QSAR TOOLBOX

The OECD QSAR Toolbox for Grouping Chemicals into Categories

OECD (Q)SAR Toolbox v.4.4.1

Example illustrating endpoint vs. endpoint correlation for apical endpoints

Outlook

- **Background**
- Objectives
- The exercise
- Workflow

Background

This presentation is designed to introduce the user to:

- Illustration of different types of endpoint vs. endpoint correlations using:
  - LLNA and GPMT skin sensitization data;
  - DPRA and LLNA skin sensitization data;
  - Skin sensitization and Ames mutagenicity data.

Outlook

- Background
- **Objectives**
- The exercise
- Workflow

Objectives

This presentation demonstrates a number of functionalities of the Toolbox:

- Illustration of endpoint vs. endpoint correlations using different types of endpoint data

Outlook

- Background
- Objectives
- **The exercise**
- Workflow

The exercise

- Illustration of different endpoint data correlations:
  - LLNA vs. GPMT skin sensitization data
  - DPRA (reactivity) vs. LLNA (skin sensitization) data
  - LLNA (skin sensitization) vs. Ames mutagenicity data

Outlook

- Background
- Objectives
- The exercise
- **Workflow**

Workflow

- **The Toolbox has six modules which are typically used in a workflow:**
  - Chemical Input
  - Profiling
  - Data
  - Category Definition
  - Filling Data Gaps
  - Report
- **In this example we will use the modules in a different order, tailored to the aims of the example.**

Outlook

- Background
- Objectives
- The exercise
- **Workflow**
  - **Correlation of data - background**

Correlation of endpoint data
Background

- This functionality introduces the user to the opportunity to analyze correlations between selected gap filling endpoint (endpoint used for prediction) and other endpoint data.
- It is applicable for correlation analysis of data presented in ordinary, interval or ratio scale.
- If correlated data are measured in interval or ratio scale they are transformed in ordinary scale and the strength of the correlation is estimated by Spearman correlation coefficient.
- Basically, this functionality provides a correlation between target endpoint (this is the initial endpoint selected by the user) displayed on ordinate axis (Y-axis) and other endpoint data displayed on abscissa (X-axis).

Correlation of endpoint data
Spearman coefficient factor

- Spearman's rank correlation coefficient is a nonparametric rank statistic proposed by Charles Spearman as a measure of the strength of an association between two variables. It assesses how well the relationship between two variables can be described using a monotonic function.
- Spearman correlation coefficient could be used for exploring the correlation between:
  - two ranked variables
  - one measurement variable and one ranked variable (in this case, the measurement variable need to be to converted to ranks)
- Spearman correlation varies from -1 to +1 and the interpretation of the coefficient factor is provided below:
  - 0.00 – 0.19 – very weak correlation
  - 0.20 – 0.39 – weak correlation
  - 0.40 – 0.59 – moderate correlation
  - 0.60 – 0.79 – strong correlation
  - 0.80 – 1.0 – very strong

Outlook

- Background
- Objectives
- The exercise
- **Workflow**
  - Correlation of data – background
  - **Types endpoint correlations**

Types of endpoint correlations

Types of endpoint correlations are as follows:

- Continuous vs. continuous*
- Categorical vs. categorical:
  - ✓Categorical vs. categorical
  - ✓Categorized continuous vs. categorical
  - ✓Categorized continuous vs. categorized continuous*

*Both type correlation is not illustrated in this presentations. They are presented in "Tutorial_4_TB 4.4_Illustrating endpoint vs. endpoint correlation using ToxCast data"

Outlook

- Background
- Objectives
- The exercise
- **Workflow**
  - Correlation of data – background
  - **Types endpoint correlations**
    - Categorical vs. categorical

Types of endpoint correlations
Categorical vs. categorical

- The aim of this type of correlation is to illustrate how categorical types of data correlate with each other.
- Categorical type data is the statistical data type consisting of categorical variables or of data that has been converted into that form. Such data is binary Ames data (dichotomic type): positive, negative or polytomic type data such as GPMT data: strong, weak and negative.
- Two examples illustrating this type correlation will be demonstrated:
  - Example 1: Correlation of two types skin sensitization data
    - LLNA (Positive, Negative) vs. GPMT (Weakly positive, Strongly positive, Negative)
  - Example 2: Correlation of skin sensitization and Ames mutagenicity data
    - LLNA (Negative, Weakly positive, Strongly positive) vs. AMES (Positive, Equivocal, Negative)
- Step by step workflow is presented on next few slides. Summary of the workflow steps are provided below:
  - *Query Tool and select FSQ file(step 1)*
  - *Gather experimental data (step 2)*
  - *Enter Gap filling (step 3)*
  - *Perform correlation between endpoints (step 4).*

Types of endpoint correlations

Categorical vs. categorical

Example 1: Correlation between LLNA and GPMT data

According to OECD Guideline 406 for testing of chemicals for skin sensitization, the LLNA test can be used as a first stage in the assessment of skin sensitization potential. If a positive result is seen, a test substance may be designated as a potential sensitizer, and it may not be necessary to conduct a further guinea pig test.[1]

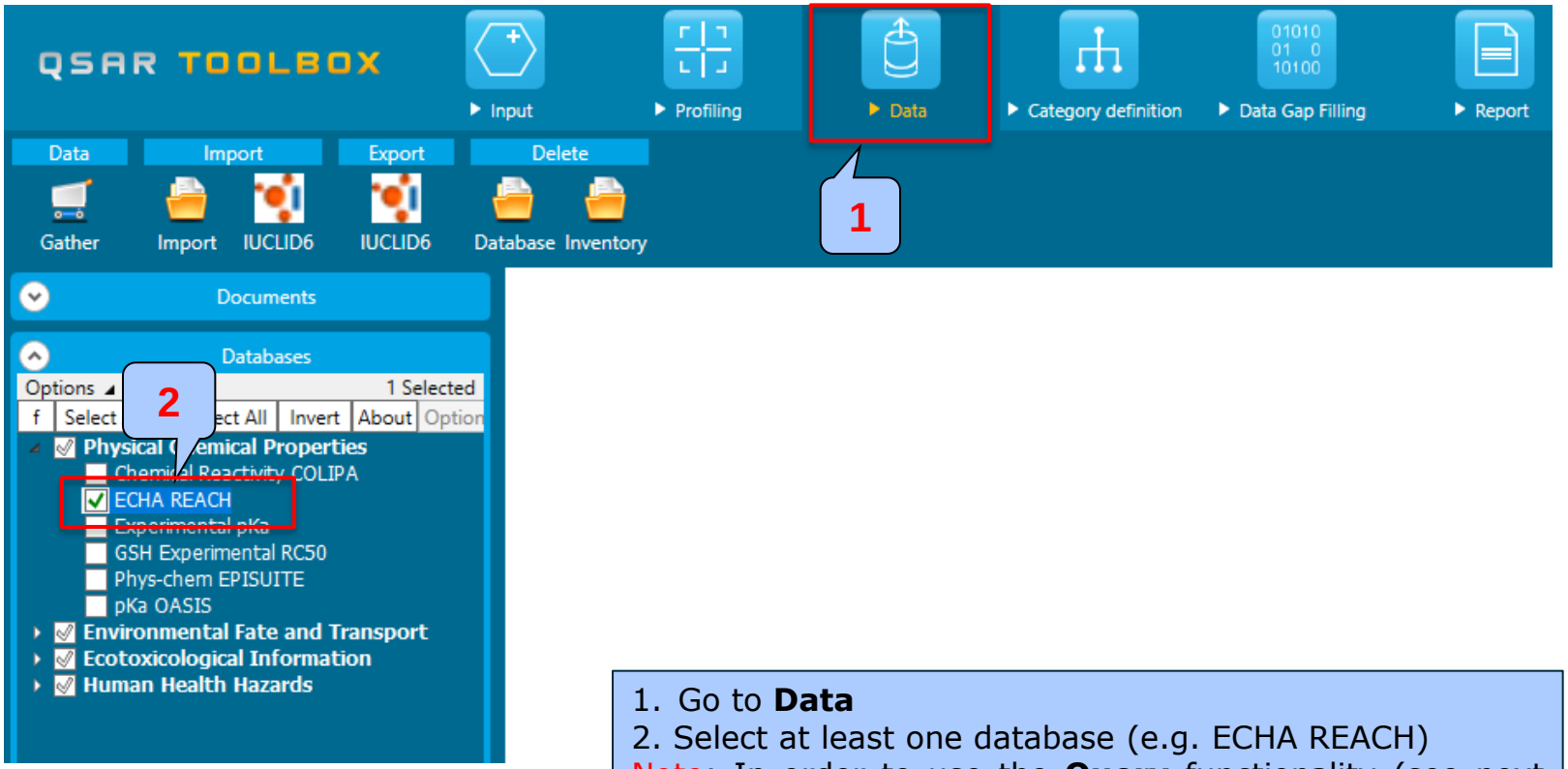
Based on that Guideline the aim of the illustrated correlation is to show how the capacity of LLNA test is compatible to that of the GPMT assay.

[1] https://www.oecd-ilibrary.org/docserver/9789264070660-en.pdf?expires=1573465805&id=id&accname=guest&checksum=D7DE1063A1EA331FCA23AC81BE1FDAC1

Types of endpoint correlations

Categorical vs. categorical

Example 1: Correlation between LLNA and GPMT data



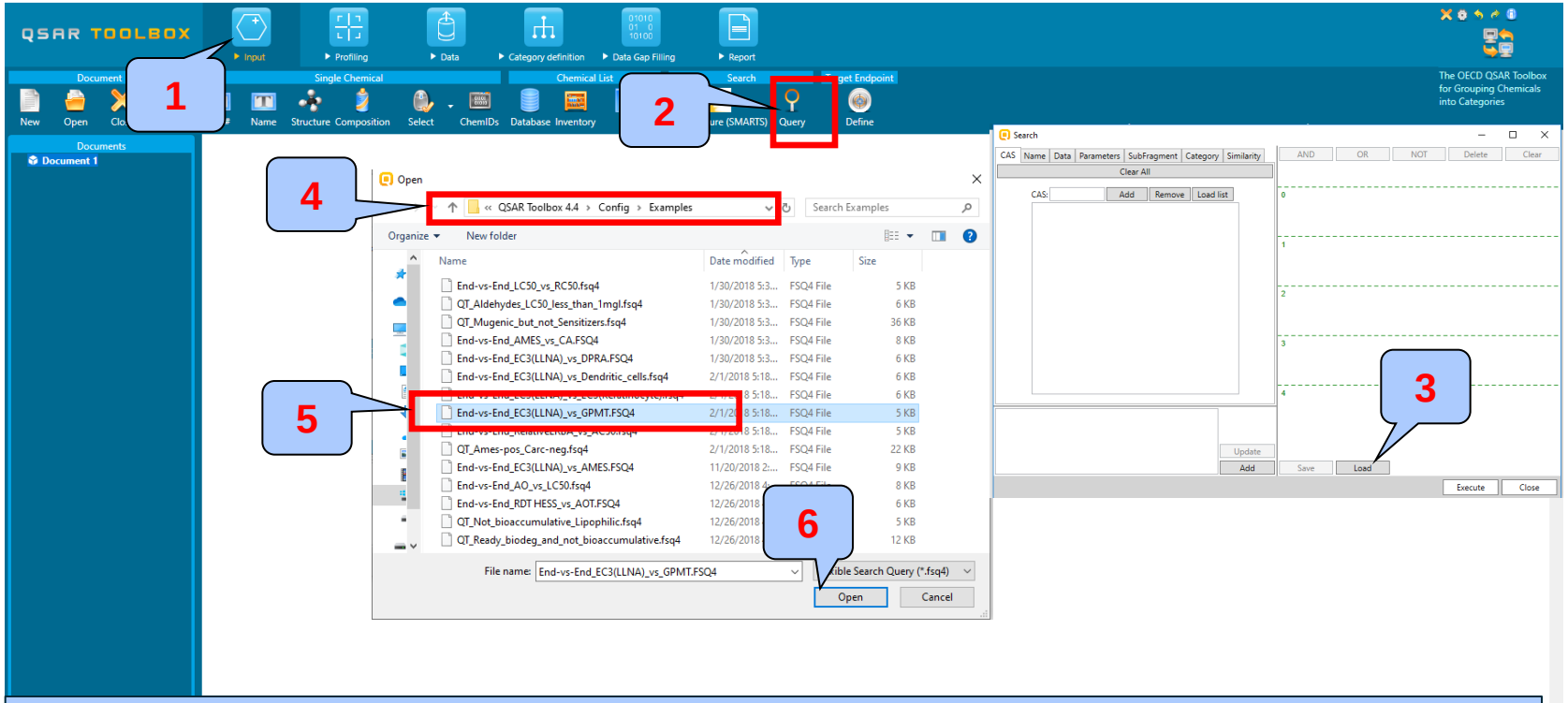
1. Go to **Data**
2. Select at least one database (e.g. ECHA REACH)

Note: In order to use the **Query** functionality (see next slide) at least one database must be selected.

Types endpoint correlations

Categorical vs. categorical

Example 1: Correlation between LLNA and GPMT data

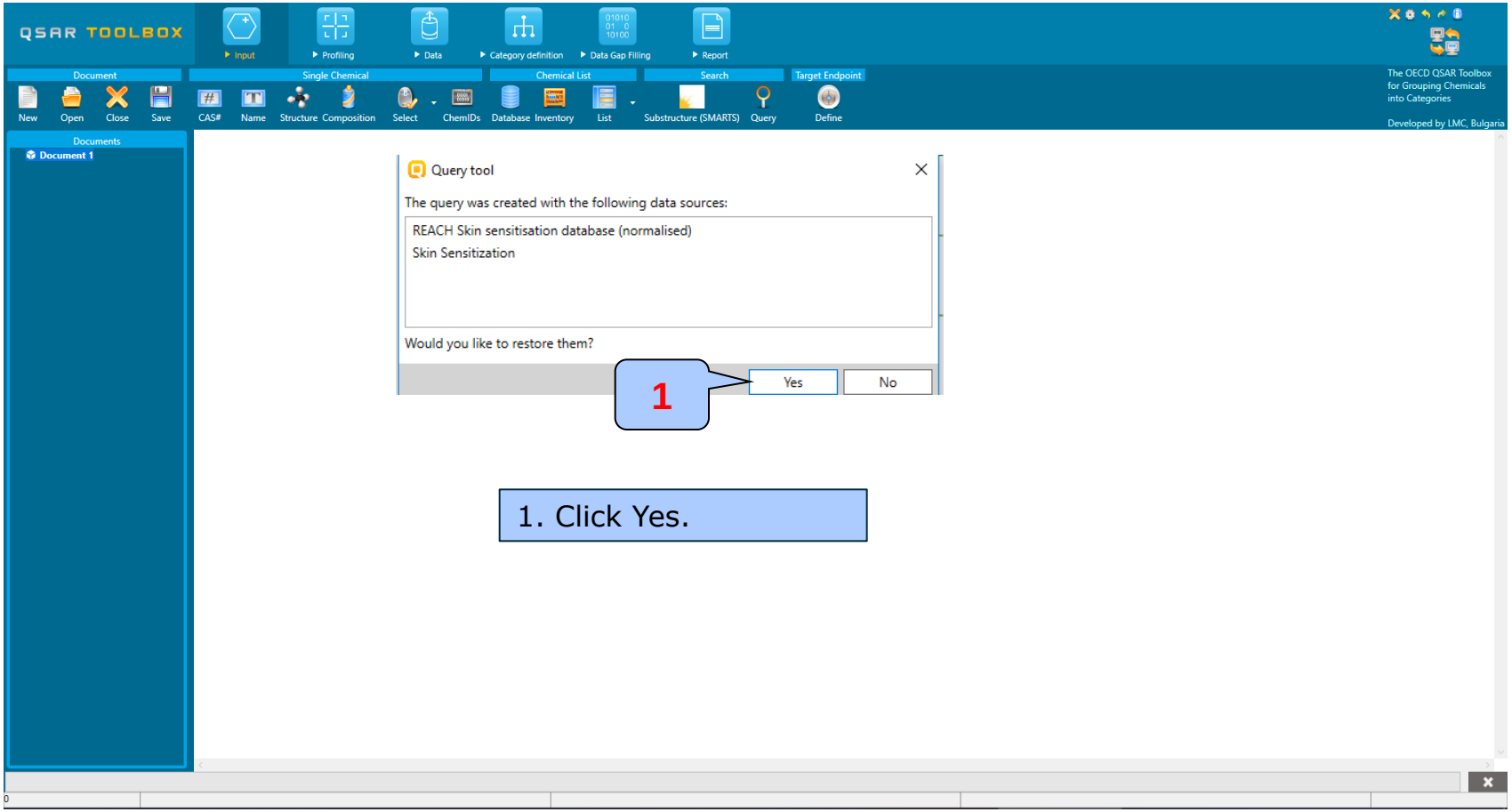


1. Go to "**Input**"; 2. Click **Query**; 3. Click **"Load"**
4. Select Example directory from TB *C:\Program Files (x86)\Common Files\(Q)SAR Toolbox 4.4\Config\Examples*;
5. End-vs-End_EC3(LLNA)_vs_GPMT.FSQ4; 6. Click **Open.**

Types of endpoint correlations

Categorical vs. categorical

Example 1: Correlation between LLNA and GPMT data

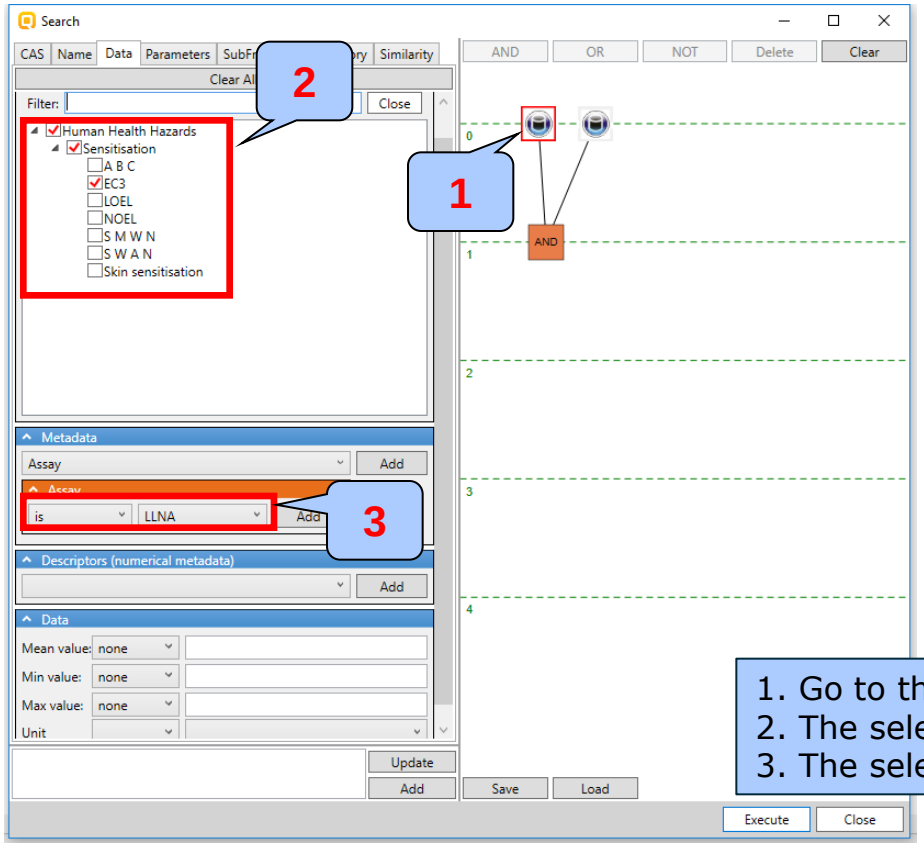


Types of endpoint correlations

Categorical vs. categorical

Gather experimental data – step 2

Example 1: Correlation between LLNA and GPMT data



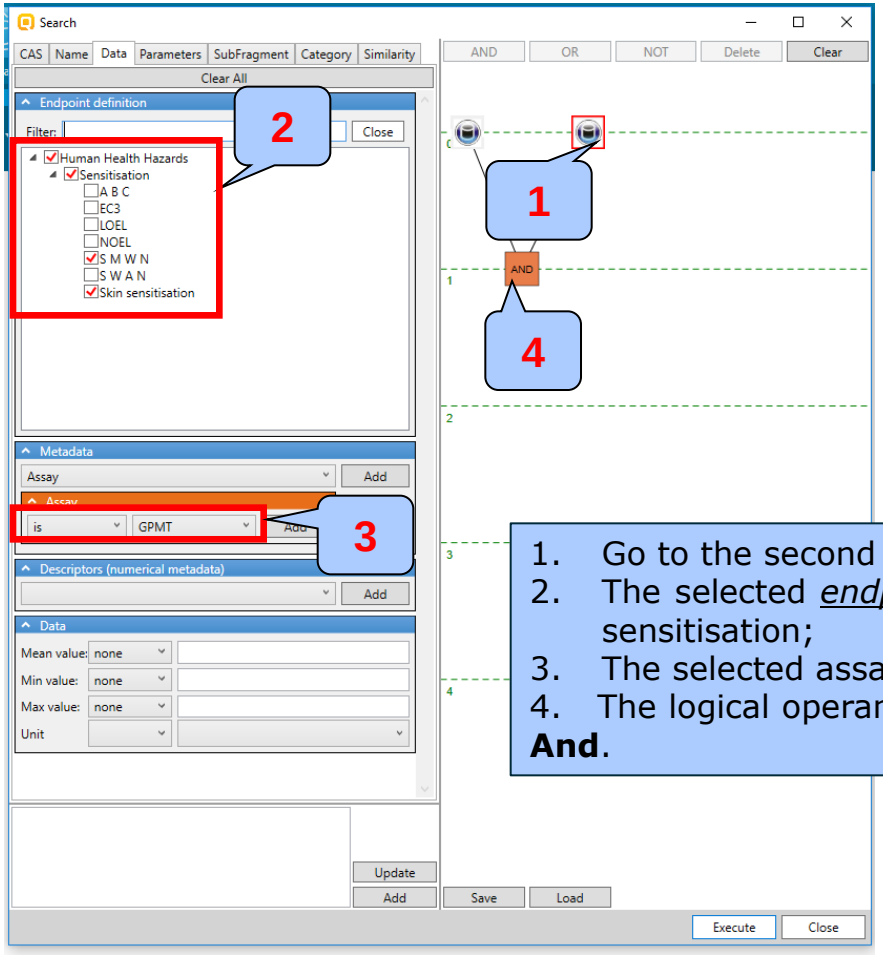
1. Go to the first query to visualized its criteria:
2. The selected *endpoint* is "Skin sensitization" – **EC 3**;
3. The selected *assay* is **LLNA**.

Types of endpoint correlations

Categorical vs. categorical

Gather experimental data – step 2

Example 1: Correlation between LLNA and GPMT data



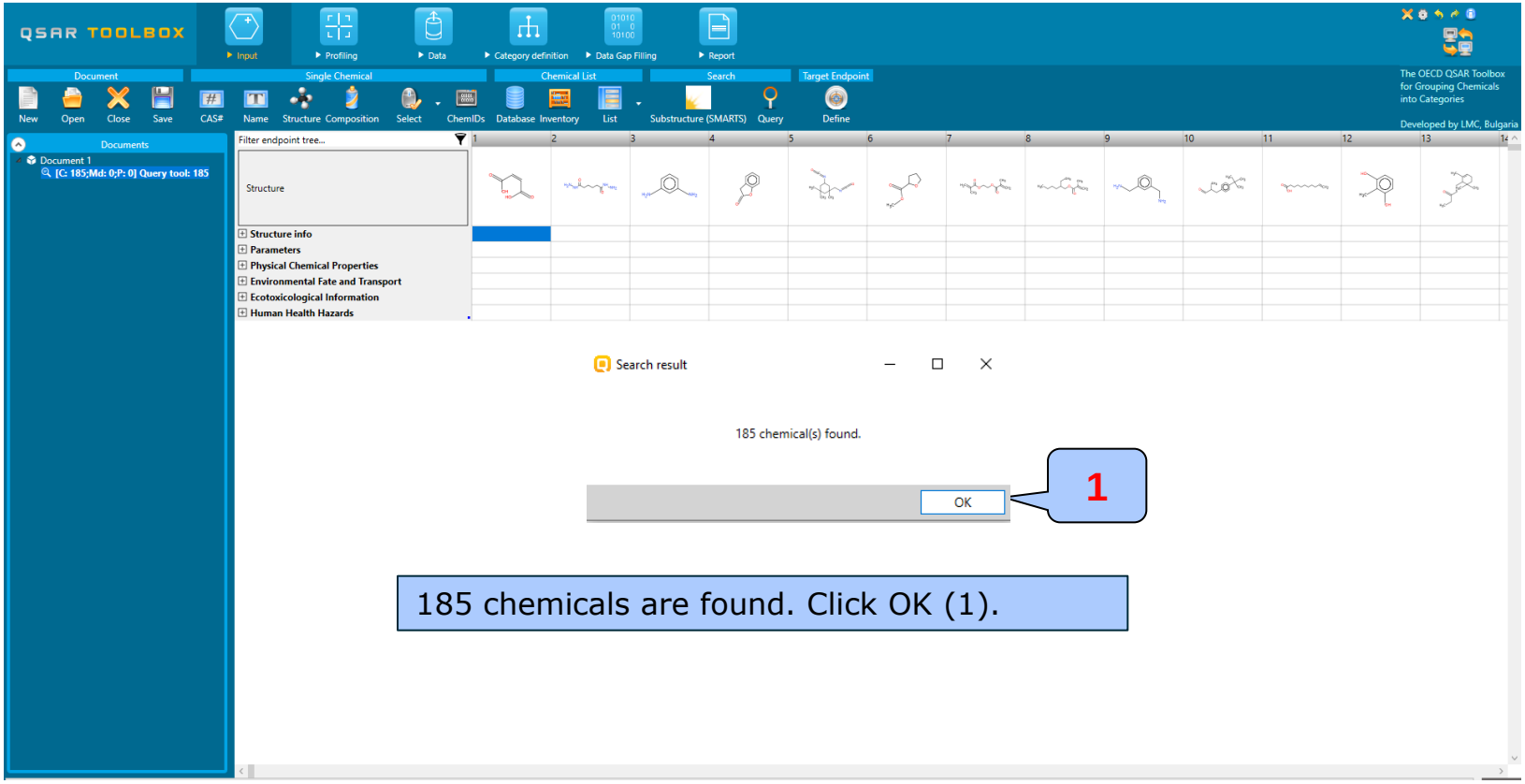
1. Go to the second Query;
2. The selected *endpoints* are "Skin sensitization" – SMWN and Skin sensitisation;
3. The selected assay is **GPMT**;
4. The logical operand that links both queries is **And**. Double-click on **And**.

Types of endpoint correlations

Categorical vs. categorical

Gather experimental data – step 2

Example 1: Correlation between LLNA and GPMT data



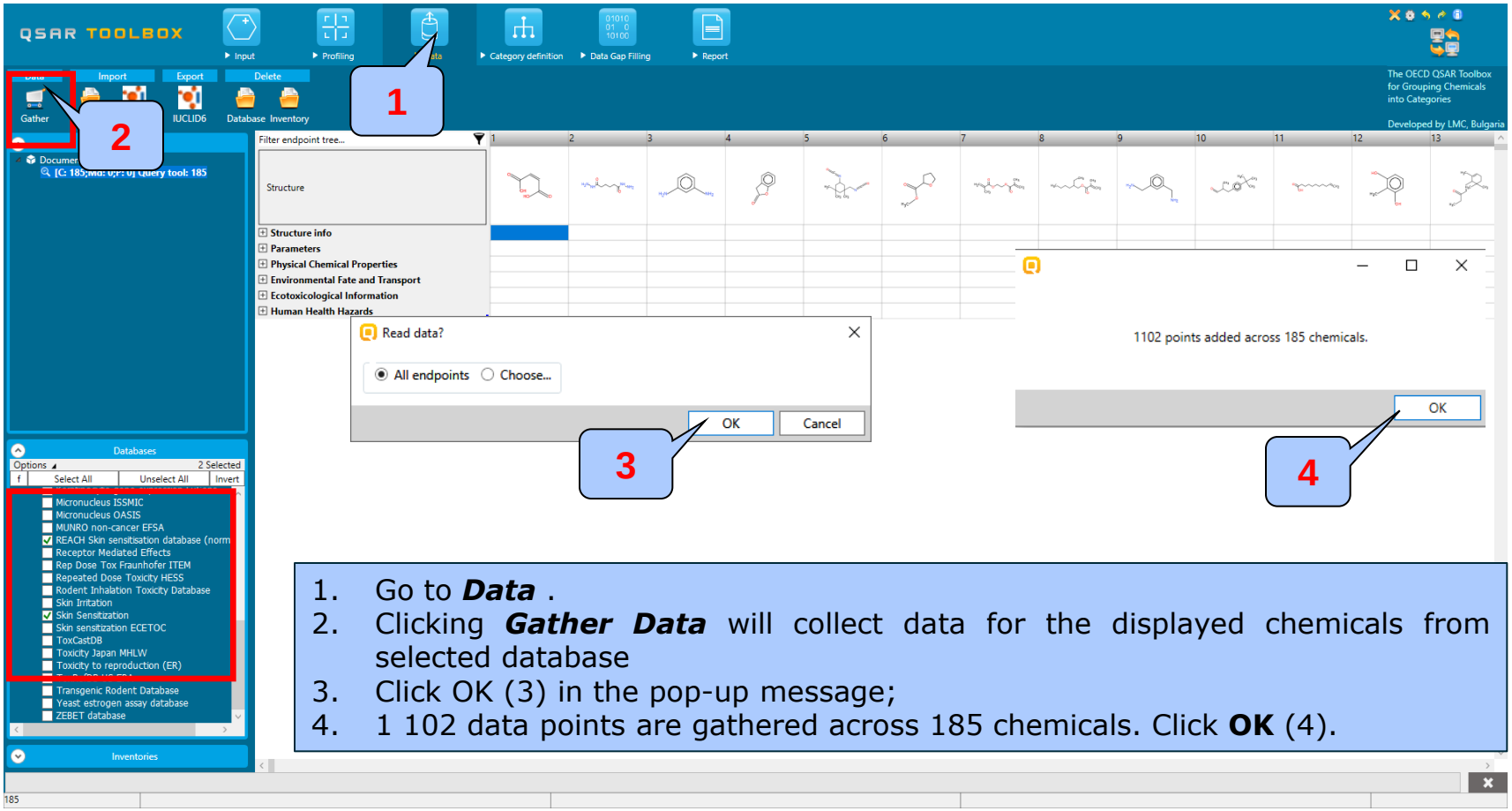
185 chemicals are found. Click OK (1).

Types endpoint correlations

Categorical vs. categorical

Gather experimental data – step 2

Example 1: Correlation between LLNA and GPMT data



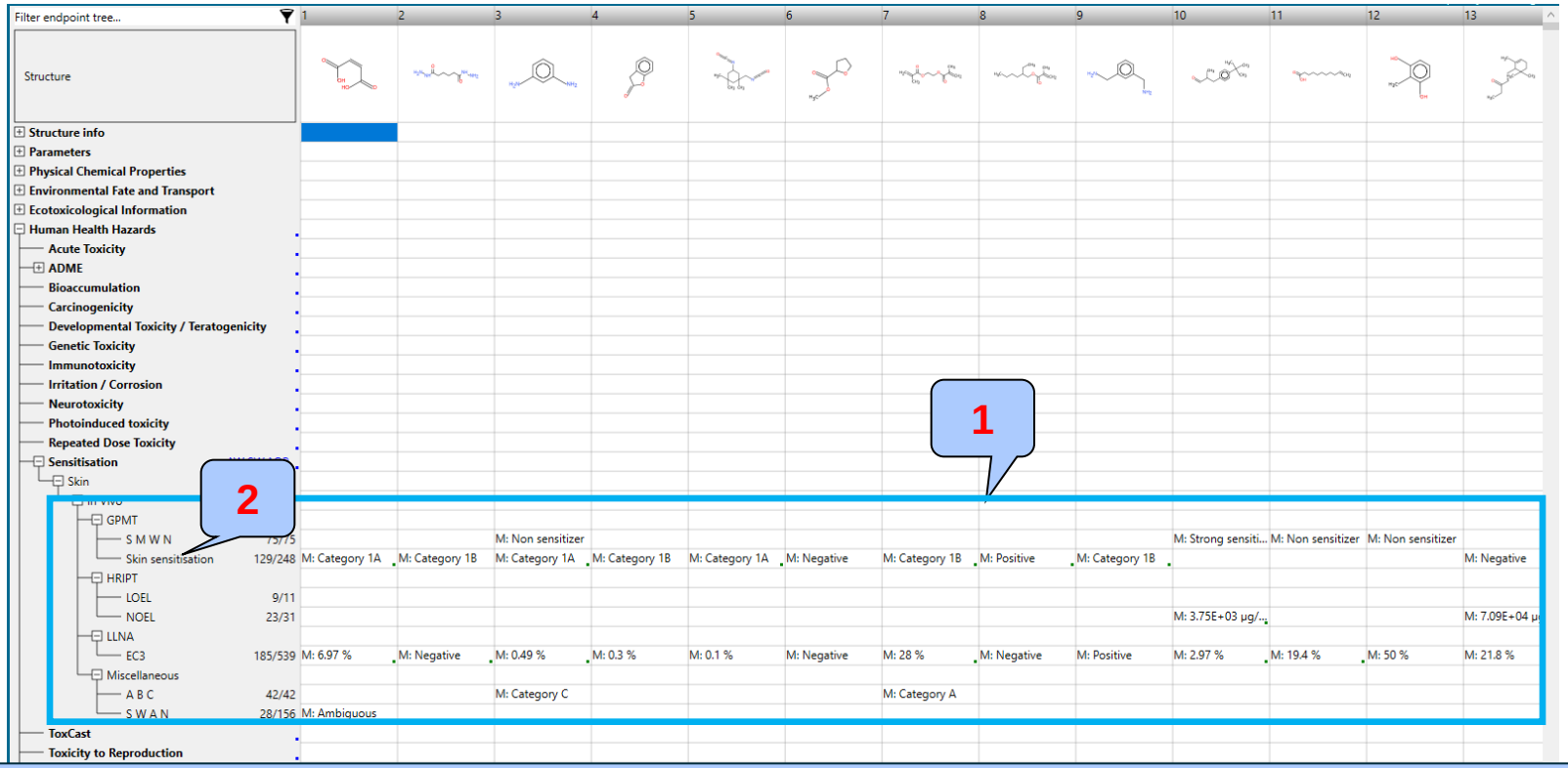
1. Go to ***Data*** .
2. Clicking ***Gather Data*** will collect data for the displayed chemicals from selected database
3. Click OK (3) in the pop-up message;
4. 1 102 data points are gathered across 185 chemicals. Click **OK** (4).

Types endpoint correlations

Categorical vs. categorical

Gather experimental data – step 2

Example 1: Correlation between LLNA and GPMT data



1. Skin sensitization data appeared on data matrix.
2. Data associated with different type assay (e.g. LLNA, GPMT, HRIPT) are distributed in separate nodes

What is "scale" and "scale conversion" ?

Reminder slide

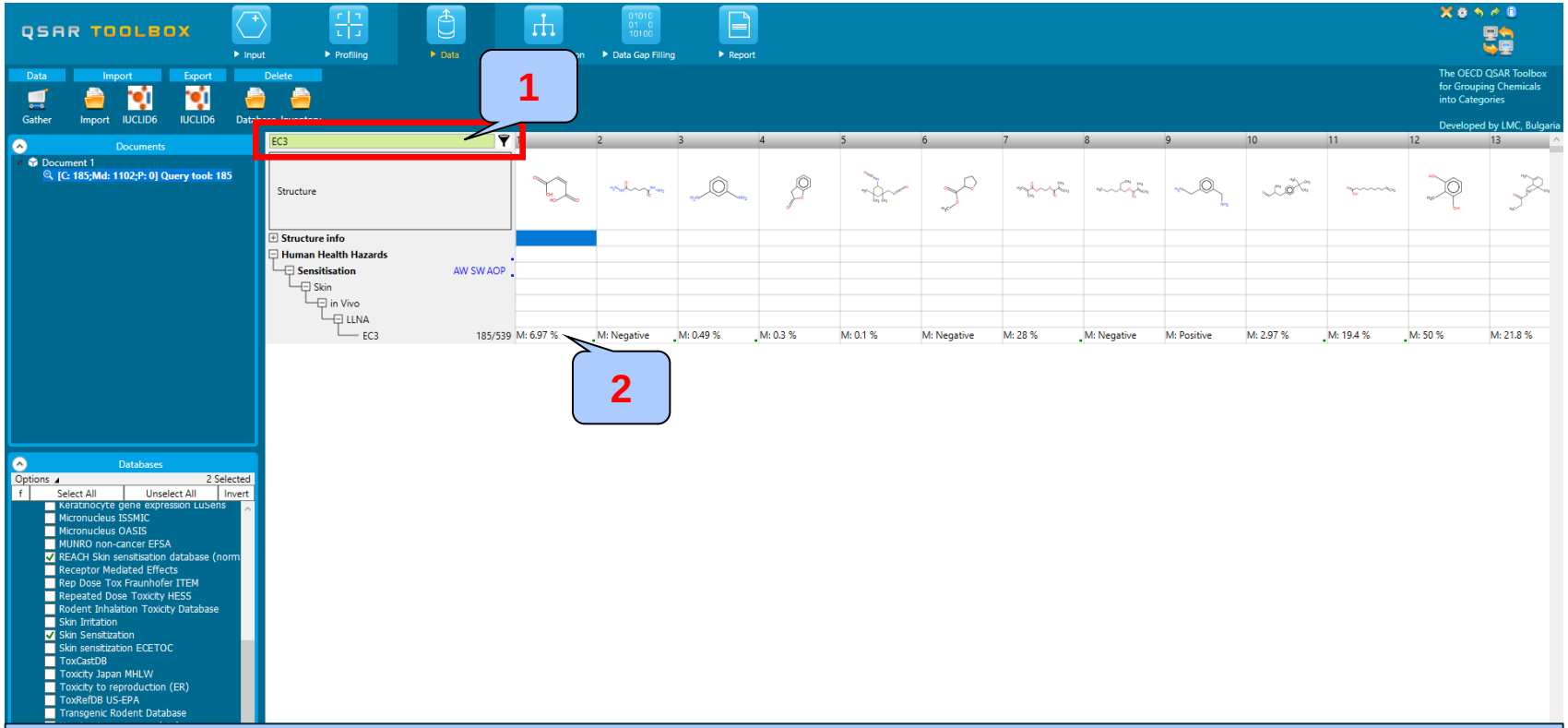
- Skin sensitisation as an example is a "qualitative" endpoint for which the results are presented with categorical type of data (for example: positive; negative; weak sensitizer; strong sensitizer, etc).
- Skin sensitisation potential data of the chemicals comes from different databases coded with different names (for example: data from John Moores University of Liverpool are: *Strongly sensitizing, Moderately sensitizing etc*.; data from European centre for Ecotoxicology and Toxicology of chemicals are: *Positive, Negative, and Equivocal*).
- The main purpose of the scales is to unify all data available in the Toolbox databases for a certain endpoint.
- "Scale conversion" is the TB instrument to create conversions between scales. It is more reasonable to convert from a more informative to less informative scale.
- The default scale for Skin Sensitisation data is "Skin Sensitisation ECETOC". It converts all skin sensitization data into: Positive and Negative. This allows skin sensitization data to be used as much as possible for gap filling purposes.

Types endpoint correlations

Categorical vs. categorical

Define target endpoint – step 3

Example 1: Correlation between LLNA and GPMT data



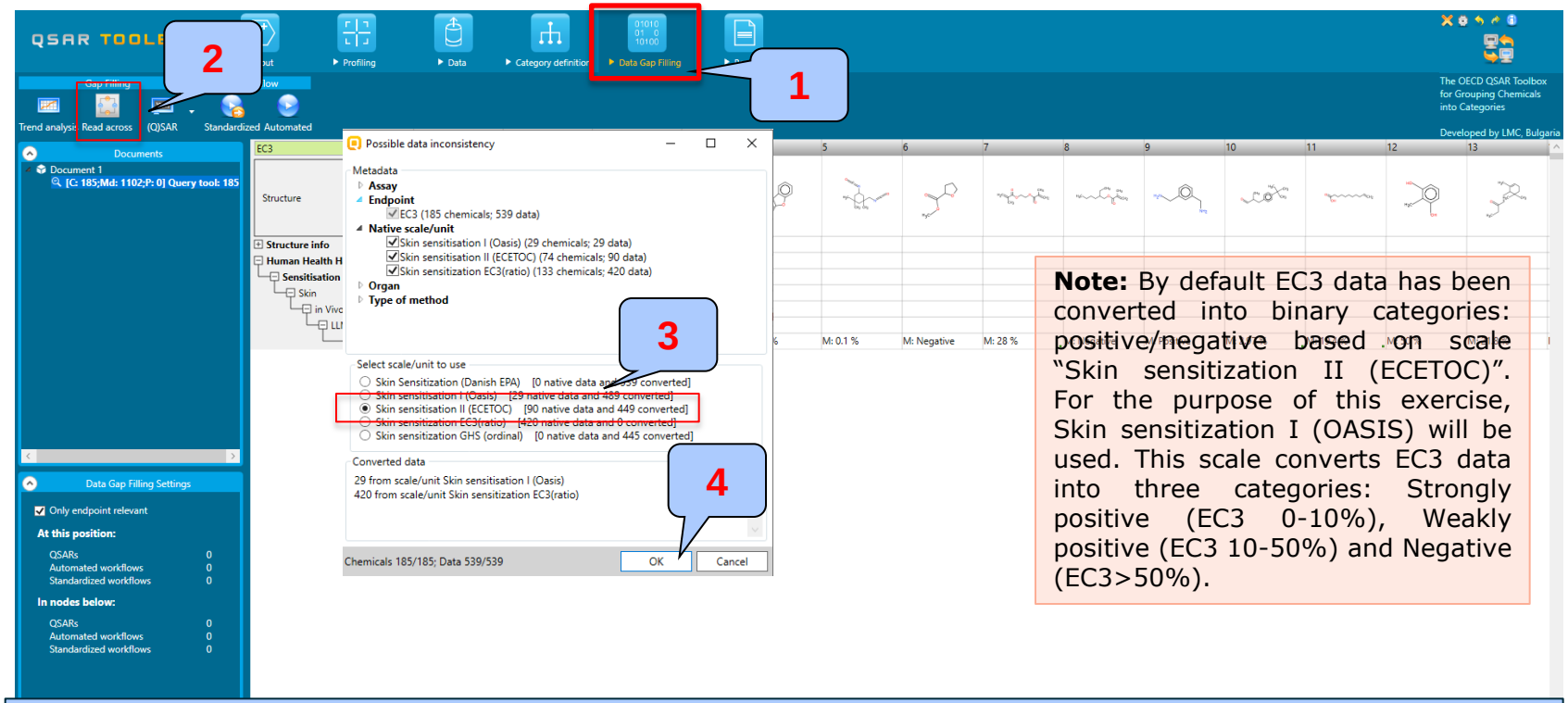
1. Type in ***EC3*** data associated with LLNA assay in the filter box, then press ***Enter*** in your keyboard which will automatically filter the tree to the target endpoint;
2. **Click** on the cell associated with target endpoint;

Types endpoint correlations

Categorical vs. categorical

Enter Gap filling – step 4

Example 1: Correlation between LLNA and GPMT data



Note: By default EC3 data has been converted into binary categories: positive/negative based on scale "Skin sensitization II (ECETOC)". For the purpose of this exercise, Skin sensitization I (OASIS) will be used. This scale converts EC3 data into three categories: Strongly positive (EC3 0-10%), Weakly positive (EC3 10-50%) and Negative (EC3>50%).

Enter Gap filling and apply read across. Read across is applied because a categorical type data is analyzed. Follow the steps:

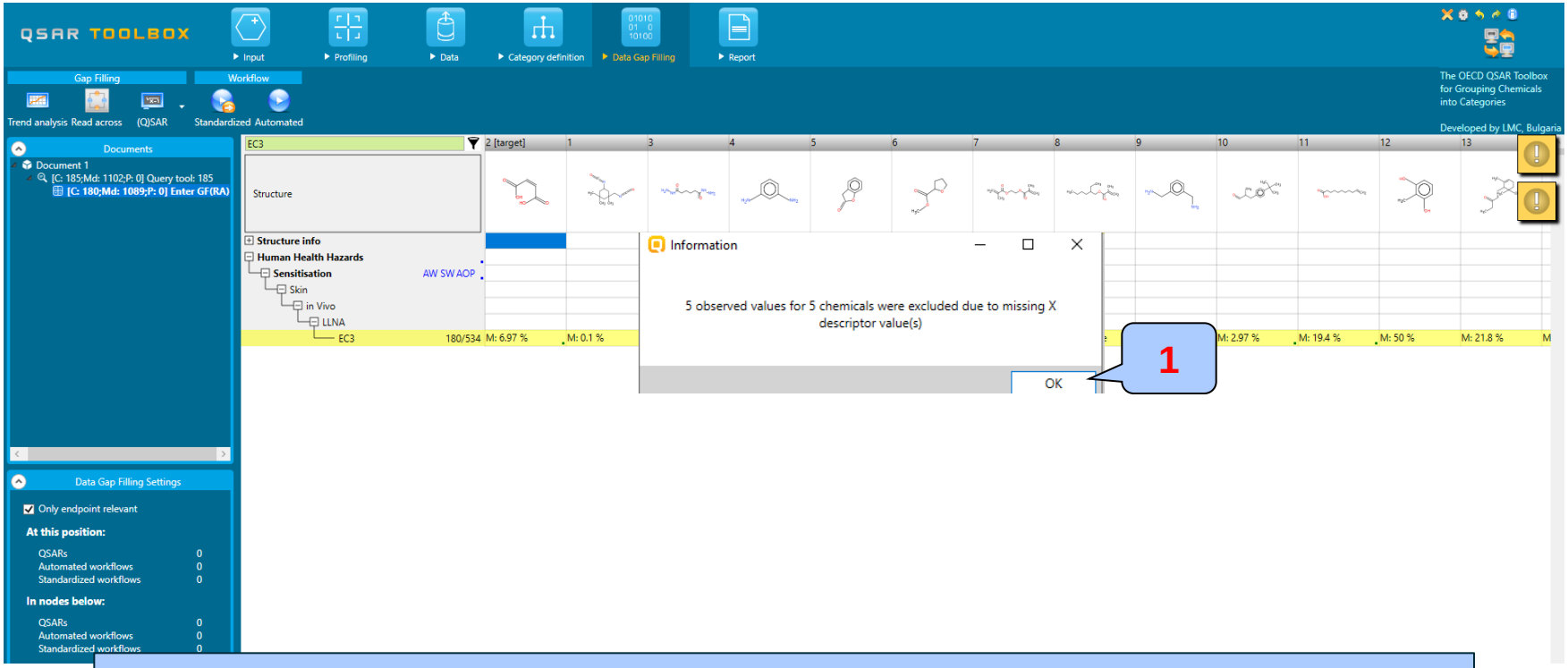
1. Go to **Data Gap filling**; 2. Select **Read-across**;
3. Select **Skin sensitization II (ECETOC)** scale (see Note); 4. Click **OK**;

Types endpoint correlations

Categorical vs. categorical

Enter Gap filling – step 4

Example 1: Correlation between LLNA and GPMT data



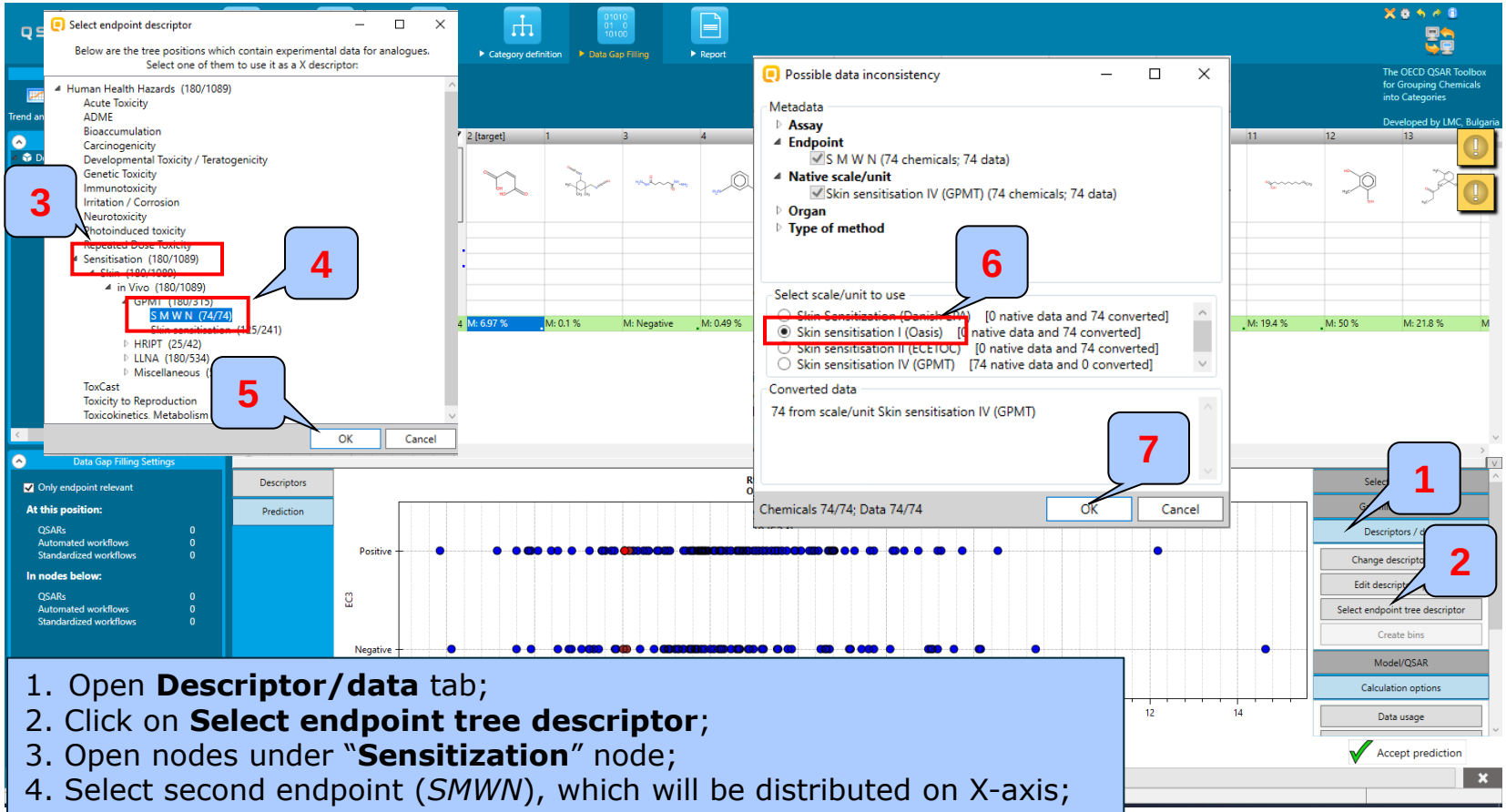
A message informs the user about the number of chemicals with experimental data that are excluded from gap filling due to missing X-descriptor value appeared. Click **OK** (1).

Types endpoint correlations

Categorical vs. categorical

Perform correlation between LLNA and GPMT data– step 5

Example 1: Correlation between LLNA and GPMT data



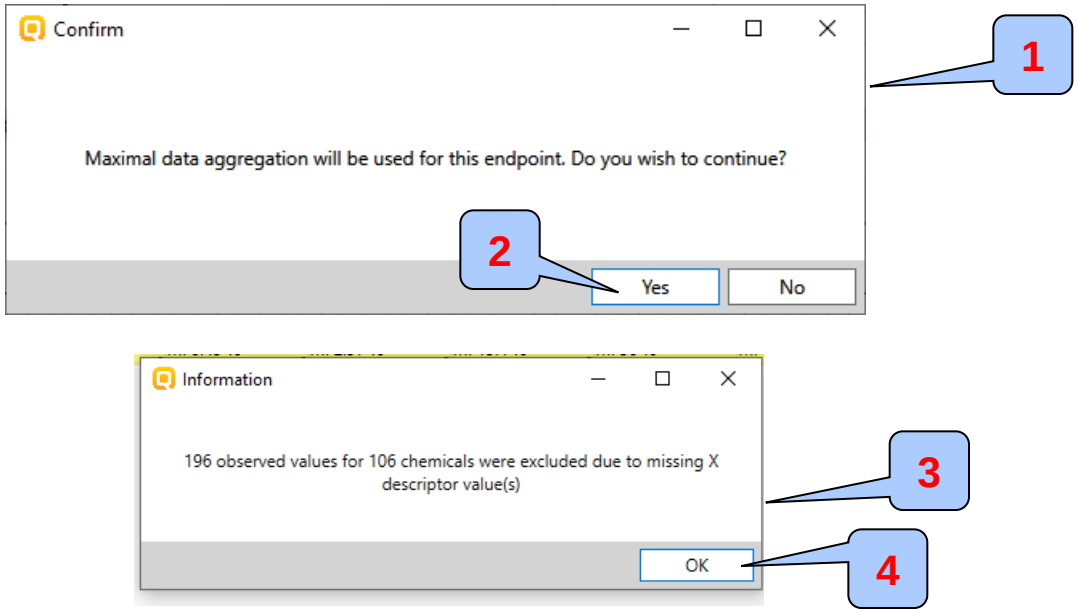
1. Open **Descriptor/data** tab;
2. Click on **Select endpoint tree descriptor**;
3. Open nodes under "**Sensitization**" node;
4. Select second endpoint (*SMWN*), which will be distributed on X-axis;
5. Click **OK** button; 6. Select Scale I OASIS; 7. Click **OK.**

Types endpoint correlations

Categorical vs. categorical

Perform correlation between LLNA and GPMT data– step 5

Example 1: Correlation between LLNA and GPMT data



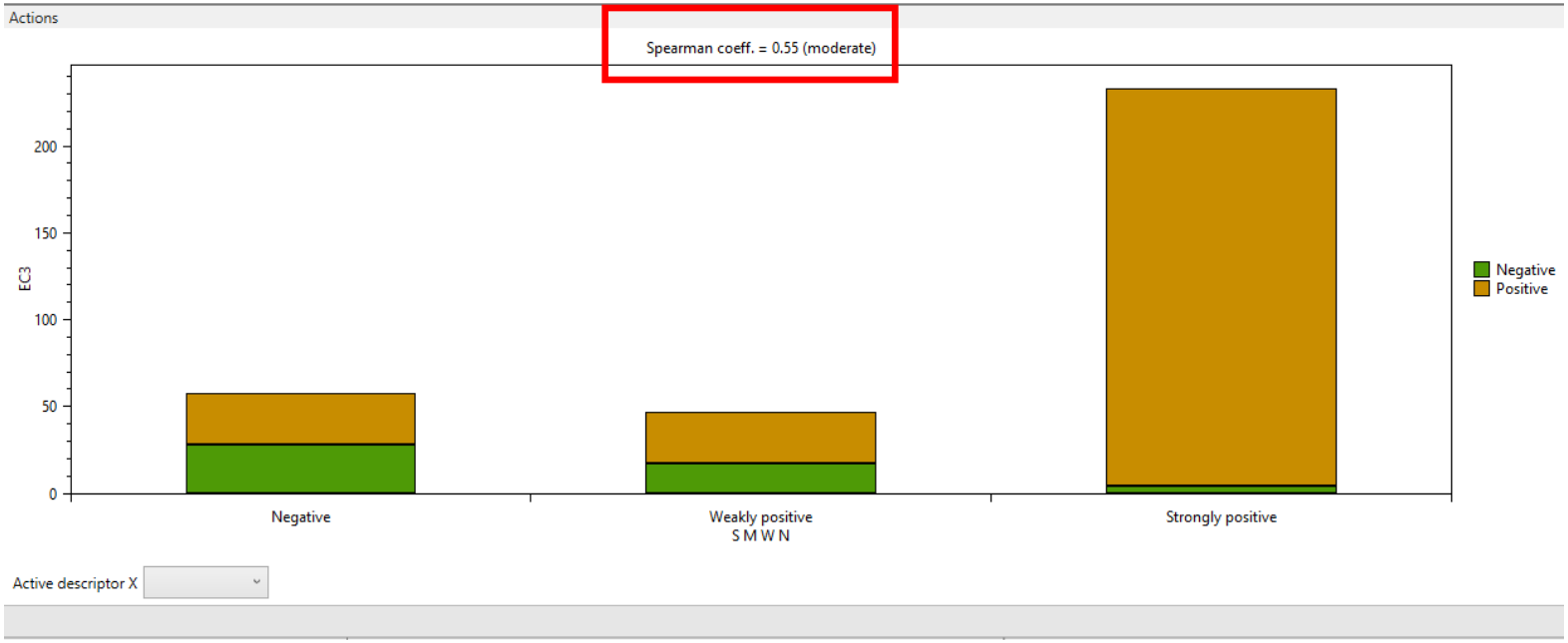
1. As only one data point per chemical is permitted in this type of correlation, the maximal data point will be considered for each chemical;
2. Click ***Yes***;
3. A message informs the user about the number of chemicals with experimental data that are excluded from gap filling due missing data for SMWN endpoint appears. This will not affect the value of correlation coefficient;
4. Click ***OK.***

Types endpoint correlations

Categorical vs. categorical

Interpretation of correlation results (LLNA vs. GPMT)

Example 1: Correlation between LLNA and GPMT data



- Correlation analysis between two categorical type skin sensitization data (LLNA and GPMT) shows moderate endpoint correlation (Spearman coefficient is 0.55).

Types endpoint correlations
Categorical vs. categorical

- The second example illustrating categorical vs. categorical type correlation is:
  - Example 2: Correlation between Skin sensitization and Ames mutagenicity data
    - LLNA (Negative, Weakly positive, Strongly positive)
    - AMES (Positive, Equivocal, Negative)

- Step by step workflow is presented on next few slides. Summary of the workflow steps are provided below:
  - *Query Tool and select FSQ file(step 1)*
  - *Gather experimental data (step 2)*
  - *Enter Gap filling (step 3)*
  - *Perform correlation between endpoints (step 4).*

Types endpoint correlations

Categorical vs. categorical

Gather experimental data – step 2

Example 2: Correlation between LLNA and AMES data

The correlation between LLNA and Ames data has been investigated in view of the proposition that mutagenicity data can be used as part of an integrated approach to testing and assessment (IATA) for skin sensitisation[1,2].

[1] Patlewicz G., C. Kuseva, A. Kesova, I. Popova, T. Zhechev, T. Pavlov, D. W. Roberts, O. Mekenyan, Towards AOP application – Implementation of an integrated approach to testing and assessment (IATA) into a pipeline tool for skin sensitization. *Regul. Toxicol. Pharmacol.* 69 (3) (2014), 529 - 545.

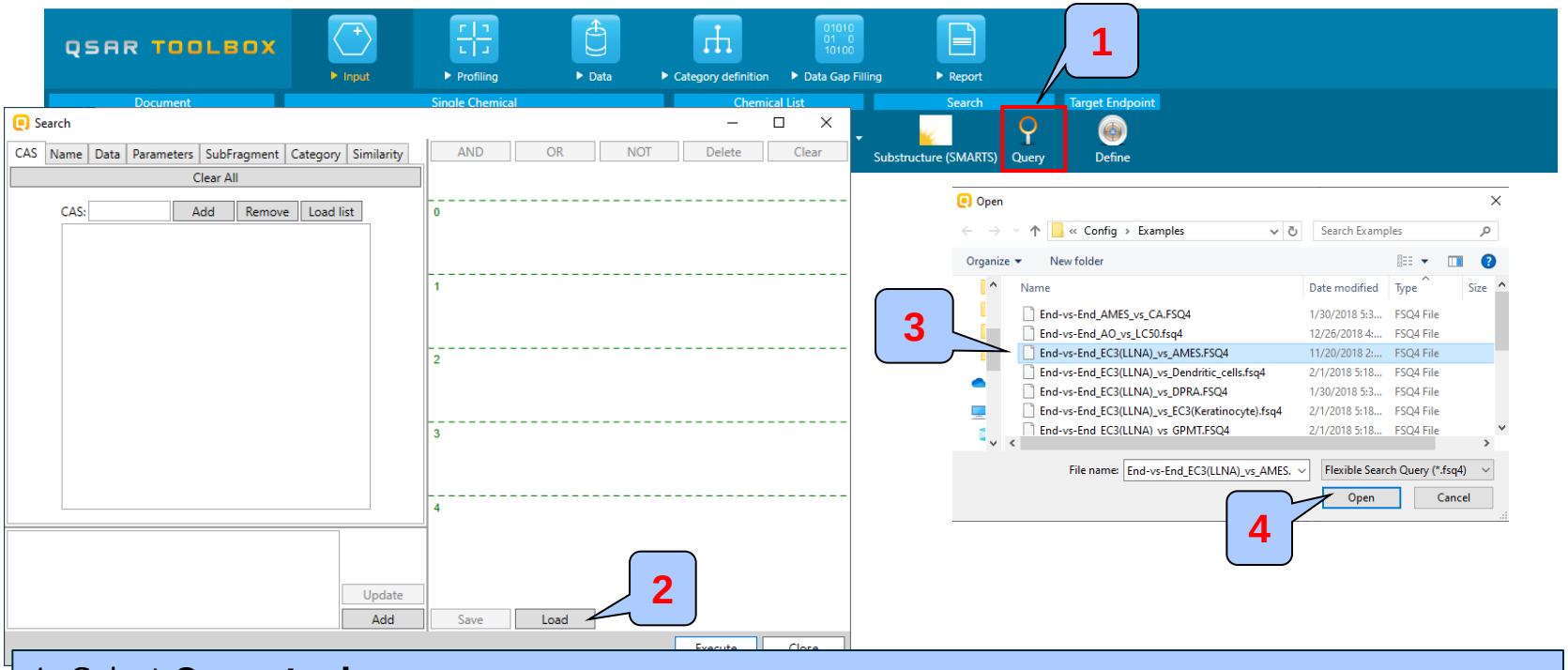
[2] Wolfreys, M,A, Basketter, A. D. Mutagens and Sensitizers—An Unequal Relationship?. Cutaneous and Ocular Toxicology. 2004

Types endpoint correlations

Categorical vs. categorical

Gather experimental data – step 2

Example 2: Correlation between LLNA and AMES data



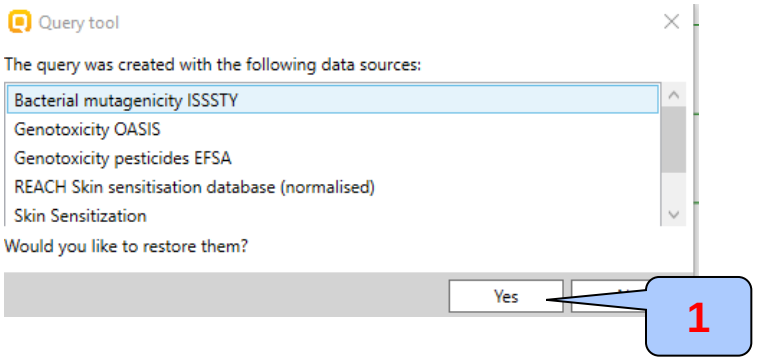
1. Select **Query tool;**
2. Click **Load**;
3. Select the file from Example directory (C:\Program Files (x86)\Common Files\(Q)SAR Toolbox 4.4\Config\Examples)
4. Click **Open.**

Types endpoint correlations

Categorical vs. categorical

Gather experimental data – step 2

Example 2: Correlation between LLNA and AMES data



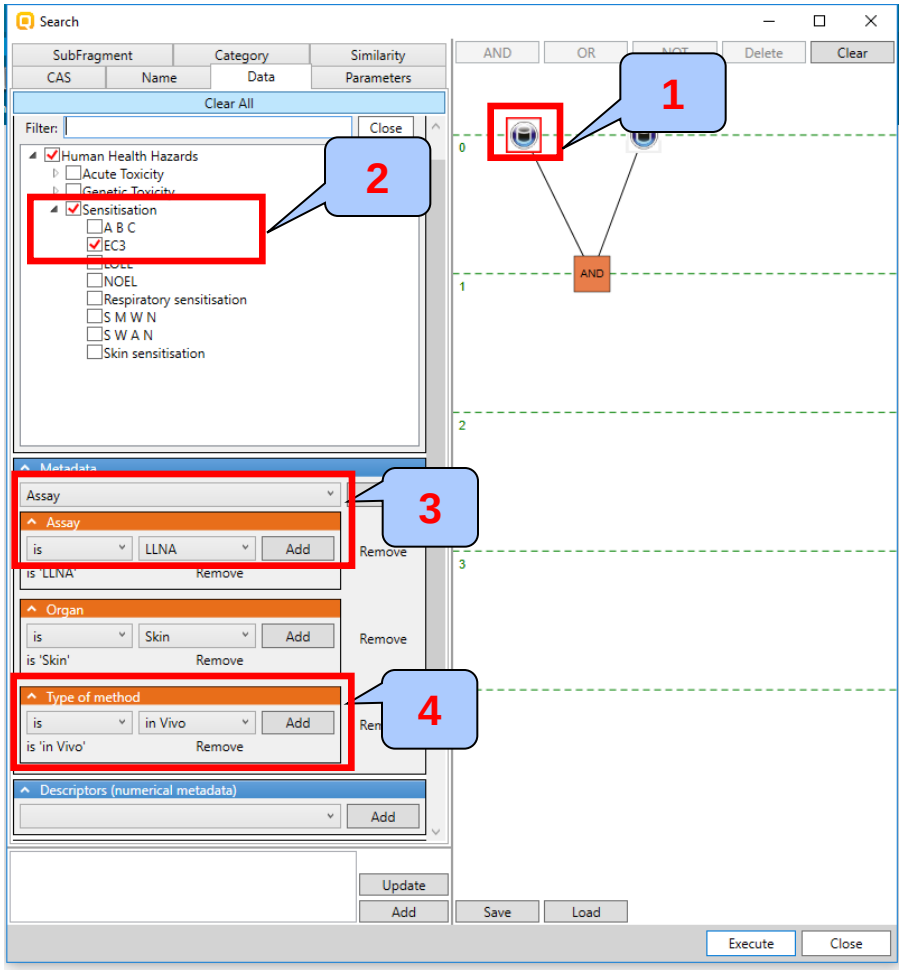
1. Click **Yes** to confirm that you want to restore the databases used during the creation of the . fsq file.

Types endpoint correlations

Categorical vs. categorical

Gather experimental data – step 2

Example 2: Correlation between LLNA and AMES data



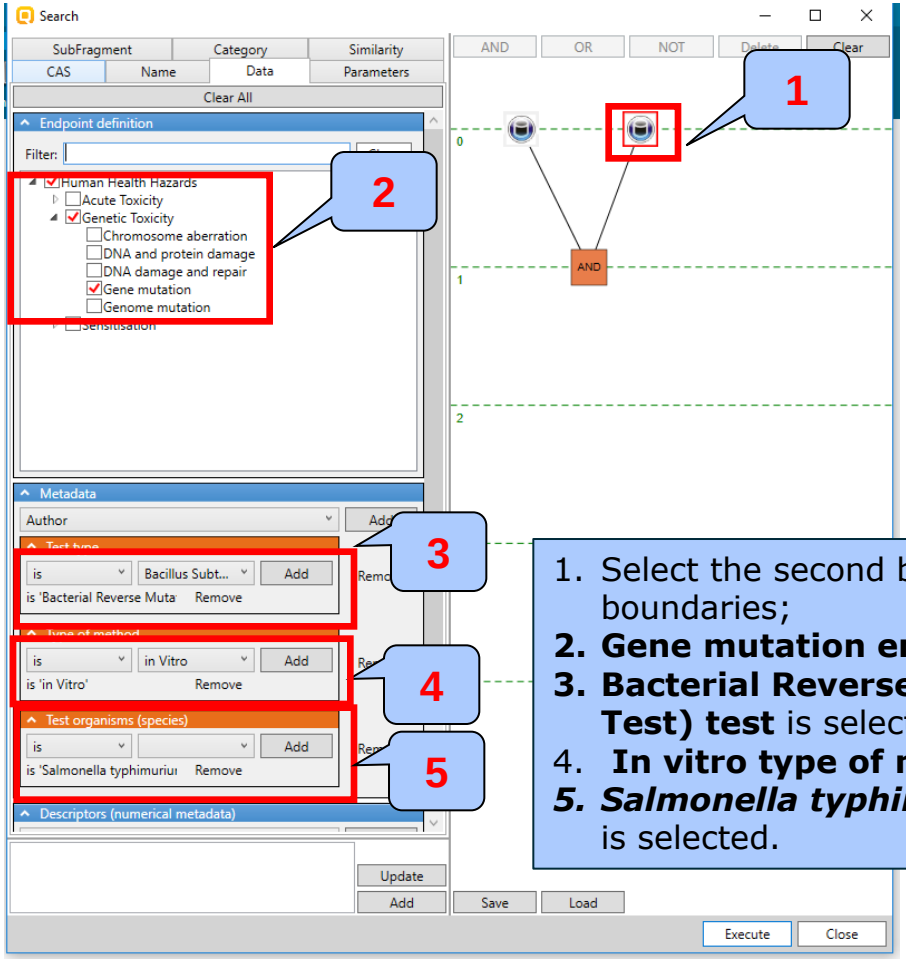
1. Select the first boundary to visualized its boundaries:
2. **EC3** is selected;
3. **LLNA** assay is selected;
4. **In vivo** type of method is selected.

Types endpoint correlations

Categorical vs. categorical

Gather experimental data – step 2

Example 2: Correlation between LLNA and AMES data



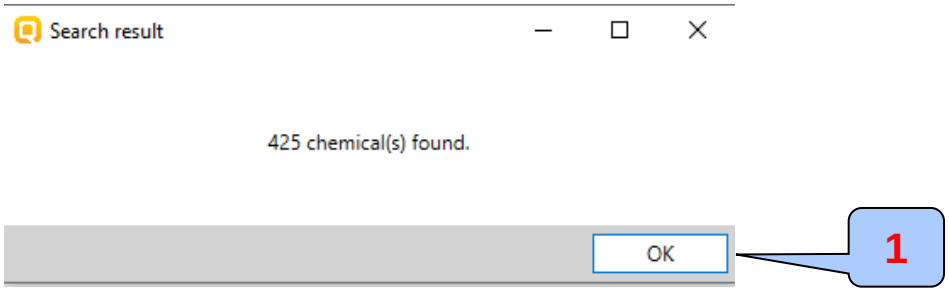
1. Select the second boundary to visualized its boundaries;
2. **Gene mutation endpoint is selected**;
3. **Bacterial Reverse Mutation Assay (e.g. Ames Test) test** is selected.;
4. **In vitro type of method** is selected;
5. ***Salmonella typhimurium test organism (species)*** is selected.

Types endpoint correlations

Categorical vs. categorical

Gather experimental data – step 2

Example 2: Correlation between LLNA and AMES data



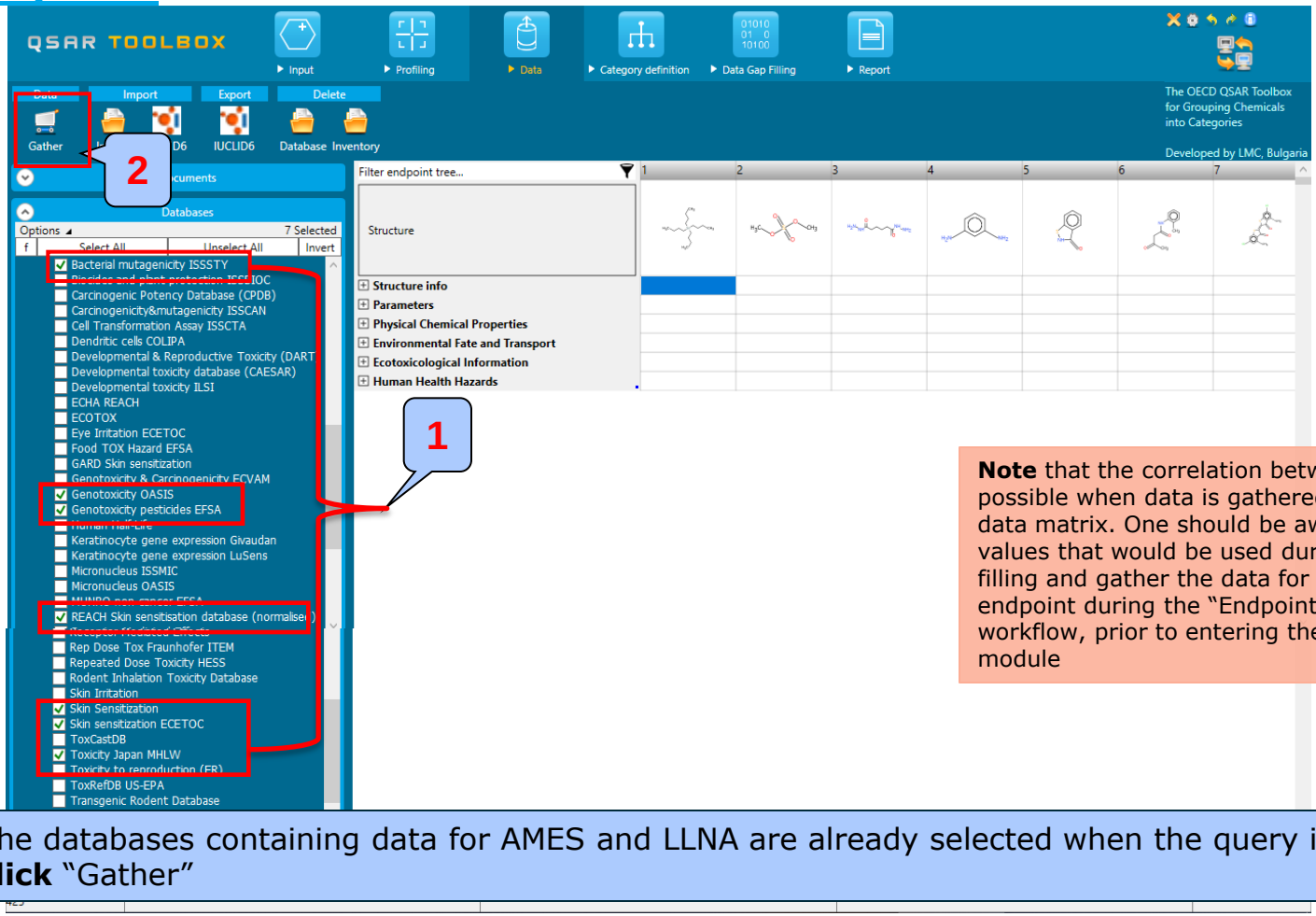
425 chemicals are found; Click OK (1).

Types endpoint correlations

Categorical vs. categorical

Gather experimental data – step 2

Example 2: Correlation between LLNA and AMES data



Note that the correlation between endpoints is possible when data is gathered and available on data matrix. One should be aware of the data values that would be used during the data gap filling and gather the data for the corresponding endpoint during the "Endpoint" stage of the workflow, prior to entering the "Data gap filling" module

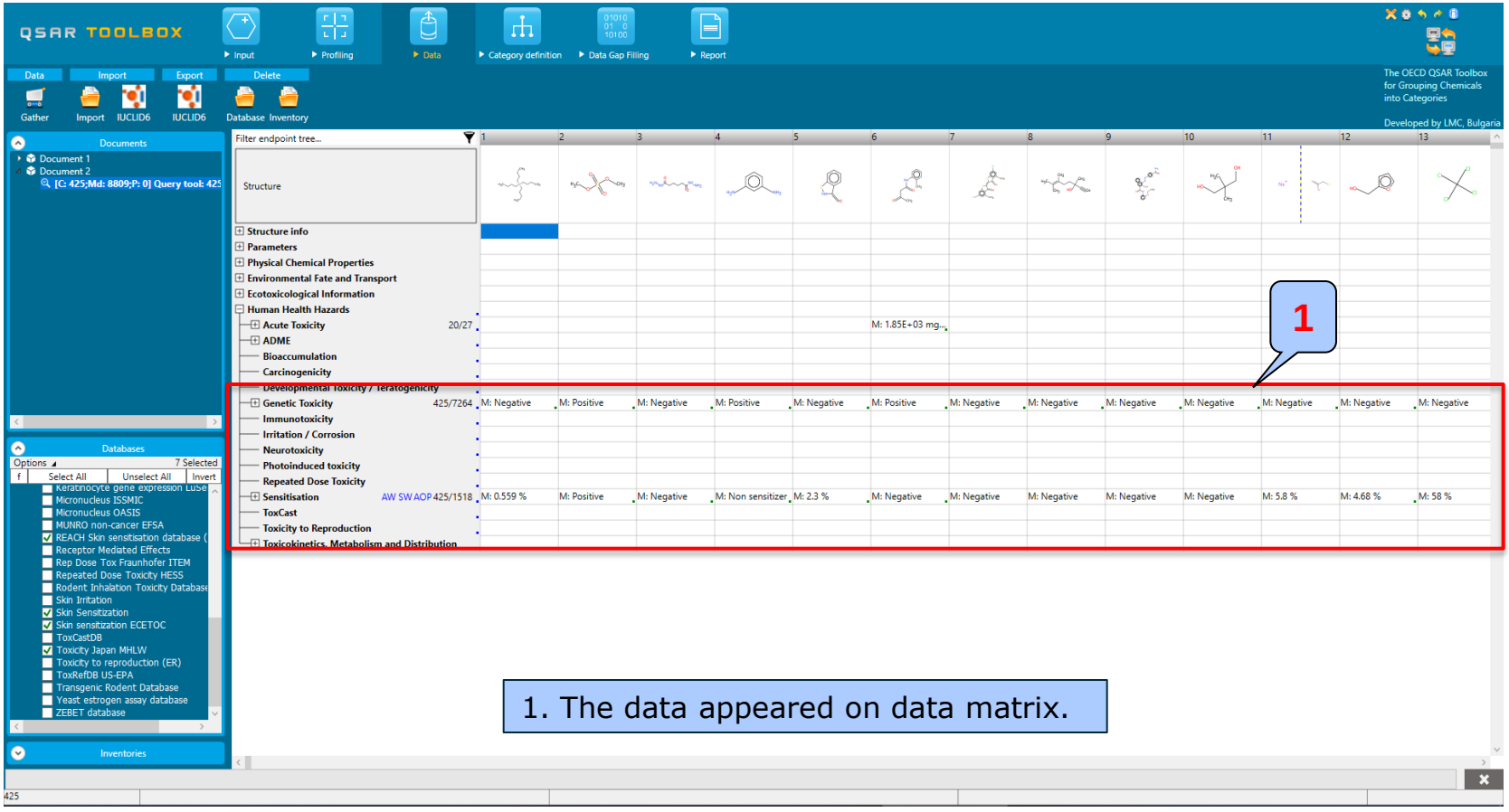
1. The databases containing data for AMES and LLNA are already selected when the query is loaded;
2. **Click** "Gather"

Types endpoint correlations

Categorical vs. categorical

Gather experimental data – step 2

Example 2: Correlation between LLNA and AMES data



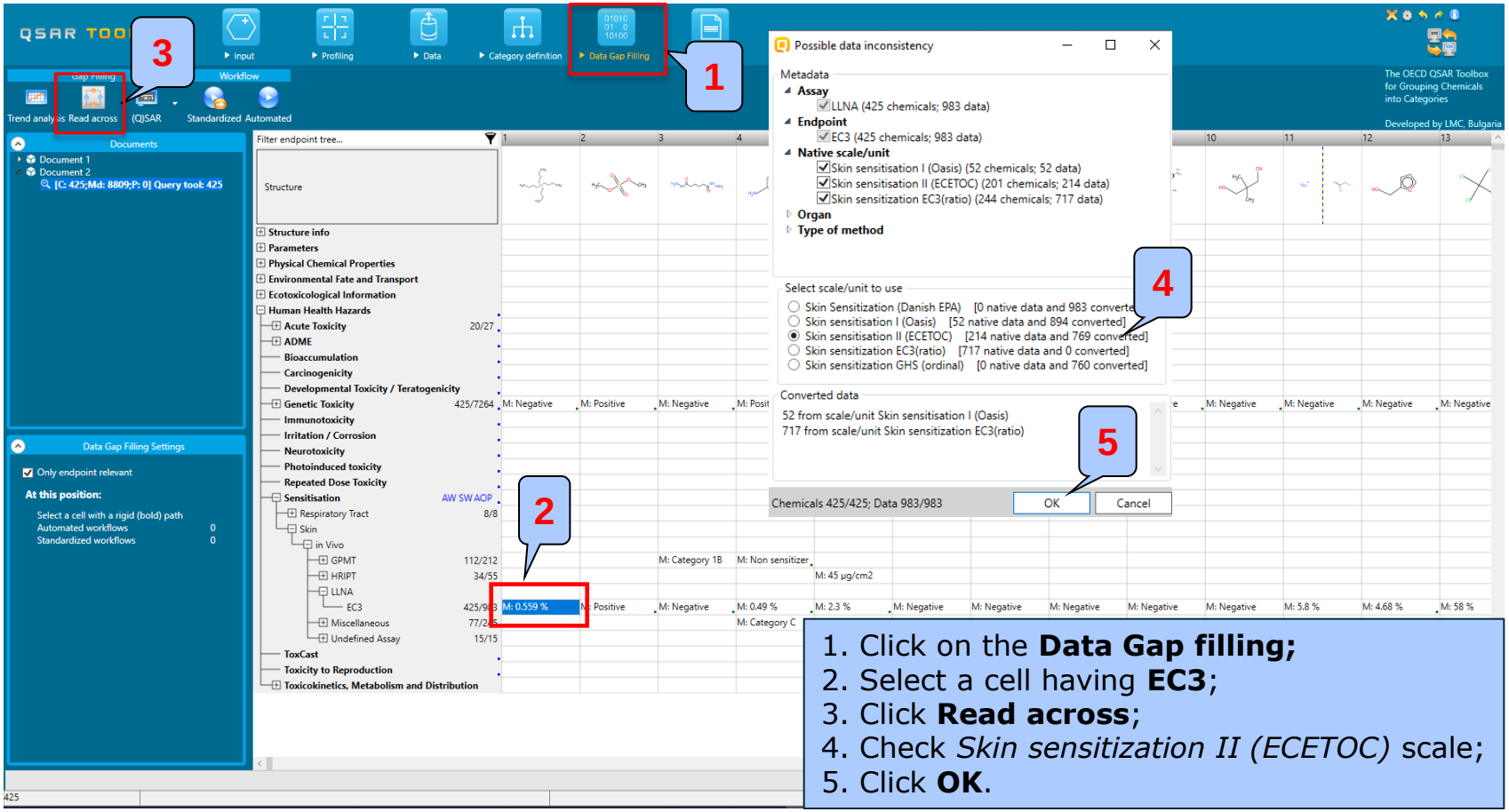
1. The data appeared on data matrix.

Types endpoint correlations

Categorical vs. categorical

Define target endpoint – step 3

Example 2: Correlation of LLNA and AMES data



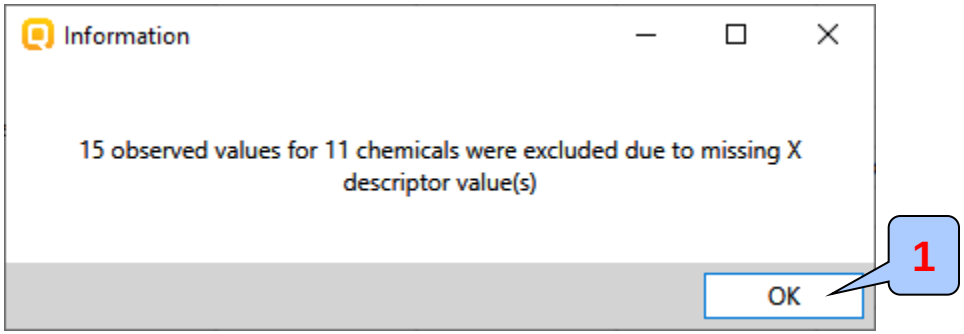
1. Click on the **Data Gap filling;**
2. Select a cell having **EC3**;
3. Click **Read across**;
4. Check *Skin sensitization II (ECETOC)* scale;
5. Click **OK**.

Types endpoint correlations

Categorical vs. categorical

Enter Gap filling – step 4

Example 2: Correlation of LLNA and AMES data



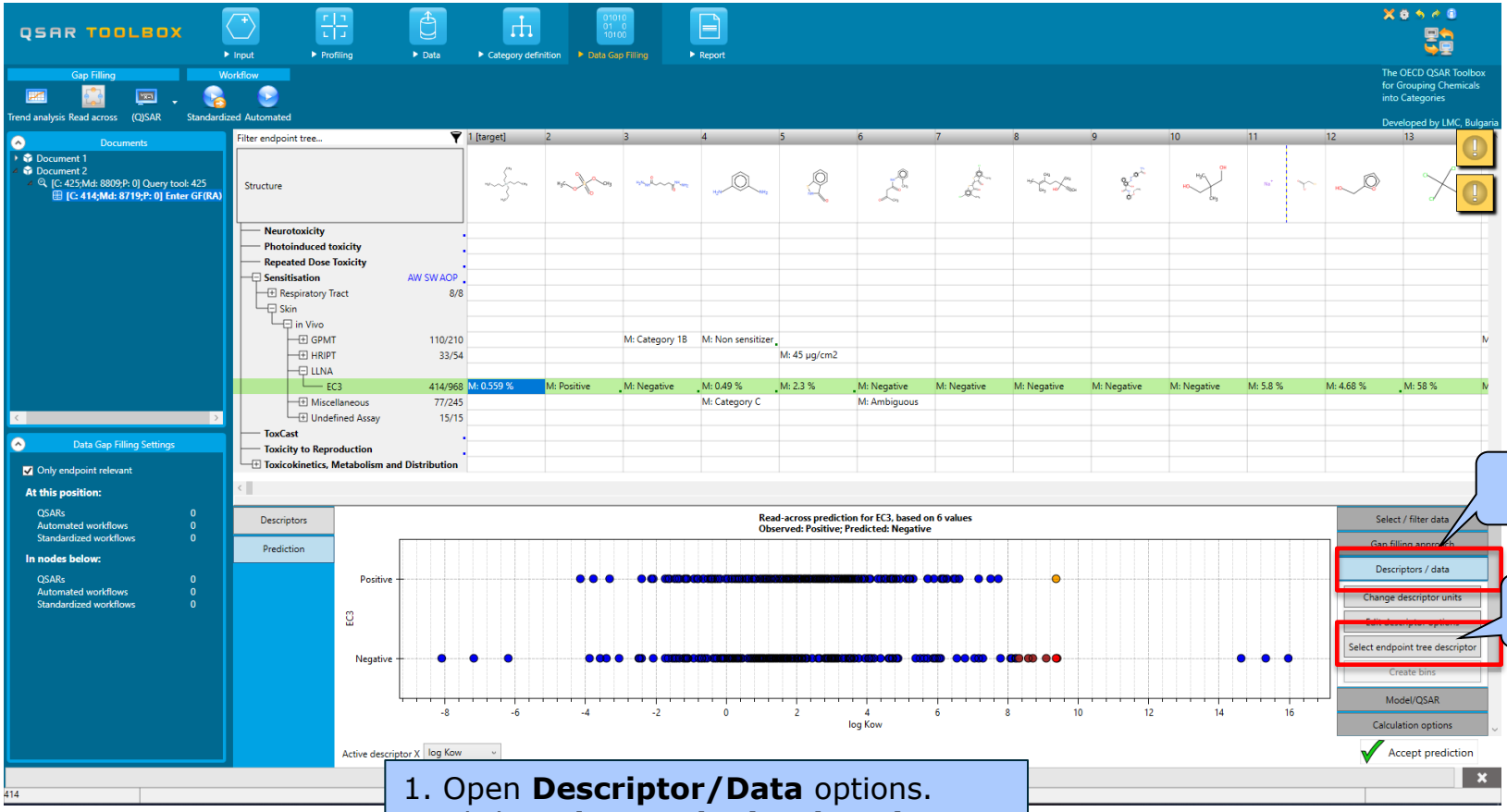
The message informs the user that some chemicals are excluded from gap filling; Click **OK** (1);

Types endpoint correlations

Categorical vs. categorical

Perform correlation between LLNA and AMES data – step 5

Example 2: Correlation between LLNA and AMES data



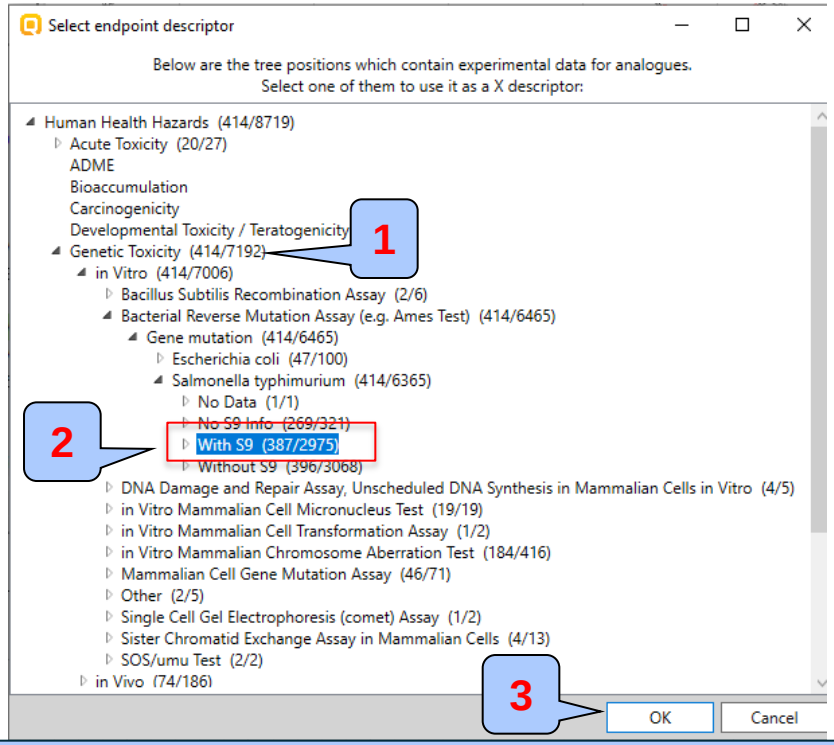
1. Open **Descriptor/Data** options.
2. Click **Select endpoint descriptor.**

Types endpoint correlations

Categorical vs. categorical

Perform correlation between LLNA and AMES data – step 5

Example 2: Correlation between LLNA and AMES data



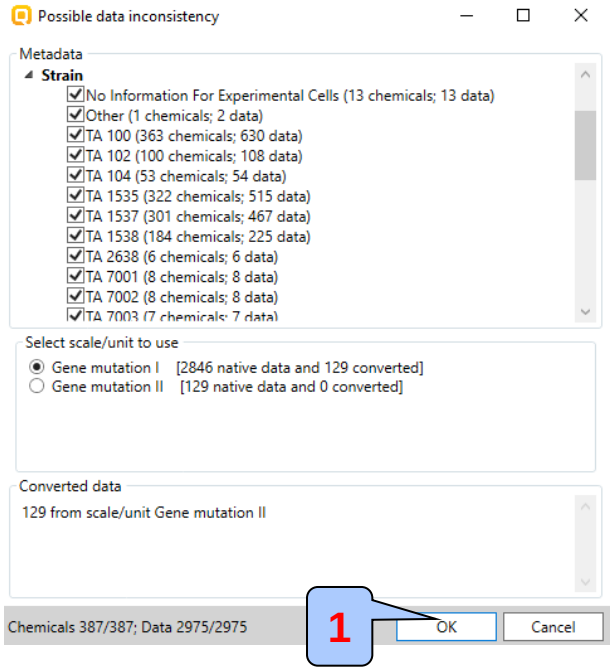
1. In the *Select endpoint descriptor* open the branches below Genetic Toxicity;
2. Select "**With S9**" under *In Vitro|Bacterial Reverse Mutation Assay (e.g. Ames Test)|Gene Mutation|Salmonella typhimurium*;
3. Click ***OK***.

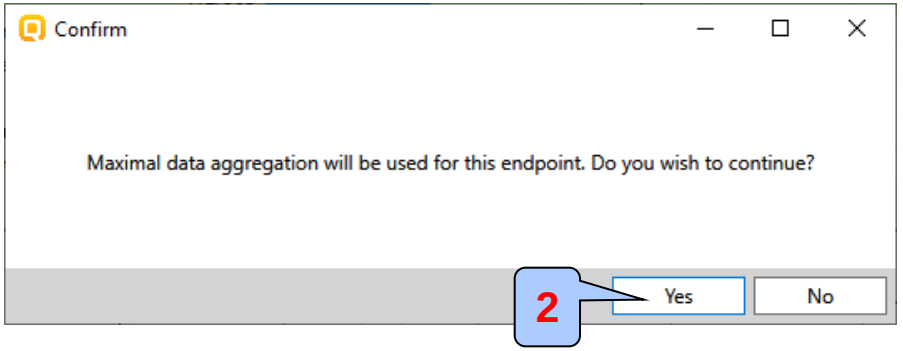
Types endpoint correlations

Categorical vs. categorical

Perform correlation between LLNA and AMES data – step 5

Example 2: Correlation between LLNA and AMES data





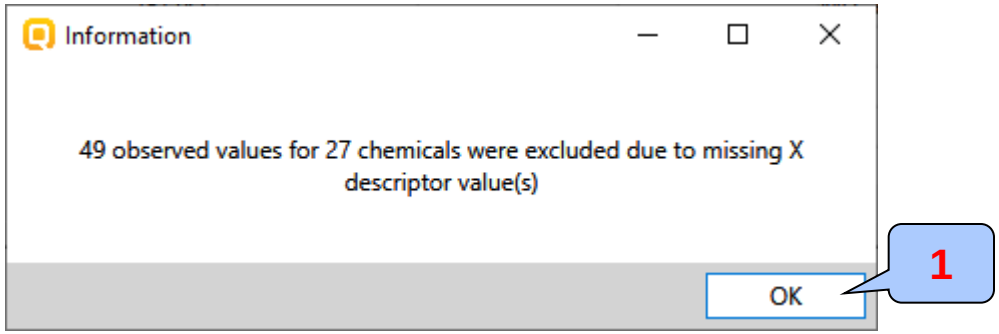
Possible data inconsistency window appears. Click **OK** (1).
As only one data point per chemical is permitted in this type of correlation, the maximal value will be considered for each chemical; Click ***Yes*** (2).

Types endpoint correlations

Categorical vs. categorical

Perform correlation between LLNA and AMES data – step 5

Example 2: Correlation between LLNA and AMES data

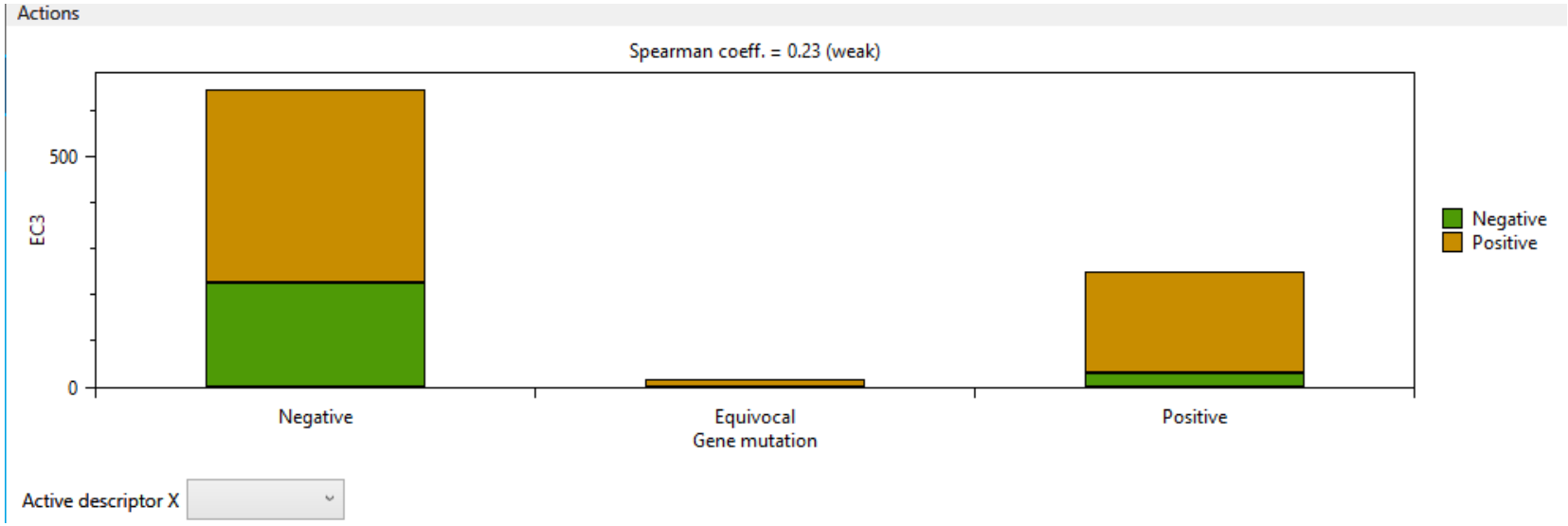


The pop-up message informs on the total number gathered data across the number chemicals that will be excluded in trend analysis due to missing X descriptor value(s).
These are analogues with no AMES data. This will not affect the value of correlation coefficient;
1. Click **OK**;

Types endpoint correlations

Categorical vs. categorical

Interpretation of correlation results (LLNA vs. AMES)



Correlation analysis between two categorical type data: LLNA and AMES shows weak correlation between two endpoints (Spearman coefficient is 0.23).

Outlook

- Background
- Objectives
- The exercise
- **Workflow**
  - Correlation of data – background
  - **Types endpoint correlations**
    - Categorical vs. categorical
    - Categorized continuous vs. categorical

Types endpoint correlations
Categorized continuous vs. categorical

- The aim of this type correlation is to illustrate how categorized continuous and categorical type of data correlate with each other.
- Categorized continuous data is the continuous type data (e.g LC50 or AC50, EC3, %) converted into categories.
- In this example we will illustrated how DPRA ratio data (%) correlates with LLNA data:
  - DPRA (ratio data expressed in % and converted in categories)
  - LLNA (categorical type: Strongly positive, Weakly positive, Negative)
- Step by step workflow is presented on next few slides. Summary of the workflow steps are provided below:
  - *Query Tool and select FSQ file(step 1)*
  - *Gather experimental data (step 2)*
  - *Define target endpoint (step 3)*
  - *Enter Gap filling (step 4)*
  - *Perform correlation between endpoints (step 5).*

Types endpoint correlations

Categorized continuous vs. categorical

Query Tool and select FSQ file - step 1

Example: Correlation between LLNA (Strongly positive, Weakly positive, Negative) and DPRA (Strongly positive, Weakly positive, Negative) data

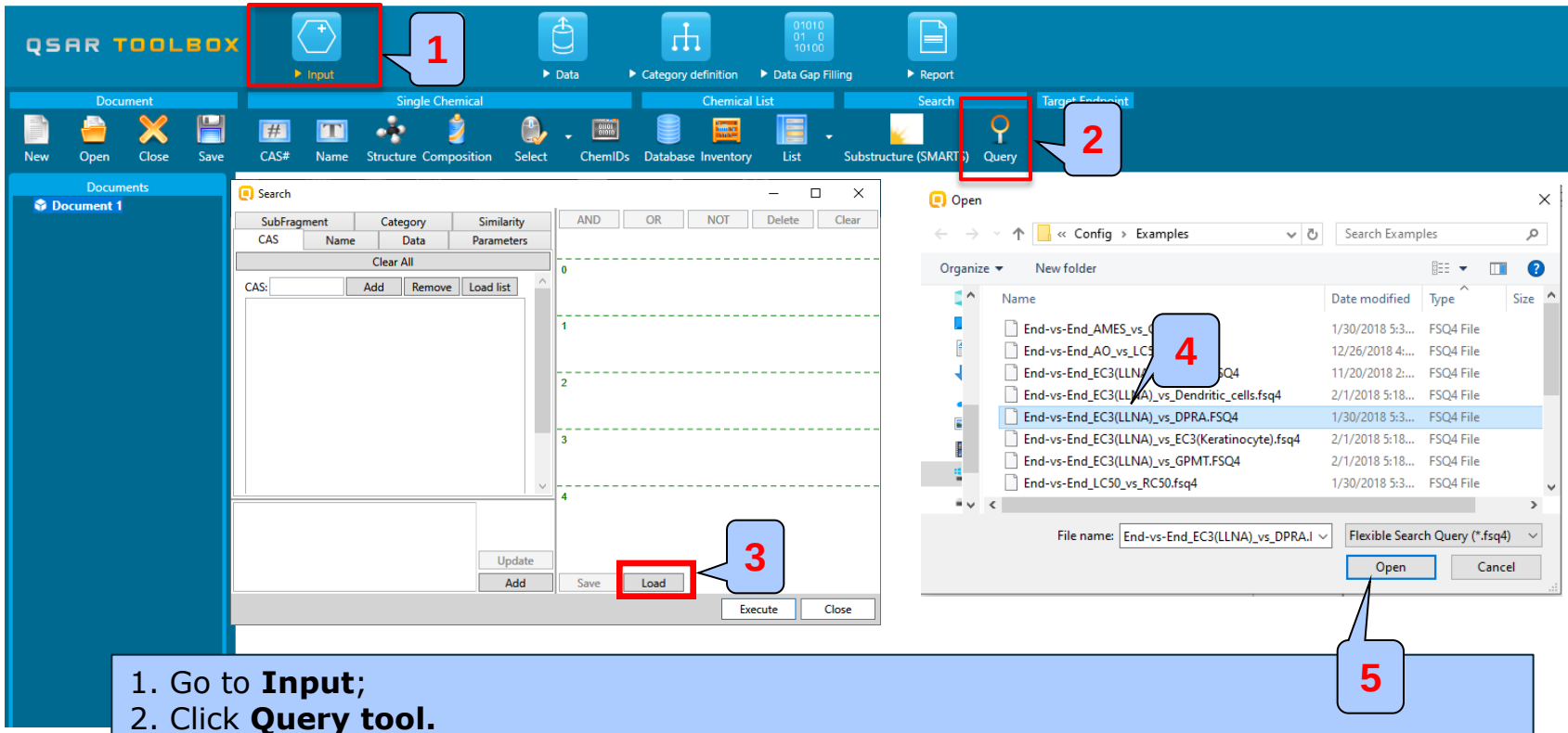
The purpose of performing of this correlation is to establish whether information from non-testing methods (DPRA, *in chemico* assay) provides sufficient evidence about a substance's skin sensitization potential as compared to that which has been elicited in an in vivo assay (LLNA).

Types endpoint correlations

Categorized continuous vs. categorical

Query Tool and select FSQ file - step 1

Example: Correlation between LLNA (Strongly positive, Weakly positive, Negative) and DPRA (Strongly positive, Weakly positive, Negative) data



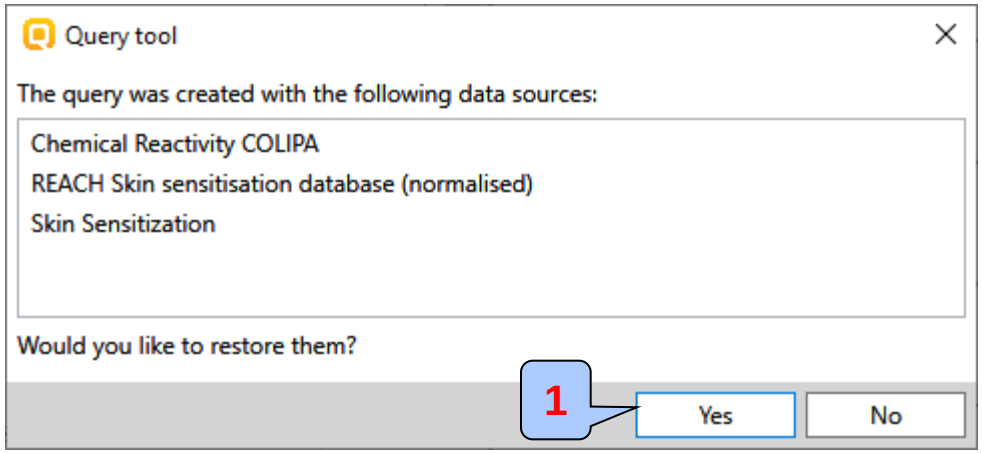
1. Go to **Input**;
2. Click **Query tool.**
3. Click **Load**
4. Select the *.fsq* file from example directory (End-vs-End_EC3(LLNA)_vs_DPRA.FSQ4)
5. Click **Open.**

Types endpoint correlations

Categorized continuous vs. categorical

Query Tool and select FSQ file - step 1

Example: Correlation between LLNA (Strongly positive, Weakly positive, Negative) and DPRA (%) data



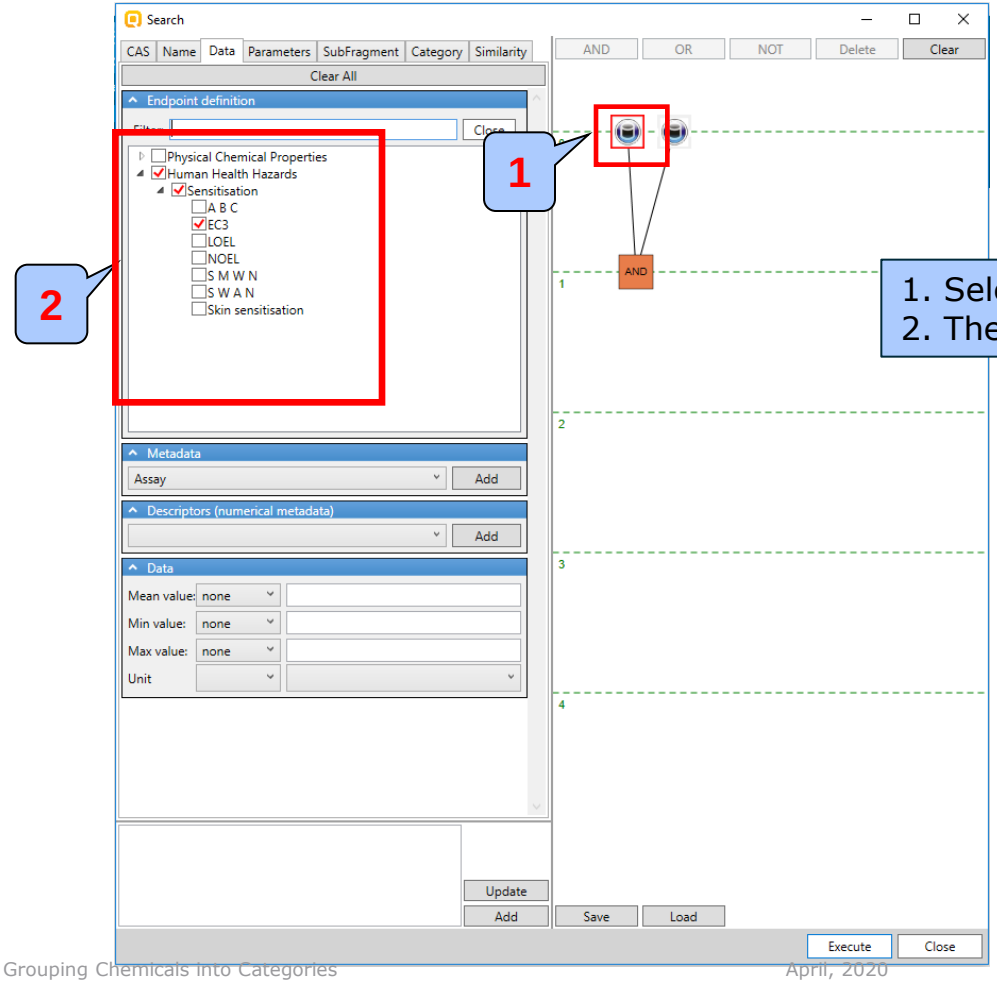
Click OK (1) in the message informing that the databases used to create the query will be restored.

Types endpoint correlations

Categorized continuous vs. categorical

Query Tool and select FSQ file - step 1

Example: Correlation between LLNA (Strongly positive, Weakly positive, Negative) and DPRA (%) data

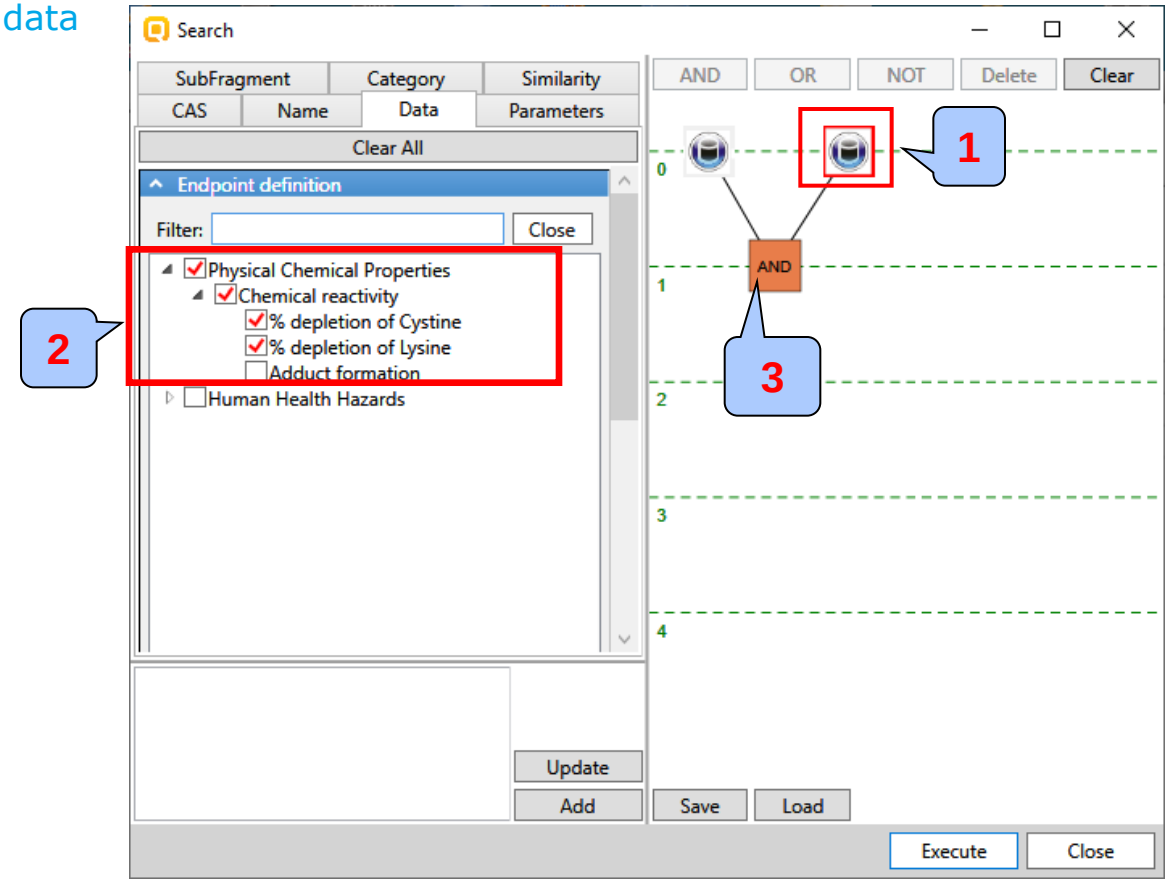


Types endpoint correlations

Categorized continuous vs. categorical

Query Tool and select FSQ file - step 1

Example: Correlation between LLNA (Strongly positive, Weakly positive, Negative) and DPRA (%) data



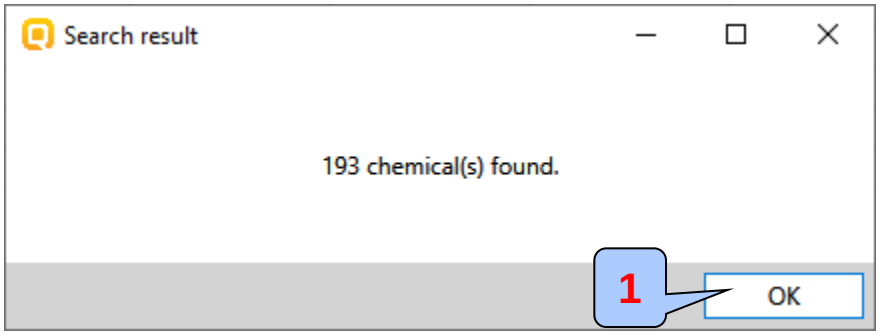
1. Select the second query;
2. The selected endpoint is *% depletion of Lysine* and *% depletion of Cystine*
3. The logical operand lining the two queries is **AND**. Double-click on it.

Types endpoint correlations

Categorized continuous vs. categorical

Query Tool and select FSQ file - step 1

Example: Correlation between LLNA (Strongly positive, Weakly positive, Negative) and DPRA (%) data



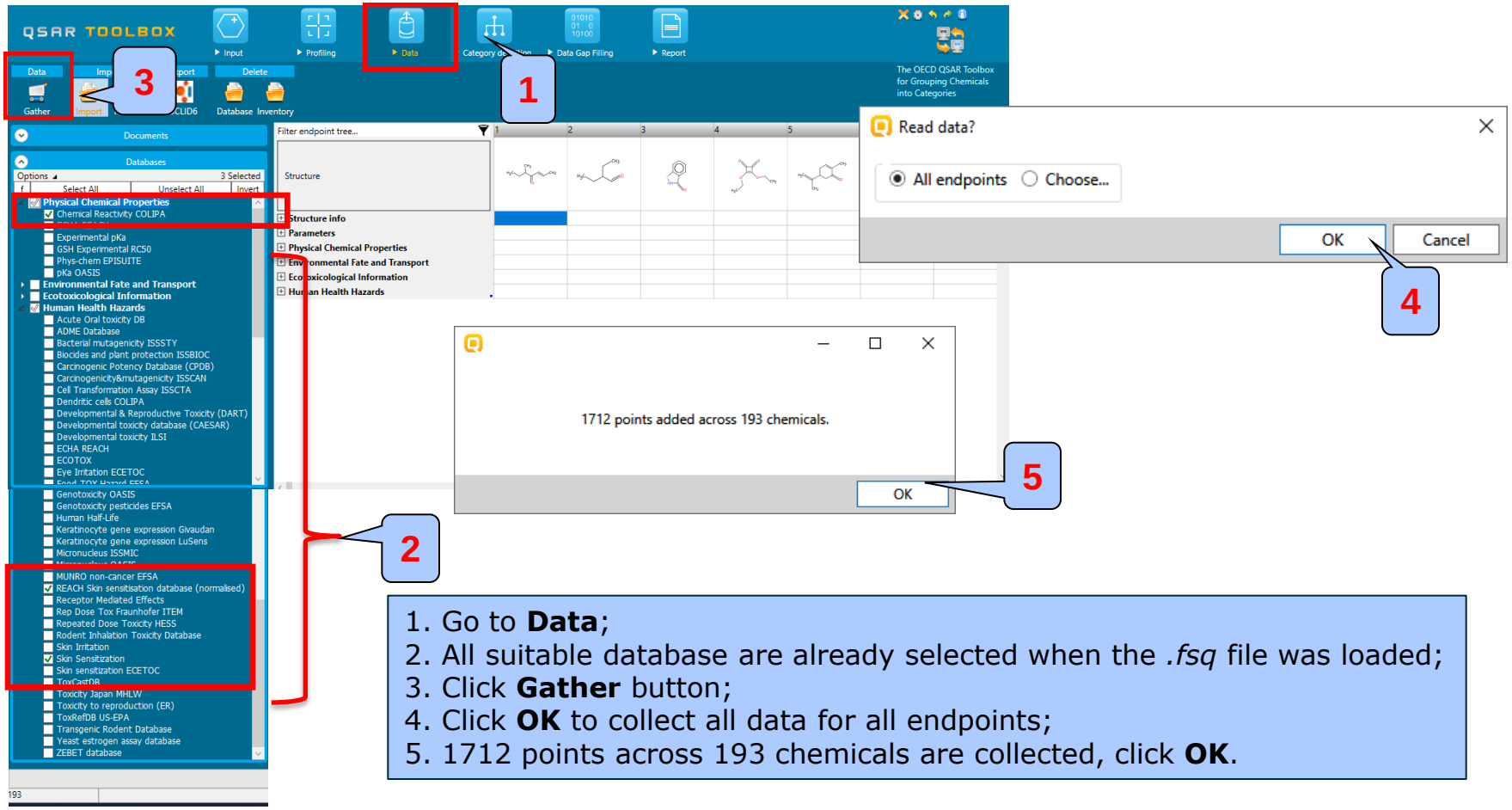
193 chemicals are found. Click OK (1).

Types endpoint correlations

Categorized continuous vs. categorical

Gather experimental data – step 2

Example: Correlation between LLNA (Strongly positive, Weakly positive, Negative) and DPRA (%) data



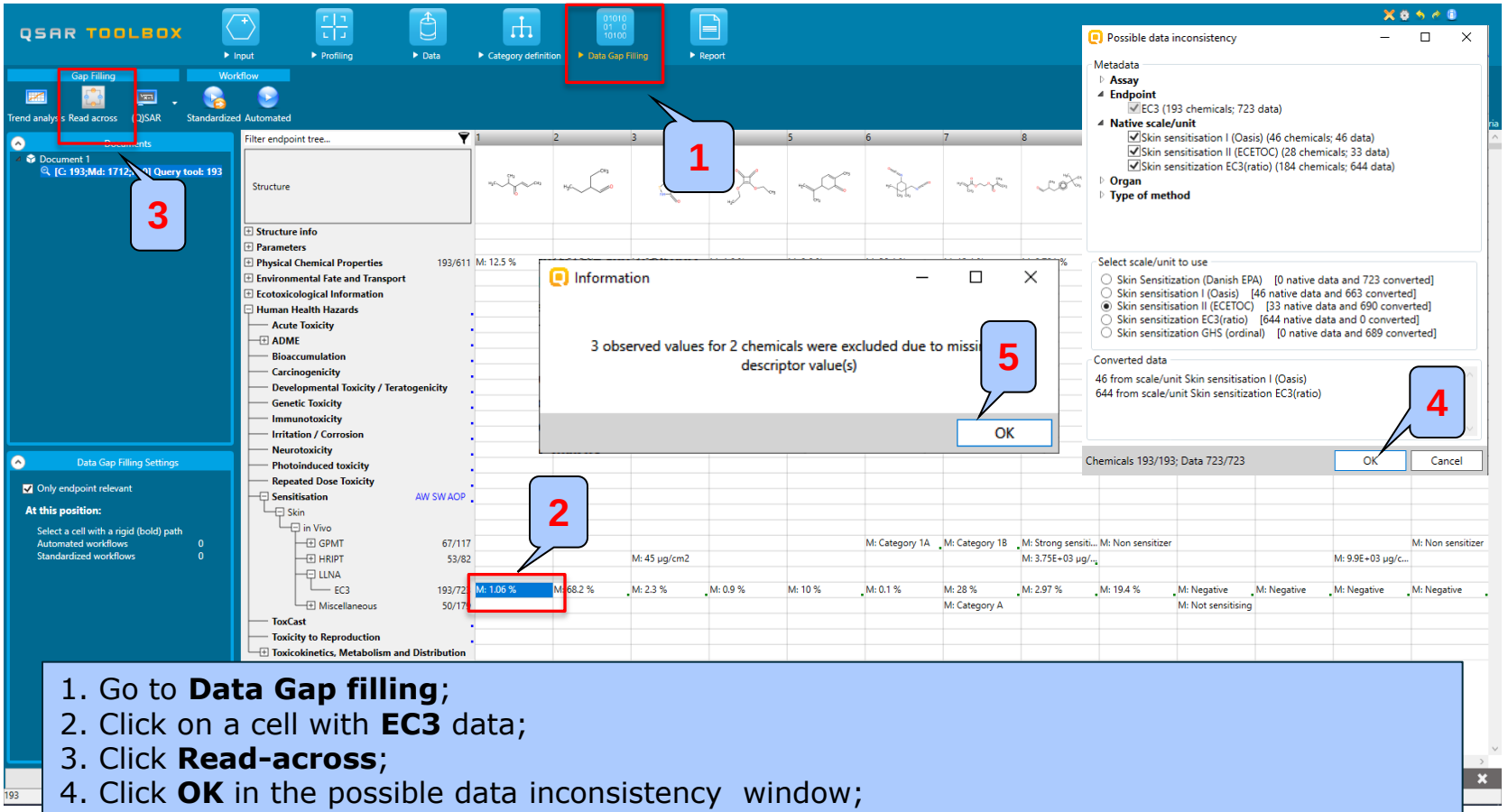
1. Go to **Data**;
2. All suitable database are already selected when the *.fsq* file was loaded;
3. Click **Gather** button;
4. Click **OK** to collect all data for all endpoints;
5. 1712 points across 193 chemicals are collected, click **OK**.

Types endpoint correlations

Categorized continuous vs. categorical

Enter Gap filling – step 4

Example: Correlation between LLNA (Strongly positive, Weakly positive, Negative) and DPRA (%) data



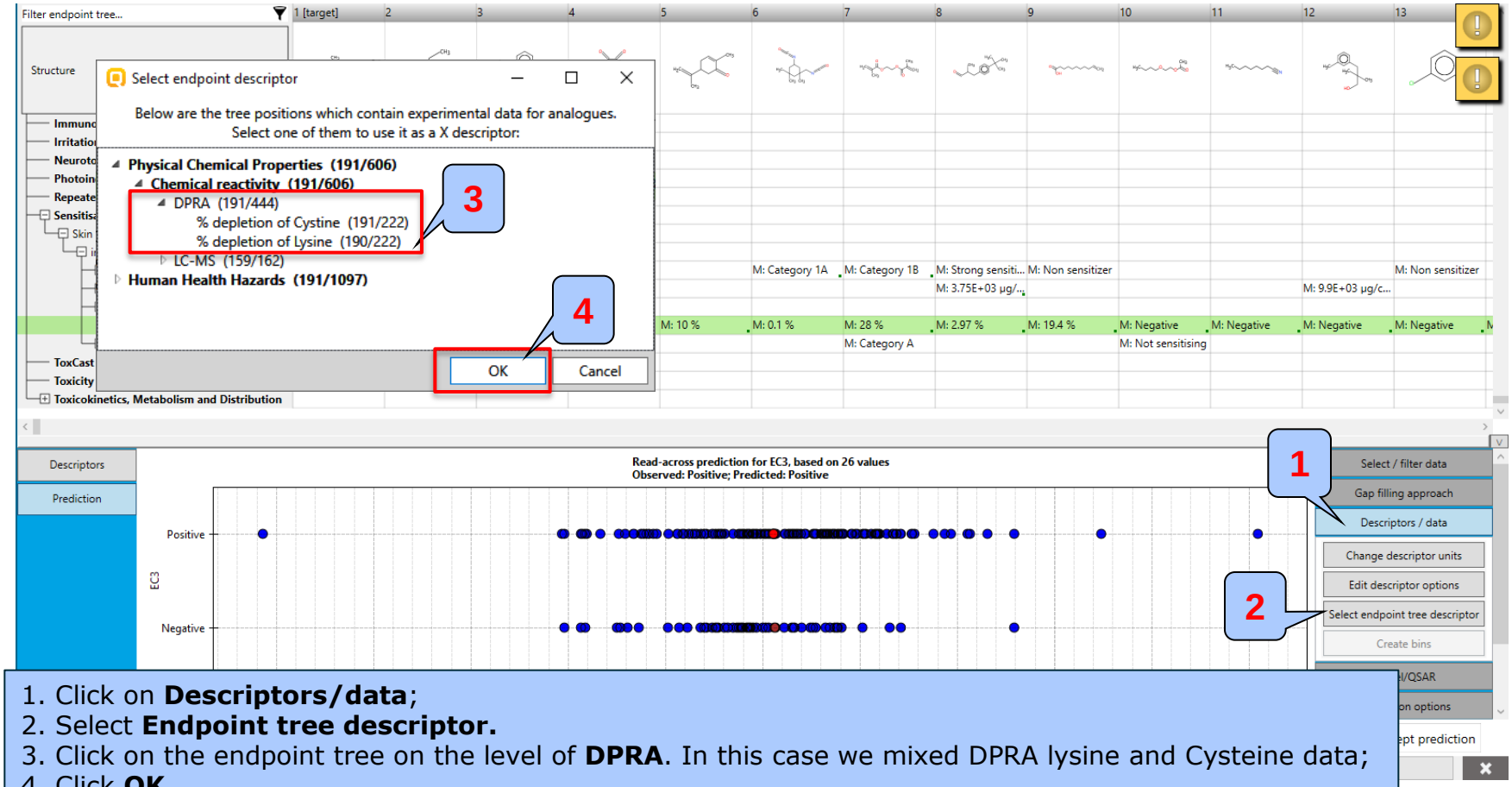
1. Go to **Data Gap filling**;
2. Click on a cell with **EC3** data;
3. Click **Read-across**;
4. Click **OK** in the possible data inconsistency window;
5. Click **OK** in the information window where the number of chemicals that will be excluded due to missing logKow value is shown.

Types endpoint correlations

Categorized continuous vs. categorical

Enter Gap filling – step 4

Example: Correlation between LLNA (Strongly positive, Weakly positive, Negative) and DPRA (%) data



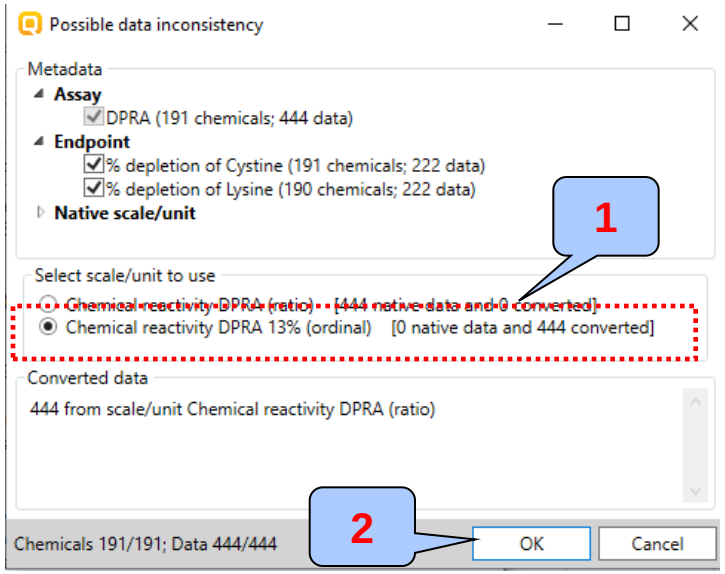
1. Click on **Descriptors/data**;
2. Select **Endpoint tree descriptor.**
3. Click on the endpoint tree on the level of **DPRA**. In this case we mixed DPRA lysine and Cysteine data;
4. Click **OK**.

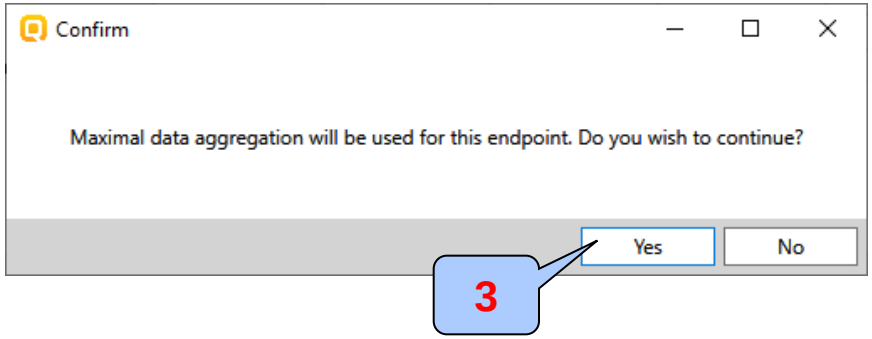
Types endpoint correlations

Categorized continuous vs. categorical

Perform correlation between LLNA and DPRA data – step 5

Example: Correlation between LLNA (Strongly positive, Weakly positive, Negative) and DPRA (%) data





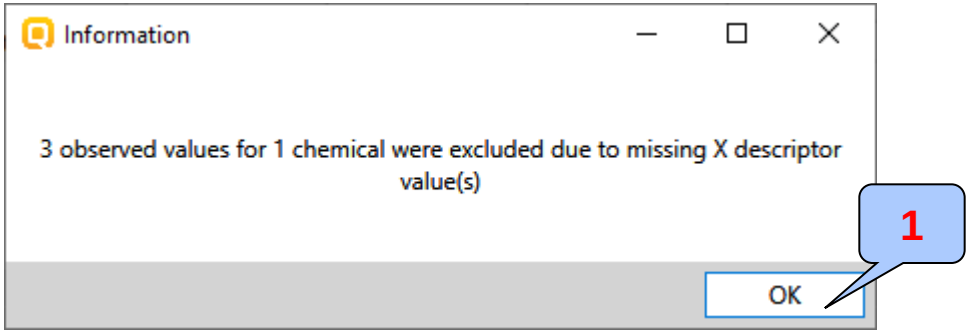
1. Select *Chemical reactivity DPRA 13% (ordinal) scale;*
2. Click **OK**;
3. Click **OK** in the pop-up message.

Types endpoint correlations

Categorized continuous vs. categorical

Perform correlation between LLNA and DPRA data – step 5

Example: Correlation between LLNA (Strongly positive, Weakly positive, Negative) and DPRA (%) data

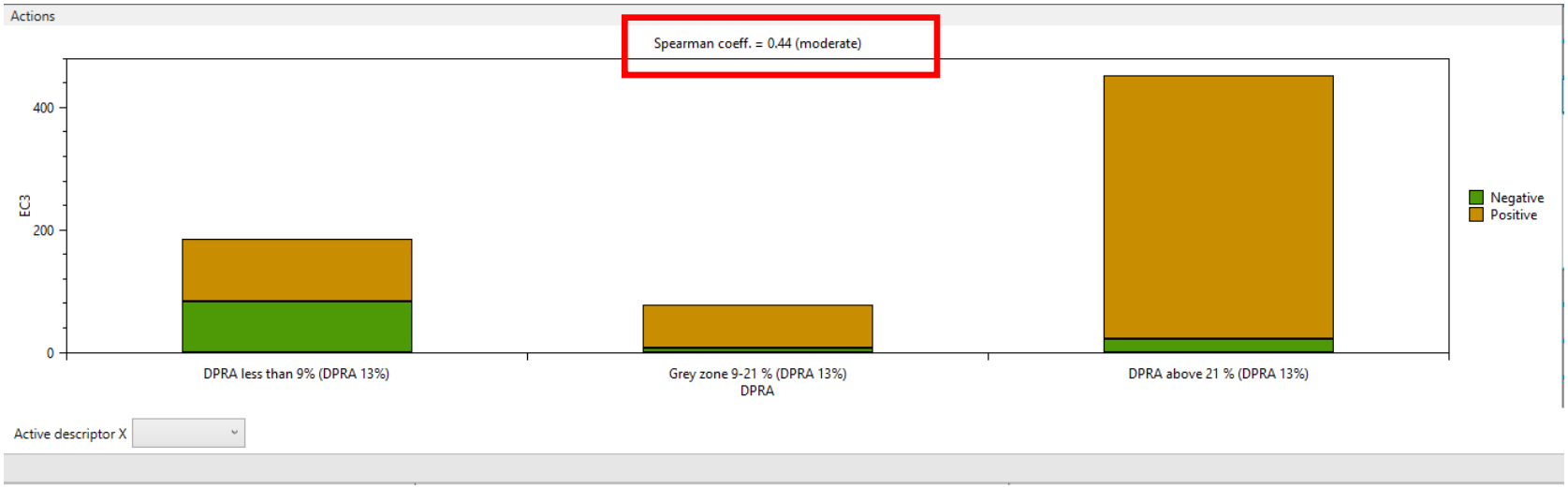


1. **Click** OK.

Types endpoint correlations

Categorized continuous vs. categorical

Interpretation of correlation results (LLNA vs. DPRA)



- In this example we have correlated continues DPRA (%) (plotted on the x axis) data distributed into 3 bins and categorical LLNA data (Strongly positive, Weakly positive, Negative)
  - Less than 9%
  - Grey zone 9 – 21%
  - Above 21%
- The high value of Spearman coefficient (0.44) shows moderate correlation between DPRA and LLNA data

Summary

- Different types of correlations have been illustrated in this tutorial based on the type of endpoint data:
  - Categorical vs. categorical:
  - Categorized continuous vs. categorical
- Correlation analysis has been evaluated by Spearman coefficient;
- Moderate endpoint correlations have been obtained for 2 out of 3 illustrated examples.