Project Outline

NCEMS – CURE

2025-08-24

Table of Contents

[(I) Examining Two “Shot-Gun” Proteomics Experiments 2](#_Toc206965070)

[Extracting key information from the publications 2](#_Toc206965071)

[What cleavage enzyme was used and what amino acids does it cut after? 2](#_Toc206965072)

[Was the dataset labeled or label-free? 3](#_Toc206965073)

[Were any specific chemical modifications or post-translational modifications (PTMs) mentioned by the authors? 4](#_Toc206965074)

[(II): Open Modification Search with Spectral Alignment Guided Engine (SAGE) 5](#_Toc206965075)

[How SAGE Works: Overview of the Pipeline 5](#_Toc206965076)

[Examine the peptide quantification file 6](#_Toc206965077)

[SAGE Output File Description 6](#_Toc206965078)

[(III) Calculate and Examine the Protein Abundance Distributions for the Soybean Datasets 7](#_Toc206965079)

[Filter the peptide tables 7](#_Toc206965080)

[Select and clean relevant columns for analysis 7](#_Toc206965081)

[Estimate Protein Abundances 8](#_Toc206965082)

[Compute Stats of Sample distributions 8](#_Toc206965083)

[Make Box Plots of each Sample Separately 8](#_Toc206965084)

[(IV): Differential Protein Expression from LFQ Mass Spectrometry Data of Soybean Dataset 8](#_Toc206965085)

[1. Protein-level abundance statistics 9](#_Toc206965086)

[2. Standard Scale Renormalization of Protein Abundance 9](#_Toc206965087)

[3. Statistical testing (Welch’s t-test) 9](#_Toc206965088)

[4. Ranking and multiple testing correction 9](#_Toc206965089)

[5. Extract significant proteins 10](#_Toc206965090)

[6. Volcano plots 10](#_Toc206965091)

[Axes: 10](#_Toc206965092)

# (I) Examining Two “Shot-Gun” Proteomics Experiments

Shotgun proteomics is a high-throughput strategy for identifying and quantifying proteins in complex biological samples. It works by enzymatically digesting proteins into peptides, separating those peptides via liquid chromatography, and analyzing them using tandem mass spectrometry (MS/MS). This approach allows researchers to characterize thousands of proteins in a single experiment without prior knowledge of the sample’s composition.

In this project, we will explore how protein abundance information can be extracted from public datasets in the [PRIDE Archive](https://www.ebi.ac.uk/pride/archive/projects/), using mass spectrometry-based shotgun proteomics. We focus on two biologically and technically distinct datasets:

* Soybean (*Glycine max*): [PXD023343](https://www.ebi.ac.uk/pride/archive/projects/PXD023343)
* Human (*Homo sapiens*): [PXD005187](https://www.ebi.ac.uk/pride/archive/projects/PXD005187)

These examples illustrate different use cases, labeling strategies, and experimental goals in proteomics.

## Extracting key information from the publications

In this section, you’ll extract important metadata from both the original publication and the experimental files associated with your dataset. This information will be used to build a \*.json\* configuration file required by SAGE to process the data.

[1] DOI: 10.1038/cddiscovery.2017.10, Des: Kathiriya JJ, Nakra N, Nixon J, Patel PS, Vaghasiya V, Alhassani A, Tian Z, Allen-Gipson D, Davé V. Galectin-1 inhibition attenuates profibrotic signaling in hypoxia-induced pulmonary fibrosis. Cell Death Discov. 2017 Apr 10;3:17010. eCollection 2017

[2] DOI: 10.3390/ijms22020920, Des: Cheng L, Min W, Li M, Zhou L, Hsu CC, Yang X, Jiang X, Ruan Z, Zhong Y, Wang ZY, Wang W. Quantitative Proteomics Reveals that GmENO2 Proteins Are Involved in Response to Phosphate Starvation in the Leaves of Glycine max L. Int J Mol Sci. 2021 22(2)

### What cleavage enzyme was used and what amino acids does it cut after?

- *Why it matters:* The enzyme determines the peptide cleavage rules which are crucial for understanding how a protein will fragment in the spectrometer.

- *Where to look:* Check the methods section of the manuscript for sample preparation or digestion steps. Common entries include trypsin, Lys-C, chymotrypsin, or Glu-C. This information may also be in PRIDE metadata under “Sample Protocol” or “Digestion.”

#### Table : Common Cleavage Enzymes

| Enzyme | Cleaves After (Residues) | Cleavage Blocked By | Terminal Side of Cleavage | Notes |
| --- | --- | --- | --- | --- |
| Trypsin | K, R | P | C-terminal | Most commonly used; highly specific |
| Lys-C | K | — | C-terminal | Stable in high denaturant (e.g., urea) |
| Arg-C | R | — | C-terminal | Less specific than trypsin |
| Glu-C | E, sometimes D | — | C-terminal | Cleaves more broadly at high pH |
| Chymotrypsin | F, Y, W, L, M | P | C-terminal | Broader specificity; more missed cleavages likely |
| Asp-N | D | — | N-terminal | N-terminal cleavage; less common |
| Pepsin | Broad (especially F, L, E) | pH-dependent | C-terminal or mixed | Non-specific, active at low pH |
| Thermolysin | L, I, V, F, M | — | N-terminal | Metalloprotease, stable at high temperature |

### Was the dataset labeled or label-free?

- *Why it matters:* Labeling strategies (e.g., SILAC, TMT) enable multiplexed, highly accurate quantification, while label-free methods are simpler and more scalable. Knowing the approach clarifies the strengths, limitations, and comparability of the quantitative results.   
- *Where to look:* Look for keywords like “TMT”, “iTRAQ”, or “LFQ” in the methods section or PRIDE metadata.

#### TMT-Labeled Data

TMT (Tandem Mass Tags) are chemical labels attached to peptides that allow multiplexed MS-based quantification using reporter ions. These are detected at MS2 or MS3 levels.

*Use this when:*

* The methods mention TMT10, TMT11, TMT16, or iTRAQ.
* Quantification is based on reporter ion intensities.

#### SILAC-Labeled Data

SILAC (Stable Isotope Labeling by Amino acids in Cell culture) incorporates heavy isotopes into amino acids (e.g., 13C615N4-Arg & 13C615N2-Lys) during protein synthesis, shifting precursor m/z but not fragment ions.

*Use this when:*

* The methods mention *SILAC*, *light/heavy*, or *metabolic labeling*.

#### Label-Free Quantification (LFQ)

LFQ uses *precursor ion intensity* without any chemical or isotopic labeling. It’s based on matching peptide features across runs.

*Use this when:*

* There is *NO mention of TMT, iTRAQ, or SILAC, sometimes the authors will explicitly state they used LFQ methods*.

### Were any specific chemical modifications or post-translational modifications (PTMs) mentioned by the authors?

- *Why it matters:* These modifications often change the mass and charge of a peptide and thus its mass spectra. Sometimes these modifications are intentional such as in the case of labeling experiments to test how protein abundances changes under different conditions.   
- *Where to look:* Usually listed under *sample preparation* or *search parameters* in the methods section.

#### Static vs. Variable Modifications

In proteomics, modifications refer to changes in the mass of a residue or terminus caused by chemical labeling, post-translational modifications (PTMs), or experimental artifacts.

#### Static Modifications

A *static modification* is a change that occurs on every instance of a given residue or terminus.

Use static mods when:

* The modification was applied to all peptides during sample preparation
* You are using chemical labels like TMT or iTRAQ
* The mass shift is consistent and universal

#### Examples of common static mods:

| Modification | Residue | Mass Shift | When It Occurs |
| --- | --- | --- | --- |
| Carbamidomethylation | C | +57.0215 | During alkylation with iodoacetamide |
| TMT16 labeling | K, ^ | +304.207 | Chemical labeling (TMT 16plex) |
| iTRAQ 8plex | K, ^ | +304.2054 | iTRAQ chemical labeling |

#### Variable Modifications

A *variable modification* is optional — the search engine will consider both the modified and unmodified versions of the peptide. This allows detection of PTMs or partial labeling.

Use variable mods when:

* You expect the modification to occur on only some peptide
* You’re searching for biologically relevant PTMs
* You want to allow missed modifications (e.g., incomplete labeling)

#### Examples of common variable mods:

| Modification | Residue | Mass Shift | When It Occurs |
| --- | --- | --- | --- |
| Oxidation | M | +15.9949 | Spontaneous or regulated (PTM) |
| Deamidation | N, Q | +0.984 | Non-enzymatic or enzymatic (PTM) |
| Pyro-glutamate formation | ^Q, ^E | −17.0265 / −18.0106 | N-terminal loss of ammonia |
| Acetylation | ^ | +42.0106 | N-terminal (PTM) |
| SILAC heavy lysine (labeling) | K | +8.0142 | Metabolic labeling |
| SILAC heavy arginine | R | +10.0083 | Metabolic labeling |

# (II): Open Modification Search with Spectral Alignment Guided Engine (SAGE)

SAGE (Spectral Alignment Guided Engine) is an open-source, high-throughput software tool for database searching and quantification of mass spectrometry (MS/MS) proteomics data. Developed by the Lazar Lab at the NIH, SAGE is designed to scale efficiently across large datasets while maintaining high sensitivity and accuracy in peptide identification.

SAGE implements modern algorithmic strategies and leverages predictive scoring models, retention time prediction, and multi-level quantification schemes to identify and quantify peptides from tandem mass spectra with high confidence.

SAGE is particularly useful for:

* High-throughput proteomics datasets
* Label-free and isobaric tag quantification (e.g., TMT, SILAC)
* Studies requiring open or semi-open modification searches
* Peptide identification across multiple samples with alignment

## How SAGE Works: Overview of the Pipeline

1. Database Generation (In Silico Digestion)
   * The input FASTA file is digested virtually using user-defined cleavage rules and specified static/variable modifications
   * All possible peptide sequences are generated, filtered by length and mass
2. Spectral Preprocessing
   * Experimental spectra are filtered to remove low-intensity noise and spectra with insufficient peaks.
3. Scoring and Peptide Matching
   * SAGE matches experimental MS/MS spectra to the in silico-generated peptide fragments using scoring algorithms based on:
     + Ion type matches (e.g., b/y ions)
     + Fragment ion intensity
     + Mass tolerance (in ppm or Da)
     + Number of matched ions
   * It computes statistical scores to rank peptide-spectrum matches (PSMs).
4. Retention Time (RT) Prediction
   * RT alignment helps improve scoring by matching predicted vs. observed elution behavior, reducing false positives.

5. Quantification of peptide abundances

## Examine the peptide quantification file

Let us examine the output from SAGE for the Soybean Samples. The initial info printed to screen summarizes the number of non-null elements and the datatype of each of your dataframes columns. We then display the dataframe in an easily scrollable table.

### SAGE Output File Description

1. peptide

* Definition: The amino-acid sequence of the identified peptide (after in silico digestion, e.g., tryptic).
* Context in Sage: Sage identifies peptides by matching MS/MS spectra to theoretical peptides from the database, including handling chimeric spectra and open modifications. ([GitHub][1], [@lazear][2])

1. charge

* Definition: The precursor ion charge state observed in MS1 (e.g., +2, +3).
* Use: Influences the theoretical fragmentation pattern and is used by Sage during candidate generation and scoring; different charge states of the same peptide are tracked separately in the PSM stage. ([GitHub][1])

1. proteins
   * Definition: The protein accession(s) the peptide maps to, in UniProt-style format (e.g., sp|P12345|PROT\_NAME).
   * Note: Shared peptides (mapping to multiple proteins) introduce ambiguity; downstream aggregation logic (e.g., razor/parsimony or simple grouping) determines assignment. Sage itself reports the peptide-to-protein mapping as part of its identification output. ([GitHub][1])
2. q\_value
   * Definition: The estimated false discovery rate (FDR)–adjusted confidence metric for the peptide-spectrum match (PSM), expressing the minimal FDR at which that identification would be accepted.
   * Sage behavior: Sage includes built-in FDR control (typically via target-decoy strategies and rescoring) to produce calibrated q-values so users can threshold identifications (e.g., ≤0.05) with an expected proportion of false positives. ([ResearchGate][3], [PubMed][4])
3. score
   * Definition (in the context of Sage): The primary peptide–spectrum match (PSM) score reflecting how well the experimental MS/MS spectrum agrees with the candidate peptide.
   * Details: Sage’s scoring pipeline includes initial matching and then optional rescoring that can incorporate additional features (e.g., retention time prediction, predicted vs observed fragment behavior, and other learned features) to improve discrimination between true and false PSMs. This composite score is used (alongside the target-decoy framework) to rank identifications before q-value estimation. ([GitHub][1], [ScienceDirect][5], [PubMed][4])
   * Teaching note: Because Sage supports rescoring, the score may reflect a learned or empirical combination of evidence beyond simple dot-product or correlation—students should inspect the particular configuration used (e.g., whether retention time prediction or other auxiliary features were enabled).
4. spectral\_angle
   * Definition: A spectral similarity metric (often the normalized spectral contrast angle) between the observed experimental MS/MS spectrum and a reference/predicted spectrum.
   * Relevance: While not a core Sage-invented metric, spectral angle is a well-established geometric measure of similarity used in spectral matching contexts; it quantifies the angle between intensity vectors in high-dimensional fragment-ion space, with smaller angles (or higher normalized similarity) indicating closer agreement. It may be computed as part of auxiliary scoring or downstream quality assessment. ([PMC][6])
   * Interpretation: Because intensity vectors are normalized before angle calculation, the spectral angle is relatively robust to absolute intensity scaling; it gives a sense of how consistent the fragmentation pattern is with the expected or library-derived pattern.
   * Teaching note: Students can contrast this with simple dot product or correlation, and discuss why normalized geometric metrics (like spectral contrast angle) can be more sensitive to subtle spectral differences. ([PMC][6])
5. \*.mzML,..
   * Definition: Quantitative intensity estimates (e.g., MS1 extracted ion currents or aggregated peptide-level abundances) for that peptide in the three biological/technical replicates of the LP condition.
   * Context: These intensities are the basis for relative quantification across conditions. Sage also provides quantification functionality (including LFQ-style approaches) that feeds into these, though the downstream script performs its own aggregation and normalization. ([ResearchGate][3], [@lazear][2])

# (III) Calculate and Examine the Protein Abundance Distributions for the Soybean Datasets

Now that you have obtain a list of peptides that have been confidently matched to proteins of interest we need to obtain estimates of the protein level abundances.

## Filter the peptide tables

Explanation:

* Filters out low-confidence identifications: only keeps peptides with q\_value ≤ 0.05, controlling peptide-level false discovery. This reduces noise downstream.

Why: Filtering by q-value ensures you are only using peptide measurements with statistical support—propagating too many false positives would pollute protein-level inference.

Teaching note: The threshold (0.05) is conventional; students can try stricter (0.01) or looser as an exercise to see how discovery changes.

## Select and clean relevant columns for analysis

Explanation:

* Constructs a subset of columns: retaining peptide, the original proteins annotation, q\_value, and all other columns except technical metadata (charge, score, spectral\_angle) which are not needed for quantification.
* The proteins field is in UniProt format like sp|P12345|PROT\_NAME;tr|P12345|PROT\_NAME;....
* For each peptide make a row copy for every protein it could possibly map to.
* Drops any peptides that failed to parse into a valid UniProt accession.

Why: Reduces clutter and focuses downstream computation on actual intensity data and needed identifiers.  
You need a clean, consistent protein identifier to collapse peptides into proteins. Dropping unmapped peptides avoids assigning ambiguous evidence.

Teaching note: Students can be asked what happens if peptides map to multiple proteins (this script doesn’t resolve shared peptides explicitly—it aggregates all evidence under each protein) and discuss how “razor” or parsimony approaches differ.

## Estimate Protein Abundances

Explanation:

* Groups all peptides assigned to each UniProt accession and collapses their intensities per sample by taking the median across peptides.

Why median? Median is robust to outliers (e.g., aberrant peptide measurements) and is a simple estimate of protein abundance from multiple peptide surrogates.

Teaching note: Alternatives include mean, weighted mean (by peptide quality), or more sophisticated models (like MaxLFQ’s ratio-based inference). Students can compare results of median vs mean as an exploration. https://www.nature.com/articles/nmeth.3901.pdf

## Compute Stats of Sample distributions

## Make Box Plots of each Sample Separately

* Do you observe any noticable differences in the distribution of protein abundances between technical replicates under the same condition (i.e. LP and HP)?
* How about between conditions?
* Do these plots make sense with the statistics you calculated previously?
* What does this tell you about the proteomes of these two organisms and does this make sense for two very different organisms?

# (IV): Differential Protein Expression from LFQ Mass Spectrometry Data of Soybean Dataset

*Using a robust, classical pipeline with fold-change inference via per-protein statistics*

Explanation:

* welch\_test: performs Welch’s t-test comparing the LP and HP sample groups for a given protein (row of the protein-level matrix). equal\_var=False allows unequal variances—a safer default in proteomics. nan\_policy='omit' means missing values don’t crash the test.

Teaching note: Welch’s t-test is generally powerful when sample sizes are modest and distributions are approximately symmetric; the Mann–Whitney test is robust to non-normality but tests for difference in distribution location in a different way. It’s good practice to examine data (e.g., via violin/box plots) to decide if assumptions are reasonable.

## 1. Protein-level abundance statistics

## 2. Standard Scale Renormalization of Protein Abundance

Explanation:

* Computes global mean and standard deviation separately for LP and HP samples (flattening all proteins × replicates).
* Z-scores each group independently: subtract mean, divide by standard deviation.

Why: This rescales each condition so that its overall distribution is standardized. It removes global scale differences that might come from batch effects or loading differences.

Caveat: Standardizing separately means the two groups are no longer directly on the same absolute scale—this could obscure true global shifts. Normally, for differential testing you’d either:

* Normalize all samples together (e.g., median normalization across all), or
* Use ratio-based methods where differences are preserved without groupwise rescaling.

Teaching exercise: Have students try skipping this step or doing a joint normalization and compare which proteins are called significant.

## 3. Statistical testing (Welch’s t-test)

Explanation:

* Applies Welch’s t-test per protein, comparing the standardized LP vs HP vectors (each of length 3).
* Stores the test statistic and raw p-value.

Why Welch’s test? It does not assume equal variance between groups, a prudent choice for biological data where heteroskedasticity is common.

Teaching note: Point out that with only 3 replicates per group, degrees of freedom are low; encourage students to visualize variance and consider the effect of small sample size on power and false negatives.

## 4. Ranking and multiple testing correction

Explanation:

* Sorts proteins by raw significance (smallest p-value first) to prioritize candidates.
* Applies Benjamini–Hochberg FDR correction (fdr\_bh) to control the expected proportion of false discoveries among calls.

Why: Thousands of proteins may be tested; without correction, the number of false positives at p<0.05 would be unacceptably high. FDR balances discovery with error control.

Teaching note: Discuss differences among correction methods (bonferroni, fdr\_bh, fdr\_by, etc.) and when each is appropriate.

## 5. Extract significant proteins

Explanation:

* Filters proteins whose adjusted p-value is below 0.05, declaring them differentially abundant between LP and HP.
* Prints their UniProt accession IDs.

Why: This yields a candidate list for biological interpretation: proteins whose standardized abundance differs beyond expected noise.

## 6. Volcano plots

A volcano plot is a scatter plot that simultaneously displays both magnitude of change and statistical significance for each protein (or peptide) when comparing two conditions. It’s a rapid visual filter to highlight candidates that are both substantially and confidently different.

### Axes:

* X-axis: Fold change (here, Fold2AbundChange). This is usually on a log2 scale (positive means higher in condition B, negative means higher in condition A). It reflects effect size — how much a protein’s abundance changes.
* Y-axis: from a statistical test (e.g., Welch’s t-test on replicate abundances). Higher values mean stronger statistical evidence against the null (i.e., more significant). Using the negative log makes small p-values (strong significance) appear at the top.