydes (LMC)"**;
2. Fields **"Version"** and **"Software implementing the model"** are automatically populated. You could add information in the empty **"Other related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

Building a new (Q)SAR
General information



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

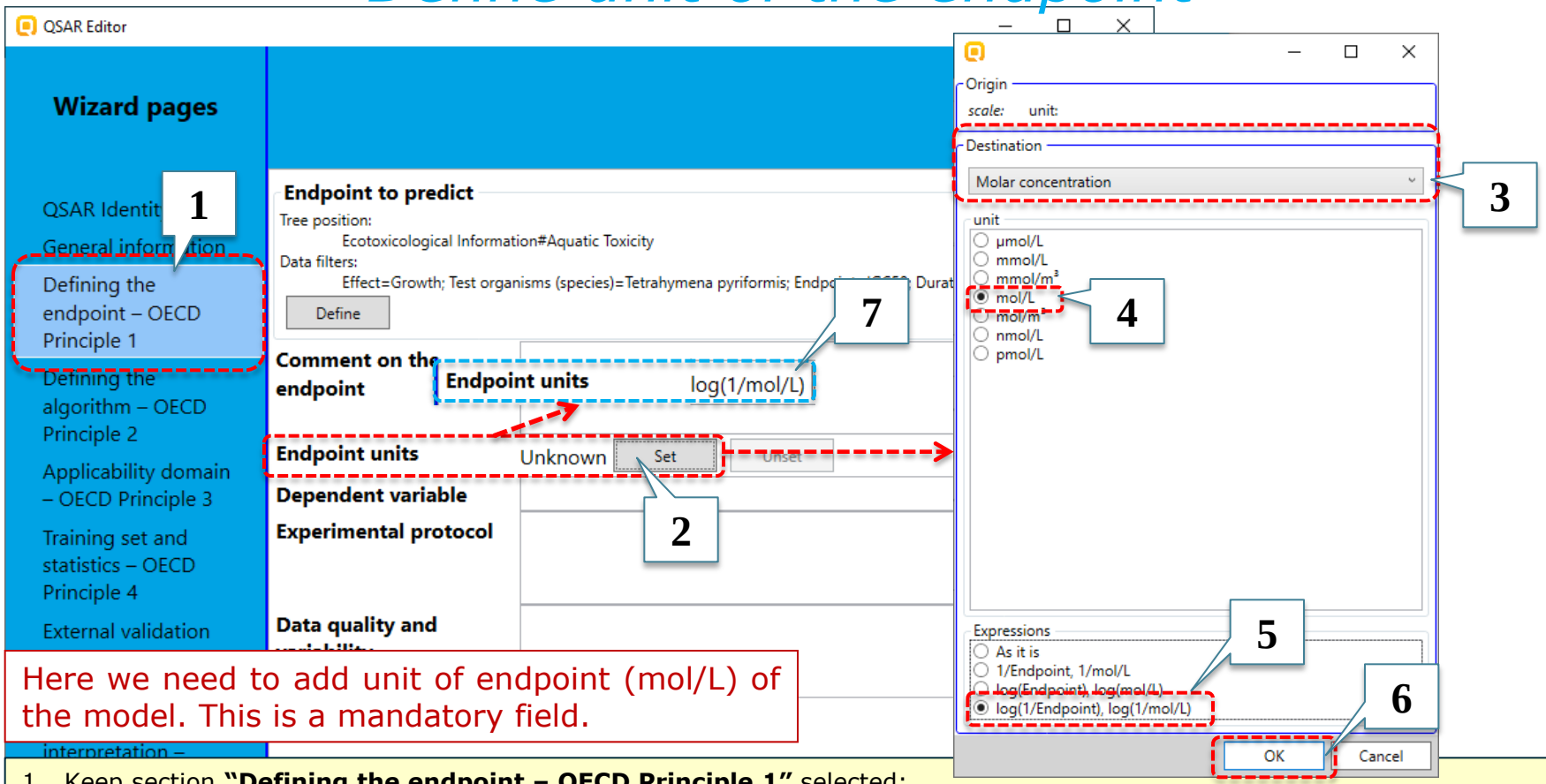


In this section we will add the endpoint of the model (IGC50). This is a mandatory step in order for the model to work properly.

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

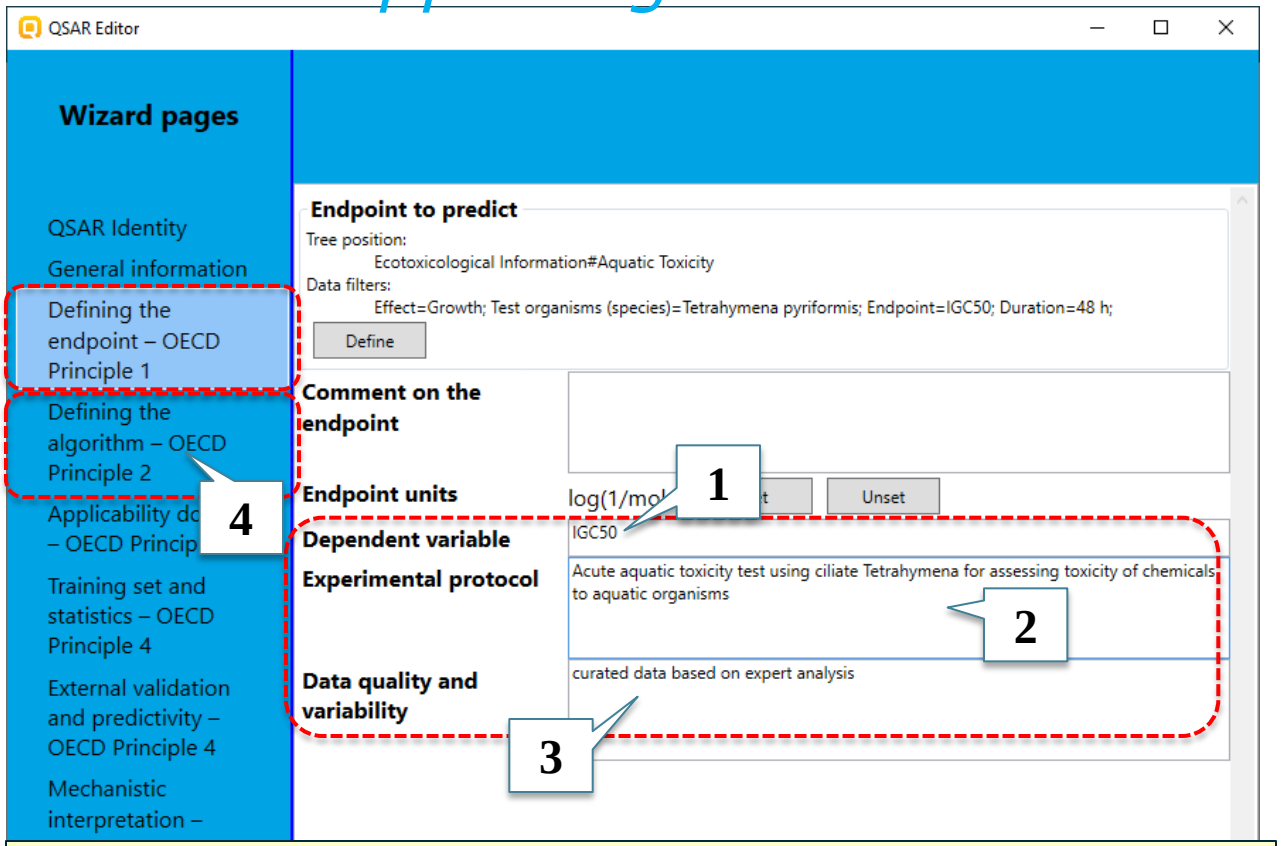


Here we need to add unit of endpoint (mol/L) of the model. This is a mandatory field.

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

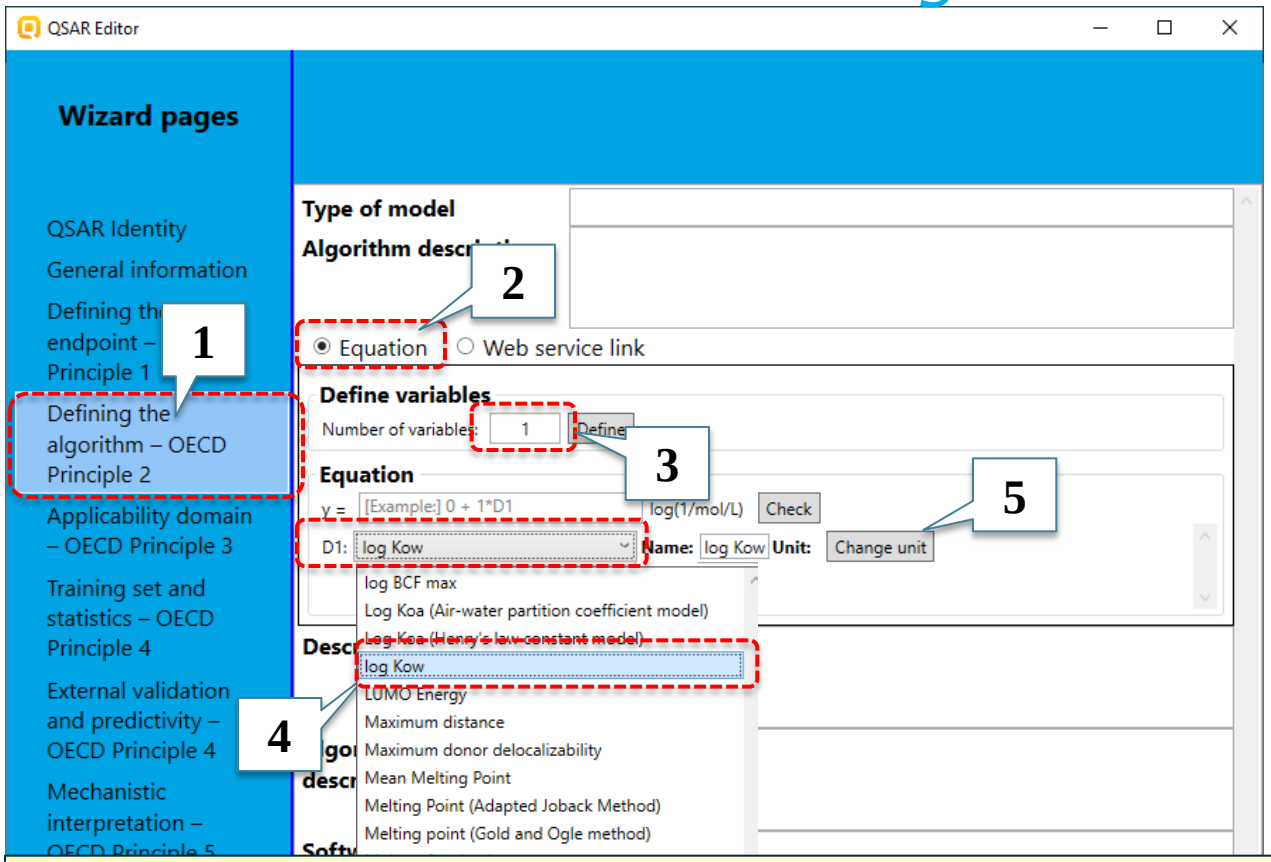
Add supporting information of the endpoint



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".**

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.

Building a new (Q)SAR
Define algorithm

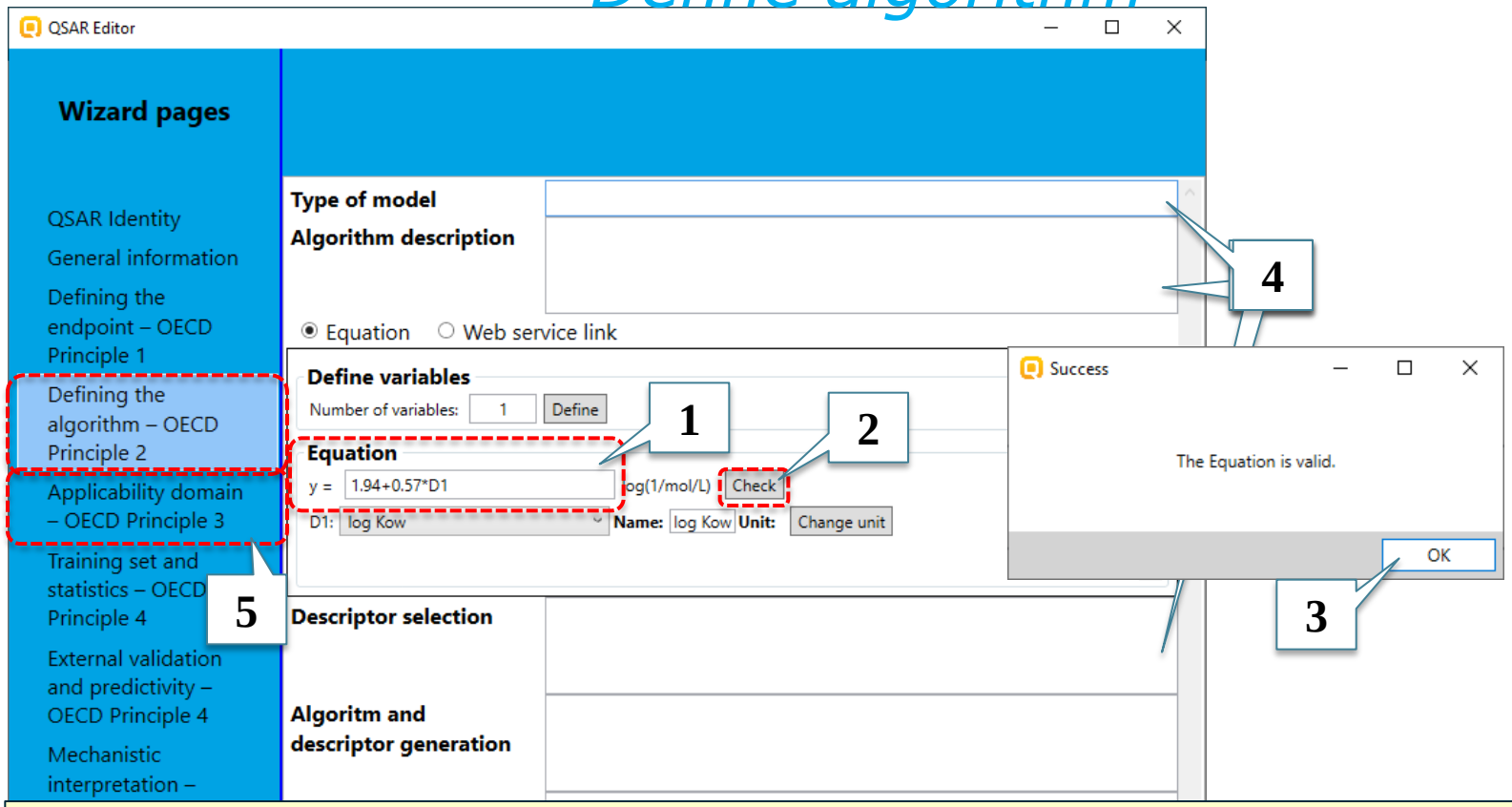


The section "Defining the algorithm" is one of the most important fields for correct configuration of the (Q)SAR model. In this section we will define the number of used variables and the equation itself.

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

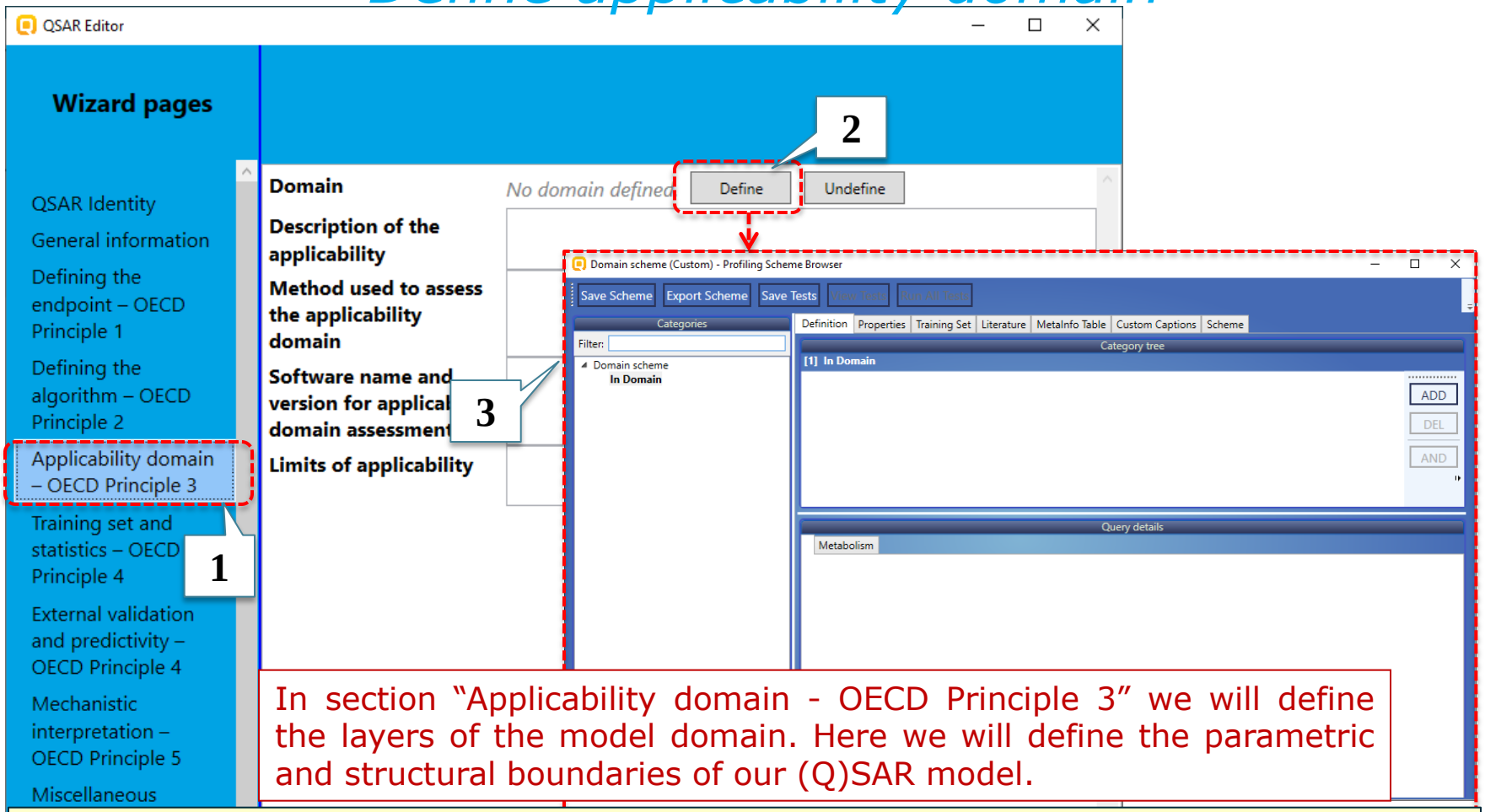


1. Enter your model equation in section **"Equation"**. In our case write "1.94+0.57*D1" in the empty field; The user is able to derive the equation (i.e. to build a model) by using the functionality "Save model" inside the Data Gap Filling stage.*
2. Click **Check** button in order system to check for 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

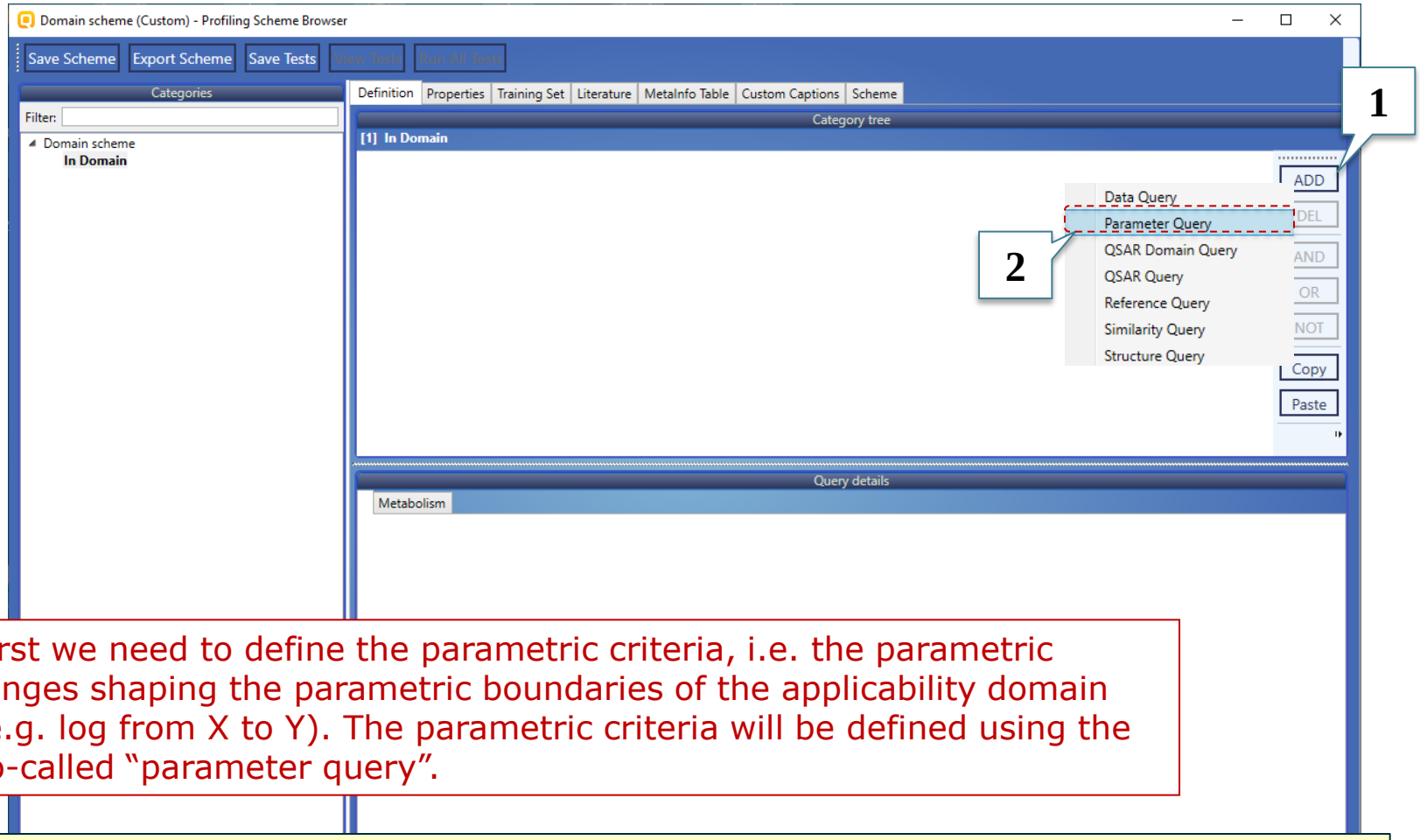


In section "Applicability domain - OECD Principle 3" we will define the layers of the model domain. Here we will define the parametric and structural boundaries of our (Q)SAR model.

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

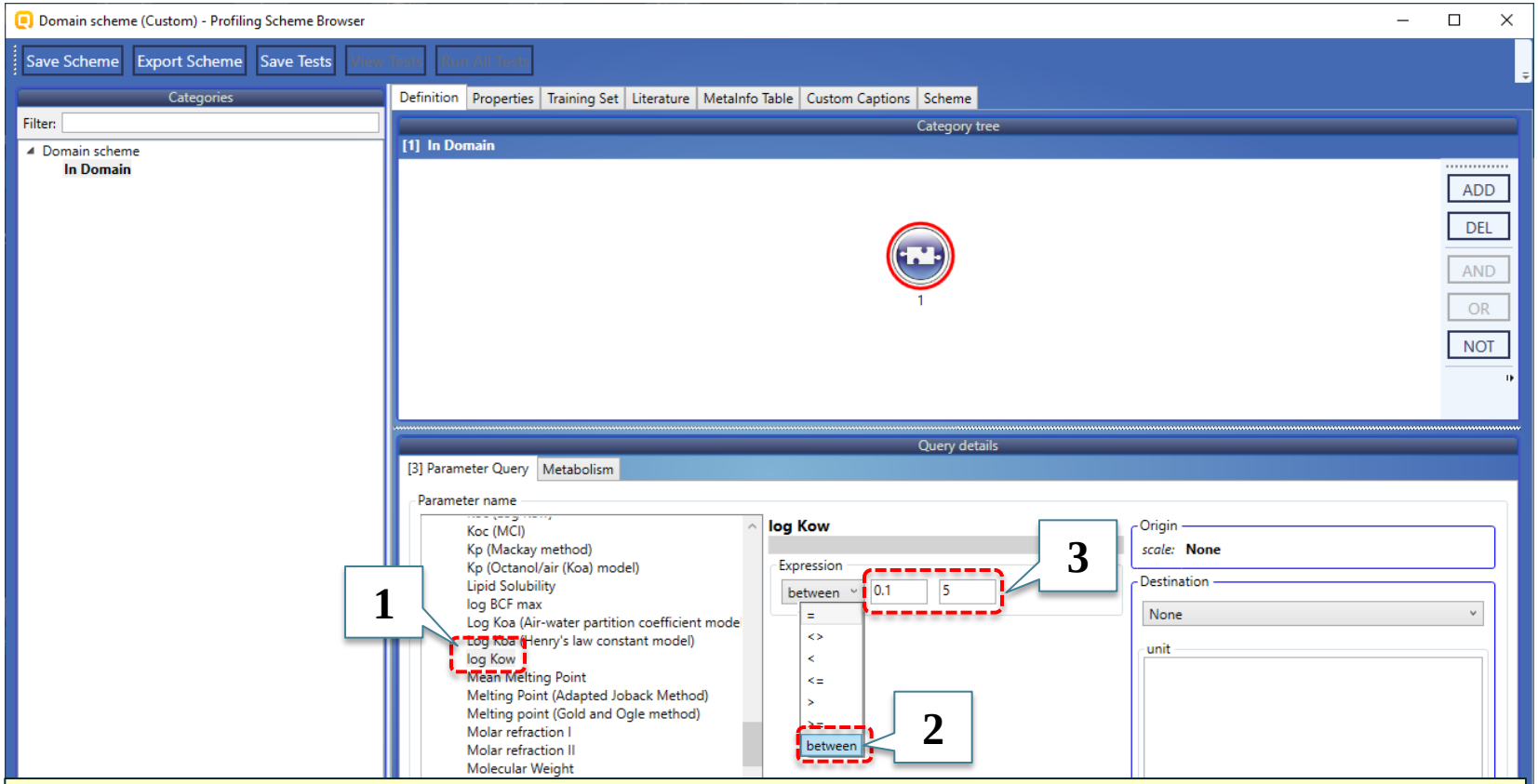


First we need to define the parametric criteria, i.e. the parametric ranges shaping the parametric boundaries of the applicability domain (e.g. log from X to Y). The parametric criteria will be defined using the so-called "parameter query".

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

Building a new (Q)SAR

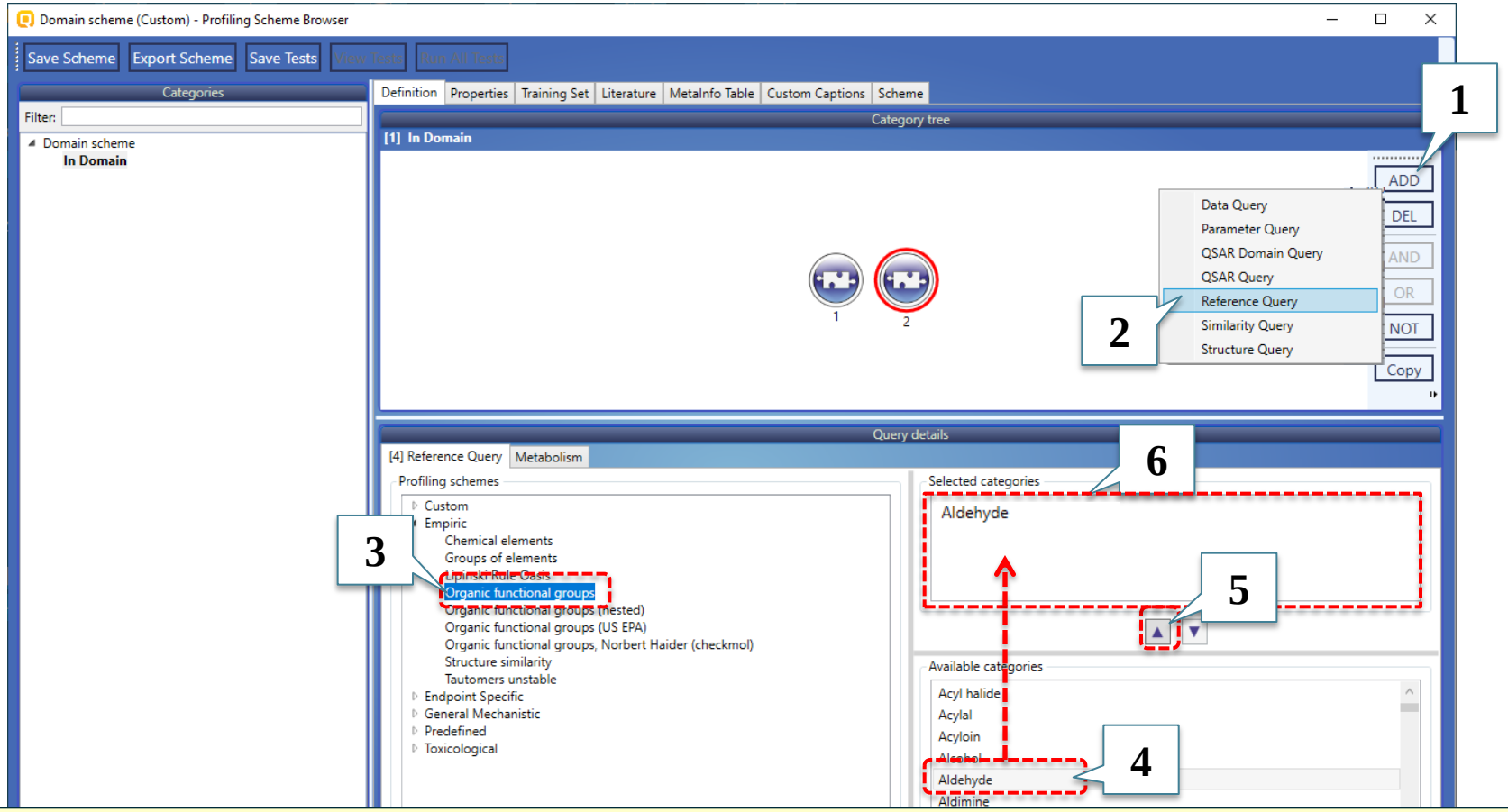
Define applicability domain – parametric boundary



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

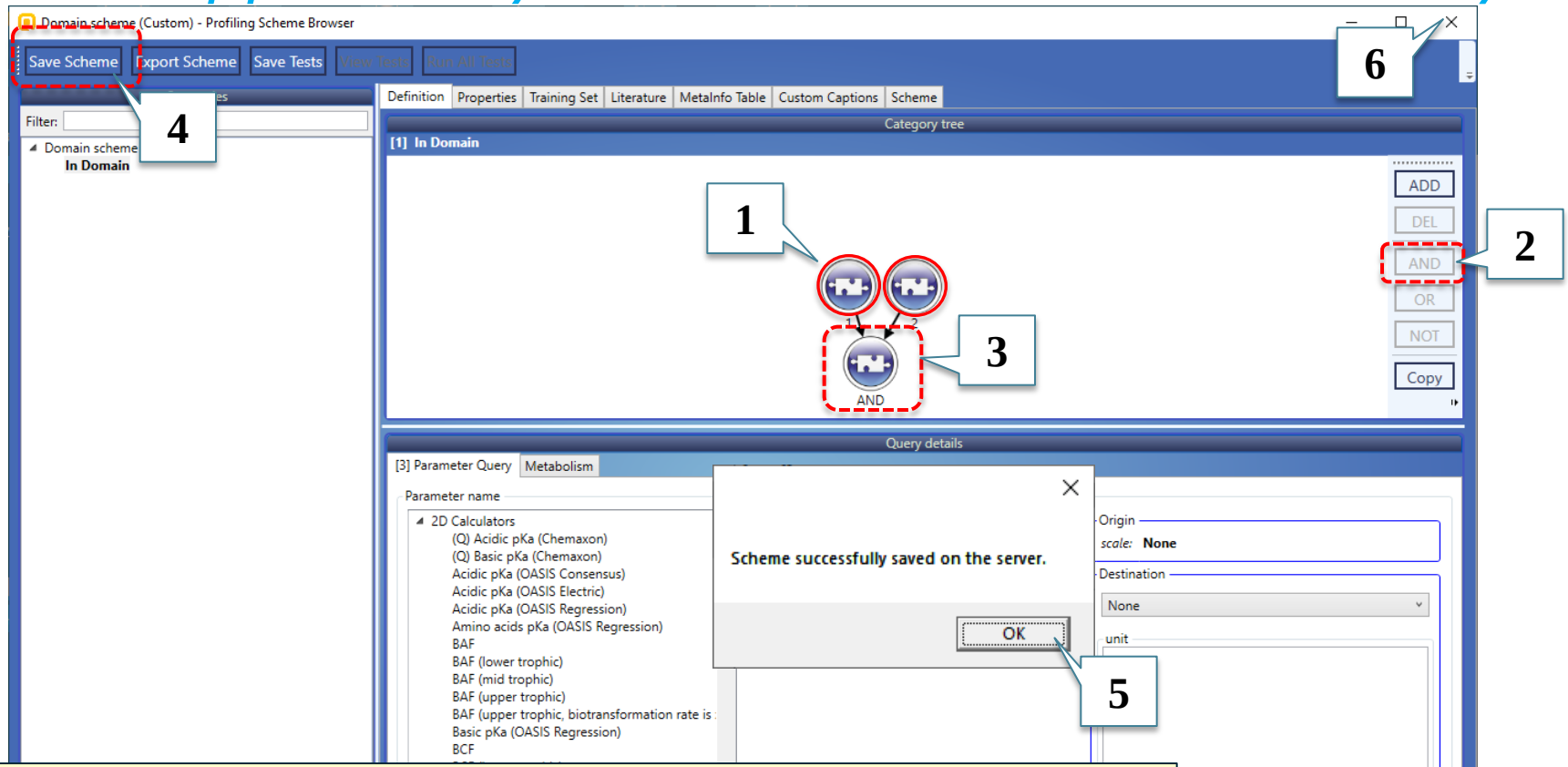
Define applicability domain – structural boundary



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.**

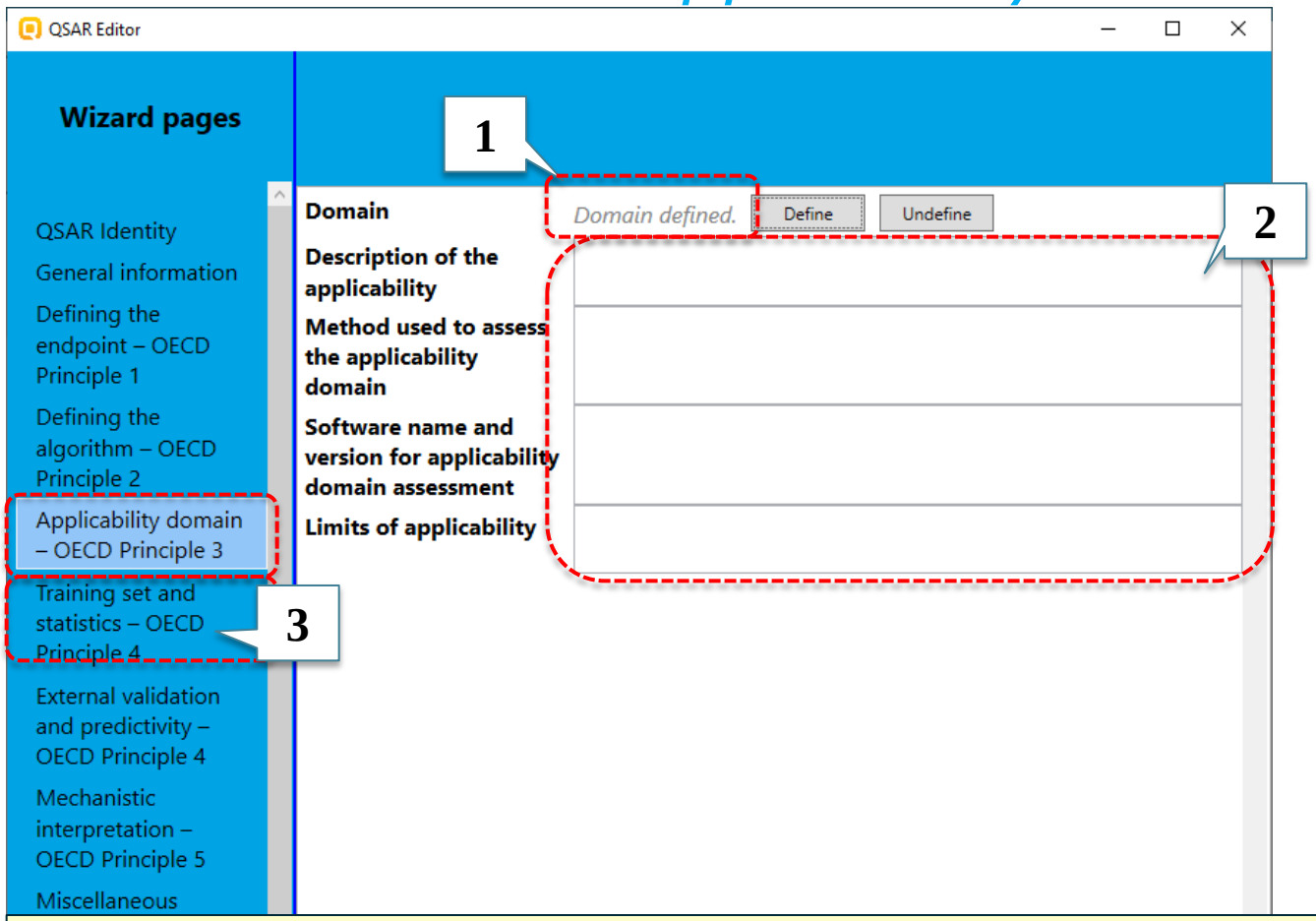
Building a new (Q)SAR

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.

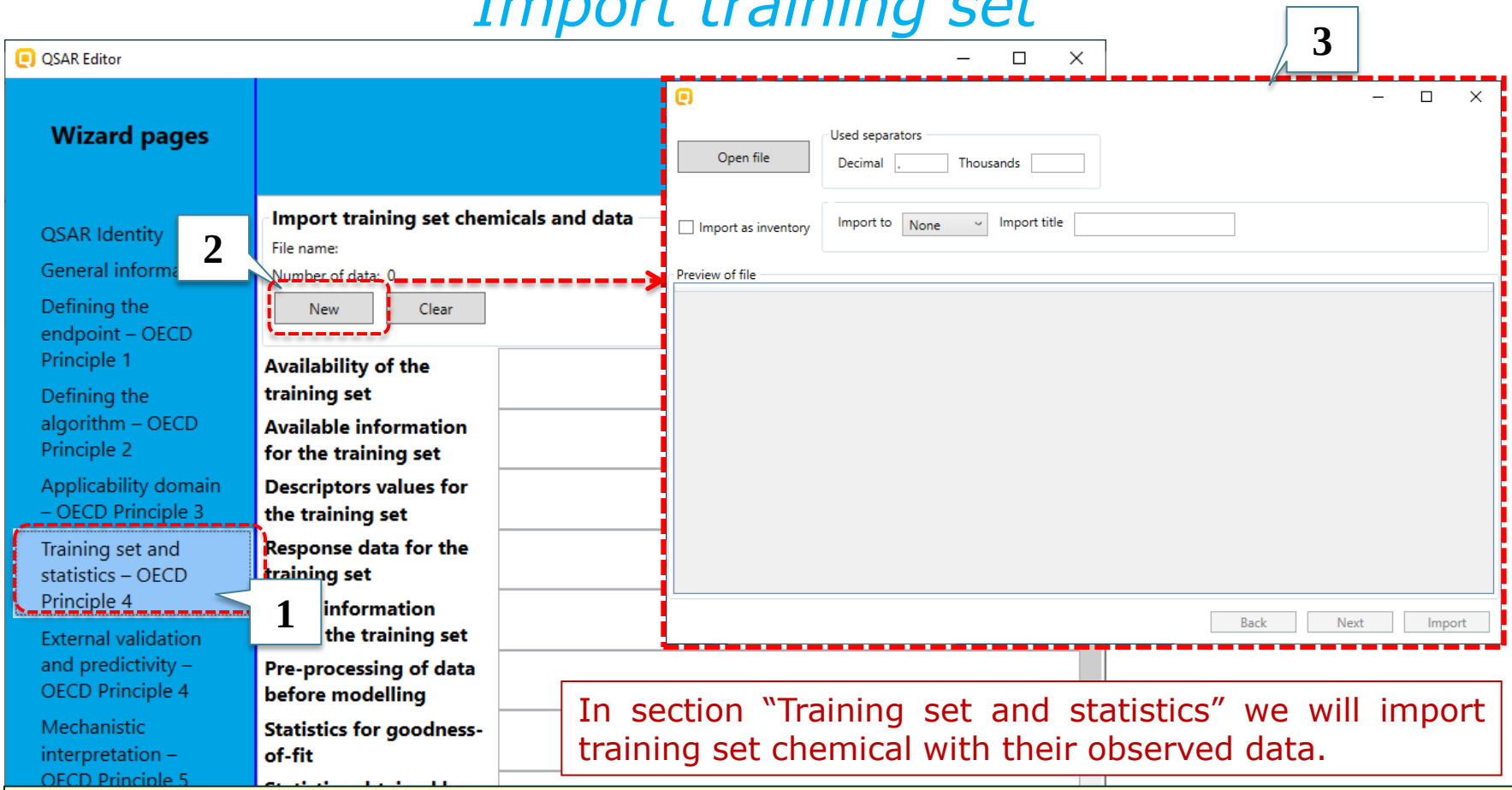
Building a new (Q)SAR
Define applicability domain



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

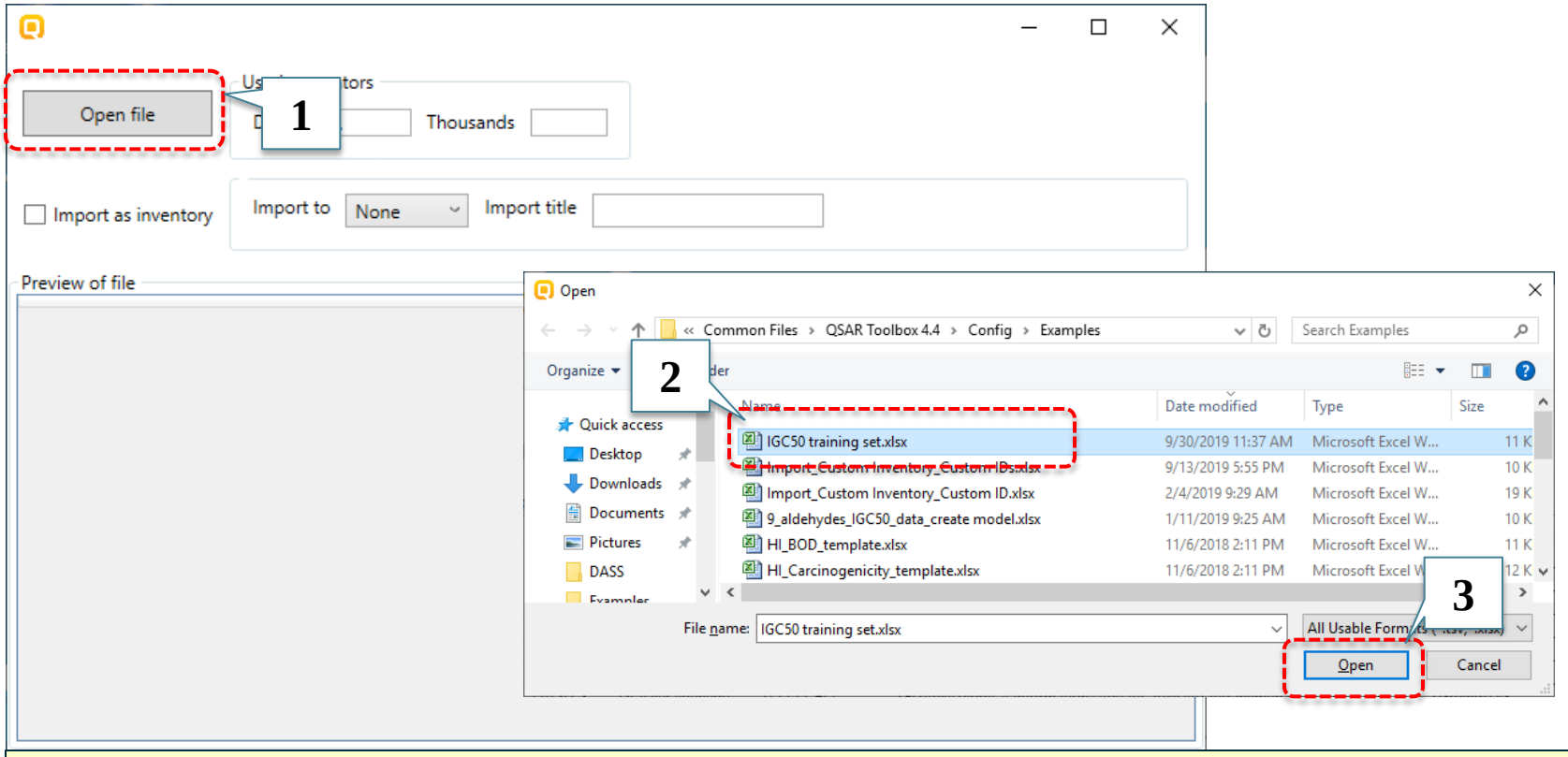
Import training set



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*

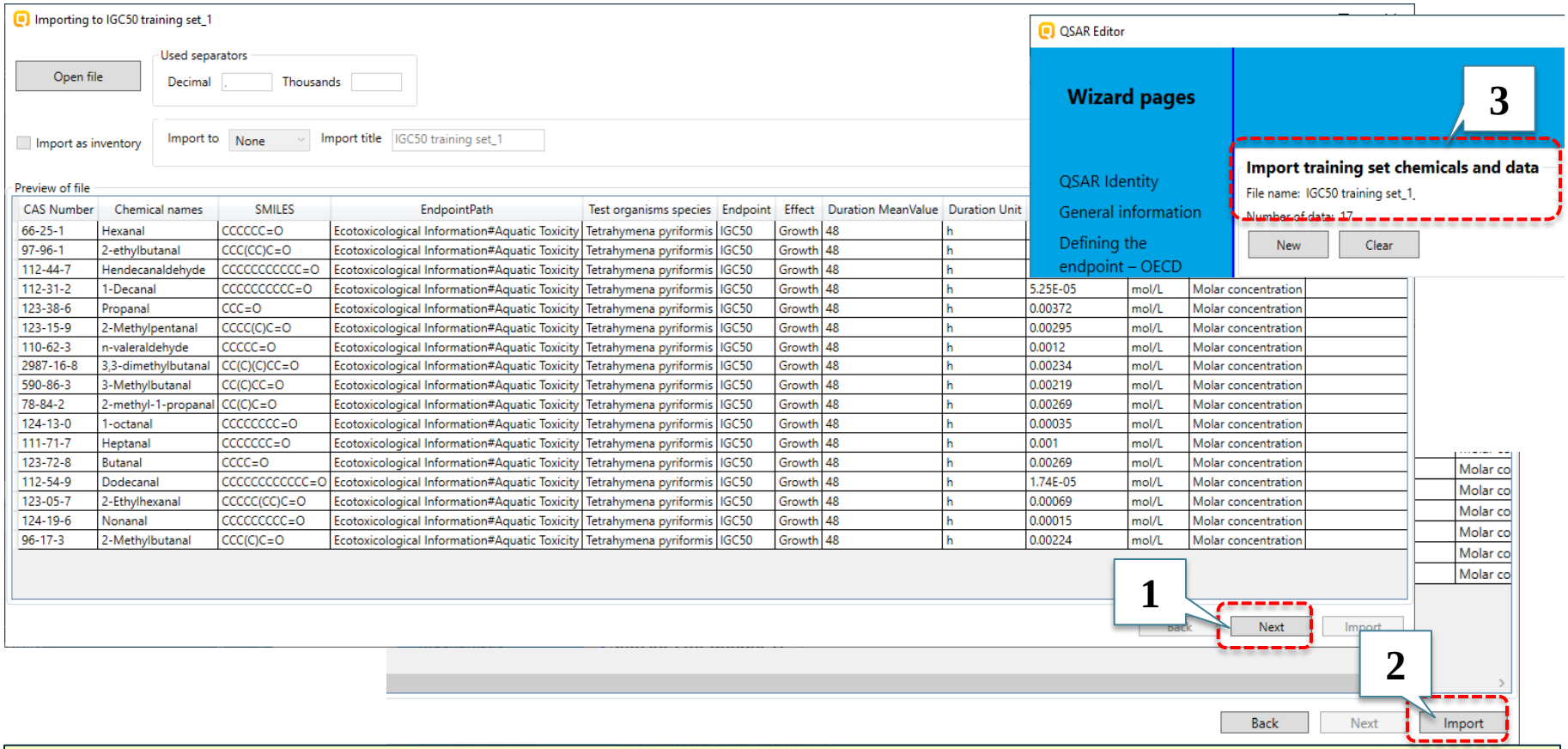
Building a new (Q)SAR
Import training set



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

Building a new (Q)SAR

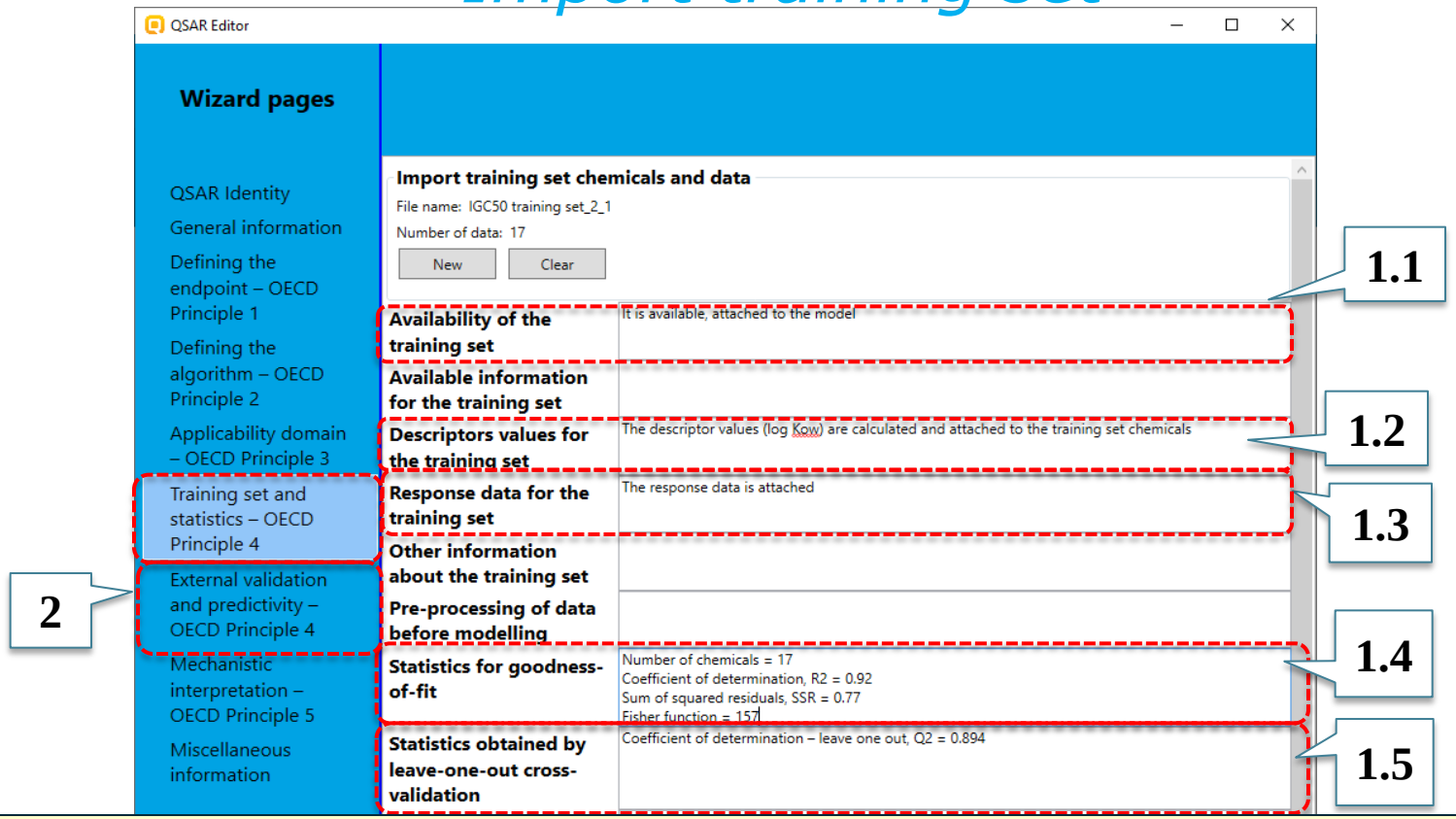
Import training set



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).

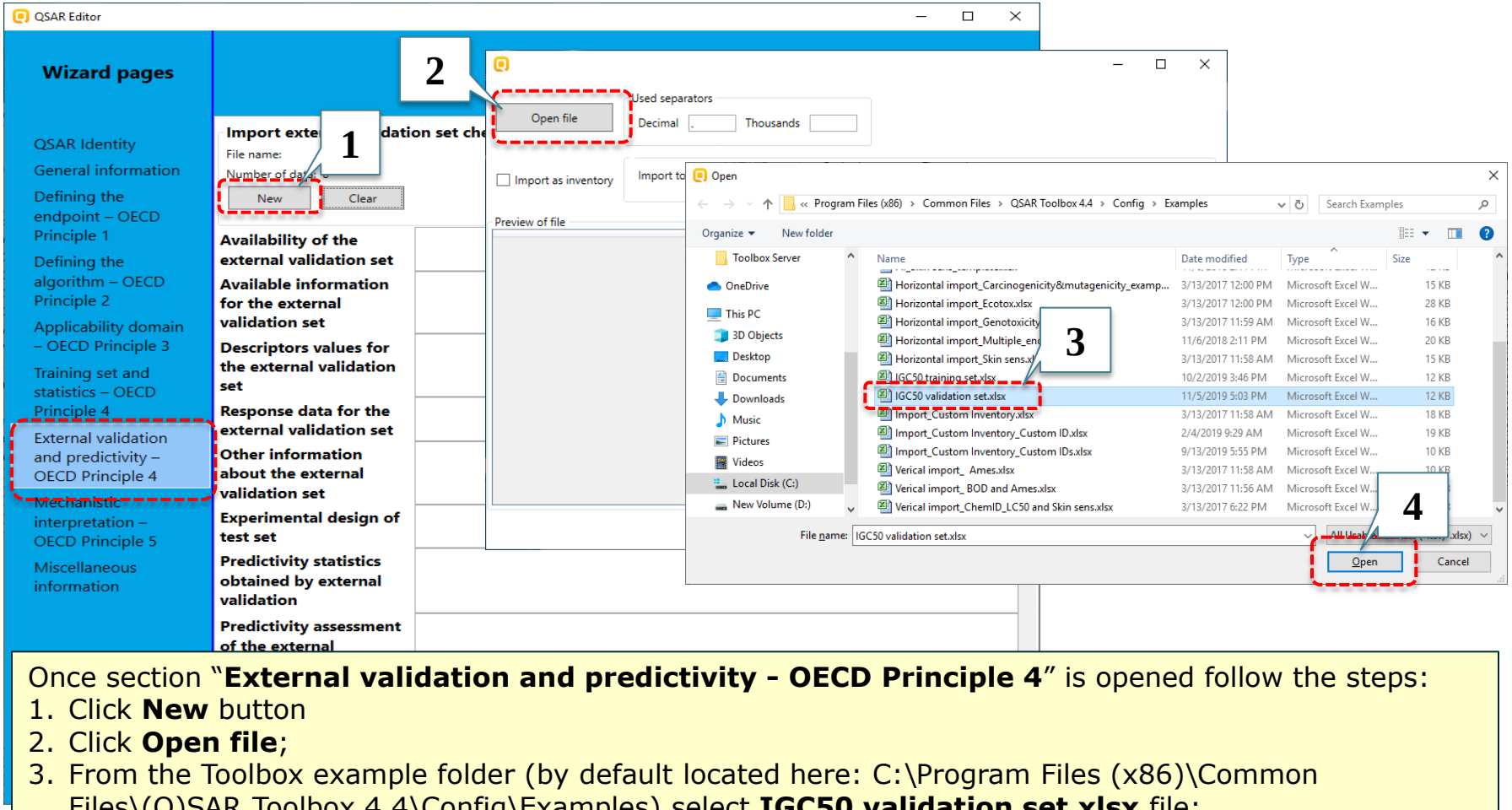
Building a new (Q)SAR

Import training set



1. Some additional information can be added to the empty fields; Please add the following text:
   - 1.1. Field "Availability of the training set" – add "It is available, attached to the model"
   - 1.2. Field "Descriptors values for the training set" - add "The descriptor values (log Kow) are calculated and attached to the training set"
   - 1.3. Field "Response data for the training set" – add "The response data is attached"
   - 1.4. Field "Statistics for goodness-of-fit" – add R2=0.92; SSR=0.77; Fisher function= 157
   - 1.5. Filed "Statistic obtained by leave-one-out cross validation" – add Q2 = 0.894
2. Move to the next section **"External validation and predictivity - OECD Principle 4"**.

Building a new (Q)SAR
Import validation set

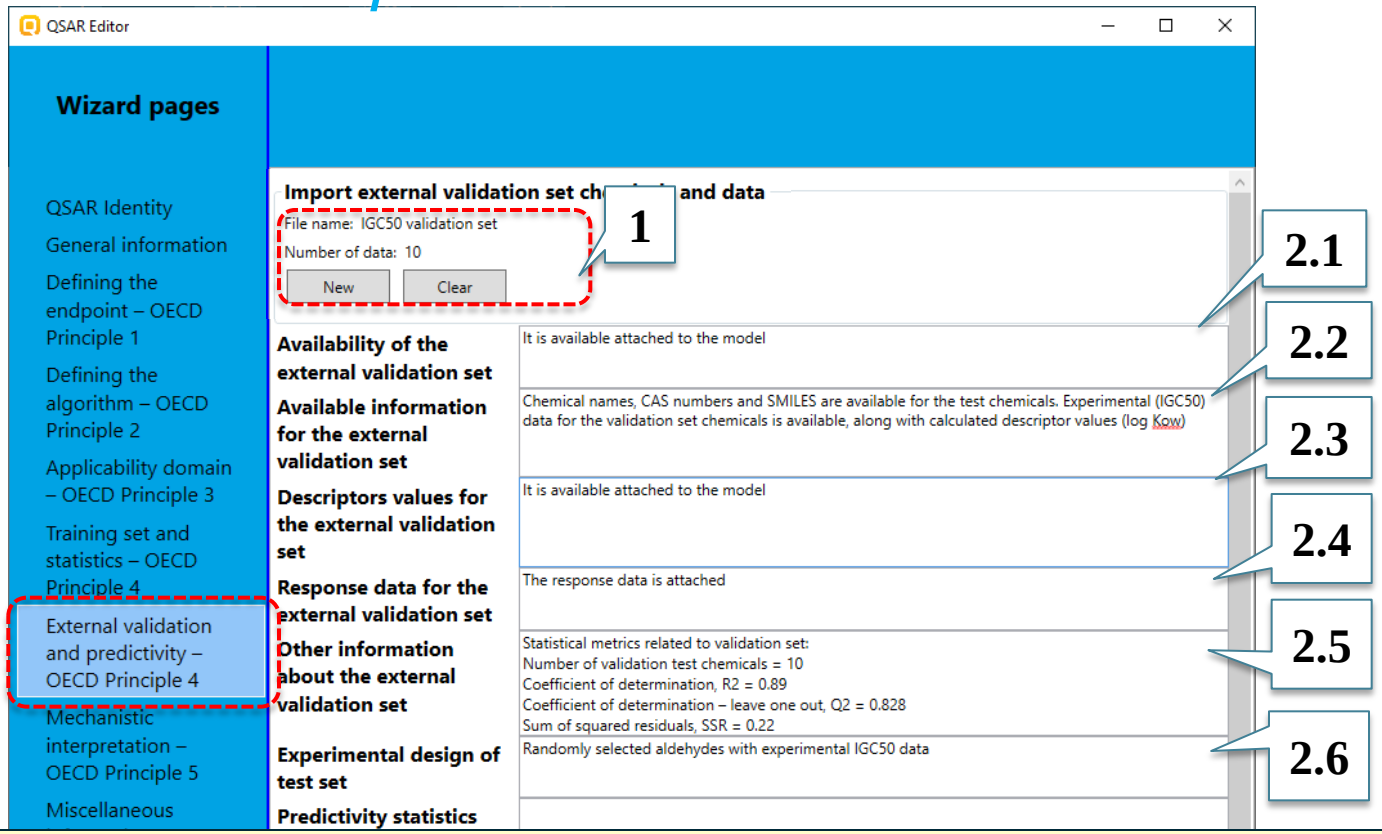


Once section "**External validation and predictivity - OECD Principle 4**" is opened follow the steps:

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

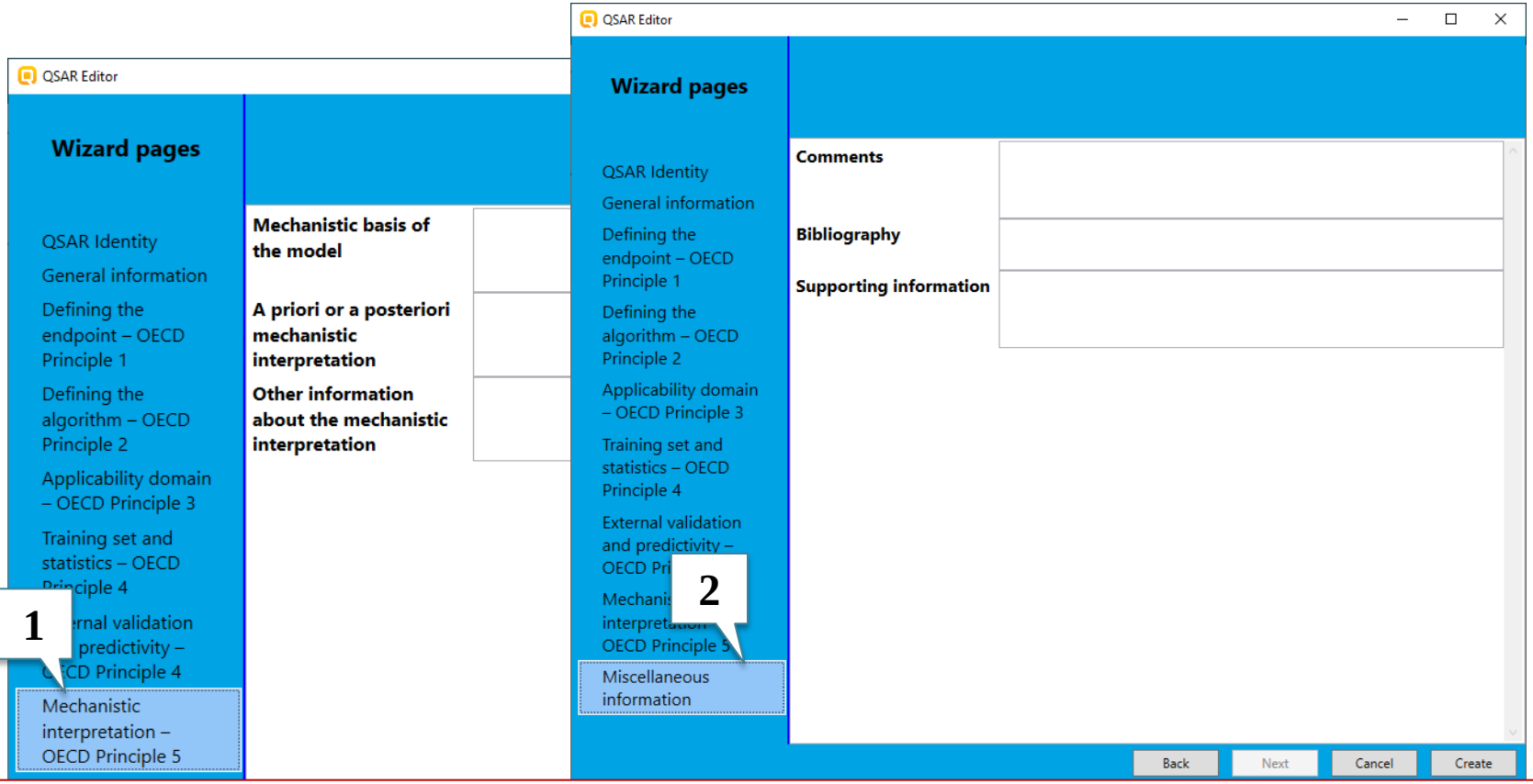
Building a new (Q)SAR

Import validation set



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. Filed "Other information about the external validation set" – additional text with statistical metrics are added;
    - 2.6. Filed "Experimental design of test set" – text is added.

Building a new (Q)SAR

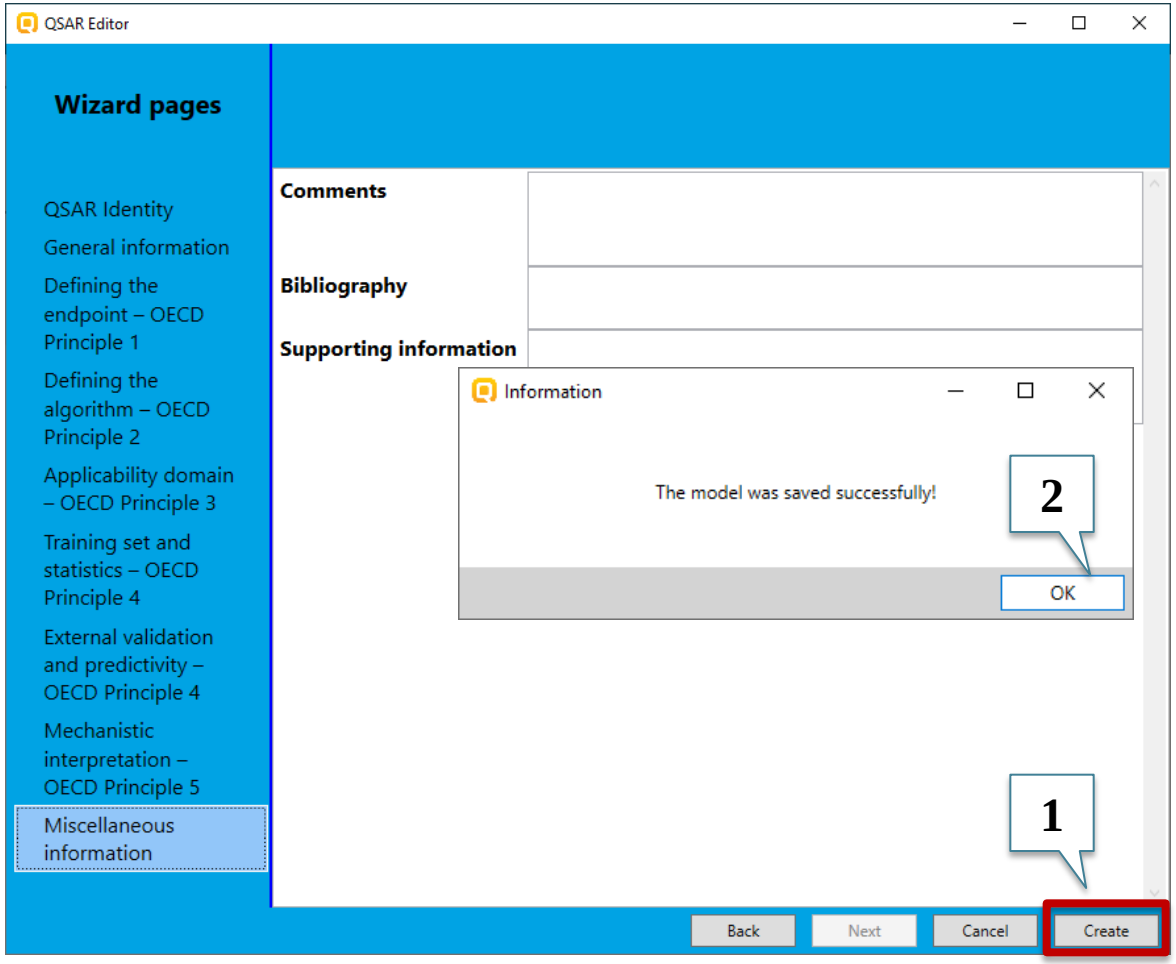


Additional information could be added to the next two sections:

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

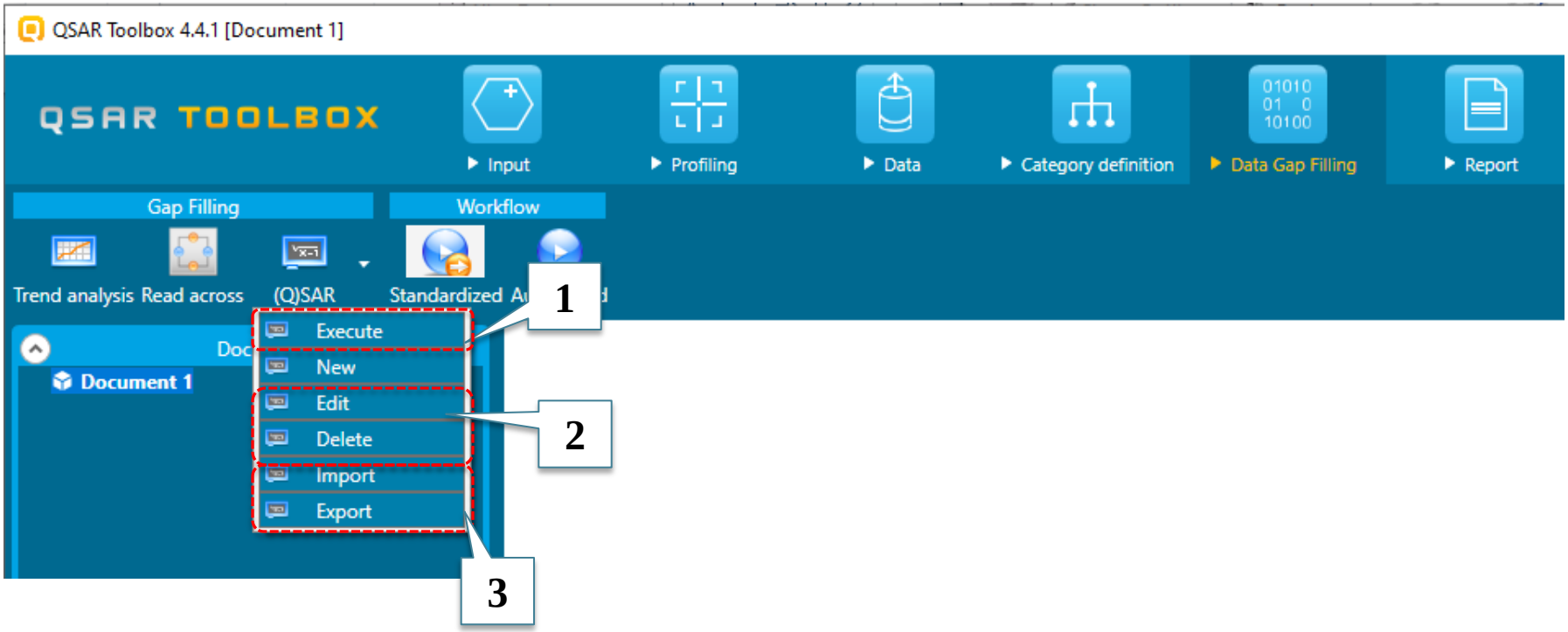
In our case these sections are left empty. As already mentioned these sections are not mandatory for correct working of the (Q)SAR model, but we recommend to fill them when you create your custom model in order for your model to follow the criteria of the five OECD Principles.

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**

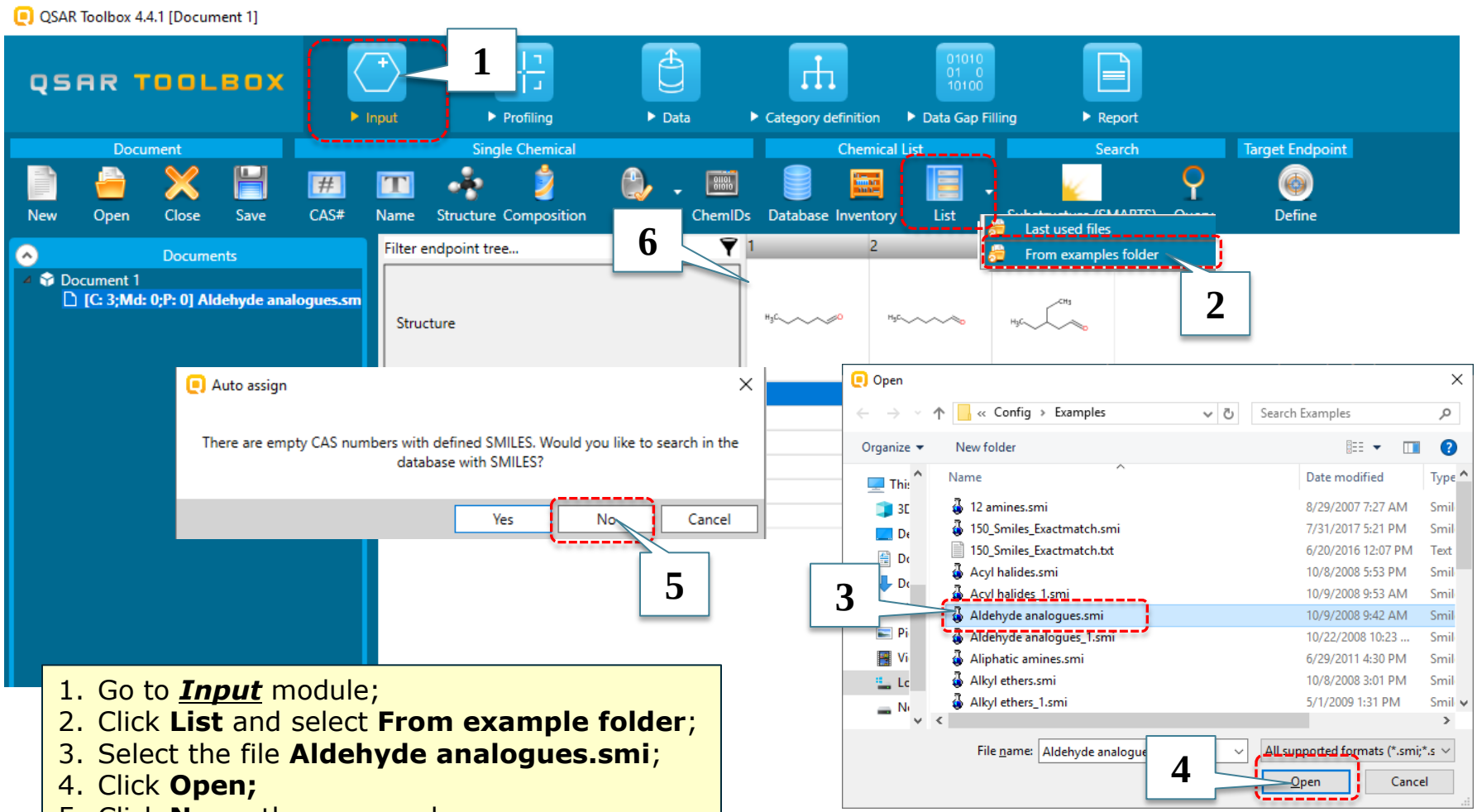
Application of the (Q)SAR



Already created (Q)SAR models could be:

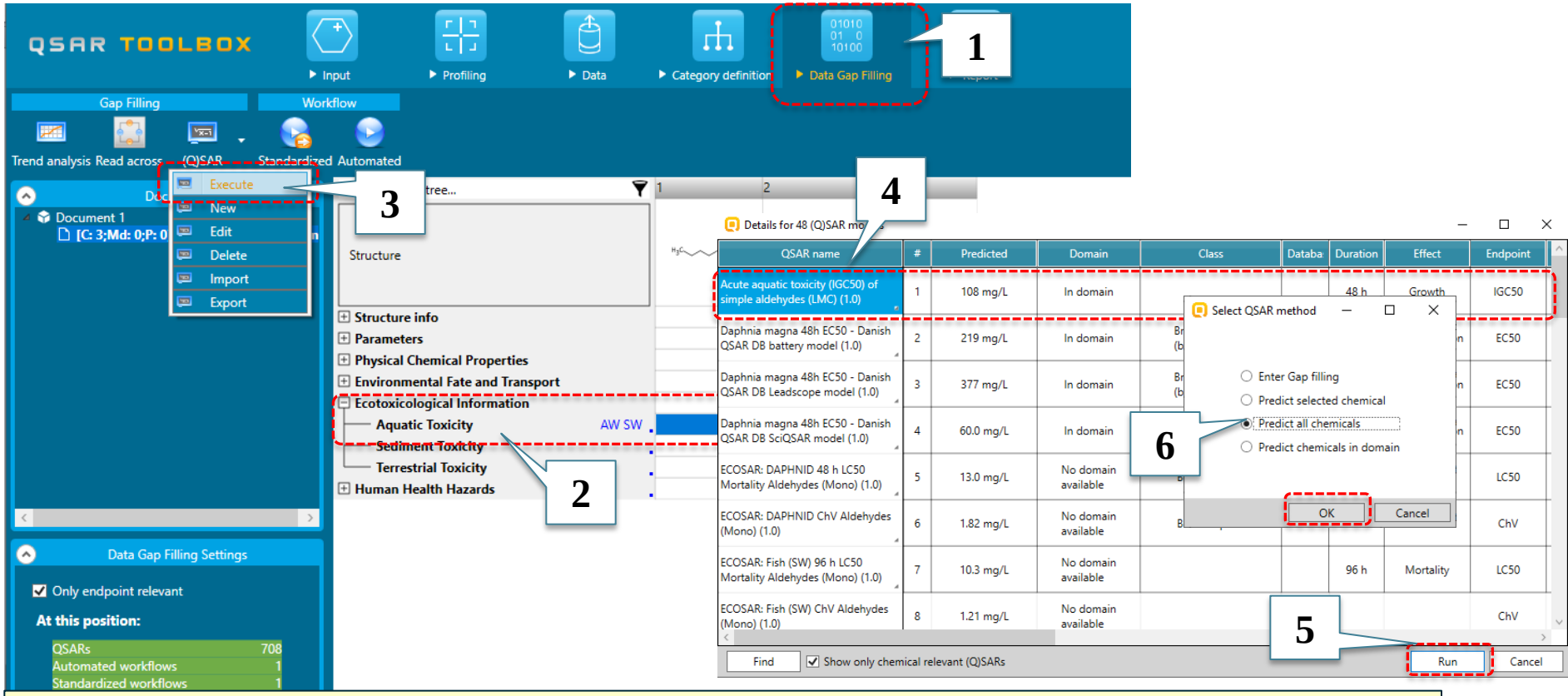
1) Applied to single or lists of chemicals (illustrated on the next slides);
2) Deleted or edited;
3) Imported or exported for use by other users (illustrated on the next slides).

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



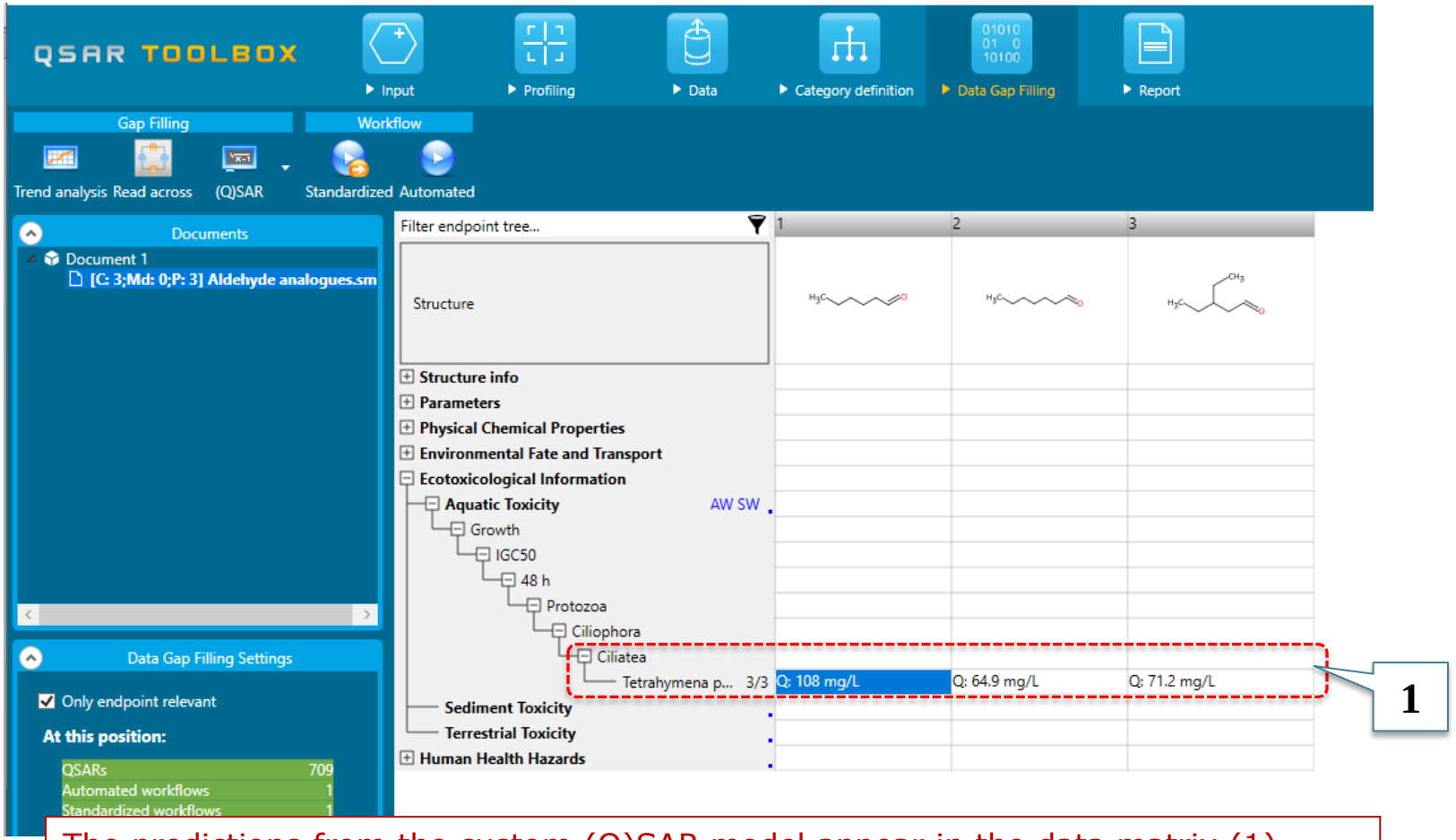
1. Go to ***Input*** module;
2. Click **List** and select **From example folder**;
3. Select the file **Aldehyde analogues.smi**;
4. Click **Open;**
5. Click **No** on the appeared message
6. The chemicals appear in the data matrix.

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



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**

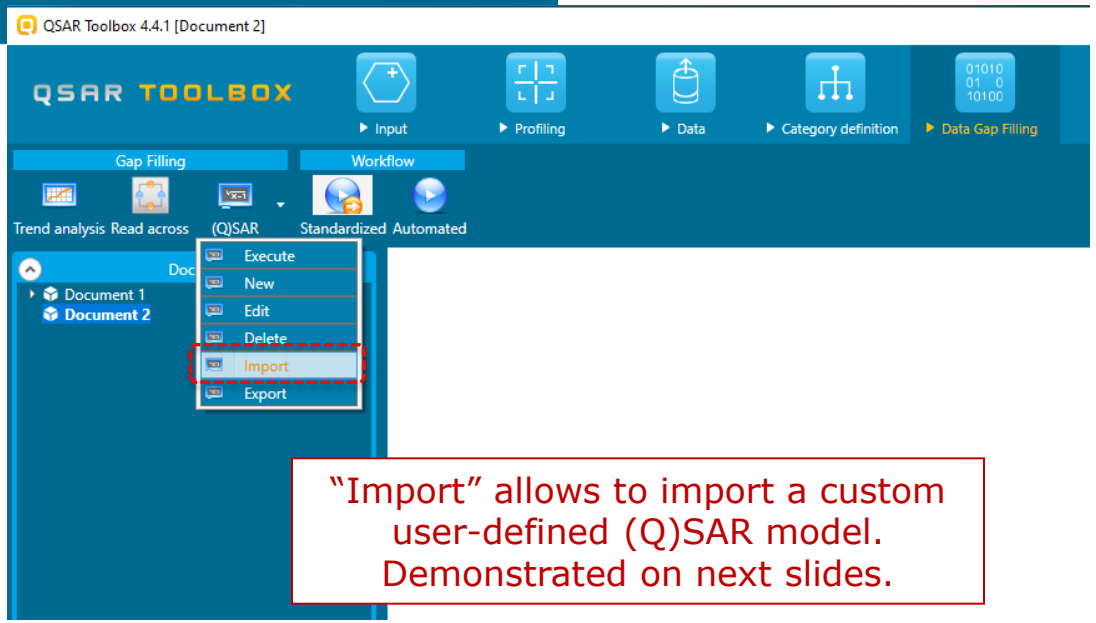
Application of the (Q)SAR to a list of chemicals



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

Import/Export of the custom (Q)SAR

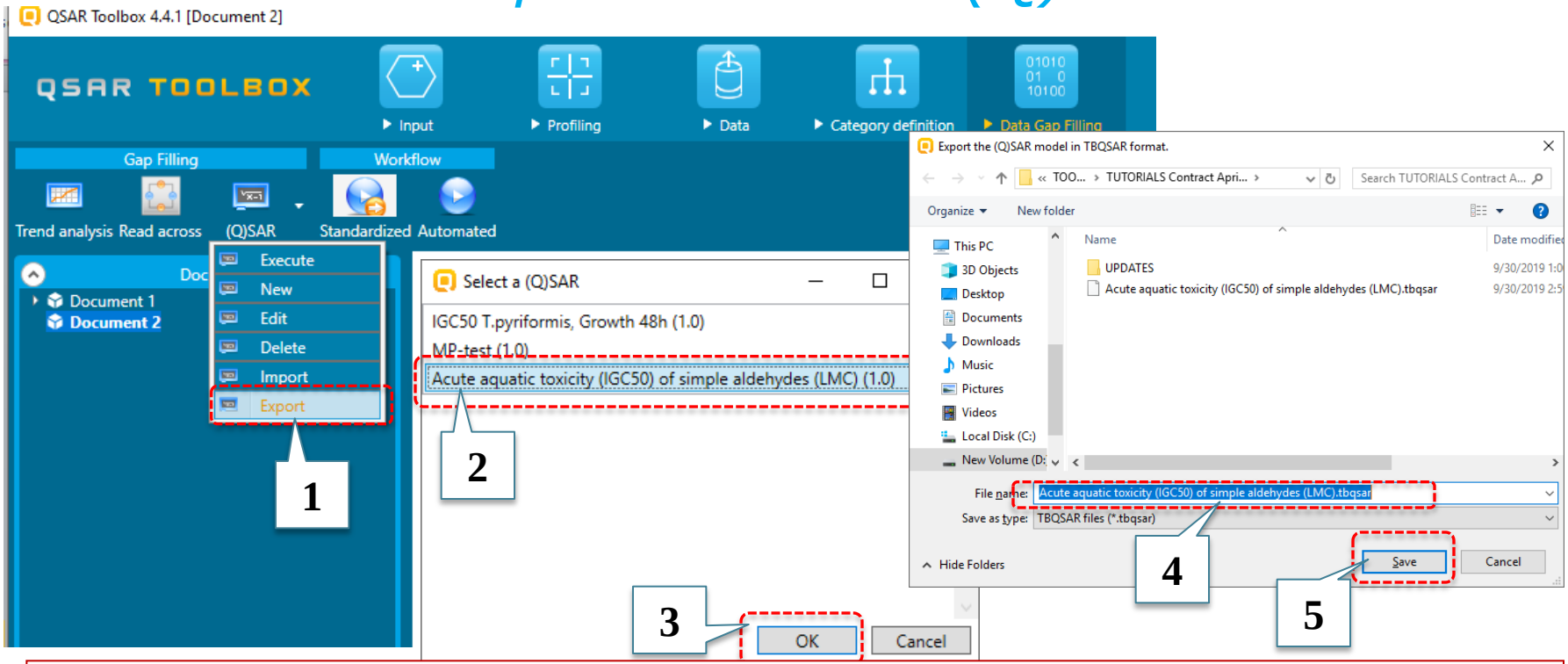
"Export" allows to export a user-defined (Q)SAR model. Demonstrated on next slides.



"Import" allows to import a custom user-defined (Q)SAR model. Demonstrated on next slides.

Import/Export of the custom (Q)SAR

Export a custom (Q)SAR

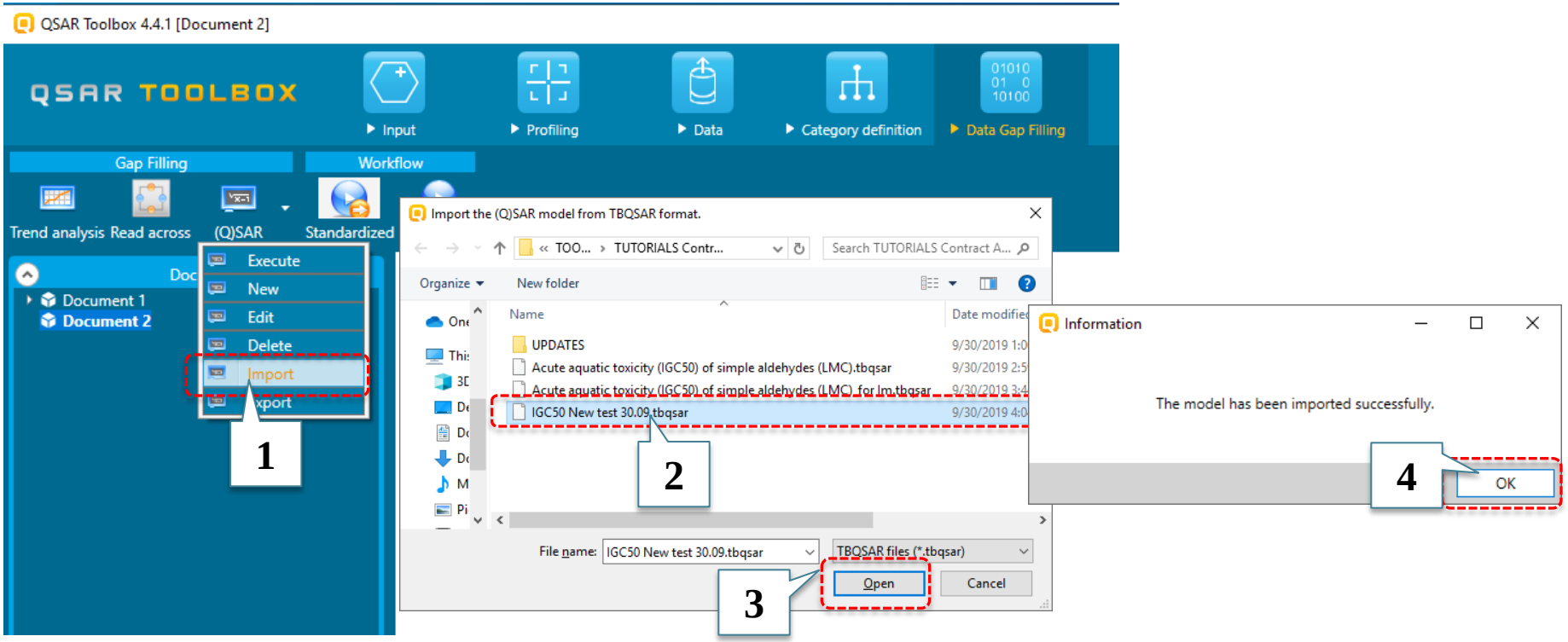


The "Export" operation packs all files associated with (Q)SAR in one file with the extension "*.tb(Q)SAR". The file can be disseminated to other users and imported by them via Import (see next slides).

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>

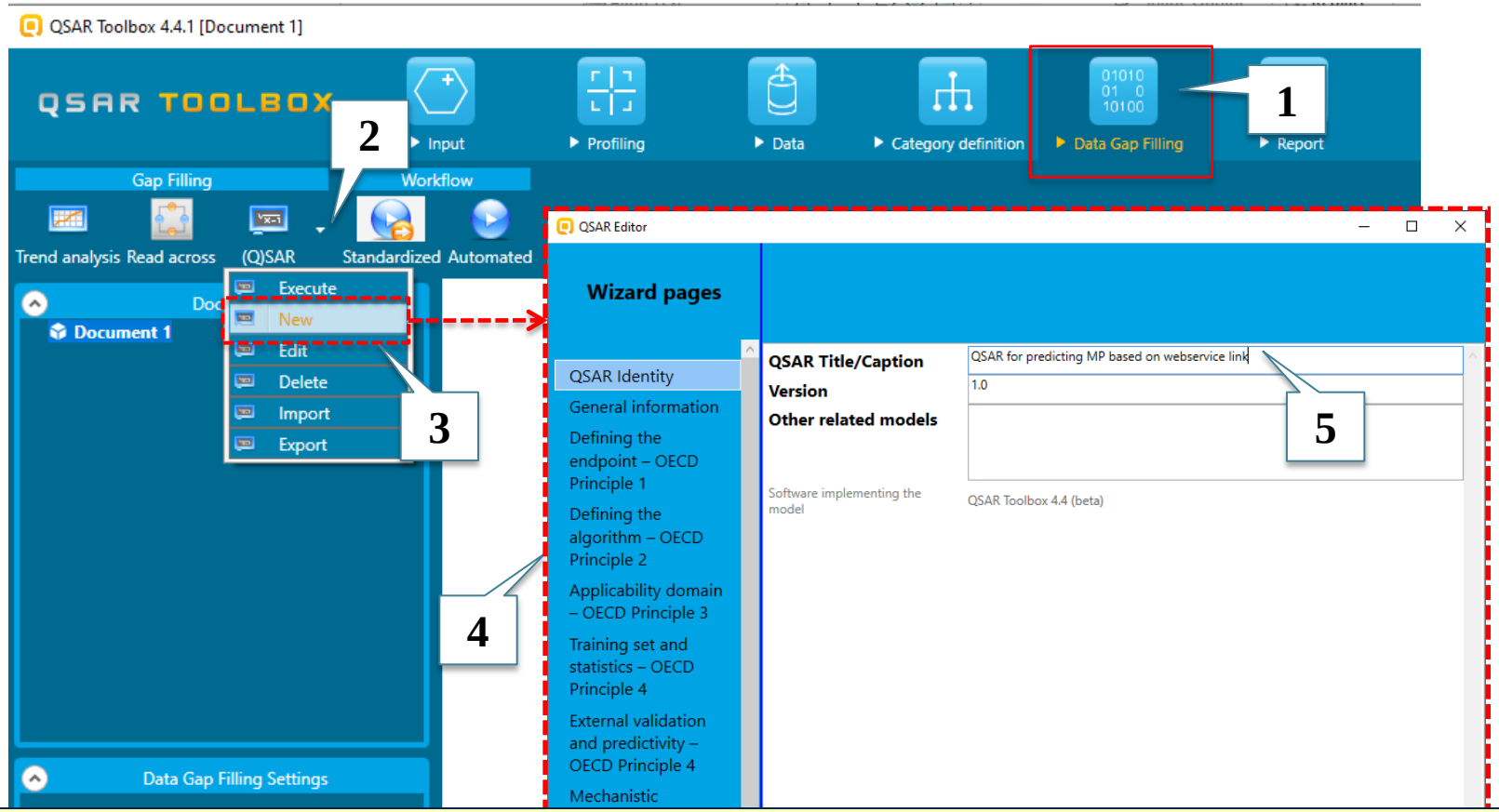
Reference: http://qsardb.org/repository/handle/10967/104

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



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:

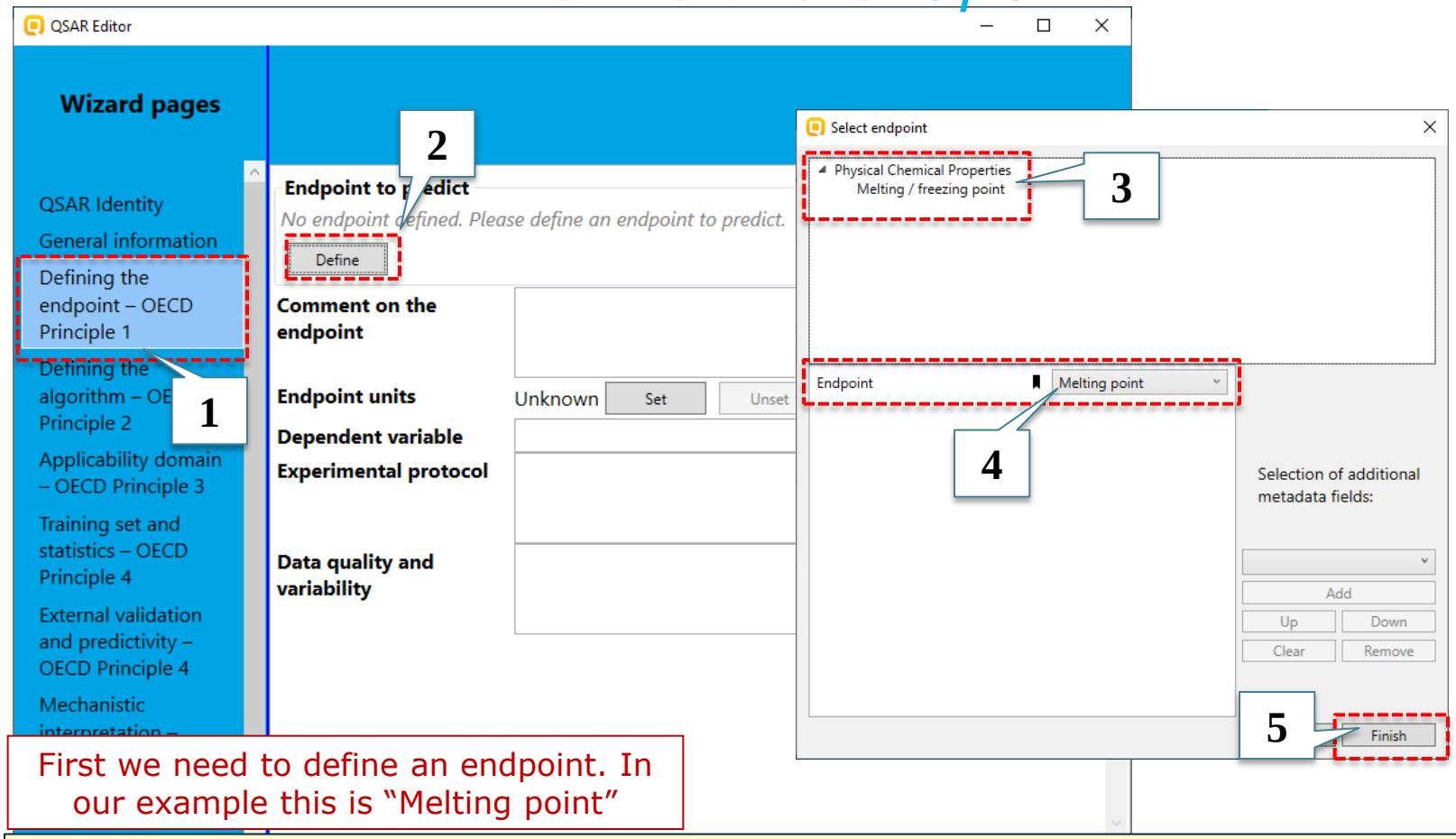
- Important sections mandatory for correct work of the (Q)SAR:
  - the (Q)SAR title **(already defined);**
  - the endpoint **(in our case MP)** and its unit **(in our case degree Celsius (°C))**
  - The web service link – Requirements for the constructing the link itself are available in the Editor 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

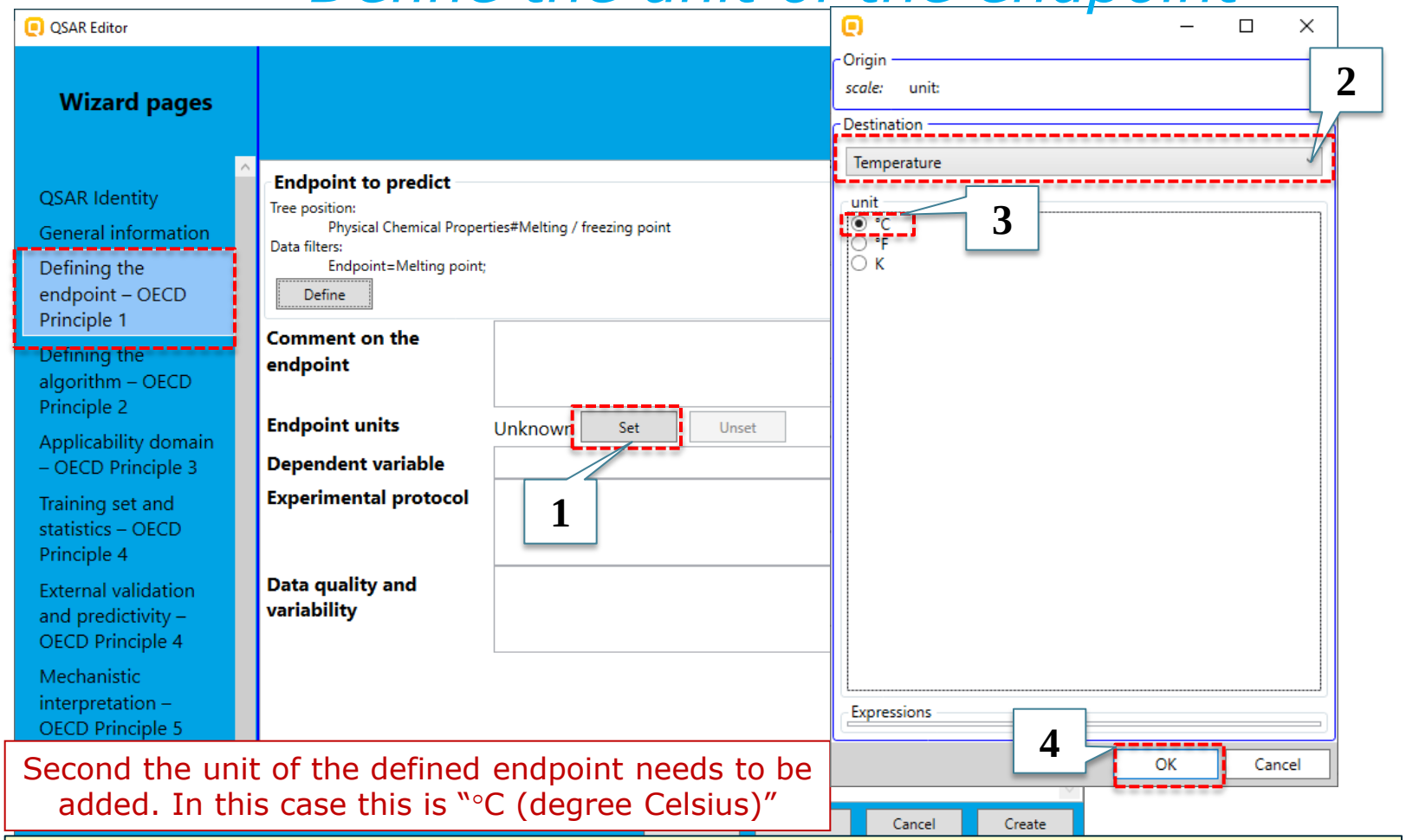


First we need to define an endpoint. In our example this is "Melting point"

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

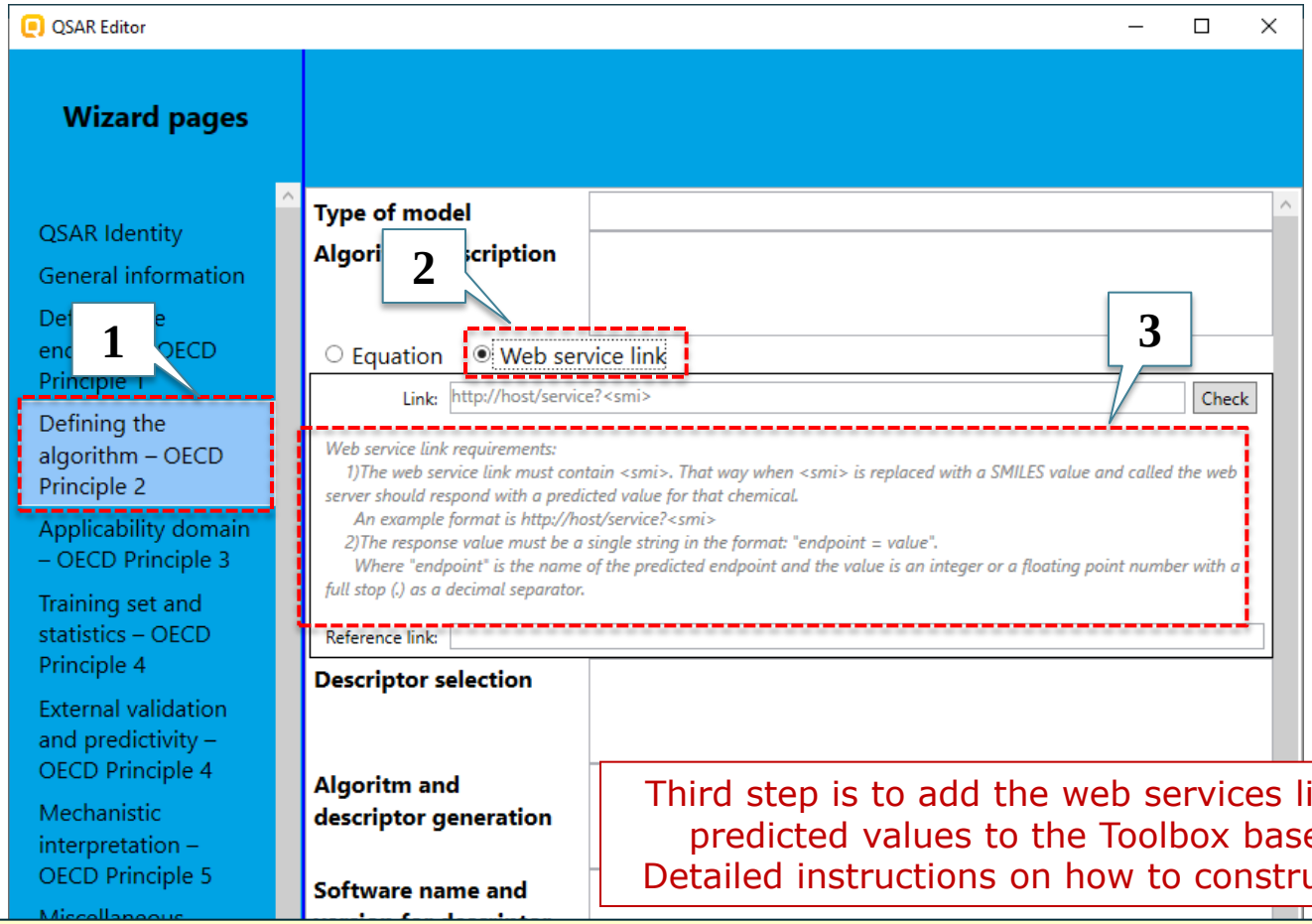


Second the unit of the defined endpoint needs to be added. In this case this is "°C (degree Celsius)"

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

Building a new (Q)SAR

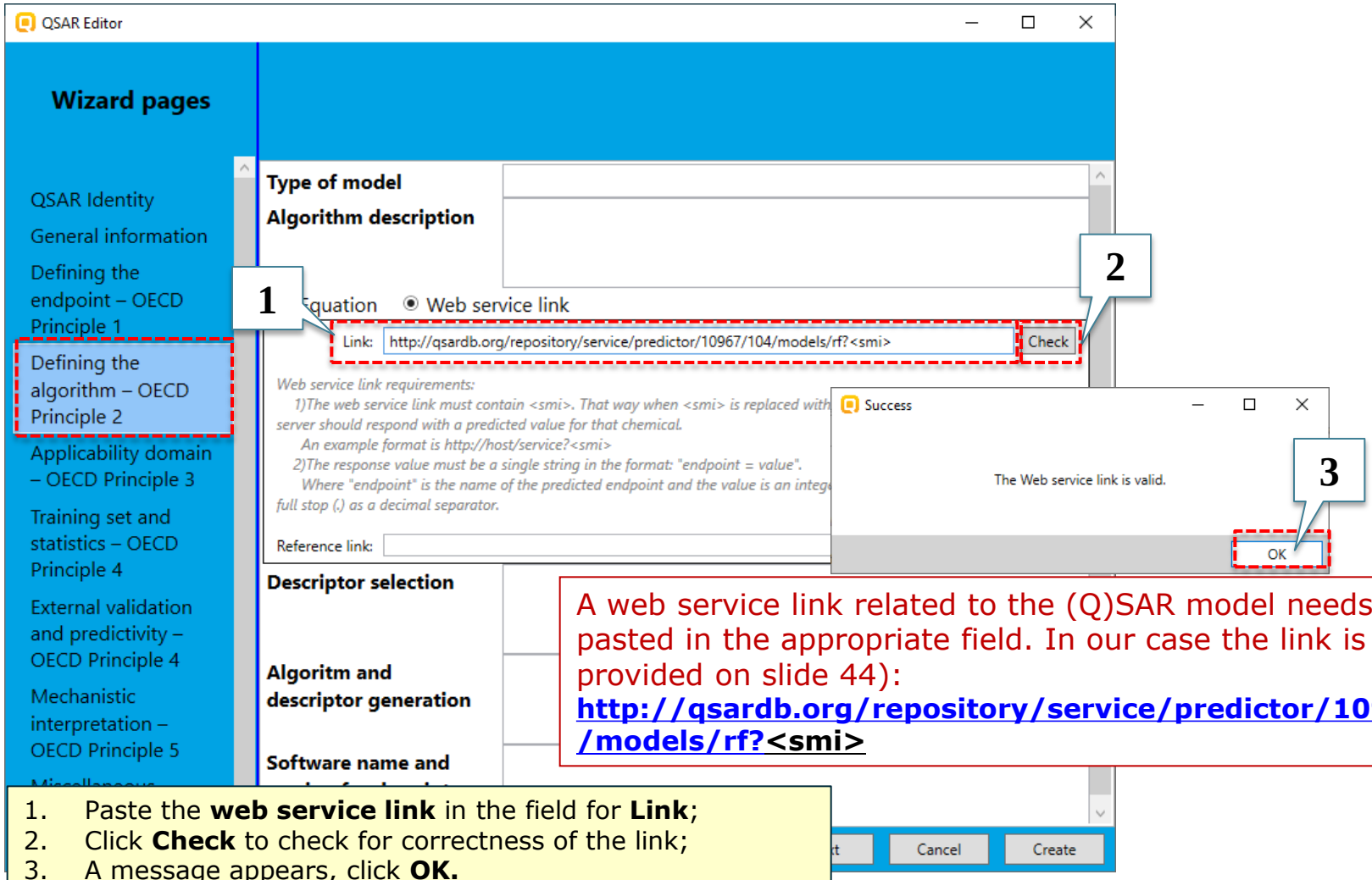
Define the web services link



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.

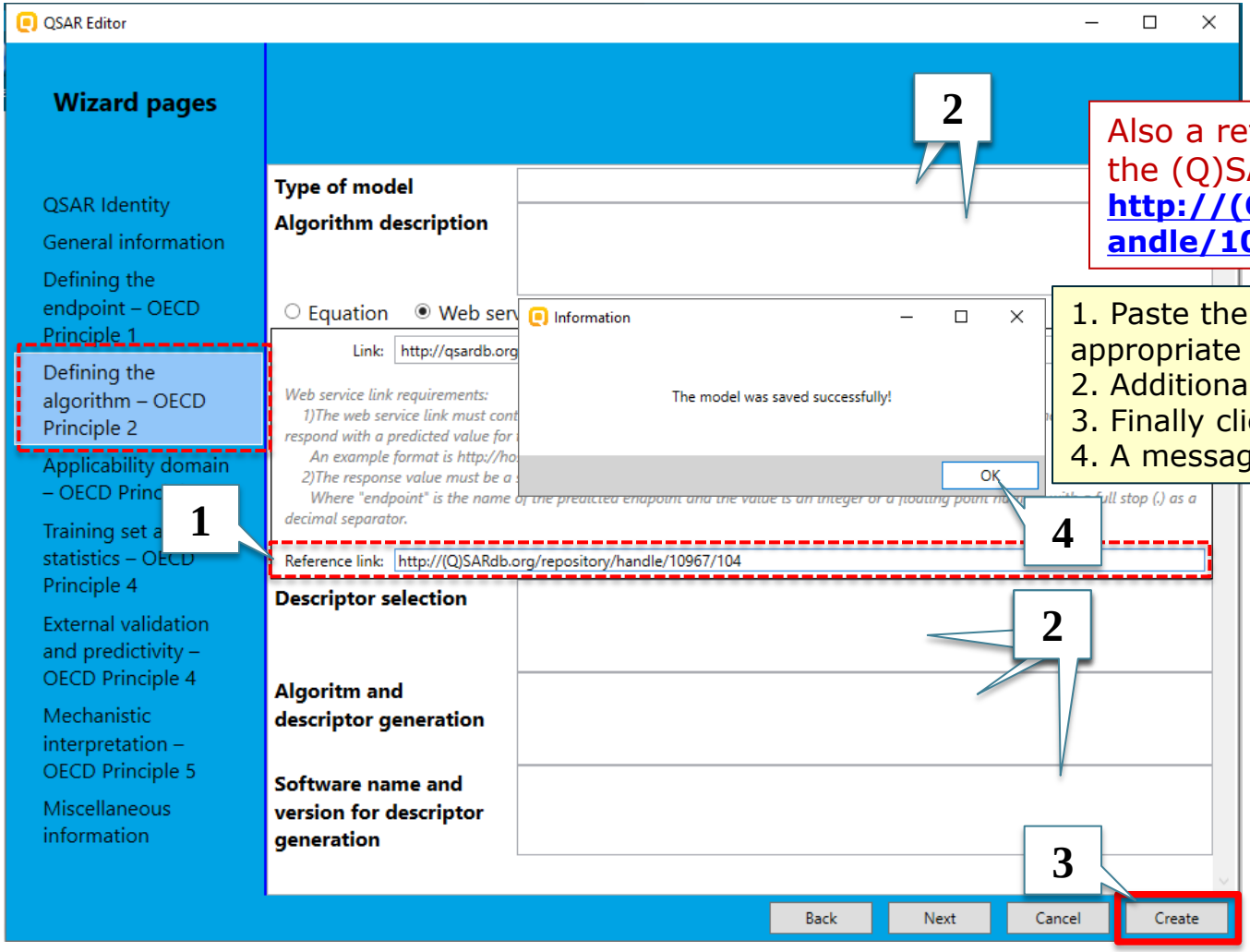
Building a new (Q)SAR
Define the web services link



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.**

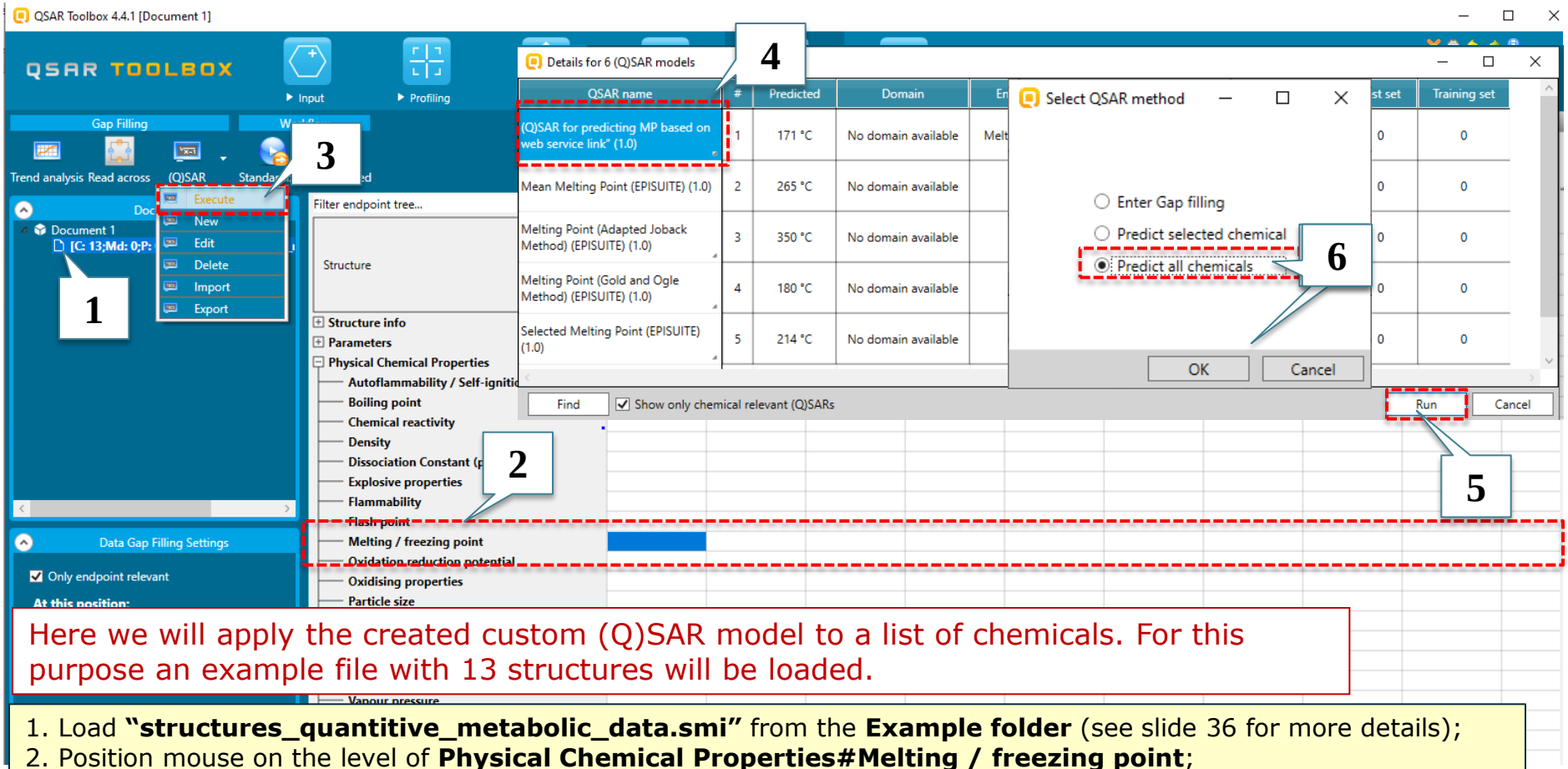
Building a new (Q)SAR
Define the reference link



Also a reference link associated with the (Q)SAR could be added: **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.**

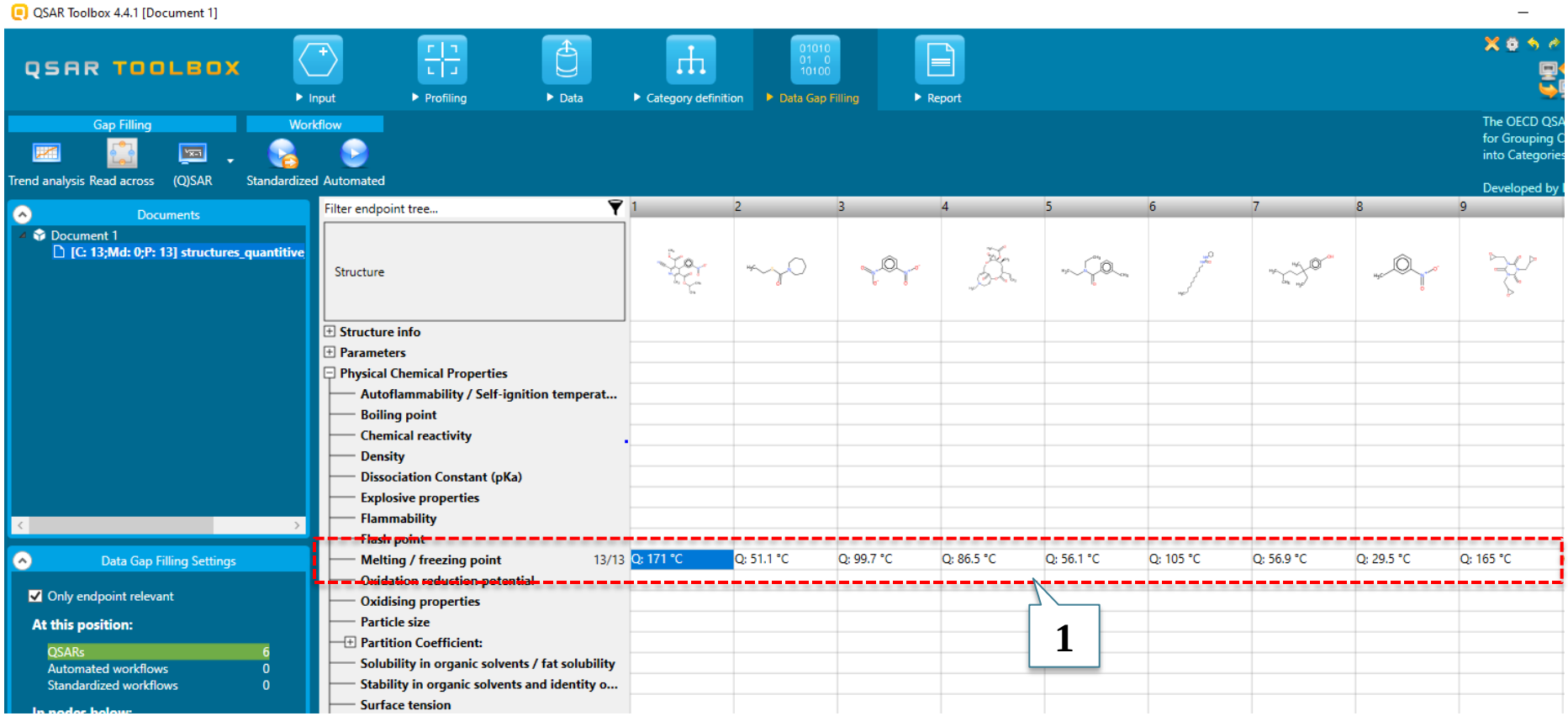
Application of the (Q)SAR to a list of chemicals



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_quantitive_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



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.