Data Analysis and Statistics

TMT quantitative mass spectrometry pipeline

To ensure consistency across all publicly available mass spectrometry datasets that used TMT isobaric mass tagging, we applied the same standardized analysis pipeline previously published in Farley et al. (2024, A), Farley et al. (2024, B), and Thomas et al. (2025)[1–3]. This approach allows for better comparability between datasets. Several of the datasets included in this repository were generated and analyzed at the European Molecular Biosciences Laboratories (EMBL) in Heidelberg, Germany, where Dr. Frank Stein developed and implemented the data analysis pipeline.

In summary, we processed the mass spectrometry (MS) data using FragPipe[4]. The protein search was performed using a UniProt FASTA database with one entry per gene, excluding isoforms and minimizing TrEMBL entries. Common contaminants and reversed sequences were also included for validation. Proteins were included in the final dataset if they contained at least two razor peptides, as FragPipe applies the Occam's Razor principle, assigning shared peptides to the protein with the highest supporting evidence.

For the database search, we included post-translational modifications (PTMs) such as oxidation (M), acetylation (N-term), and carbamidomethylation (C). The raw files were analyzed using FragPipe's protein.tsv output, which was used as the basis for downstream quantification.

To account for potential technical variations across experiments, we applied a normalization step using the 'removeBatchEffect' function from the limma package to correct for batch effects in log2-transformed TMT reporter ion intensities[5,6]. Further normalization was carried out using the 'normalizeVSN' function of the same package. We identified differentially expressed proteins using the moderated t-test within limma, incorporating replicate information into the statistical model through the 'lmFit' function.

For the protein-lipid probe interaction experiments (+/- UV exposure), proteins were categorized based on their enrichment in the +UV condition. Proteins showing a log2 fold-change of at least 1 and a p-value below 0.05 were classified as "enriched hits," while those showing similar enrichment trends but with p-values above 0.05 were labeled as "enriched candidates."

To better understand the biological significance of the detected proteins, we conducted gene ontology (GO) enrichment analysis using the 'compareCluster' function from the 'clusterProfiler' package[7]. This method evaluates whether specific GO terms are overrepresented in the dataset compared to a background gene set. We performed enrichment analysis for three GO categories: Cellular Component (CC), Molecular Function (MF), and Biological Process (BP). The reference database used was 'org.Pf.plasmo.db.' To quantify enrichment, we calculated the odds ratio by comparing the proportion of genes associated with each GO term in our dataset ('GeneRatio') with the proportion in the reference background set ('BgRatio'). A value greater than 1 suggests that a given GO term is more prevalent in our dataset than expected by chance.

SILAC quantitative and mass spectrometry pipeline

In this repository, only one dataset utilized SILAC to quantify protein interactors, Chiu et al. 2025, and so an *ad hoc* method was applied to coerce the data into the framework of the standardized data visualizations of the site. The results are as reported in Chiu et al. 2025[8].

Visualizing proteomics data

Volcano Plots

- Visualize enrichment data versus the statistical p value associated with every protein identified in a dataset.
- In each of the datasets depicted on this site, data points to the right of the x-axis (logFC $>$ 0) denote proteins which were enriched when the lipid probe was irradiated with UV light – stimulating chemical crosslinking.
- The p value of each protein was -log10 transformed, meaning that the smaller the p value, the higher the point is.
- In the datasets depicted on this site, there is little meaning to points with logFC $<$ 0; as such, these proteins are ignored regardless of their statistical significance

Rank-ordered Plots

- Line up each protein identified in the data set from lowest logFC to highest.
- Enables the viewer to see the shape of the dataset – for example, some datasets will have only a few proteins substantially enriched, and this is reflected by a sharp "elbow" at the edges of the plot.
- This can be helpful in selecting logFC cutoffs when determining significance cutoffs.

MA Plots

- Depict the logFC of each protein (**M**) versus the average m/z intensity of the peptides (**A**) that contributed to that proteins identification.
- Gives greater insight into the quality of a datapoint. For example, a point on the far left of the MA plot had relatively low total abundance and may be more sensitive to dramatic changes in enrichment – whereas points to the right were substantially more abundant in the sample and will be less sensitive to change.
- Thus, "significantly enriched" proteins with higher **A** intensity may be more trustworthy than those with lower **A** values.

Heat Maps

Gene Ontology Overview

- Gene Ontology (GO) analyses take a list of proteins and determines whether there is a greater than random enrichment of a biological feature. These features can include pathways, molecular functions, cellular compartments or biological processes.
- In each of the GO analyses presented on this site, we filtered according to a significance cutoff – these cutoffs are explained on the respective data pages (either "enriched candidates" or "enriched hits", as in the TMT-based datasets, or unique subsets of probe enrichment, as in the PSM-based datasets).
- Enriched pathways are given statistical significance by assessing the likelihood that a pathway would be overrepresented through random chance.

1. Farley, S., Stein, F., Haberkant, P., Tafesse, F. G. & Schultz, C. Trifunctional Sphinganine: A New Tool to Dissect Sphingolipid Function. *ACS Chemical Biology* **19**, 336–347 (2024).
2. Farley, S. E. *et al.* Trifunctional fatty acid derivatives: The impact of diazirine placement. *Chemical Communications* **60**, 6651–6654 (2024).
3. Thomas, A. *et al.* Trifunctional lipid derivatives: PE's mitochondrial interactome. *Chemical Communications* 10.1039.D4CC03599B (2025) doi:10.1039/D4CC03599B.
4. Yu, F. *et al.* Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform. *Nature Communications* **14**, 4154 (2023).
5. Huber, W., Von Heydebreck, A., Sültmann, H., Poustka, A. & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18**, S96–S104 (2002).
6. Ritchie, M. E. *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**, e47–e47 (2015).
7. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology* **16**, 284–287 (2012).

8. Chiu, D.-C., Lin, H. & Baskin, J. M. Photoaffinity Labeling Reveals a Role for the Unusual Triply Acylated Phospholipid N-Acylphosphatidylethanolamine in Lactate Homeostasis. (2025) doi:10.26434/chemrxiv-2025-j8nqz.