Meta4P

MetaProteins-Peptides-PSMs Parser

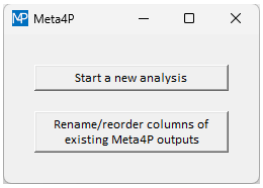
User guide

v.1.4.4 (April 2023)

0. Launch the application and start a new analysis

Double-click on the Meta4P_1.4.4.exe file to launch the application. The first time the program is launched, the operating system will ask for the user's 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 the user 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 5).

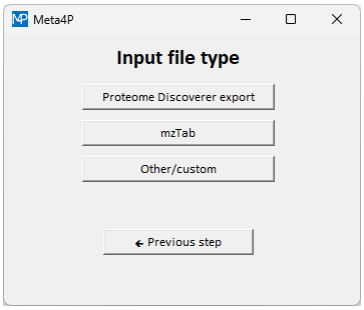


1. Identification/quantification input

As a first input, the user has to provide a file containing identification and quantification information. Meta4P can handle different file types and data levels, for a total of nine combinations.

1.1. Identification/quantification input: file type

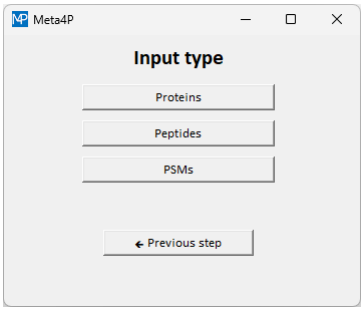
This window allows the user to select the desired file type, clicking on one of the following buttons: "**Proteome Discoverer export**", "**mzTab**" or "**Other/custom**" (see the image below).



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

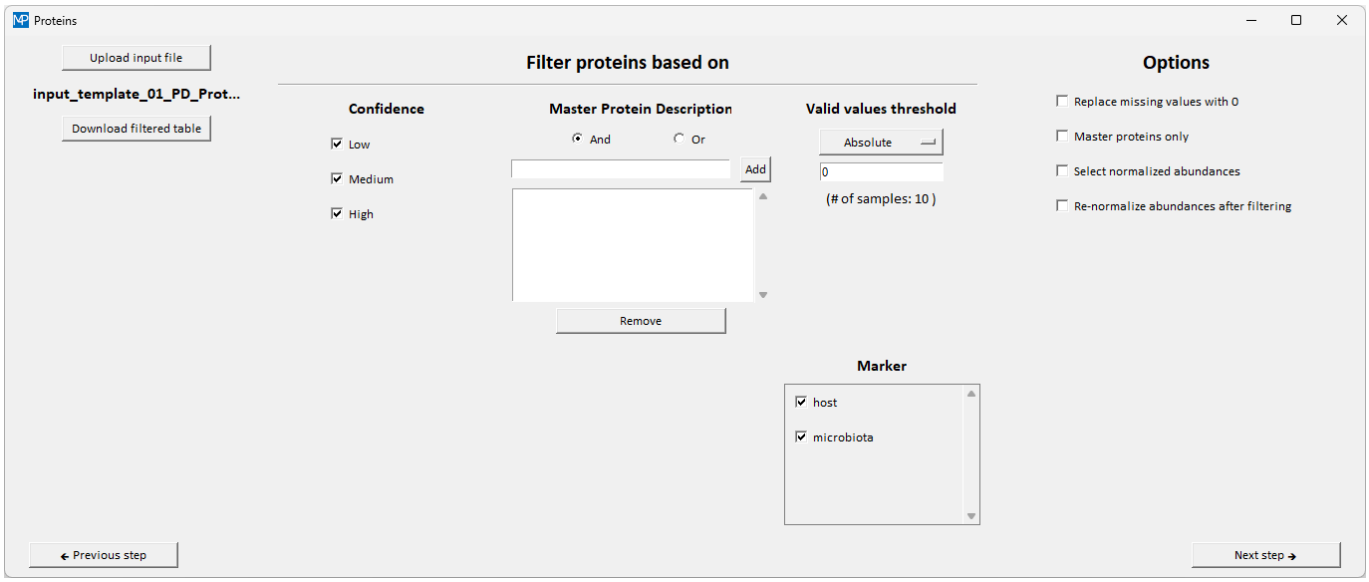
Once the file type has been selected, another window allows the user to choose the data level (see the image below). The following buttons are shown, each corresponding to a different level of identification and quantification data: "**Proteins**" (protein identifications with MS1-based quantitative 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 and 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 the user tries 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 also available). Once the file is uploaded, the window is populated with all the filtering options based on the file content (see the image below).



Proteins can be always 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. The user 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, the user 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, the user can choose between the following visualization and calculation options:

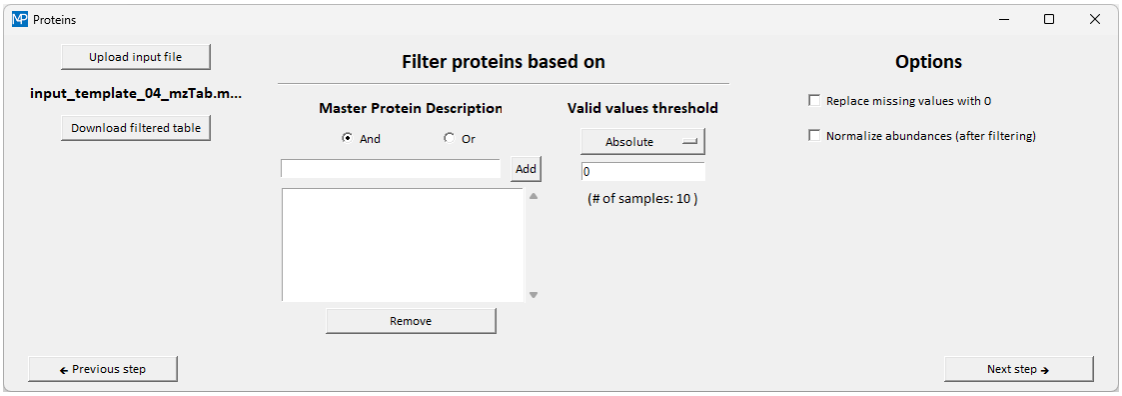
- ***Replace missing values with 0***: if selected, missing values (empty cells) are replaced by 0; this selection will be applied to all the following output tables.
- ***Master proteins only***: if selected, only proteins designated as "Master Protein" – i.e., a protein identified by a set of peptides 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. mzTab - Proteins

When the "mzTab" and "Proteins" options are sequentially selected, protein identification and quantification data are retrieved from a mzTab input file. If the user tries 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 also available). Once the file is uploaded, the window is populated with all the filtering options based on the file content (see the image below).



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. The user 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, the user 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, the user can select the following options:

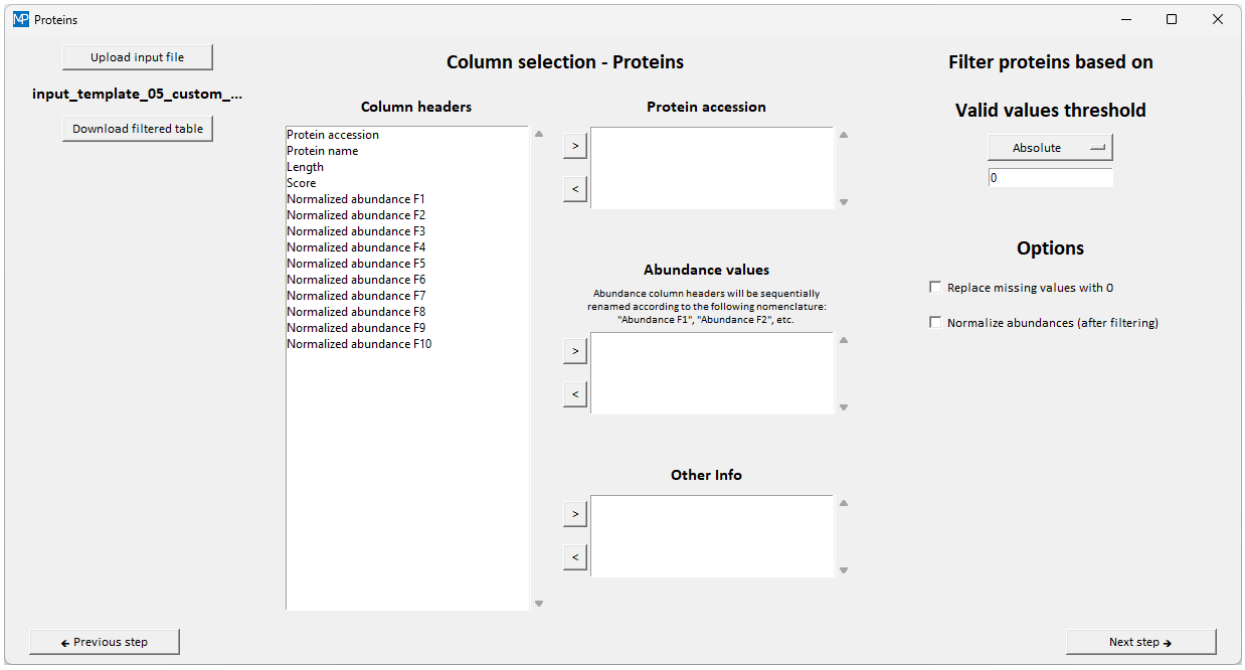
- ***Normalize abundances (after filtering)***: if selected, once the chosen filters are applied and the filtered protein list is obtained, the abundance value measured for a protein in a given sample is divided by the total protein abundance measured in that sample and multiplied by 10^{10}.
- ***Replace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

1.5. 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. If the user tries 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 also available). Once the file is uploaded, the file column headers are listed in the "Column headers" box (see the image below).



The user can 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 ad 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 moved to the "***Other info***" box (in case the user wants them to be kept in the output table).

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. The user 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 user can select the following options:

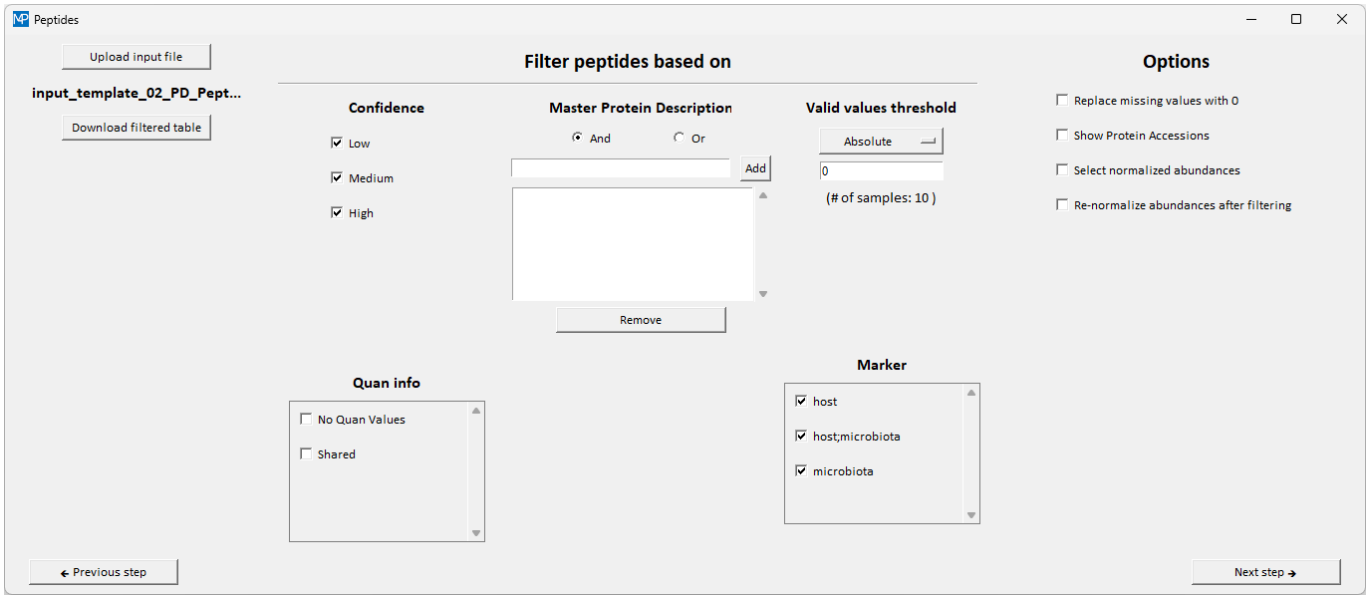
- ***Replace missing values with 0***: if selected, missing values (empty cells) are replaced by 0; this selection will be applied to all the following output tables.
- ***Normalize abundances (after filtering)***: if selected, once the chosen filters are applied and the filtered protein list is obtained, 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**".

1.6. 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 and 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. If the user tries 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 also available). Once the file is uploaded, the window is populated with all the filtering options based on the file content (see the image below).



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. The user 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, the user 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, the user can choose between the following visualization and calculation options:

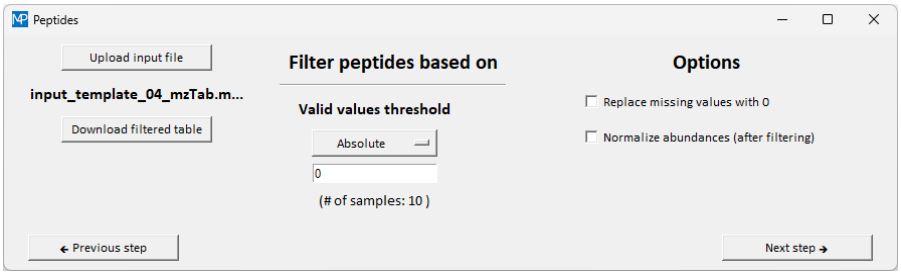
- ***Replace missing values with 0***: if selected, missing values (empty cells) are replaced by 0; this selection will be applied to all the following output tables.
- ***Show Protein Accessions***: if selected, the "Protein Accessions" column (i.e., the column indicating the accession number of all 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**".

1.7. mzTab - Peptides

When the "mzTab" and "Peptides" options are sequentially selected, peptide identification and quantification data are retrieved from a mzTab input file. If the user tries 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 also available). Once the file is uploaded, the window is populated with all the 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. The user 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 user can select the following options:

- ***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. Other/custom - Peptides

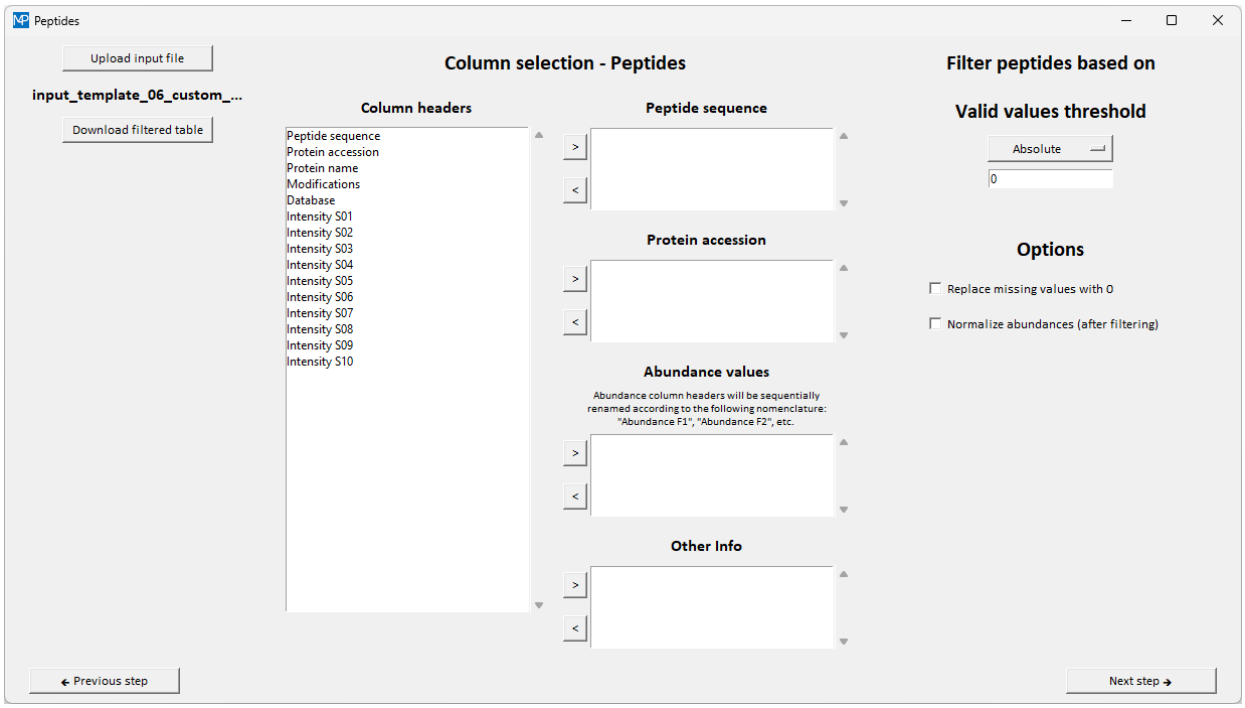
When the "Other/custom" and "Peptides" options are sequentially selected, peptide identification and quantification data can be retrieved from any tabular input file, in one of the following formats: xlsx, txt or generic tab-separated format. If the user tries to upload an input file with a wrong format, 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 also available). Once the file is uploaded, the file column headers are listed in the "Column headers" box (see the image below). The user can 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 ad 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 moved to the "***Other info***" box (in case the user wants them to be kept in the output table).

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. The user 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 user can select the following 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.

- ***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}.

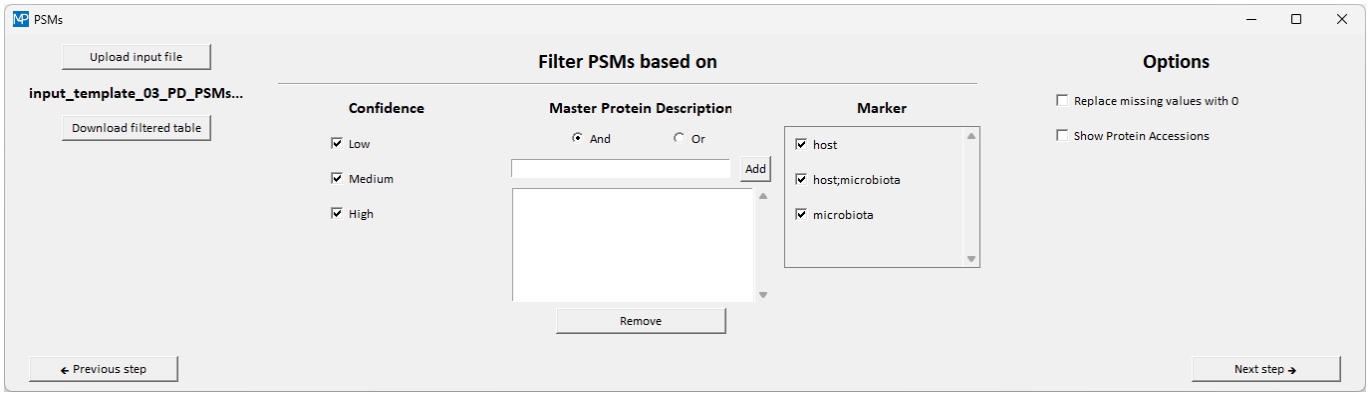


At the end, the identification and quantification table can be downloaded in xlsx, txt or generic tab-separated format by clicking on "**Download filtered table**".

1.9. 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 and 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 the user tries 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 also available). Once the file is uploaded, the window is populated with all the 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, the user 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, the user can choose between the following visualization and calculation options:

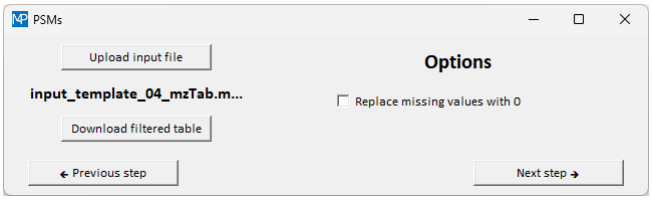
- ***Replace missing values with 0***: if selected, missing values (empty cells) are replaced by 0; this selection will be applied to all the following output tables.
- ***Show Protein Accessions***: if selected, the "Protein Accessions" column (i.e., the column indicating the accession number of all 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**".

1.10. mzTab - PSMs

When the "mzTab" and "PSMs" options are sequentially selected, PSM data are retrieved from a mzTab input file. If the user tries 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 also available). Once the file is uploaded, the following window is shown.



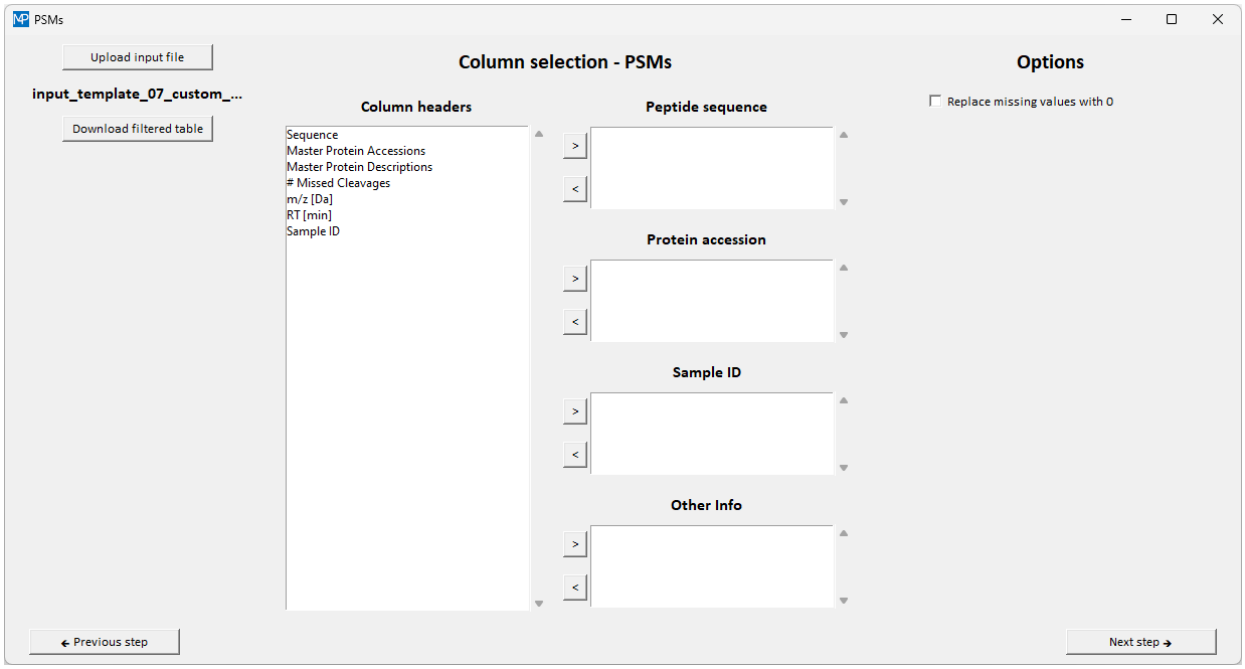
If the "***Replace missing values with 0***" option is selected, missing values (empty cells) are replaced by 0. This selection will be applied to all the following output tables.

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**".

1.11. 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. If the user tries 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 also available). Once the file is uploaded, the file column headers are listed in the "Column headers" box (see the image below).



The user can 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 ad 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 moved to the "***Other info***" box (in case the user wants them to be kept in the output table).

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**".

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.

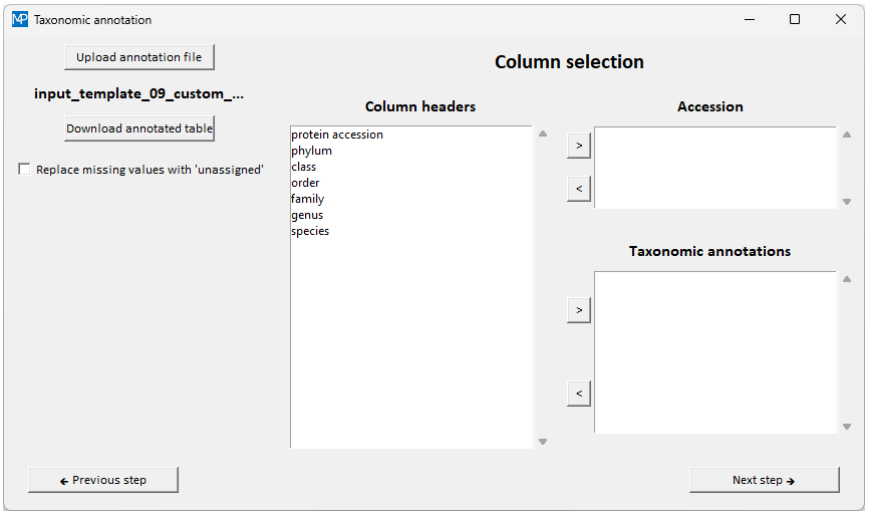
2.1. Proteins

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). Here, the user can go ahead with the upload of a taxonomic annotation file (clicking on "**Other/custom taxonomic annotation**") or skip this step if no taxonomic annotation is available.

[screenshot of the new window]

2.1.1. Other/custom taxonomic annotation

This window allows the user to retrieve taxonomic annotation data by uploading any tabular input file, in one of the following formats: xlsx, txt or generic tab-separated format. If the user tries 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 also available). Once the file is uploaded, the file column headers are listed in the "Column headers" box (see the image below).



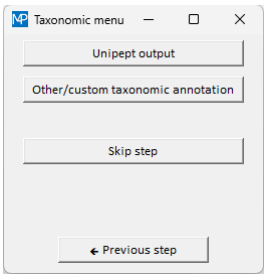
The user can select the columns of interests and move them to the corresponding box by using the arrows next to the boxes. At least one column listing protein accessions and one column listing taxonomic annotations must be selected and moved to the "***Accession***" and "***Taxonomic annotations***"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

In addition, an option named "***Replace missing values with 'unassigned'***" allows the user 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**".

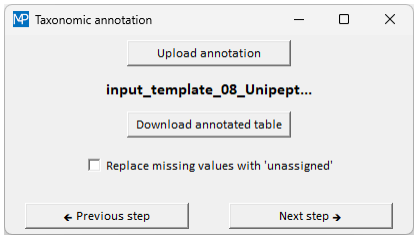
2.2. Peptides/PSMs

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). Here, the user can choose between two file types: standard Unipept output table ("**Unipept output**" button) or generic taxonomic annotation tabular file ("**Other/custom taxonomic annotation**" button). Alternatively, if no taxonomic annotation is available, this step can be skipped.



2.2.1. Unipept output

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



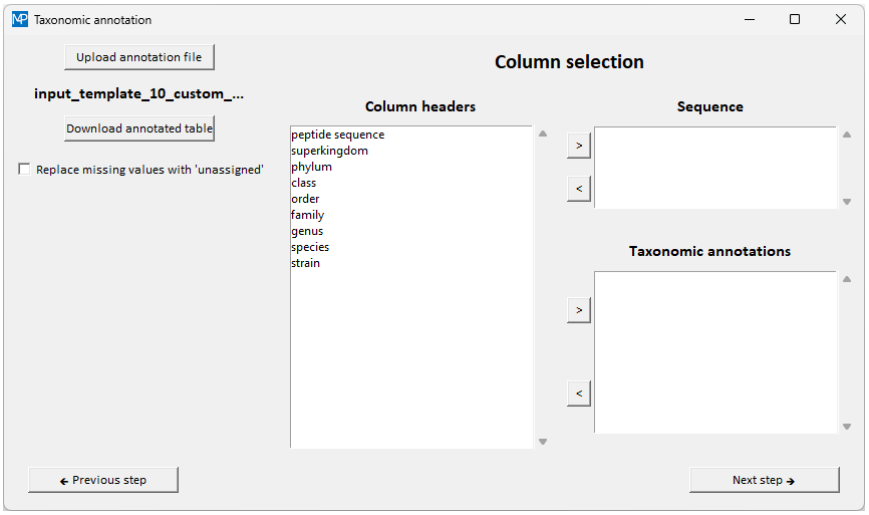
Click on "**Upload annotation**" to select and upload the file containing protein taxonomic annotation data (a template file named *input_template_08_Unipept_peptide_annotation.csv* is also available). Columns containing the main taxonomic annotations (LCA, superkingdom, phylum, class, order, family, genus, species) will be retrieved.

In addition, an option named "***Replace missing values with 'unassigned'***" allows the user 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**".

2.2.2. Other/custom taxonomic annotation

This window allows the user to retrieve taxonomic annotation data by uploading any tabular input file, in one of the following formats: xlsx, txt or generic tab-separated format. If the user tries 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_10_custom_peptide_taxonomic_annotation.tab* is also available). Once the file is uploaded, the file column headers are listed in the "Column headers" box (see the image below).



The user can 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.

In addition, an option named "***Replace missing values with 'unassigned'***" allows the user 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.

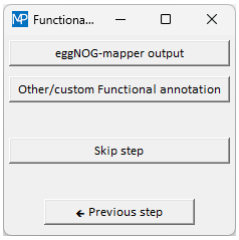
An annotated output table, combining taxonomic annotations with identification and quantification data, can be downloaded in xlsx, txt or generic tab-separated format by clicking on "**Download annotated table**".

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.

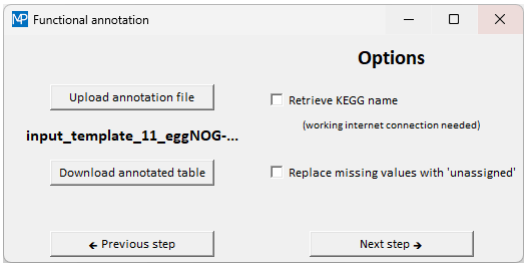
3.1. Proteins

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, the user 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.



3.1.1 eggNOG-mapper output

This window (see the image below) allows the user to retrieve taxonomic annotation data from a standard eggNOG-mapper output. If the user tries 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 also available). 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, the user can choose to retrieve and include in the table (as supplementary columns) the annotation names provided by the KEGG database for all KEGG categories, by clicking 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. 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).

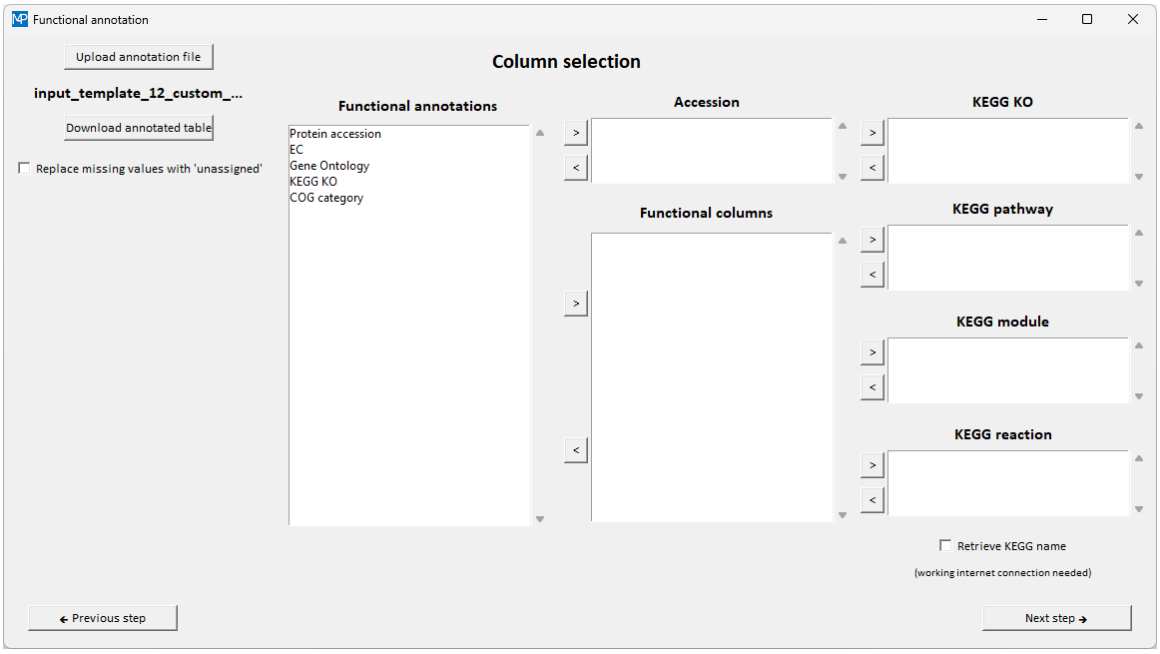
A further option, named "***Replace missing values with 'unassigned'***", allows the user to keep missing annotations as empty cells or, if selected, to denominate them as "unassigned"; only in the latter case the quantitative values related to missing annotations are considered in the following data 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 the user to retrieve functional annotation data by uploading any tabular input file, in one of the following formats: xlsx, txt or generic tab-separated format. If the user tries 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 also available). Once the file is uploaded, the file column headers are listed in the "Column headers" box (see the image below).



The user can 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***" and "***Functional columns***" boxes, respectively; when columns listing KEGG 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, the user can choose to retrieve and include in the table (as supplementary columns) the annotation names provided by the KEGG database for all KEGG categories, by clicking on "***Retrieve KEGG name***". As this information is retrieved from the KEGG website, a working internet connection is needed in

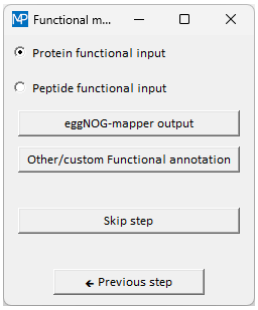
order that this operation is performed. In case a protein has multiple functional annotations, their names will be separated by a comma as well as their codes (except for the "KEGG Module" annotation names, for which a vertical bar is used to avoid mistakes). In case a peptide is associated to multiple 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).

A further option, named "***Replace missing values with 'unassigned'***", allows the user 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, the user has to 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.

Furthermore, a choice between two file types must be done: 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.

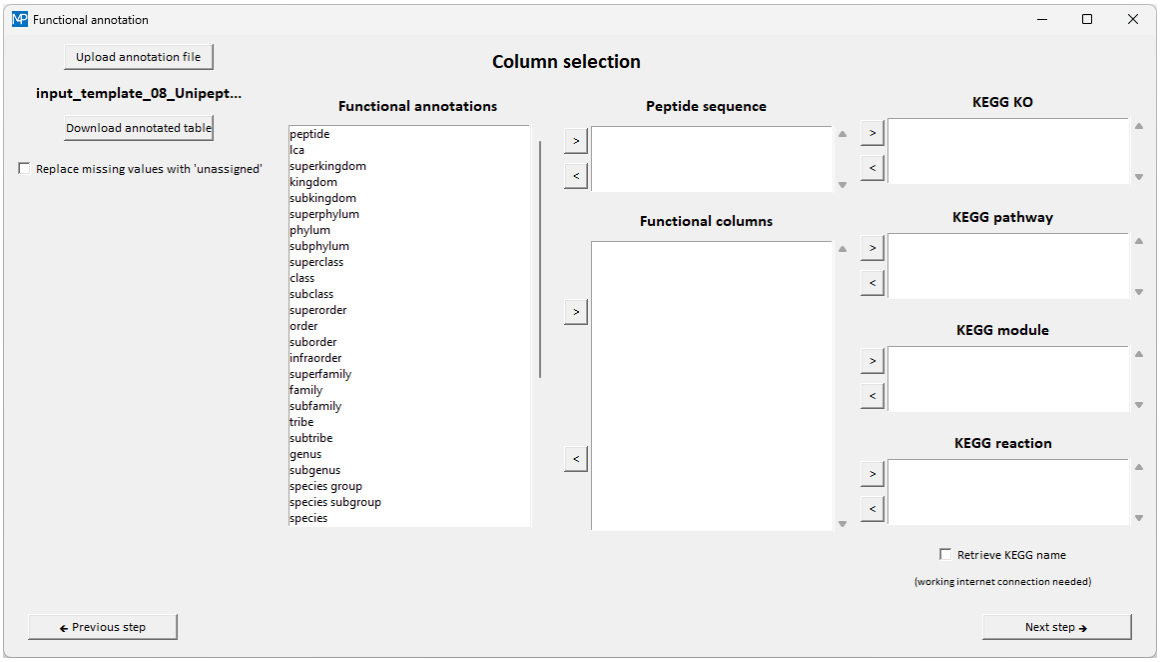
3.2.1 eggNOG-mapper output

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 the user to retrieve functional annotation data by uploading any tabular input file, in one of the following formats: xlsx, txt or generic tab-separated format. If the user tries 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 information). Once the file is uploaded, the file column headers are listed in the "Column headers" box (see the image below).



The user can 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***" and "***Functional columns***" boxes, respectively; when columns listing KEGG 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.

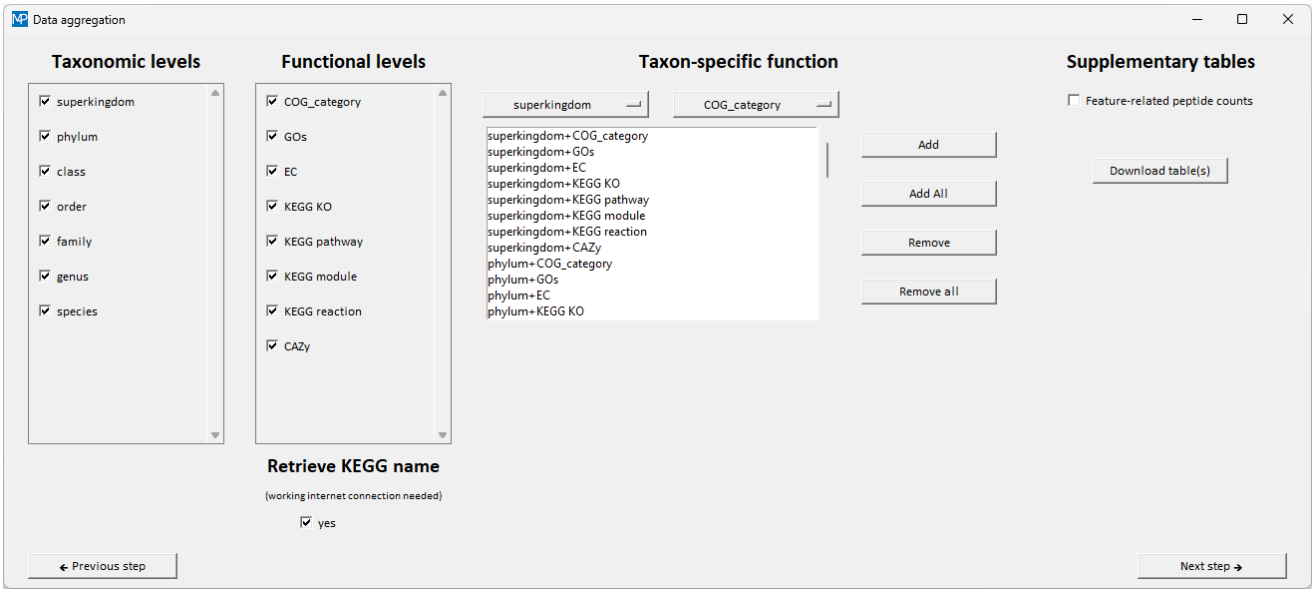
As an option, the user can choose to retrieve and include in the table (as supplementary columns) the annotation names provided by the KEGG database for all KEGG categories, by clicking on "***Retrieve KEGG name***". As this information is retrieved from the KEGG website, a working internet connection is needed in order that this operation is performed. 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).

A further option, named "***Replace missing values with 'unassigned'***", allows the user 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**".

4. 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. 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.

As an option, the user can choose 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, by checking the "**Retrieve KEGG name**" checkbox.

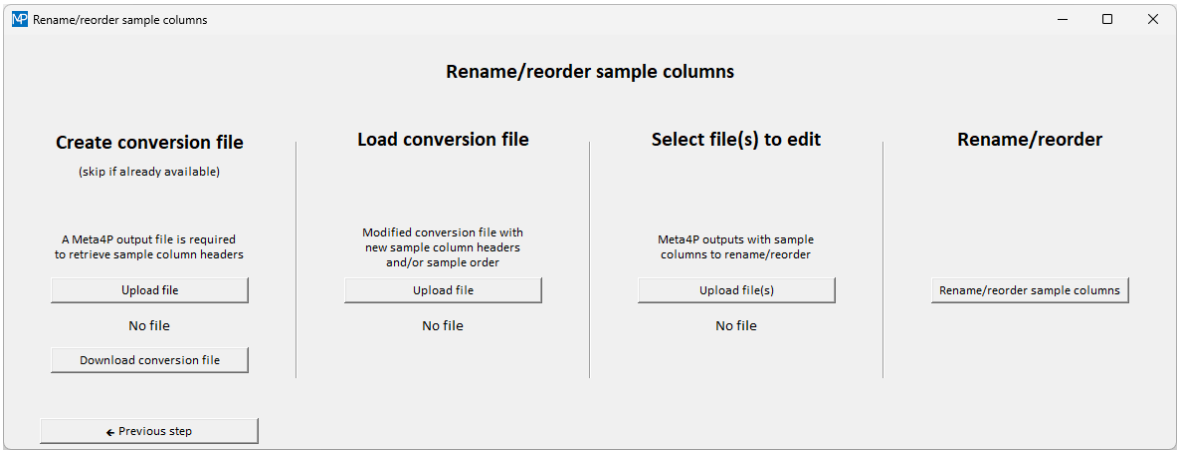
As a further option, by checking the corresponding checkbox, a supplementary table (for each output table selected) can be generated containing the **feature-related peptide (or protein) counts**, i.e., the number of peptides (or proteins) per sample for which an abundance value was measured (contributing to the summed abundance showed as aggregated value in the main table).

At the end, the list(s) of all annotations (taxa, functions and/or taxon-specific functions) belonging to the selected annotation level(s), together with their aggregated abundance values, can be downloaded in xlsx, txt or generic tab-separated format by clicking on by clicking on "**Download tables**". The user can also add a prefix and/or a suffix the all the output file names.

5. Rename/reorder sample columns

In this step, the user 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; in particular, sample column headers must not have been modified in any way by the user and must correspond to the original headers of the Proteome Discoverer output. Based on this input, the software creates a conversion file, i.e., a tabular file containing a first column with the current list of samples (based on the original input file, one sample per row) and a second column to be completed by the user with the customized sample names once the conversion file has been downloaded. If useful, the sample order in the first column can also be changed and Meta4P will change the column order in the output tables accordingly.



Once the conversion file has been filled in, upload it 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).

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