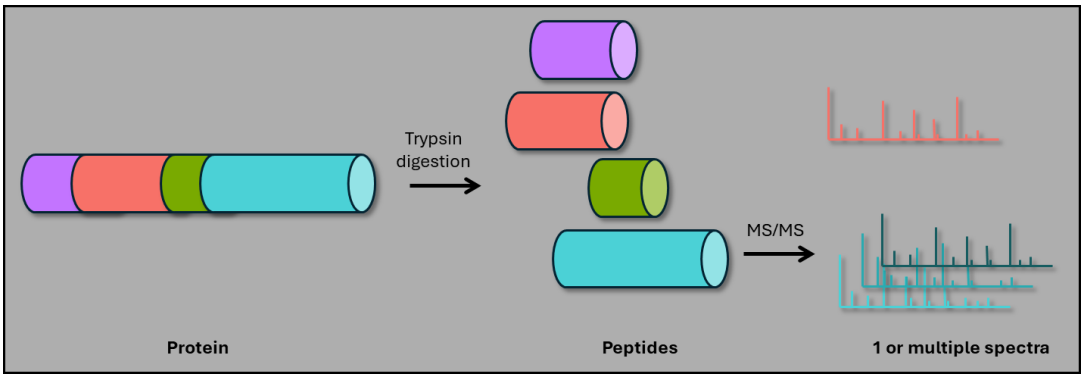
A quick guide to interpret MS results with Mascot & Scaffold

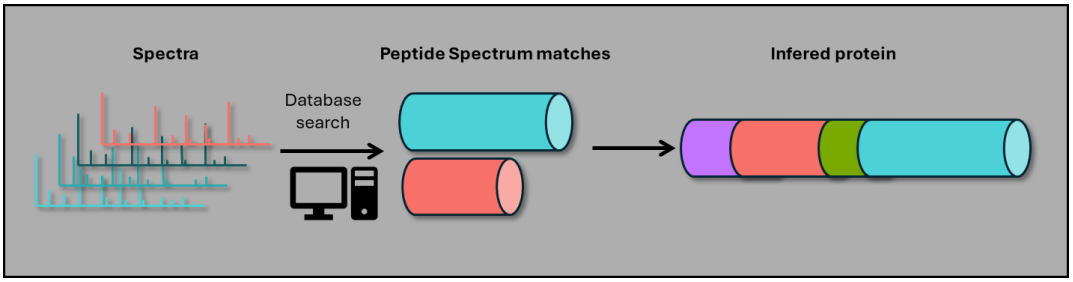
To keep in mind

Proteins are digested into peptides, which after ionization and fragmentation, can produce one or more MS/MS spectra depending on their charge state and MS/MS acquisition redundancy.



Proteomics is not about sequencing but matching in the database.
MASCOT, one of the software used for protein identification, matches MS/MS spectra to predicted spectra derived from protein sequence databases. When an experimental MS/MS spectrum matches a theoretical peptide sequence from the database, this is referred to a peptide–spectrum match (PSM).
Because this approach relies on **database matching**, proteins that are absent from the database—or lack a sufficiently similar homolog—cannot be identified, resulting in no match and therefore no output. This limitation is generally minimal when working with model organisms, for which protein databases are almost comprehensive.

Strictly speaking, proteins are not directly identified but rather infered.
MASCOT infers protein identity from the peptides identified during database searching. In principle, a single unique peptide is sufficient to support the identification of a protein, but having multiple matches greatly increases confidence.
However, different proteins may share identical peptide sets, particularly when they belong to highly homologous families that differ by only a few amino acids (e.g., tubulins). For this reason, results are often reported as **protein groups** rather than as single proteins. Importantly, the software applies the principle of parsimony, reporting the minimal set of protein sequences required to explain the maximum number of observed peptides.



Which kind of results do you get?

The **raw data** files will not be provided, unless they are requested, due to data size and format constraints.
However, you will receive the processed results in an **Excel table exported from Scaffold**, along with **a link** to download the complete dataset in a format readable with **Scaffold**. This software is free, when used as a viewer, user-friendly and can be downloaded from the developer's website.

If you have any questions about how to analyze your data, please do not hesitate to contact us.

Now, the most important: how to read the excel table

	A	B	C	D	E	F	G	H	I
1	**Samples report created on 11/06/2025**								
2	**Experiment: 251106-XXX-20546-49**								
3									
4	Database Name: the contaminants_PAF_20251028_1555.fasta; RefProt_Homo_sapiens_20250206 database								
5	Version: unknown								
6	Taxonomy: All Entries								
7	Number of Proteins: 83744								
8	Does database contain common contaminants?: Yes								
9									
10	Search Engine: Mascot								
11	Version: 3.1.0								
12	Samples: All Samples								
13	Fragment Tolerance: 0.020 Da (Monoisotopic)								
14	Parent Tolerance: 10.0 PPM (Monoisotopic)								
15	Fixed Modifications: +57 on C (Carbamidomethyl)								
16	Variable Modifications: +16 on M (Oxidation), +42 on Peptide N-Terminal (Acetyl)								
17	Digestion Enzyme: stricttrypsin								
18	Max Missed Cleavages: 2								
19	Probability Model:								
20	251020_XXX_20546_130min_8ul.raw (F017488.msr): Percolator [all charge states]								
21	251020_XXX_20547_130min_8ul.raw (F017487.msr): Percolator [all charge states]								
22	251020_XXX_20548_130min_4ul.raw (F017490.msr): Percolator [all charge states]								
23	251020_XXX_20549_130min_15ul.raw (F017491.msr): Percolator [all charge states]								
24									
25	**Scaffold: Version: Scaffold_5.3.4**				Target protein				
26	**Protein Grouping Strategy: Experiment-wide grouping with protein cluster analysis**								
27	**Peptide Thresholds: 90.0% minimum**								
28	**Protein Thresholds: 95.0% minimum and 2 peptides minimum**								
29	**Peptide FDR: 0.1% (Decoy)**								
30	**Protein FDR: 0.4% (Decoy)**								
31	Total Spectrum Count								
32	**Alternate ID Source(s): FASTA:UniProt/Swiss-Prot Alt. Accession (UniProtKB)**								
33									
34						20546	20547	20548	20549
35	#	Identified Proteins	Accession Number	Alternate ID	Molecular Weight	Band 1	Band 2	Band 3	Band 4
36	1.1	Unconventional myosin-IXb OS=Homo sa	Q13459	MYO9B	243 kDa	46	13	0	9
37	2.1	Clathrin heavy chain OS=Homo sapiens O	A0A8V8TQ18 (+1)	CLTC	196 kDa	57	178	34	299
38	2.2	Clathrin heavy chain OS=Homo sapiens O	A0A8V8TQK1	CLTC	189 kDa	56	176	34	297
39	3.1	Myosin-9 OS=Homo sapiens OX=9606 GI	P35579	MYH9	227 kDa	391	148	839	431
40	3.2	Myosin-10 OS=Homo sapiens OX=9606 (	P35580	MYH10	229 kDa	137	46	236	150

The lines above the tables list the MASCOT parameters used for peptide and protein identification and serve mainly as a record of the analysis. The sequence database used and the minimum number of peptides used for protein identification are specified.

Column 1-4: " #, Identified proteins, accession number, alternate ID"
As previously mentioned, in cases where **two or more proteins share all of their peptides**, there is no basis for discrimination among them and the proteins are treated as a unit called a **protein group**. These proteins appear in the Samples Table as a single line with the accession number of one of them followed by a plus sign and the number of other proteins in the group (example: (+1) in the line 37 of the above sample). The "preferred" or named protein is selected arbitrarily and may be changed by the user".
Scaffold adds an additional hierarchical level, the **protein cluster**, which appears as distinct subentries in the "#" column (e.g., 1.1, 1.2, 1.3, and 1.4 in the example above). This is a group of protein groups that share some peptides but not all of them. Protein Clusters are by default represented by the protein that shows the highest associated probability.

Column 5: "molecular weight"
The fifth column lists the **theoretical molecular weight** calculated from the database sequence, typically assuming the protein is unprocessed. This value does not necessarily reflect the actual mass of the protein in the sample. If the detected protein represents only a fragment of the full sequence, this can be hypothesized only by examining the sequence coverage in detail, specifically the positions of the matched peptides within the sequence. For this purpose, the complete dataset (Scaffold file) should be reviewed. Since sequence coverage is often incomplete, the lack of peptide recovery from a certain region *per se* should not be taken automatically as suggestion that the region is missing.

Following columns
The **following columns** report the **number of spectra assigned to each identified protein**. This number provides an indication of the confidence of the identification. Note that the values shown in this table represent the number of assigned spectra, not the number of unique peptides, as the same peptide may be matched by multiple spectra. To determine the exact number of distinct peptides, the complete dataset (Scaffold file) must be examined. As a practical guideline when evaluating the Excel table, identifications supported by four or more spectra can generally be considered reliable, whereas those supported by fewer spectra should be interpreted with caution.

A roughly linear relationship exists between the number of spectra assigned to a given protein and its concentration in a sample. Therefore, the number of matched spectra can be used to generate **semi-quantitative** estimates of protein abundance — an approach known as ***spectral counting***. This linearity is reasonably reliable when the number of matched spectra is sufficiently high (generally >10). However, at lower spectral counts, the relationship becomes much less dependable.

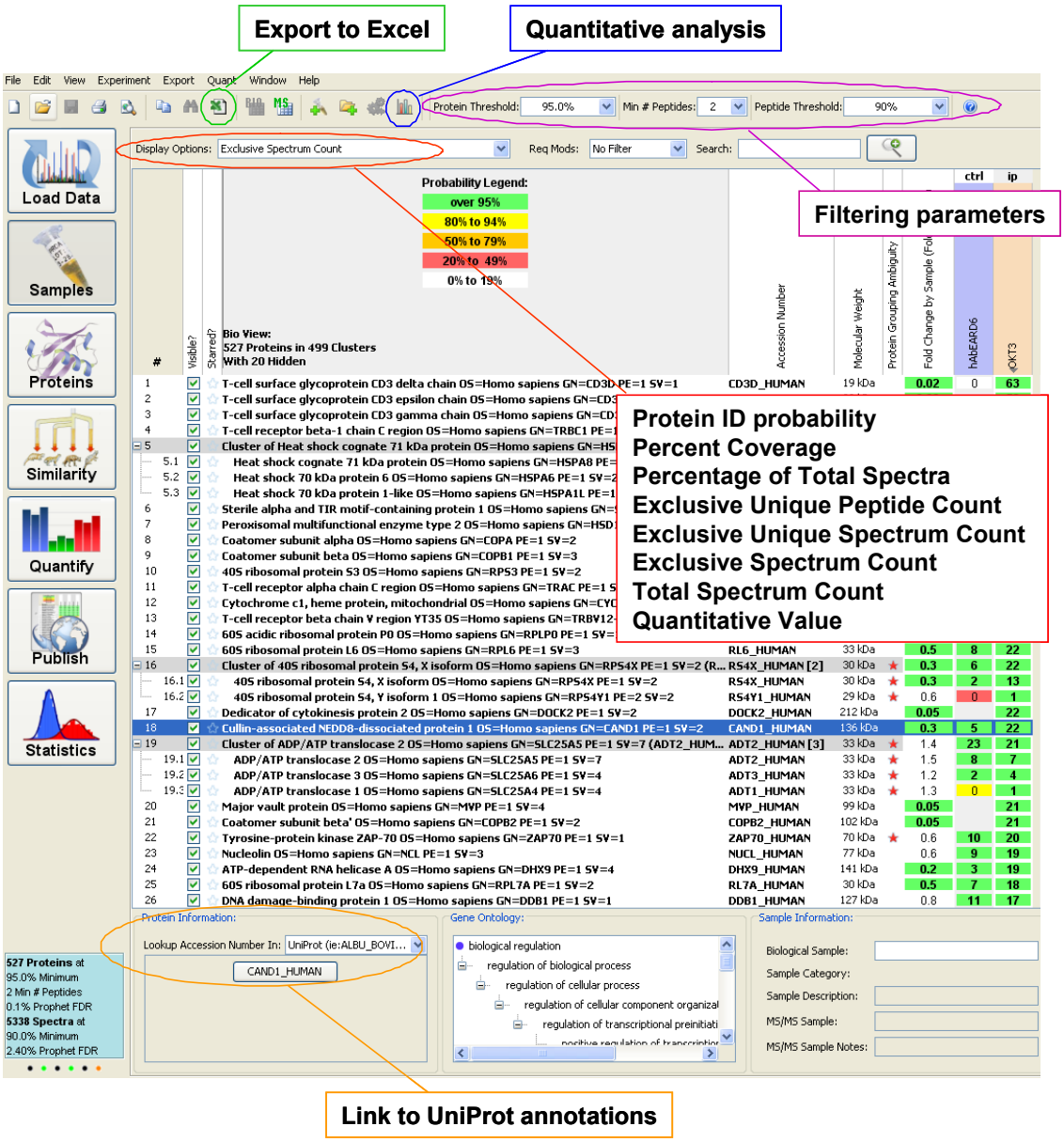
For example, if tubulin is identified with 100 spectra in sample A and 300 spectra in sample B, it is reasonable to infer an approximate 2–3 fold difference in abundance between the two samples. In contrast, if the insulin receptor is identified with 3 spectra in sample A and 1 spectrum in sample B, no meaningful conclusion about a difference in concentration can be drawn. The same limitation applies when comparing 3 spectra versus 0 spectra — the absence of detected spectra in sample B does not necessarily indicate that the protein is absent; it may simply fall below the detection threshold.
Importantly, spectral counting is a lot less reliable for comparing the abundance of different proteins within the same sample, as factors such as protein size, peptide detectability, and ionization efficiency can strongly influence spectral counts.

A specific question regarding protein pull-down (IP for instance)
A protein appearing in the list does not automatically mean it interacts with your target protein. Pull-down experiments can generate various artifacts. The comparison of a suitable negative control is essential to draw conclusions about specificity of interactions. A list of common contaminants can be found at https://reprint-apms.org/. Nevertheless, this cannot replace an internal negative control prepared in parallel to your positive sample.

How does a Scaffold file look like?

Scaffold enables exploration of your data, validation of peptides & protein matches, and quantitative comparison of samples based on spectral counting. Standard filtering criteria are 95% protein probability, 90% peptide probability, and a minimum of two peptides. These thresholds can be adjusted when targeting low-abundance proteins or peptides. Identifications obtained under relaxed criteria should be treated with caution. Please ask us if you have any questions.

We cannot provide a comprehensive introduction to Scaffold's capabilities here, but the software is user-friendly and includes a well-developed Help menu. It can be downloaded as a free viewer. Ensure that you are using the latest version of the software to open Scaffold (.sf3) files. For large files, a reasonably powerful computer with at least 16 GB of RAM is recommended.



The table can be sorted by any column by clicking its header. The option “search” option is very useful when you are looking at a specific protein in a list of a few thousands hits! It is also easy to sort the list by protein name, molecular weight, number of spectra, etc.

Several display options are available for examining the list of identified proteins (see below). “**Exclusive Spectrum Count**” and “**Total Spectrum Count**” are commonly used, as they provide the conventional basis for (semi-)quantitative comparison of proteins among samples.

- **Protein ID Probability -** Scaffold estimates the probability that each protein identification is correct in a given MS sample. The results are color-coded to indicate significant variations in identification confidence.

- **Percentage Coverage -** Percentage of the protein sequence covered by identified peptides.

- **Percentage of Total Spectra** – Percentage of spectra matched to a particular protein over all spectra in the sample.

• **Exclusive Unique Peptide Count** - Number of distinct peptides that map to exactly one protein group, providing high confidence for protein identification.

• **Exclusive Unique Spectrum Count** - Number of distinct spectra associated with one protein group only. Spectra are considered distinct when i) they identify different sequences of amino acids of peptides or ii) they identify different charge states/modified form of the same peptide.

• **Exclusive Spectrum Count** - Number of spectra associated exclusively with a single protein group.

• **Total Spectrum Count** - Total number of spectra assigned to a protein group, including spectra shared with other proteins.

• **Quantitative Value** - Results of the Quantitative Method selected from the Quantitative Analysis Dialog Box.

For more details on the sequence coverage, assigned peptides or spectra for a particular protein, select it by clicking and use "Proteins" window in the left menu bar. GO annotations for all identified proteins can be downloaded into the Scaffold file via **Menu bar → Experiment → Add annotations**. You can also view alternative accession numbers and detailed UniProt annotations for a selected protein using the link in the "Protein Information" frame at the bottom of the Scafffold table.

Data can be exported in different formats, with Excel being the most convenient - and the format we provide to you.

Samples can be grouped in categories (e.g. in the case of replicate samples) and more sophisticated semi-quantitative analyses and statistical tests can be performed in the "Quantify" window in the left menu bar.

Links & how to get more infos

- EXPASY as a general ressources for proteins and proteomes.
- UNIPROT, THE reference for protein and proteomes.
- Reactome for biological pathways
- Eucaryotic Linear Motif (ELM) for the prediction of protein functional sites.
- Protein-protein interaction databases :
    - INTACT for molecular interaction data
    - HPRD for Human proteins database
    - BIND for molecular interactions
    - Database of Interaction Proteins (DIP)
    - STRING for protein-protein interaction networks.
- **About Scaffold**:
    - Searle, B. C. (2010). Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics*, *10* (6), 1265–9.
    - Scaffold User's Guide, available from the Help menu in Scaffold or in this link.