

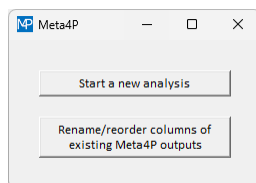
## User guide

v.1.5.5 (February 2024)

### 0. Launch the application and start a new analysis

Double-click on the Meta4P.exe (Windows) or Meta4P.app (MacOS) file to launch the application. The first time the program is launched, the operating system will ask for your explicit approval to execute the application. Click on "More info" to proceed with security checks and then on "Run anyway".

The opening window (see the image below) allows you to "**Start a new analysis**". Alternatively, if you have previously downloaded Meta4P outputs and just want to rename and/or reorder sample column headers in them, click on "**Rename/reorder sample columns of existing Meta4P outputs**" to go directly to the last Meta4P window (see section 7).

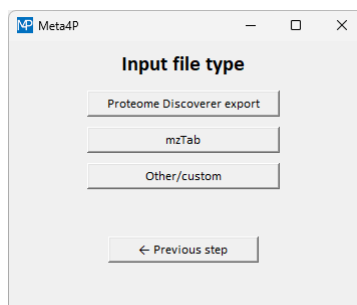


### 1. Identification/quantification input

As a first input, you have to provide a file containing identification and quantification information. Meta4P can handle different file types and data levels, for a total of nine combinations, as detailed below.

#### 1.1. Identification/quantification input: file type

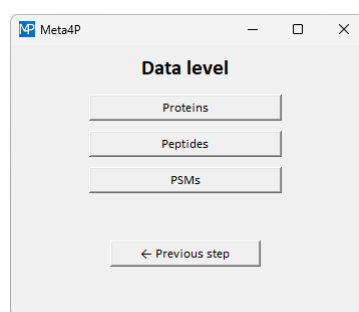
This window allows you to select the desired file type, clicking on one of the following buttons: "**Proteome Discoverer export**", "**mzTab**" or "**Other/custom**" (see the image below).



The three types of files accepted are: Proteome Discoverer files (in xlsx or txt format), mzTab files (one of the standard formats for proteomic data exchange) and generic tabular files (in xlsx, txt or another tab-separated format).

#### 1.2. Identification/quantification input: data level

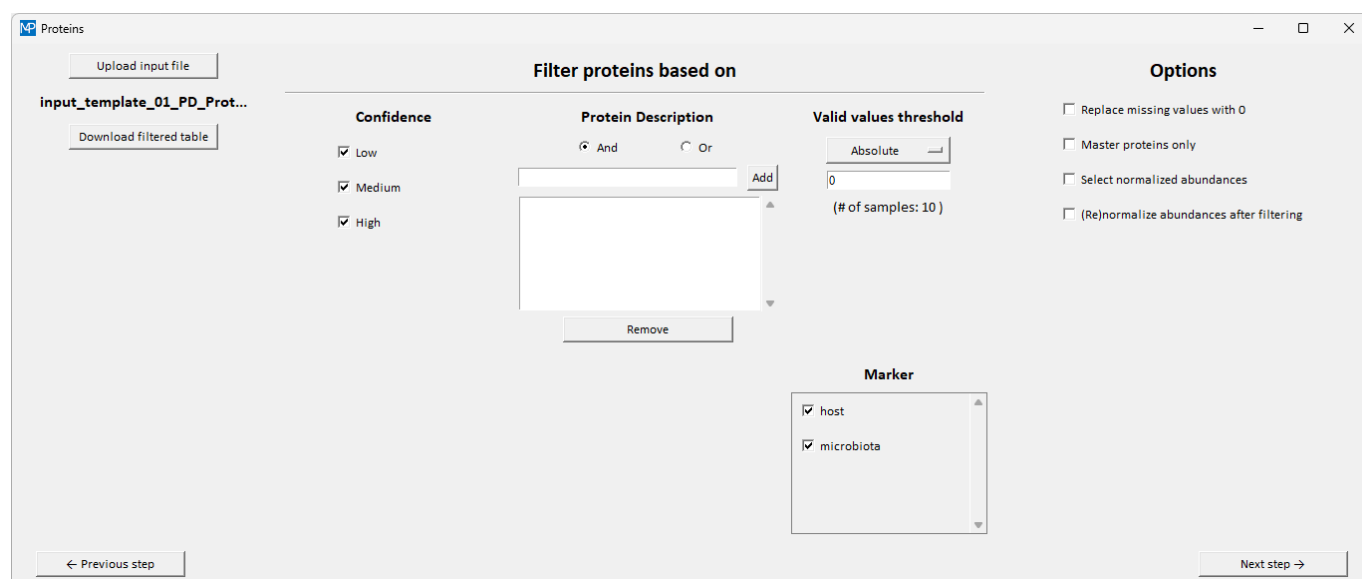
Once the file type has been selected, another window allows you to choose the data level (see the image below). The following buttons are shown, each corresponding to a different level of identification and quantification data: "**Proteins**" (protein identifications with MS1-based quantitative data), "**Peptides**" (peptide identifications with MS1-based quantitative data) and "**PSMs**" (peptide-spectrum matches to be used for spectral counting quantification).



### 1.3. Proteome Discoverer export - Proteins

When the "Proteome Discoverer export" and "Proteins" options are sequentially selected, protein identification and quantification data are retrieved from a "Proteins" file exported from Proteome Discoverer, available in one of the following formats: xlsx (Microsoft Excel) or txt (tab-separated values). The input file must contain the "Accession" column and at least one "Abundance" column. If you try to upload an input file with a wrong structure/format, an error message will be shown.

Click on "**Upload input file**" to select and upload the file containing protein identification and quantification data (a template file named *input\_template\_01\_PD\_Proteins.xlsx* is available for download). Once the file is uploaded, the file name is shown under the upload button and the window is populated with several filtering options based on the file content (see the image below).



Proteins can always be filtered based on the number/percentage of valid values ("**Valid values threshold**"), so that only proteins with a number of valid values (i.e., non-missing values) greater than or equal to the selected threshold are kept. You can choose between indicating absolute or percentage values by selecting the corresponding option in the drop-down menu. Leaving the default value (0) means that all proteins pass the filter. The total number of samples in the dataset is shown in brackets under the textbox.

Other optional filters might also be available (if the corresponding columns are present in the input file):

- **Confidence:** only proteins with the selected level(s) of statistical confidence (low, medium and/or high) are kept in the output, based on the checkbox(es) checked.
- **Protein Description:** to select only those proteins which contain a specific text (e.g., a protein name or an organism name) in their "Description" column, type the text of interest in the textbox (be aware that the filter is case sensitive) and click on "**Add**". Multiple texts can be typed and added sequentially;

in this case, you can choose between two boolean operators, "**And**" and "**Or**", to determine whether all the texts added or only one of them must be present in the string, respectively, so that a protein passes the filter.

- **Marker:** marker names included in the "Marked as" column of the input file, usually indicating which of the protein database(s) used for identification contained that protein sequence, are retrieved by the software and shown next to their respective checkboxes; only proteins annotated with the checked marker names are kept in the output.

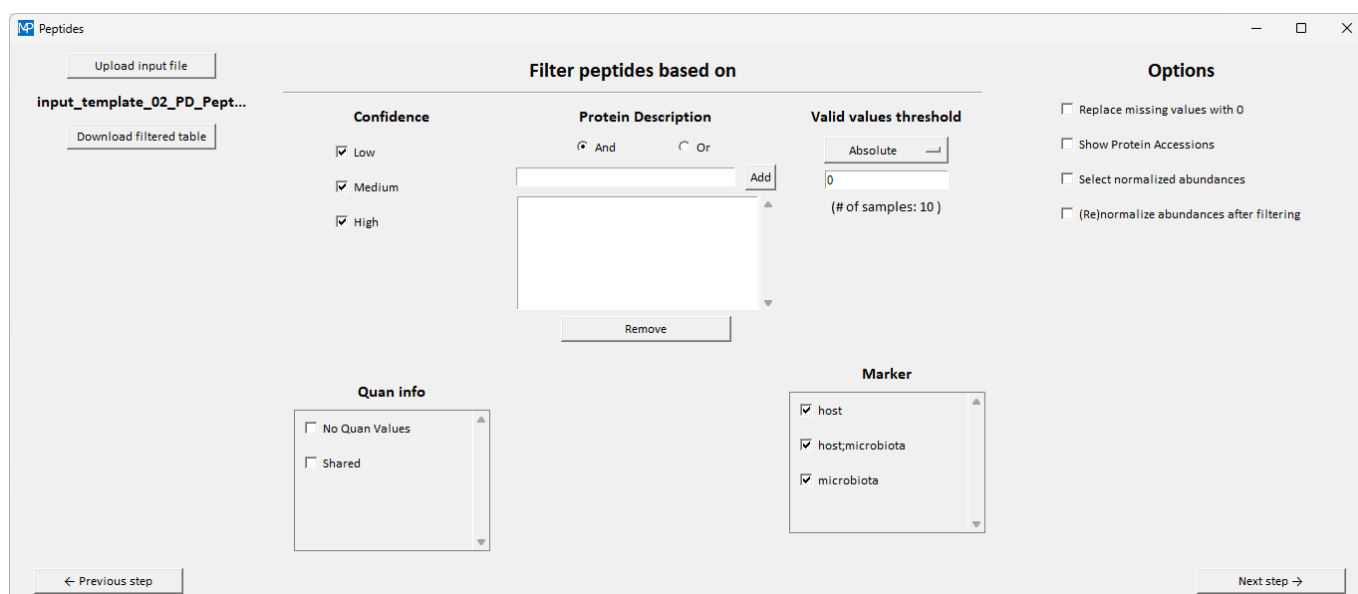
Furthermore, you can choose between the following visualization and calculation options:

- **Replace missing values with 0:** if selected, missing values (empty cells) are replaced by 0; this selection will be applied to all the following output tables.
- **Master proteins only:** if selected, only proteins designated as "Master Protein" – i.e., a protein identified by a set of peptides that are not included (all together) in any other protein group – are kept in the output (only available when the corresponding column is present in the input file).
- **Select normalized abundances:** if selected, normalized abundance values will be reported in the table (only available when also normalized abundance values are included in the input file).
- **(Re)normalize abundances after filtering:** if selected, once the chosen filters are applied and the filtered protein list is obtained, the abundance value measured for a protein in a given sample is divided by the total protein abundance measured in that sample and multiplied by  $10^{10}$ .

At the end, the (filtered) identification and quantification table can be downloaded in xlsx, txt or generic tab-separated format by clicking on "**Download filtered table**".

#### 1.4. Proteome Discoverer export - Peptides

When the "Proteome Discoverer export" and "Peptides" options are sequentially selected, peptide identification and quantification data are retrieved from a "Peptide Groups" file exported from Proteome Discoverer, available in one of the following formats: xlsx (Microsoft Excel) or txt (tab-separated values). The input file must contain the "Sequence" column, the "Master Protein Accessions" column and at least one "Abundance" column.



The screenshot shows the "Peptides" configuration window in the Meta4P software. The interface is divided into several sections:

- Input:** Includes an "Upload input file" button and a text field showing "input\_template\_02\_PD\_Pept...". Below it is a "Download filtered table" button.
- Filter peptides based on:**
  - Confidence:** Three checkboxes are checked: "Low", "Medium", and "High".
  - Protein Description:** Two radio buttons are present: "And" (selected) and "Or". Below them is a list box for adding filters, currently empty, with an "Add" button to the right and a "Remove" button below.
  - Valid values threshold:** A dropdown menu is set to "Absolute", and a text input field contains the value "0". Below this, it says "(# of samples: 10)".
- Options:** Four checkboxes are listed on the right:
  - ☐ Replace missing values with 0
  - ☐ Show Protein Accessions
  - ☐ Select normalized abundances
  - ☐ (Re)normalize abundances after filtering
- Quan info:** A section with two checkboxes: "No Quan Values" (unchecked) and "Shared" (unchecked).
- Marker:** A list box contains three items: "host", "host;microbiota", and "microbiota", all of which are checked.

At the bottom of the window, there are two navigation buttons: "← Previous step" on the left and "Next step →" on the right.

Note that when the same peptide sequence is listed more than once (e.g., when presenting different modifications or charge states), Meta4P will report it once and sum its related abundance values. If you try to upload an input file with a wrong structure/format, an error message will be shown.

Click on "**Upload input file**" to select and upload the file containing protein identification and quantification data (a template file named *input\_template\_02\_PD\_PeptideGroups.xlsx* is available for download). Once the file is uploaded, the file name is shown under the upload button and the window is populated with several filtering options based on the file content (see the image above).

Peptides can be always filtered based on the number/percentage of valid values ("**Valid values threshold**"), so that only peptides with a number of valid values (i.e., non-missing values) greater than or equal to the selected threshold are kept. You can choose between indicating absolute or percentage values by selecting the corresponding option in the drop-down menu. Leaving the default value (0) means that all peptides pass the filter. The total number of samples in the dataset is shown in brackets under the textbox.

Other optional filters might also be available (if the corresponding columns are present in the input file):

- **Confidence:** only peptides with the selected level(s) of statistical confidence (low, medium and/or high) are kept in the output, based on the checkbox(es) checked.
- **Master Protein Description:** to select only those peptides which belong to a Master Protein containing a specific text in its description (e.g., a protein name or an organism name), type the text of interest in the textbox (be aware that the filter is case sensitive) and click on "**Add**". Multiple texts can be typed and added sequentially; in this case, you can choose between two boolean operators, "**And**" and "**Or**", to determine whether all the texts added or only one of them must be present in the string, respectively, so that the peptide passes the filter.
- **Quan info:** only peptides belonging to the selected quantification categories are kept in the output.
- **Marker:** marker names included in the "Marked as" column of the input file, usually indicating which of the protein database(s) used for identification contained that peptide sequence, are retrieved by the software and shown next to their respective checkboxes; only peptides annotated with the checked marker names are kept in the output.

Furthermore, you can choose between the following visualization and calculation options:

- **Replace missing values with 0:** if selected, missing values (empty cells) are replaced by 0; this selection will be applied to all the following output tables.
- **Show Protein Accessions:** if selected, the "Protein Accessions" column (i.e., the column indicating the accession number of all the protein entries matching with a peptide, including non-master proteins) is included in the filtered table (only available when this column is present in the input file).
- **Select normalized abundances:** if selected, normalized abundance values will be reported in the table (only available when also normalized abundance values are included in the input file).
- **(Re)normalize abundances after filtering:** if selected, once the chosen filters are applied and the filtered peptide list is obtained, the abundance value measured for a peptide in a given sample is divided by the total peptide abundance measured in that sample and multiplied by  $10^{10}$ .

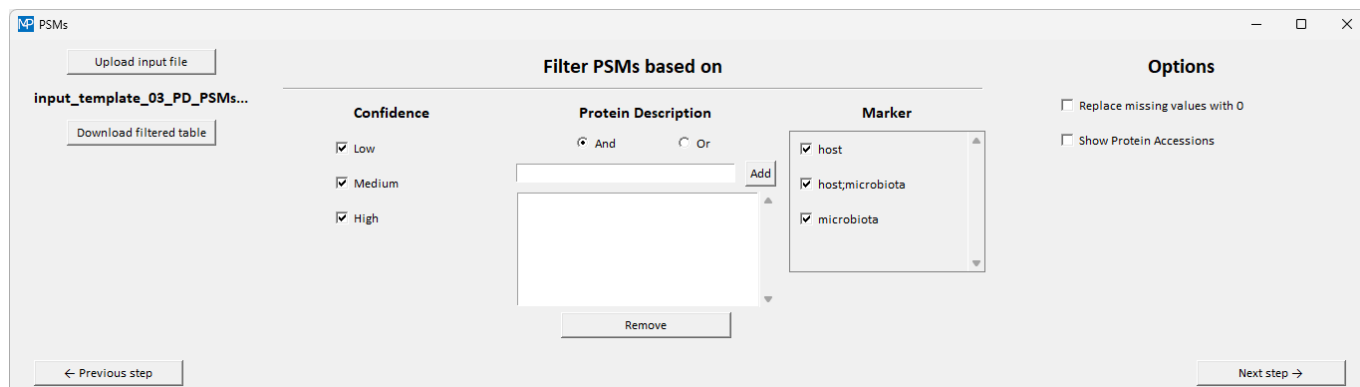
At the end, the (filtered) identification and quantification table can be downloaded in xlsx, txt or generic tab-separated format by clicking on "**Download filtered table**".

### 1.5. Proteome Discoverer export - PSMs

When the "Proteome Discoverer export" and "PSMs" options are sequentially selected, PSM data are retrieved from a "PSMs" file exported from Proteome Discoverer, available in one of the following formats: xlsx

(Microsoft Excel) or txt (tab-separated values). The input file must contain the "Sequence" column, the "Master Protein Accessions" column and the "File ID" column. If you try to upload an input file with a wrong structure/format, an error message will be shown.

Click on "**Upload input file**" to select and upload the file containing PSM data (a template file named *input\_template\_03\_PD\_PSMs.xlsx* is available for download). Once the file is uploaded, the file name is shown under the upload button and the window is populated with several filtering options based on the file content (see the image below).



Optional filters might be available (if the corresponding columns are present in the input file):

- **Confidence:** only PSMs with the selected level(s) of statistical confidence (low, medium and/or high) are kept in the output, based on the checkbox(es) checked.
- **Master Protein Description:** to select only those PSMs which belong to a Master Protein containing a specific text in its description (e.g., a protein name or an organism name), type the text of interest in the textbox (be aware that the filter is case sensitive) and click on "**Add**". Multiple texts can be typed and added sequentially; in this case, you can choose between two boolean operators, "**And**" and "**Or**", to determine whether all the texts added or only one of them must be present in the string, respectively, so that the PSM passes the filter.
- **Marker:** marker names included in the "Marked as" column of the input file, usually indicating which of the protein database(s) used for identification contained that PSM sequence, are retrieved by the software and shown next to their respective checkboxes; only PSMs annotated with the checked marker names are kept in the output.

Furthermore, you can choose between the following visualization and calculation options:

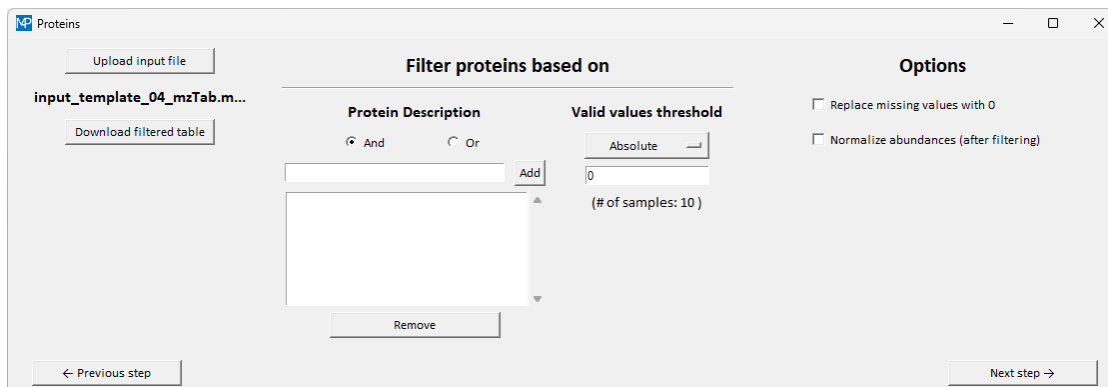
- **Replace missing values with 0:** if selected, missing values (empty cells) are replaced by 0; this selection will be applied to all the following output tables.
- **Show Protein Accessions:** if selected, the "Protein Accessions" column (i.e., the column indicating the accession number of all the protein entries matching with a PSM, including non-master proteins) is included in the filtered table (only available when this column is present in the input file).

At the end, Meta4P calculates the number of PSMs detected per sample and reports the corresponding values (next to the corresponding peptide sequence) in a tabular format. The (filtered) PSM table can be downloaded in xlsx, txt or generic tab-separated format by clicking on "**Download filtered table**".

## 1.6. mzTab - Proteins

When the "mzTab" and "Proteins" options are sequentially selected, protein identification and quantification data are retrieved from a standard mzTab input file. If you try to upload an input file with a wrong format, an error message will be shown.

Click on "**Upload input file**" to select and upload the file containing protein identification and quantification data (a template file named *input\_template\_04\_mzTab.mzTab* is available for download). Once the file is uploaded, the file name is shown under the upload button and the window is populated with several filtering options based on the file content (see the image below).



The screenshot shows the 'Proteins' window of the Meta4P application. It has three main sections:
 

- Upload:** An 'Upload input file' button with the filename 'input\_template\_04\_mzTab.m...' displayed below it. A 'Download filtered table' button is also present.
- Filter proteins based on:**
  - Protein Description:** Includes radio buttons for 'And' (selected) and 'Or', a text input field, an 'Add' button, and a 'Remove' button.
  - Valid values threshold:** A dropdown menu set to 'Absolute', a text input field with '0', and a note '(# of samples: 10)'.
- Options:** Two checkboxes: 'Replace missing values with 0' and 'Normalize abundances (after filtering)', both currently unchecked.

 Navigation buttons '← Previous step' and 'Next step →' are at the bottom.

Proteins can be filtered based on:

- the number/percentage of valid values ("**Valid values threshold**"), so that only proteins with a number of valid values (i.e., non-missing values) greater than or equal to the selected threshold are kept. You can choose between indicating absolute or percentage values by selecting the corresponding option in the drop-down menu. Leaving the default value (0) means that all proteins pass the filter. The total number of samples in the dataset is shown in brackets under the textbox.
- the presence of a specific text within the protein name/description ("**Protein Description**"), so that only proteins containing that specific text are kept. To do so, type the text of interest in the textbox (be aware that the filter is case sensitive) and click on "**Add**". Multiple texts can be typed and added sequentially; in this case, you can choose between two boolean operators, "**And**" and "**Or**", to determine whether all the texts added or only one of them must be present in the string, respectively, so that a protein passes the filter.

Furthermore, you can select the following options:

- Normalize abundances (after filtering):** if selected, once the chosen filters are applied and the filtered protein list is obtained, the abundance value measured for a protein in a given sample is divided by the total protein abundance measured in that sample and multiplied by  $10^{10}$ .
- Replace missing values with 0:** if selected, missing values (empty cells) are replaced by 0; this selection will be applied to all the following output tables.

At the end, the identification and quantification table can be downloaded in *xlsx*, *txt* or generic tab-separated format by clicking on "**Download filtered table**".

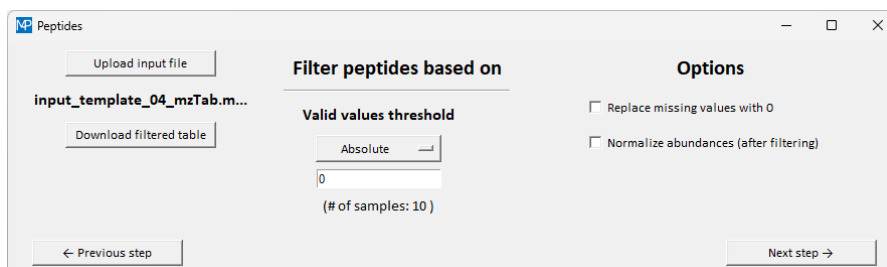
### 1.7. *mzTab* - Peptides

When the "*mzTab*" and "Peptides" options are sequentially selected, peptide identification and quantification data are retrieved from a standard *mzTab* input file.

Note that when the same peptide sequence is listed more than once (e.g., when presenting different modifications or charge states), Meta4P will report it once and sum its related abundance values. If you try to upload an input file with a wrong format, an error message will be shown.

Click on "**Upload input file**" to select and upload the file containing protein identification and quantification data (a template file named *input\_template\_04\_mzTab.mzTab* is available for download). Once the file is

uploaded, the file name is shown under the upload button and the window is populated with several filtering options based on the file content (see the image below).



Peptides can be filtered based on the number/percentage of valid values ("**Valid values threshold**"), so that only peptides with a number of valid values (i.e., non-missing values) greater than or equal to the selected threshold are kept. You can choose between indicating absolute or percentage values by selecting the corresponding option in the drop-down menu. Leaving the default value (0) means that all peptides pass the filter. The total number of samples in the dataset is shown in brackets under the textbox.

Furthermore, the following options can be selected:

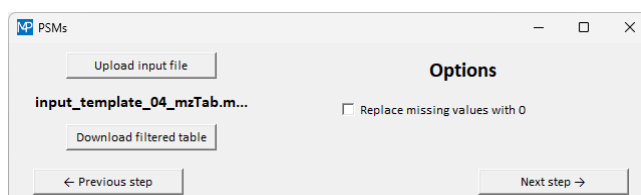
- **Normalize abundances (after filtering):** if selected, once the chosen filters are applied and the filtered peptide list is obtained, the abundance value measured for a peptide in a given sample is divided by the total peptide abundance measured in that sample and multiplied by  $10^{10}$ .
- **Replace missing values with 0:** if selected, missing values (empty cells) are replaced by 0; this selection will be applied to all the following output tables.

At the end, the identification and quantification table can be downloaded in xlsx, txt or generic tab-separated format by clicking on "**Download filtered table**".

### 1.8. mzTab - PSMs

When the "mzTab" and "PSMs" options are sequentially selected, PSM data are retrieved from a standard mzTab input file. If you try to upload an input file with a wrong format, an error message will be shown.

Click on "**Upload input file**" to select and upload the file containing PSM data (a template file named *input\_template\_04\_mzTab.mzTab* is available for download). Once the file is uploaded, the following window is shown:



If the "**Replace missing values with 0**" option is selected, missing values (empty cells) are replaced by 0. This selection will be applied to all the following output tables.

At the end, Meta4P calculates the number of PSMs detected per sample and reports the corresponding values (next to the corresponding peptide sequence) in a tabular format. The (filtered) PSM table can be downloaded in xlsx, txt or generic tab-separated format by clicking on "**Download filtered table**".

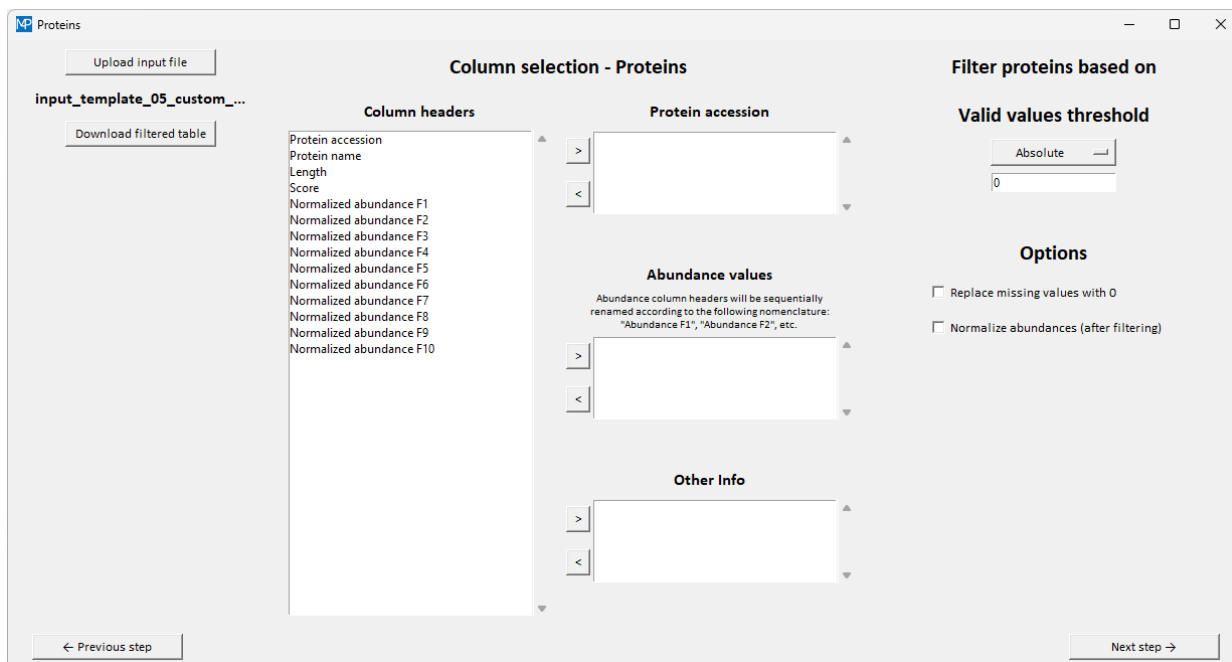
### 1.9. Other/custom - Proteins

When the "Other/custom" and "Proteins" options are sequentially selected, protein identification and quantification data can be retrieved from any tabular input file, in one of the following formats: xlsx, txt or



generic tab-separated format. Column headers must be contained in the first row. If you try to upload an input file with a wrong format, an error message will be shown.

Click on **"Upload input file"** to select and upload the file containing protein identification and quantification data (a template file named *input\_template\_05\_custom\_proteins.txt* is available for download). Once the file is uploaded, the file name is shown under the upload button and the file column headers are listed in the **"Column headers"** box (see the image below).



The screenshot shows the 'Proteins' interface of the Meta4P parser. It features a 'Column selection - Proteins' section with three main areas: 'Column headers', 'Protein accession', and 'Abundance values'. The 'Column headers' list includes 'Protein accession', 'Protein name', 'Length', 'Score', and ten 'Normalized abundance' columns (F1-F10). Arrows allow moving these headers to the other two sections. The 'Protein accession' section has a text input field. The 'Abundance values' section has a text input field and a note: 'Abundance column headers will be sequentially renamed according to the following nomenclature: "Abundance F1", "Abundance F2", etc.' Below these is an 'Other Info' section with another text input field. To the right, the 'Filter proteins based on' section includes a 'Valid values threshold' with a dropdown set to 'Absolute' and a text box containing '0'. At the bottom right, there are two checkboxes: 'Replace missing values with 0' and 'Normalize abundances (after filtering)'. Navigation buttons 'Previous step' and 'Next step' are at the bottom.

Select the columns of interests and move them to the corresponding box by using the arrows next to the boxes. At least one column listing protein accession numbers and one or more columns with abundance values must be selected and moved to the **"Protein accession"** and **"Abundance values"** boxes, respectively; other possible columns might remain unselected or be moved to the **"Other info"** box (in case you want them to be kept in the output table). Column reporting information useful for grouping (see "Marked as" column of Proteome Discoverer output) must have "Marked as" as header and semicolon as separator between multiple annotations.

Proteins can be filtered based on the number/percentage of valid values (**"Valid values threshold"**), so that only proteins with a number of valid values (i.e., non-missing values) greater than or equal to the selected threshold are kept. You can choose between indicating absolute or percentage values by selecting the corresponding option in the drop-down menu. Leaving the default value (0) means that all proteins pass the filter. The total number of samples in the dataset is shown in brackets under the textbox.

Furthermore, the following options can be selected:

- **Replace missing values with 0:** if selected, missing values (empty cells) are replaced by 0; this selection will be applied to all the following output tables.
- **Normalize abundances (after filtering):** if selected, once the chosen filters are applied and the filtered protein list is obtained, the abundance value measured for a protein in a given sample is divided by the total protein abundance measured in that sample and multiplied by  $10^{10}$ .

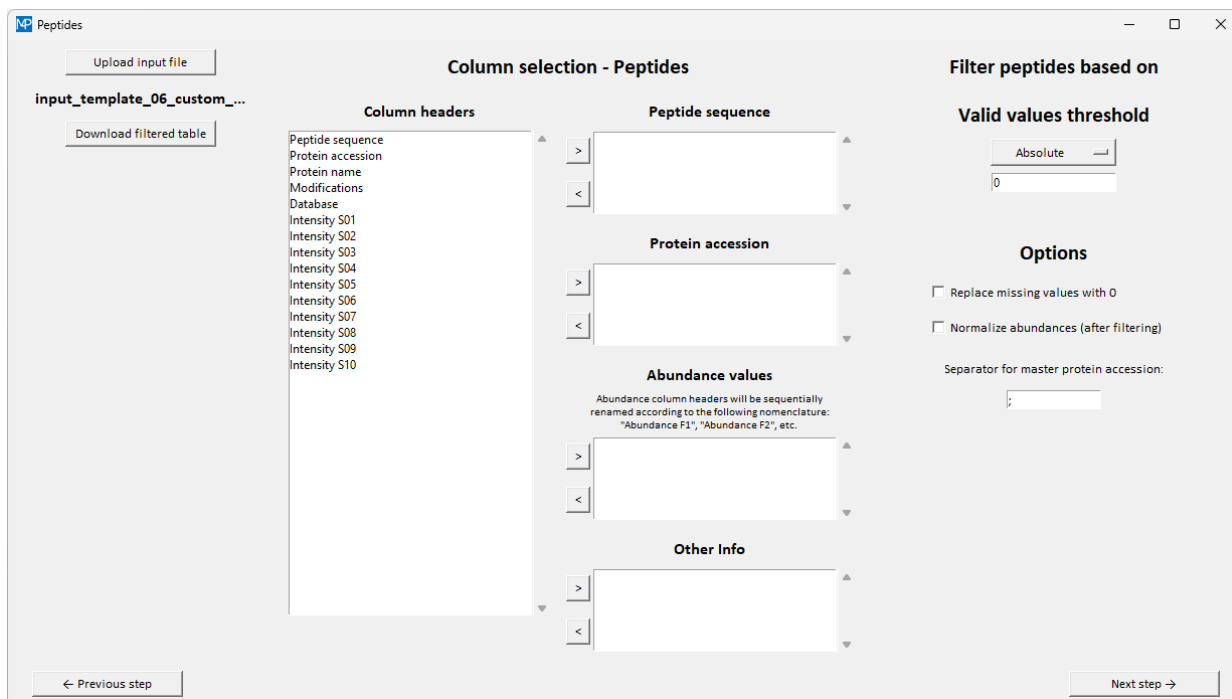
At the end, the identification and quantification table can be downloaded in xlsx, txt or generic tab-separated format by clicking on **"Download filtered table"**.



### 1.10. Other/custom - Peptides

When the "Other/custom" and "Peptides" options are sequentially selected, peptide identification and quantification data can be retrieved from any tabular input file, in one of the following formats: xlsx, txt or generic tab-separated format. Column headers must be contained in the first row. If you try to upload an input file with a wrong format, an error message will be shown.

Click on "**Upload input file**" to select and upload the file containing protein identification and quantification data (a template file named *input\_template\_06\_custom\_peptides.txt* is available for download). Once the file is uploaded, the file name is shown under the upload button and the file column headers are listed in the "**Column headers**" box (see the image below).



Select the columns of interests and move them to the corresponding box by using the arrows next to the boxes. At least one column listing peptide sequences, one column listing protein accessions (the separator used in case of multiple master protein accession and one or more columns with abundance values must be selected and moved to the "**Peptide sequence**", "**Protein accession**" and "**Abundance values**" boxes, respectively; other possible columns might remain unselected or be moved to the "**Other info**" box (in case you want them to be kept in the output table). Column reporting information useful for grouping (similarly to the "Marked as" column of Proteome Discoverer output) must have "Marked as" as header and semicolon as separator between multiple annotations.

Peptides can be filtered based on the number/percentage of valid values ("**Valid values threshold**"), so that only peptides with a number of valid values (i.e., non-missing values) greater than or equal to the selected threshold are kept. You can choose between indicating absolute or percentage values by selecting the corresponding option in the drop-down menu. Leaving the default value (0) means that all peptides pass the filter. The total number of samples in the dataset is shown in brackets under the textbox.

Furthermore, the following options can be selected:

- **Replace missing values with 0:** if selected, missing values (empty cells) are replaced by 0; this selection will be applied to all the following output tables.

- **Normalize abundances (after filtering):** if selected, once the chosen filters are applied and the filtered peptide list is obtained, the abundance value measured for a peptide in a given sample is divided by the total peptide abundance measured in that sample and multiplied by  $10^{10}$ .

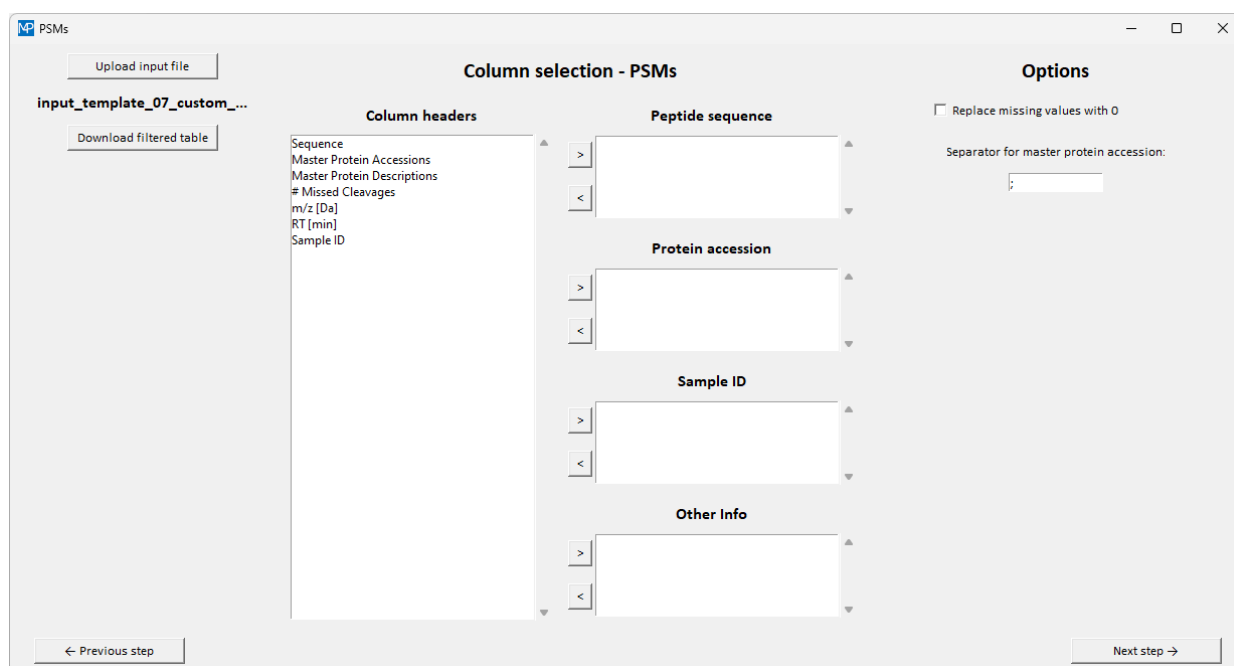
Moreover, it is possible to specify which separator (including comma, semicolon and space) has been used in the "Protein accession" column of the input file in case of multiple protein accessions.

At the end, the identification and quantification table can be downloaded in xlsx, txt or generic tab-separated format by clicking on **"Download filtered table"**.

### 1.11. Other/custom - PSMs

When the "Other/custom" and "PSMs" options are sequentially selected, PSM data can be retrieved from any tabular input file, in one of the following formats: xlsx, txt or generic tab-separated format. Column headers must be contained in the first row. If you try to upload an input file with a wrong format, an error message will be shown.

Click on **"Upload input file"** to select and upload the file containing PSM data (a template file named *input\_template\_07\_custom\_PSMs.txt* is available for download). Once the file is uploaded, the file name is shown under the upload button and the file column headers are listed in the **"Column headers"** box (see the image below).



The screenshot shows the 'PSMs' window of the Meta4P software. It features three main sections: 'Upload input file', 'Column selection - PSMs', and 'Options'.

- Upload input file:** Contains a button labeled 'Upload input file' and a text field showing the filename 'input\_template\_07\_custom\_...'. Below it is a button labeled 'Download filtered table'.
- Column selection - PSMs:** This section allows users to select columns from an input file. On the left, under 'Column headers', is a list of available columns: Sequence, Master Protein Accessions, Master Protein Descriptions, # Missed Cleavages, m/z [Da], RT [min], and Sample ID. On the right, there are four boxes for selecting columns: 'Peptide sequence', 'Protein accession', 'Sample ID', and 'Other Info'. Each box has arrows to move columns between the list and the box.
- Options:** Contains a checkbox 'Replace missing values with 0' and a text field for 'Separator for master protein accession:'.

Navigation buttons 'Previous step' and 'Next step' are located at the bottom left and right respectively.

Select the columns of interests and move them to the corresponding box by using the arrows next to the boxes. At least one column listing peptide sequences, one column listing protein accessions and one column with sample IDs must be selected and moved to the **"Peptide sequence"**, **"Protein accession"** and **"Sample ID"** boxes, respectively; other possible columns might remain unselected or be moved to the **"Other info"** box (in case you want them to be kept in the output table). Column reporting information useful for grouping (similarly to the "Marked as" column of Proteome Discoverer output) must have "Marked as" as header and semicolon as separator between multiple annotations.

If the **"Replace missing values with 0"** option is selected, missing values (empty cells) are replaced by 0. This selection will be applied to all the following output tables.

Moreover, it is possible to specify which separator has been used in the "Protein accession" column of the input file in case of multiple protein accessions.

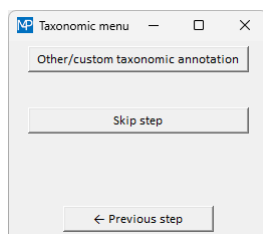
At the end, Meta4P calculates the number of PSMs detected per sample and reports the corresponding values (next to the corresponding peptide sequence) in a tabular format. The (filtered) PSM table can be downloaded in xlsx, txt or generic tab-separated format by clicking on "**Download filtered table**".

## 2. Taxonomic annotation

In this (optional) step, a second input file can be uploaded to retrieve taxonomic annotation data and include them in the table containing identification and quantification data.

### 2.1. Proteins

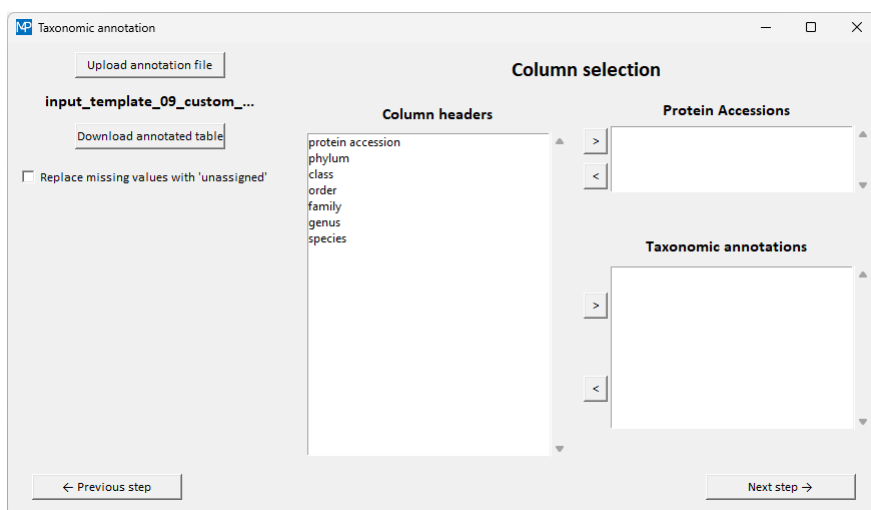
After loading protein identification and quantification data as first input file and clicking on "**Next step**", a "Taxonomic menu" window is shown (see the image below).



Go ahead with the upload of a taxonomic annotation file clicking on "**Other/custom taxonomic annotation**". Alternatively, click on "**Skip step**" if no taxonomic annotation is available.

#### 2.1.1. Other/custom taxonomic annotation

This window allows you to retrieve taxonomic annotation data by uploading any tabular input file, in one of the following formats: xlsx, txt or generic tab-separated format. Column headers must be contained in the first row. If you try to upload an input file with a wrong format, an error message will be shown.



Click on "**Upload annotation file**" to select and upload the file containing protein taxonomic annotation data (a template file named *input\_template\_09\_custom\_protein\_taxonomic\_annotation.xlsx* is available for download). Once the file is uploaded, the file name is shown under the upload button and the file column headers are listed in the "Column headers" box (see the image above).

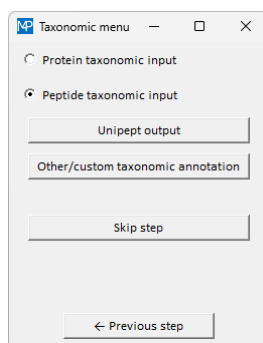
Select the columns of interests and move them to the corresponding box by using the arrows next to the boxes. At least one column listing protein accessions and one column listing taxonomic annotations must be selected and moved to the "**Accession**" and "**Taxonomic annotations**" boxes, respectively. Unnecessary columns, if any, should remain unselected.

In addition, an option named "**Replace missing values with 'unassigned'**" allows you to keep missing annotations as empty cells or, if selected, to denominate them as "unassigned"; only in the latter case the quantitative values related to missing annotations are considered in the following data aggregation step.

An annotated output table, combining taxonomic annotations with identification and quantification data, can be downloaded in xlsx, txt or generic tab-separated format by clicking on "**Download annotated table**".

## 2.2. Peptides/PSMs

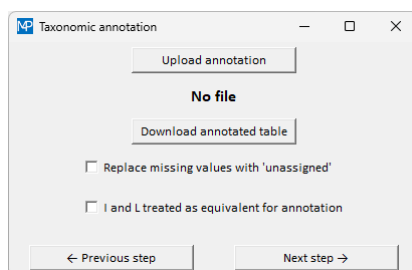
After loading peptide/PSM identification and quantification data as first input file and clicking on "**Next step**", a "Taxonomic menu" window is shown (see the image below).



First, select if the taxonomic input is at protein or peptide level. Then, choose the file type: standard Unipept output table ("**Unipept output**" button, which is available only when the peptide level has been selected) or generic functional annotation tabular file ("**Other/custom functional annotation**" button). Alternatively, if no taxonomic annotation is available, this step can be skipped by clicking on "**Skip step**".

### 2.2.1. Unipept output

This window (see the image below) allows you to retrieve taxonomic annotation data from a standard Unipept tabular output. If you try to upload an input file with a wrong format, an error message will be shown.



Click on "**Upload annotation**" to select and upload the file containing peptide taxonomic annotation data (a template file named *input\_template\_08\_Unipept\_peptide\_annotation.csv* is available for download). Columns containing the main taxonomic annotations (LCA, superkingdom, phylum, class, order, family, genus, species) will be retrieved.

An option named "**Replace missing values with 'unassigned'**" allows you to keep missing annotations as empty cells or, if selected, to denominate them as "unassigned"; only in the latter case the quantitative values related to missing annotations are considered in the following data aggregation step. Another option, named "**I and L**

**treated as equivalent for annotation"**, must be checked if the "Equate I and L" option was selected before starting the Unipept analysis.

An annotated output table, combining taxonomic annotations with identification and quantification data, can be downloaded in xlsx, txt or generic tab-separated format by clicking on **"Download annotated table"**.

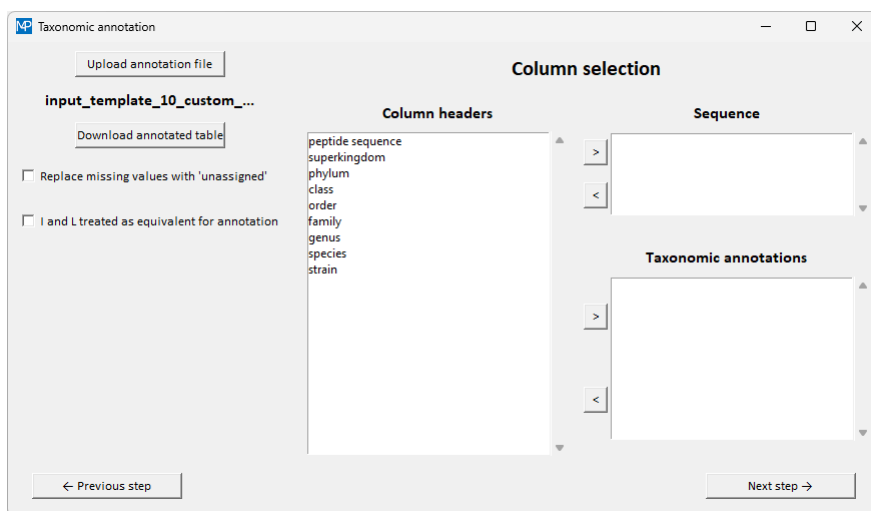
### 2.2.2. Other/custom taxonomic annotation

This window allows you to retrieve protein taxonomic annotation data by uploading any tabular input file, in one of the following formats: xlsx, txt or generic tab-separated format. Column headers must be contained in the first row. If you try to upload an input file with a wrong format, an error message will be shown.

If you selected "Protein taxonomic input", windows and options are those described in sections 2.1.1.

If you selected "Peptide taxonomic input", a window is shown that allows you to retrieve peptide taxonomic annotation data by uploading any tabular input file, in one of the following formats: xlsx, txt or generic tab-separated format. Column headers must be contained in the first row. If you try to upload an input file with a wrong format, an error message will be shown.

Click on **"Upload annotation file"** to select and upload the file containing peptide taxonomic annotation data (a template file named *input\_template\_10\_custom\_peptide\_taxonomic\_annotation.tab* is available for download). Once the file is uploaded, the file name is shown under the upload button and file column headers are listed in the "Column headers" box (see the image below).



Select the columns of interests and move them to the corresponding box by using the arrows next to the boxes. At least one column listing peptide sequences and one column listing taxonomic annotations must be selected and moved to the **"Sequence"** and **"Taxonomic annotations"** boxes, respectively. Unnecessary columns, if any, should remain unselected.

An option named **"Replace missing values with 'unassigned'"** allows you to keep missing annotations as empty cells or, if selected, to denominate them as "unassigned"; only in the latter case the quantitative values related to missing annotations are considered in the following data aggregation step. Another option, named **"I and L treated as equivalent for annotation"**, must be checked if isoleucine and leucine were considered as equivalent for taxonomic annotation.

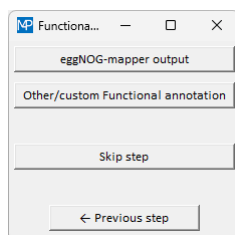
An annotated output table, combining taxonomic annotations with identification and quantification data, can be downloaded in xlsx, txt or generic tab-separated format by clicking on **"Download annotated table"**.

### 3. Functional annotation

In this (optional) step, another input file can be uploaded to retrieve functional annotation data and include them in the table containing identification and quantification (and optionally taxonomic annotation) data.

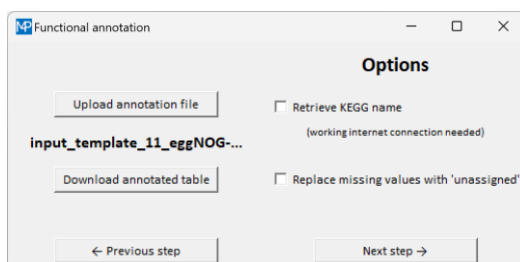
#### 3.1. Proteins

After loading taxonomic annotation and clicking on "**Next step**" (or after skipping taxonomic annotation), a "Functional menu" window is shown (see the image below). Here, you can choose between two file types: standard eggNOG-mapper output ("**eggNOG-mapper output**" button) or generic functional annotation tabular file ("**Other/custom functional annotation**" button). Alternatively, if no functional annotation is available, this step can be skipped by clicking on "**Skip step**".



##### 3.1.1 eggNOG-mapper output

This window (see the image below) allows you to retrieve taxonomic annotation data from a standard eggNOG-mapper output. If you try to upload an input file with a wrong format, an error message will be shown.



Click on "**Upload annotation**" to select and upload the file containing protein functional annotation data (a template file named *input\_template\_11\_eggNOG-mapper\_protein\_functional\_annotation* is available for download). Columns containing the functional levels provided by the input file (COG category, GO category, EC number, CAZy code, as well as KEGG KO, Pathway, Module and Reaction annotations) will be retrieved.

As an option, to retrieve and include in the table (as supplementary columns) the annotation names provided by the KEGG database for all KEGG categories, click on "**Retrieve KEGG name**". As this information is retrieved from the KEGG website, a working internet connection is needed in order that this operation is performed; this operation may take up to a few minutes. In case a protein has multiple functional annotations, their names will be separated by a comma as well as their codes (except for the "KEGG Module" annotation names, for which a vertical bar is used to avoid mistakes). In case a peptide is associated to multiple Master Proteins (usually separated by a semicolon), codes and names of functional annotations assigned to different Master Proteins are consistently separated by a semicolon (except for the "KEGG Reaction" annotation names, for which a vertical bar is used to avoid mistakes). When two analogous "KEGG Pathway" annotations are reported (i.e., having the same numeric code but two different prefixes, namely 'map' and 'ko'), the code with the 'ko' prefix is removed.

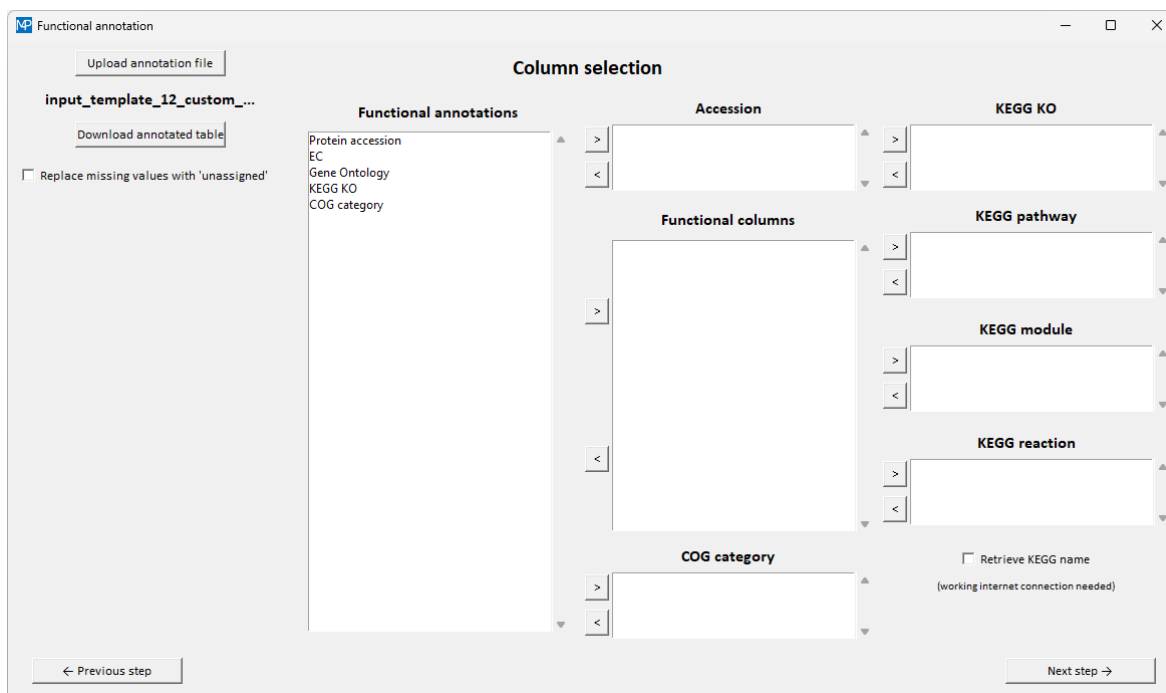
A further option, named "**Replace missing values with 'unassigned'**", allows you to keep missing annotations as empty cells or, if selected, to denominate them as "unassigned"; only in the latter case the quantitative values related to missing annotations are considered in the following data aggregation step.

An annotated output table, combining functional annotations with identification and quantification (and possibly taxonomic annotation) data, can be downloaded in xlsx, txt or generic tab-separated format by clicking on "**Download annotated table**".

### 3.1.2. Other/custom functional annotation

This window allows you to retrieve functional annotation data by uploading any tabular input file, in one of the following formats: xlsx, txt or generic tab-separated format. Column headers must be contained in the first row. Multiple functional annotations must be separated by a comma. If you try to upload an input file with a wrong format, an error message will be shown.

Click on "**Upload annotation file**" to select and upload the file containing protein functional annotation data (a template file named *input\_template\_12\_custom\_protein\_functional\_annotation.tsv* is available for download). Once the file is uploaded, the file name is shown under the upload button and the file column headers are listed in the "Column headers" box (see the image below).



Select the columns of interests and move them to the corresponding box by using the arrows next to the boxes. At least one column listing protein accessions and one column listing functional annotations must be selected and moved to the "**Accession**" box and to one of the functional annotation boxes, respectively; when columns listing KEGG or COG annotations are present in the input file, these have to be moved to the specific box corresponding to their category. Unnecessary columns, if any, should remain unselected.

As an option, to retrieve and include in the table (as supplementary columns) the annotation names provided by the KEGG database for all KEGG categories, click on "**Retrieve KEGG name**". As this information is retrieved from the KEGG website, a working internet connection is needed in order that this operation is performed; this operation may take up to a few minutes. In case a protein has multiple functional annotations, their names will be separated by a comma as well as their codes (except for the "KEGG Module" annotation names, for which a vertical bar is used to avoid mistakes). In case a peptide is associated to multiple Master Proteins



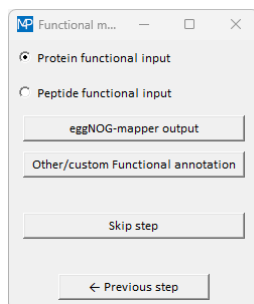
(usually separated by a semicolon), codes and names of functional annotations assigned to different Master Proteins are consistently separated by a semicolon (except for the "KEGG Reaction" annotation names, for which a vertical bar is used to avoid mistakes). When two analogous "KEGG Pathway" annotations are reported (i.e., having the same numeric code but two different prefixes, namely 'map' and 'ko'), the code with the 'ko' prefix is removed.

A further option, named "**Replace missing values with 'unassigned'**", allows you to keep missing annotations as empty cells or, if selected, to denominate them as "unassigned"; only in the latter case the quantitative values related to missing annotations are considered in the following data aggregation step.

An annotated output table, combining functional annotations with identification and quantification (and possibly taxonomic annotation) data, can be downloaded in xlsx, txt or generic tab-separated format by clicking on "**Download annotated table**".

### 3.2. Peptides/PSMs

After loading taxonomic annotation and clicking on "**Next step**" (or after skipping taxonomic annotation), a "Functional menu" window is shown (see the image below). First, select if the functional input is at protein or peptide level (in the former case, windows and options are those described in sections 3.1.1 and 3.1.2; for the latter case, see sections 3.2.1 and 3.2.2). Then, choose the file type: standard eggNOG-mapper output ("**eggNOG-mapper output**" button) or generic functional annotation tabular file ("**Other/custom functional annotation**" button).



Alternatively, if no functional annotation is available, this entire step can be skipped by clicking on "**Skip step**".

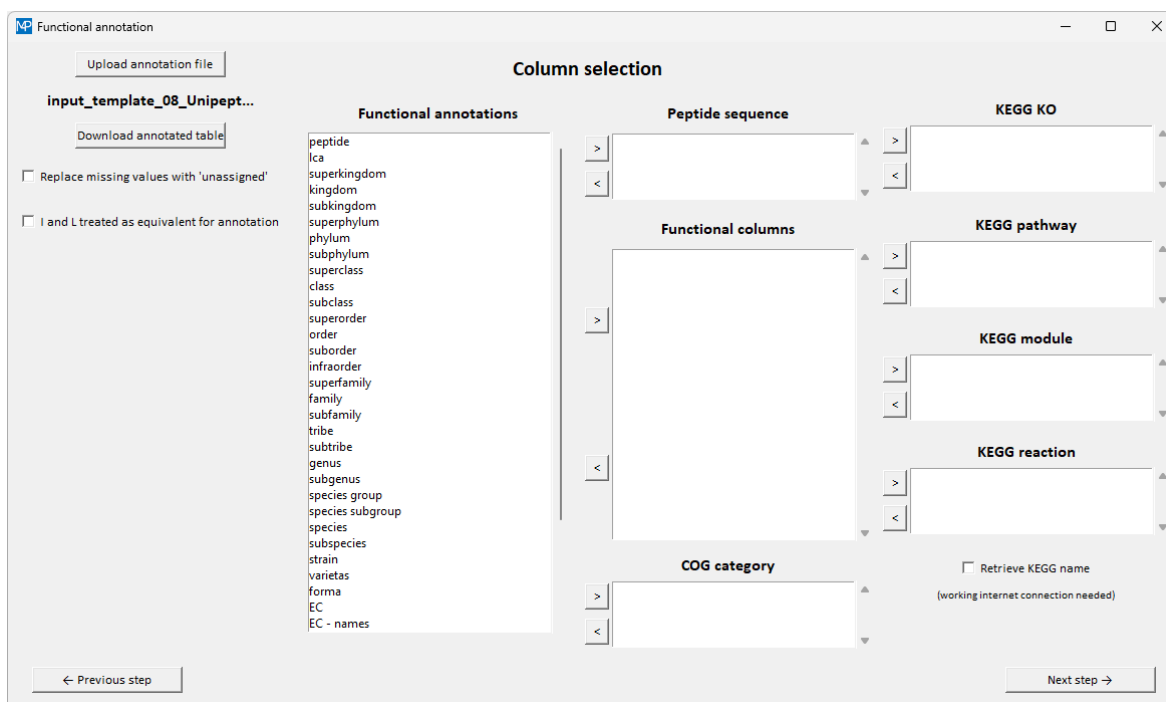
#### 3.2.1 eggNOG-mapper output

Window and options are identical to those described in section 3.1.1. The only difference lies in the presence of peptide instead of protein sequences in the "query" column of the input.

#### 3.2.2 Other/custom functional annotation

This window allows you to retrieve functional annotation data by uploading any tabular input file, in one of the following formats: xlsx, txt or generic tab-separated format. Column headers must be contained in the first row. Multiple functional annotations must be separated by a comma. If you try to upload an input file with a wrong format, an error message will be shown.

Click on "**Upload annotation file**" to select and upload the file containing peptide functional annotation data (the template file named *input\_template\_08\_Unipept\_peptide\_annotation.csv* also contains this kind of information). Once the file is uploaded, the file name is shown under the upload button and the file column headers are listed in the "Column headers" box (see the image below).



Select the columns of interests and move them to the corresponding box by using the arrows next to the boxes. At least one column listing peptide sequences and one column listing functional annotations must be selected and moved to the "**Peptide sequence**" box and to one of the functional boxes, respectively; when columns listing KEGG or COG annotations are present in the input file, these have to be moved to the specific box corresponding to their category. Unnecessary columns, if any, should remain unselected.

As an option, to retrieve and include in the table (as supplementary columns) the annotation names provided by the KEGG database for all KEGG categories, click on "**Retrieve KEGG name**". As this information is retrieved from the KEGG website, a working internet connection is needed in order that this operation is performed; this operation may take up to a few minutes. In case a protein has multiple functional annotations, their names will be separated by a comma as well as their codes (except for the "KEGG Module" annotation names, for which a vertical bar is used to avoid mistakes). In case a peptide is associated to multiple Master Proteins (usually separated by a semicolon), codes and names of functional annotations assigned to different Master Proteins are consistently separated by a semicolon (except for the "KEGG Reaction" annotation names, for which a vertical bar is used to avoid mistakes). When two analogous "KEGG Pathway" annotations are reported (i.e., having the same numeric code but two different prefixes, namely 'map' and 'ko'), the code with the 'ko' prefix is removed.

Another option, named "**Replace missing values with 'unassigned'**", allows you to keep missing annotations as empty cells or, if selected, to denominate them as "unassigned"; only in the latter case the quantitative values related to missing annotations are considered in the following data aggregation step. In addition, the option "**I and L treated as equivalent for annotation**" must be checked if isoleucine and leucine were considered as equivalent for peptide functional annotation.

An annotated output table, combining functional annotations with identification and quantification (and possibly taxonomic annotation) data, can be downloaded in xlsx, txt or generic tab-separated format by clicking on "**Download annotated table**".

## 4. Protein/Peptide/PSM metrics

The next "Summary metrics" windows allows you to download a summary table listing the main metrics of the identification, quantification and annotation data. More specifically:

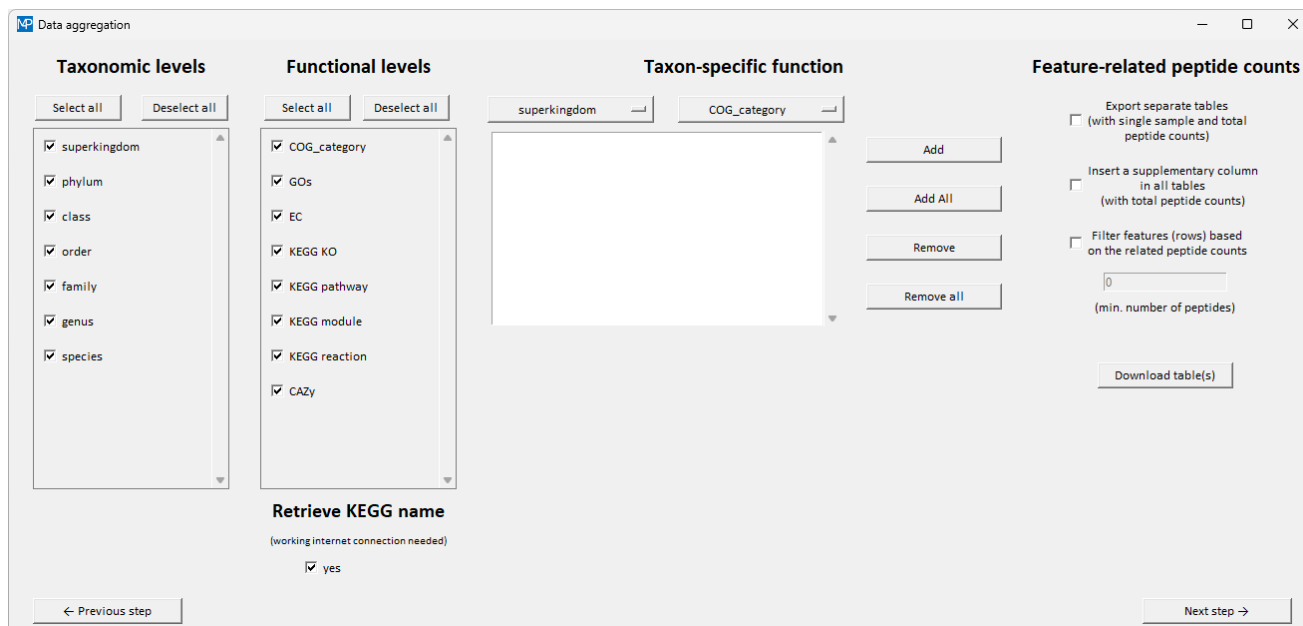
- when "Proteins" data are concerned ("**Download protein metrics**" button), the summary table reports the number of quantified proteins and their total abundance for each annotation category, for each sample and for the whole dataset;
- when "Peptides" data are concerned ("**Download peptide metrics**" button), the summary table reports the number of quantified peptides and their total abundance for each annotation category, for each sample and for the whole dataset;
- when "PSMs" data are concerned ("**Download PSM metrics**" button), the summary table reports the number of identified peptides and PSMs for each annotation category, for each sample and for the whole dataset.

Note that a protein/peptide with a missing annotation in a specific category does not contribute to the summary metrics for that particular annotation category. This also applies when the "**Replace missing values with 'unassigned'**" option has been selected for taxonomic and/or functional annotations.

## 5. Data aggregation

In this step, abundance data can be aggregated based on taxonomic, functional and/or taxon-specific functional annotations; in other words, the abundances of all proteins/peptides/PSMs sharing the same annotation are summed for each sample.

The layout of the "Data aggregation" windows is illustrated in the image below. Taxonomic and functional levels of interest can be selected, if available, by checking their respective checkbox ("**Select all**" and "**Deselect all**" buttons are also available on top). In addition, taxon-specific functions can be customized by combining a taxonomic level (drop-down menu on the left) with a functional level (drop-down menu on the right) and clicking on "**Add**"; click on "**Add all**" to select all possible combinations between taxa and functions.



The screenshot shows the "Data aggregation" window with the following sections:

- Taxonomic levels:** Includes "Select all" and "Deselect all" buttons. A list of taxonomic levels is shown with checkboxes: superkingdom, phylum, class, order, family, genus, and species. All are checked.
- Functional levels:** Includes "Select all" and "Deselect all" buttons. A list of functional levels is shown with checkboxes: COG\_category, GOs, EC, KEGG KO, KEGG pathway, KEGG module, KEGG reaction, and CAZy. All are checked.
- Taxon-specific function:** Includes two drop-down menus: "superkingdom" and "COG\_category". Below them is a large empty box for combinations. To the right are buttons: "Add", "Add All", "Remove", and "Remove all".
- Feature-related peptide counts:** Includes three checkboxes:
  - ☐ Export separate tables (with single sample and total peptide counts)
  - ☐ Insert a supplementary column in all tables (with total peptide counts)
  - ☐ Filter features (rows) based on the related peptide counts
 Below the third checkbox is a text input field with "0" and the label "(min. number of peptides)". At the bottom is a "Download table(s)" button.

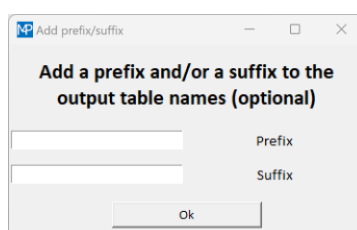
At the bottom of the window, there is a "Retrieve KEGG name" section with the note "(working internet connection needed)" and a checked checkbox "yes". Navigation buttons "← Previous step" and "Next step →" are at the very bottom.

As an option, to retrieve and include in the table(s) (as a supplementary column) the annotation names provided by the KEGG database for all the selected KEGG categories, check the **"Retrieve KEGG name"** checkbox.

For each taxonomic, functional or taxon-specific functional aggregated table selected, **feature-related peptide (or protein) counts** can also be retrieved, i.e., the number of peptides (or proteins) contributing to the summed abundance (showed as aggregated value in the table) for which an abundance value was measured for each measured feature. By checking the corresponding checkbox(es), this type of information can be **exported in separate tables** (having the same name of the corresponding tables reporting the aggregated abundance data, plus the suffix "\_proteincounts" for protein level inputs and "\_peptidecounts" for peptide/PSM level inputs) and/or **inserted as a supplementary column** (named "Total peptide count") in all aggregated tables. Furthermore, by checking the third checkbox, it is possible to **filter out** all features that do not reach a minimum peptide (or protein) count (the threshold must be typed in the box).

At the end, for each annotation level (taxa, functions and/or taxon-specific functions) selected, a table is generated listing the annotation features detected in the study, along with their total abundance values measured in each sample. These output files can be downloaded in xlsx, txt or generic tab-separated format by clicking on by clicking on **"Download tables"**; each table is saved with the name of the corresponding annotation level.

A further window (see the image below) will then appear enabling the addition of a prefix and/or a suffix to the names of all output files. Type the desired prefix and/or suffix in the corresponding box; boxes may be left blank if no prefix/suffix is needed.



Finally, in the "Save as" window, you will read the following reminder instead of the file name: "Filenames will be generated automatically, just choose the folder and file extension". In case you select the generic tab-separated format, the desired file extension must be specified at the end of the reminder.

Output template files (aggregated tables at various taxonomic, functional and taxon-specific functional levels) are also available for download.

## 6. Annotation metrics

The next "Summary metrics" windows allows you to download a summary table listing the main metrics of the annotation data, based on the previously downloaded aggregated tables. Note that this option is only available when at least one table has been downloaded in the previous step and that its metrics are specifically referred to the previously downloaded tables. More specifically:

- when "Proteins" data are concerned, the summary table reports the number of quantified features and their total abundance for each annotation level (taxa, functions and/or taxon-specific functions) selected;
- when "Peptides" data are concerned, the summary table reports the number of quantified features and their total abundance for each annotation level (taxa, functions and/or taxon-specific functions) selected;

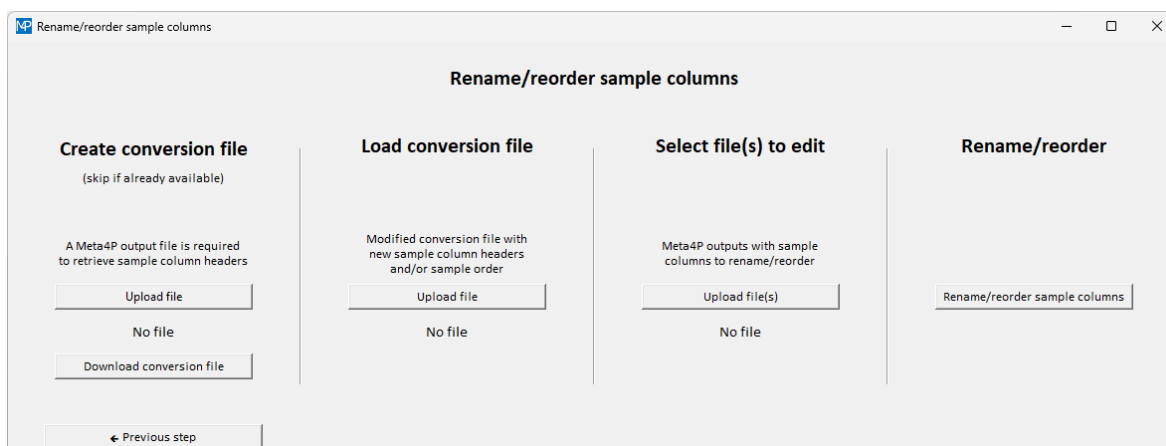
- when "PSMs" data are concerned, the summary table reports the number of identified features and related PSMs for each annotation level (taxa, functions and/or taxon-specific functions) selected.

Note that unassigned features do not contribute to the annotation metrics, even when the **"Replace missing values with 'unassigned'"** option has been selected for taxonomic and/or functional annotations; in the case of a taxon-specific feature, the presence of a single unassigned term (taxonomic or functional) implies that that given feature will not be taken into account for the calculation of the annotation metrics.

## 7. Rename/reorder sample columns

In this step, you can customize name and order of sample columns, based on a conversion file.

To create a conversion file (**"Create conversion file"** section; see the image below), upload one of the output tables generated by Meta4P in the previous steps (sample column headers must not have been modified in any way) by clicking on **"Upload file"**. Based on this input, the software creates a conversion file, i.e., a tabular file containing a first column (header "Old Name") reporting the sample list of the original input file (one sample per row) and a second, empty column (header "New name"). Click on **"Download conversion file"**, open the downloaded file and type the new sample names in the "New name" column. If useful, sample order can also be changed and Meta4P will change the column order in the output tables accordingly.



**Rename/reorder sample columns**

**Create conversion file**  
(skip if already available)

A Meta4P output file is required to retrieve sample column headers

Upload file

No file

Download conversion file

**Load conversion file**

Modified conversion file with new sample column headers and/or sample order

Upload file

No file

**Select file(s) to edit**

Meta4P outputs with sample columns to rename/reorder

Upload file(s)

No file

**Rename/reorder**

Rename/reorder sample columns

← Previous step

Once the conversion file has been filled in, upload it by clicking on "Upload file" in the **"Load conversion file"** section.

Then, select which of the output table(s) generated and downloaded in the previous steps need(s) to be subjected to renaming/reordering of sample columns (**"Select file(s) to edit"** section, **"Upload file(s)"** button).

Finally, click on **"Rename/reorder sample columns"** (**"Rename/reorder"** section) to complete the process.