The Microbiologist's Guide to Metaproteomics

Tim Van Den Bossche[1,2], Jean Armengaud[3], Dirk Benndorf[4,5,6], Jose Alfredo Blakeley-Ruiz[7], Madita Brauer[8,9], Kai Cheng[10], Marybeth Creskey[11], Daniel Figeys[10,12,13], Lucia Grenga[3], Timothy J Griffin[14], Céline Henry[15], Robert L Hettich[16], Tanja Holstein[1,2,17], Pratik D Jagtap[14], Nico Jehmlich[18], Manuel Kleiner[7], Benoit J Kunath[7,19], Xuxa Malliet[1,2], Lennart Martens[1,2], Subina Mehta[14], Bart Mesuere[20], Zhibin Ning[10], Alessandro Tanca[21,22], Sergio Uzzau[21,22], Pieter Verschaffelt[1,2,20], Jing Wang[23], Paul Wilmes[19,24], Xu Zhang[11], Xin Zhang[23], and Leyuan Li[23,], on behalf of the Metaproteomics Initiative*

(1) Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University, 9052 Ghent, Belgium
(2) VIB-UGent Center for Medical Biotechnology, VIB, 9052 Ghent, Belgium
(3) Département Médicaments Et Technologies Pour La Santé (DMTS), Université Paris-Saclay, CEA, INRAE, SPI, 30200, Bagnols-Sur-Cèze, France
(4) Applied Biosciences and Process Engineering, Anhalt University of Applied Sciences, Köthen, Germany
(5) Bioprocess Engineering, Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany
(6) Bioprocess Engineering, Otto von Guericke University, Magdeburg, Germany
(7) Department of Plant and Microbial Biology, North Carolina State University, Raleigh, USA
(8) Molecular Disease Mechanisms Group, Department of Life Sciences and Medicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg
(9) Institute for Advanced Studies, University of Luxembourg, Esch-sur-Alzette, Luxembourg
(10) School of Pharmaceutical Sciences, Faculty of Medicine, University of Ottawa, Ottawa, Canada
(11) Biologic and Radiopharmaceutical Drugs Directorate, Health Products and Food Branch, Health Canada, Ottawa, Canada
(12) Quadram Institute Bioscience, Norwich Research Park, Norwich, Norfolk, UK
(13) University of East Anglia, Norwich, Norfolk, UK
(14) Department of Biochemistry, Molecular Biology, and Biophysics, University of Minnesota. Minneapolis, MN 55455
(15) Université Paris-Saclay, INRAE, AgroParisTech, Micalis Institute, PAPPSO, 78350, Jouy-en-Josas, France
(16) Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA
(17) Data Competence Center MF2, Robert-Koch-Institut, Berlin, Germany
(18) Helmholtz-Centre for Environmental Research - UFZ GmbH, Department of Molecular Toxicology, Leipzig, Germany
(19) Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg
(20) Department of Applied Mathematics, Statistics and Computer Science, Faculty of Sciences, Ghent University, 9000 Ghent, Belgium

(21) Department of Biomedical Sciences, University of Sassari, 07100 Sassari, Italy
(22) Unit of Microbiology and Virology, University Hospital of Sassari, 07100 Sassari, Italy
(23) State Key Laboratory of Medical Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Lifeomics, 102206 Beijing, China
(24) Department of Life Sciences and Medicine, Faculty of Science, Technology and Medicine, University of Luxembourg, Belvaux, Luxembourg

* Correspondence should be addressed to Leyuan Li (lileyuan@ncpsb.org.cn)

Abstract

Metaproteomics is an emerging approach for studying microbiomes, offering the ability to characterize proteins that underpin microbial functionality within diverse ecosystems. As the primary catalytic and structural components of microbiomes, proteins provide unique insights into the active processes and ecological roles of microbial communities. By integrating metaproteomics with other omics disciplines, researchers can gain a comprehensive understanding of microbial ecology, interactions, and functional dynamics. This review, developed by the Metaproteomics Initiative (www.metaproteomics.org), serves as a practical guide for both microbiome and proteomics researchers, presenting key principles, state-of-the-art methodologies, and analytical workflows essential to metaproteomics. Topics covered include, among others, experimental design, sample preparation, mass spectrometry techniques, data analysis strategies, and statistical approaches.

Table of Contents

1. Why metaproteomics?

The importance of microbiomes in nearly all processes within the biosphere is increasingly clear. Composed of bacteria, bacteriophages, archaea, yeasts, fungi, protozoa, and viruses, microbiomes are highly diverse in taxonomic composition. A microbiome and its theater of activity—including microbial elements such as genes, transcripts, proteins, and metabolites—together form a microbiome (Berg et al. 2020). Microbiomes are, in most cases, highly structured in both membership and function. This underscores the need to understand microbiomes and their interactions with their environment or eukaryotic hosts, whether beneficial or harmful. However, the complexity of these systems challenges traditional research tools, particularly cultivation-dependent approaches, which, given the wealth of intra-organism interactions, are not scalable for large-scale microbiome studies.

The rapid advancement of omics-based approaches has opened new avenues for systems biology-based research into the complexity of microbiomes. Shotgun metagenomics, in particular, has proven to be a powerful tool, offering much deeper insights than older techniques such as 16S rRNA gene amplicon sequencing. Metagenomics enables the discovery of complete genomic inventories, even for uncultured microorganisms, revealing the metabolic and physiological capabilities of a microbiome. However, it is limited to predicting functions rather than identifying active processes. To overcome this limitation, omics approaches such as metatranscriptomics, metaproteomics, and metabolomics provide essential insights into actual gene expression and activity under specific conditions. Together, these techniques bridge the gap from taxonomic structure to genomic potential and dynamic, context-dependent functions.

Among these tools, metaproteomics enables the comprehensive analysis of the proteins expressed and functional in a microbiome, quantifies their abundances, and characterizes their modifications, interactions, and localizations **(Figure 1)**. Proteins serve as the primary catalytic units and structural elements of microbiomes, making metaproteomics a direct reflection of the microbiome's phenotype. This approach provides a detailed functional description and examines specific protein changes associated with structure, homeostasis, and enzymatic activity. Differences in protein sequences allow researchers to determine the taxonomic origins of particular enzyme sets, linking functions to taxonomic units.

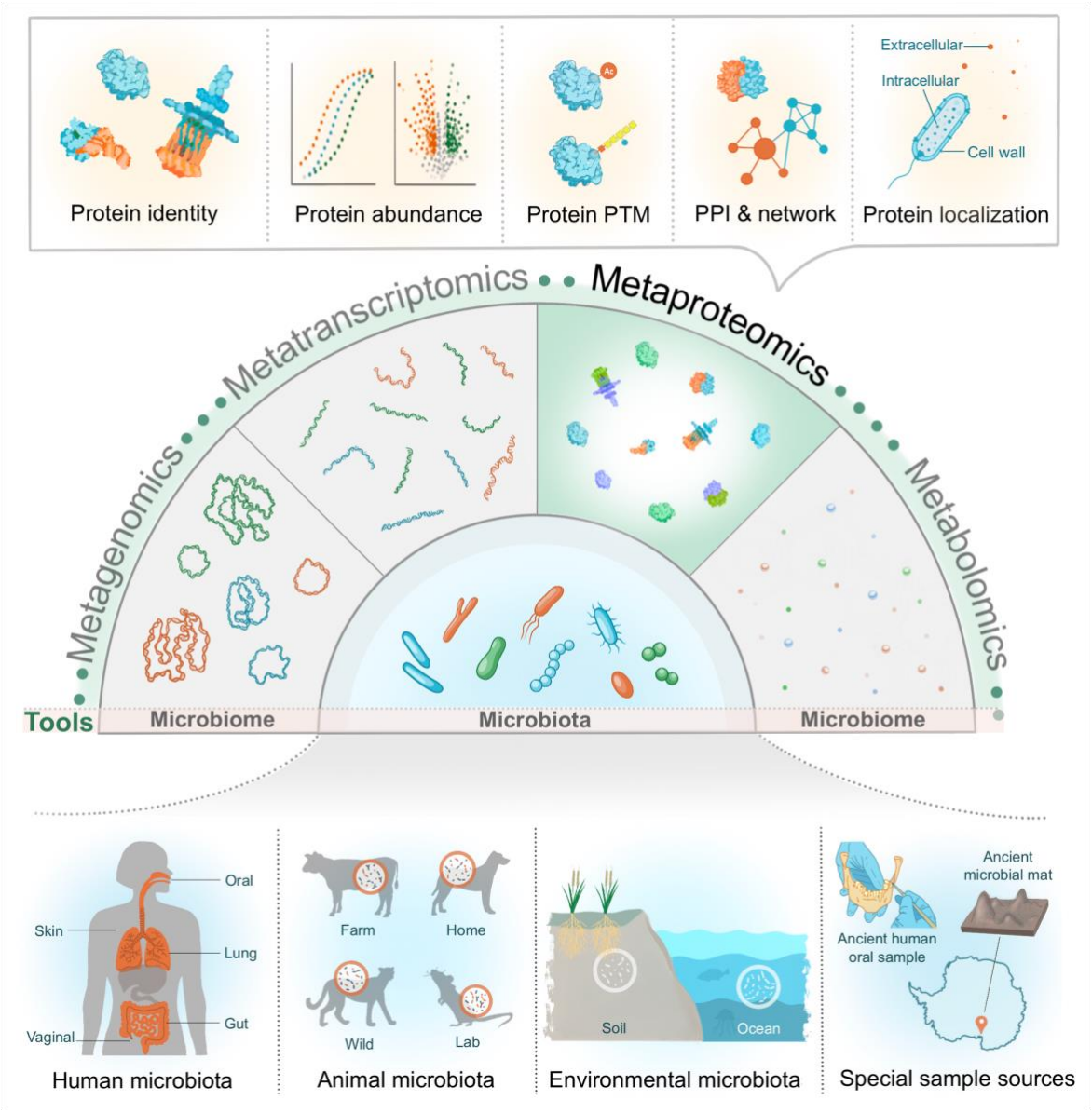
Metaproteomics can address several important questions such as:

- What are the metabolic and physiological processes of microorganisms in diverse habitats, including environmental, technical, and host-associated systems?

- How do microbiomes respond to changing conditions, as reflected by differential protein expression?
- How do microbes interact with their environment, including extracellular and intracellular protein dynamics?
- What post-translational modifications (PTMs) regulate protein activity and structure?
- How do microbiome phenotypes change over time or across spatial scales?
- How can stable isotope information from metaproteomes represent microbial activity and substrate utilization (Justice et al. 2014; Kleiner et al. 2023)?

While ongoing technological advancements are driving rapid progress, metaproteomics has already been successfully applied in the context of many impactful studies. It has contributed to fundamental understanding of microbial ecology, host-microorganism interactions, and disease mechanisms (Wolf et al. 2023). It has also improved biotechnological processes such as anaerobic digestion and wastewater treatment (Kleikamp et al. 2023; Justice et al. 2014; Heyer et al. 2024; Francesco Delogu et al. 2024), supported environmental monitoring (Pan, Wattiez, and Gillan 2024), and improved agricultural productivity (Andersen et al. 2021; Xue et al. 2024). Furthermore, it has applications in describing historical heritage and solving forensic questions (Jarman et al. 2018). Readers interested in further details on the benefits of metaproteomics can explore several recommended reviews (Heintz-Buschart and Wilmes 2018; Sun, Ning, and Figeys 2024; Herbst et al. 2016; Hettich et al. 2013; Kleiner 2019) and perspectives on its future (Van Den Bossche, Arntzen, et al. 2021; Wilmes, Heintz-Buschart, and Bond 2015; X. Zhang and Figeys 2019; Armengaud 2023).

This review, prepared by the Metaproteomics Initiative, aims to serve as a practical and accessible guide to metaproteomics. A detailed overview of the organization and presentation of this collaborative work is provided in **Section 5**, highlighting our dedication to delivering a comprehensive and valuable resource for the microbiome research community.



187

188 **Figure 1. Overview of metaproteomics within the multi-meta-omics toolbox applied to diverse**
189 **microbiome research domains.** This figure highlights the role of metaproteomics in identifying
190 proteins, quantifying their abundances, detecting post-translational modifications (PTMs), mapping
191 protein-protein interactions (PPIs), and determining protein localizations. Metaproteomics
192 complements other omics approaches, including metagenomics, metatranscriptomics, and
193 metabolomics, to provide a comprehensive understanding of microbial systems. Examples of
194 microbiome research domains include the human microbiome (oral, skin, gut, lung, vaginal), animal
195 microbiomes (farm, wild, and laboratory animals), environmental microbiomes (soil, ocean), and
196 special sample sources (e.g., ancient microbiome samples).

2. Basics of proteomics

Proteins are the essential structures and machinery that execute the instructions encoded in DNA, performing tasks ranging from catalyzing biochemical reactions to providing structural support. The term "proteome" refers to the complete set of proteins expressed in a cell, tissue, or organism (Wilkins et al. 1996). Proteomics, as a field, seeks to uncover the identities, quantities, structures, interactions, and modifications of proteins to better understand their roles in biological systems.

Although the term "proteome" was coined in the mid-1990s, its foundations lie in decades of protein biochemistry research that continues to shape modern proteomics. One of the earliest applications of proteomics combined gel electrophoresis (1D and 2D) with mass spectrometry techniques such as MALDI and ESI-MS/MS (James et al. 1993). Initially, protein samples were separated on a combination of 1D and 2D gels. One gel was electro-blotted onto a nitrocellulose membrane and stained using amido black, while the other gel was silver-stained for higher sensitivity. Protein bands or spots were excised from the nitrocellulose membrane, digested with trypsin, and identified using mass spectrometry. Aligning the nitrocellulose membrane with the silver-stained gel allowed researchers to locate bands that were difficult to visualize on the less-sensitive stain. Subsequent improvements, such as in-gel digestion, eliminated the need for electro-blotting. Early proteomics efforts also gave rise to software tools that automated protein identification, and therefore replacing manual annotation of peptide sequences. Many of these early innovations however, formed the basis for modern proteomics workflows.

The development of gel-free proteomics marked a significant advancement in the field. This approach bypasses gel-based separation, proceeding directly from protein extraction to digestion and mass spectrometry. Gel-free methods catalyzed a wave of new techniques, reagents (e.g., SILAC, ICAT, ITRAQ), and software, which collectively improved protein identification, PTM analysis, quantitation, and multiplexing. Tasks that were once labor-intensive with 2D gel MS became faster and more accessible through gel-free workflows. Moreover, mass spectrometers, which were initially optimized for small molecule research, were adapted for proteomics. Over the past 15 years, proteomics-dedicated mass spectrometers have been developed, offering greater speed, sensitivity, and accuracy in peptide identification and quantitation.

Proteomics today falls into two broad methodological categories: shotgun (or bottom-up) proteomics (Diz and Sánchez-Marín 2021) and top-down proteomics (Habeck et al. 2024). Shotgun proteomics, the more widely used approach, involves enzymatic digestion of

proteins into peptides, which are analyzed by mass spectrometry. This method is robust and effective for protein identification and quantification. In contrast, top-down proteomics directly analyzes intact proteins, providing insights into sequences, structures, and modifications. Although top-down proteomics offers unique advantages, it is technically demanding, less commonly used in single-species proteomics, and not currently applied in metaproteomics.

A typical bottom-up proteomics workflow begins with the enzymatic digestion of proteins, most commonly using trypsin, into smaller peptides. These peptides are separated through liquid chromatography and analyzed by tandem mass spectrometry (LC-MS/MS). In the mass spectrometer, the peptides are ionized, and their intact forms are detected to generate MS1 spectra. The peptides are further fragmented to produce MS2 spectra, which are analyzed by proteomics software. In most cases, database searches match these spectra to theoretical spectra derived from protein databases. This approach enables the identification and quantification of peptides and their corresponding proteins. For those seeking a deeper understanding of proteomics, numerous resources and reviews provide detailed insights into the field (Matthiesen and Bunkenborg 2013; Shuken 2023; Jiang et al. 2024; Sinitcyn, Rudolph, and Cox 2018).

3. Experimental methods in metaproteomics

Metaproteomics expands upon proteomics techniques, leveraging high-resolution LC-MS/MS instruments (Gómez-Varela et al. 2023; Dumas et al. 2024) and accompanying software tools for mass spectra identification. However, metaproteomics goes beyond the straightforward application of proteomics to microbiome research. Its added complexity arises from the requirement to consider both species-specific and functional annotations for each protein. Additionally, the presence of protein homologs across phylogenetically related species within a single sample further complicates protein inference.

The key distinctions between proteomics and metaproteomics lie in the taxonomic and functional complexity of microbiomes, the vast size of microbiome databases, and the challenges associated with sample processing, as well as the identification and quantitation of peptides and proteins. Additionally, specialized bioinformatic and statistical tools are required to track both the taxonomic and functional annotations of peptides and proteins. These aspects, which are unique to metaproteomics, will be discussed in detail throughout the remainder of this article.

This section provides an essential foundational guide to start with metaproteomics studies (**Figure 2**). We outline the basic principles for each step, starting with experimental design (**Section 3.1**), followed by sample collection, preservation, and preprocessing (**Sections 3.2–3.3**). Protein sample preparation is then described, covering both manual workflows (**Sections 3.4–3.5**) and automated workflows (**Section 3.6**). Next, we explain the basics of MS data acquisition (**Section 3.7**), before delving into the detailed bioinformatics workflows used in metaproteomics (**Sections 4.1–4.3**).

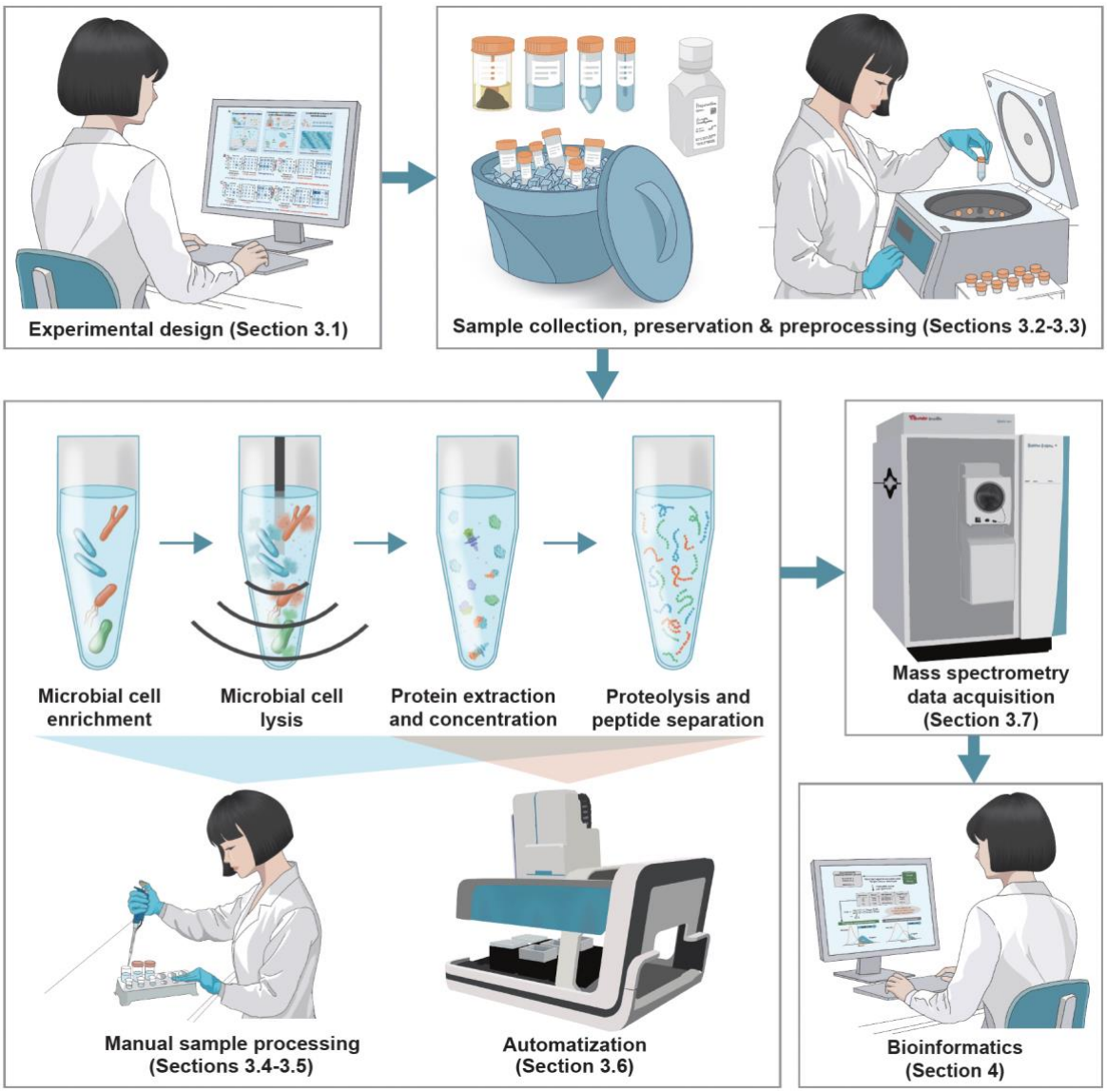


Figure 2. Overview of key principles and workflows in metaproteomics, aligned with corresponding subsections in this review.

3.1. Experiment design

3.1.1. Aligning experimental design with the scientific question

A well-designed metaproteomics experiment forms the basis for generating meaningful insights that directly address the scientific question being studied. Most importantly, the experimental design must align with the specific scientific question being addressed and the resources available to answer that question. Broadly, three experimental scenarios can be outlined (**Figure 3A**):

i) Unique sample without a control: The goal here is to provide a comprehensive description of the taxonomic and functional units present in the sample, although comparisons with a control are not possible. Examples include desiccated material from a historical Antarctic ice core (Lezcano et al. 2022), a unique biofilm from an industrial storage pool (Pible et al. 2023), residues from an ancient tomb (Charlier et al. 2024), or medieval dental calculus (Jersie-Christensen et al. 2018) were analyzed using metaproteomics. Differential functional abundances among the identified microorganisms can reveal their metabolic specialization.

ii) Comparison of microbiomes under different conditions: This common approach highlights differences between conditions. Comparisons may involve two conditions (i.e., condition A vs. condition B) or more complex setups with multiple conditions. Specific cases include dose-response analyses, where a single parameter such as stress intensity is modified, or spatial comparisons. Examples include characterizing microbial communities along a 5,000 km Pacific Ocean transect (Saunders et al. 2022) or analyzing microbiome responses to various xenobiotics *in vitro* (L. Li et al. 2020).

iii) Longitudinal analysis of a single microbiome or multiple microbiomes: This strategy captures temporal dynamics within a microbial community, and potentially the host's response, by analyzing the same microbiome at different time points. A more complex approach examines temporal changes across multiple conditions or sampling sites. Examples include monitoring gut microbiomes in Crohn's disease patients post-resection surgery over one year (Blakeley-Ruiz et al. 2019) or monthly analyses of specialized microbiomes in a two-stage anaerobic digester for lignocellulose breakdown, tracking the dynamics between hydrolytic and methanogenic subsystems (Heyer et al. 2024).

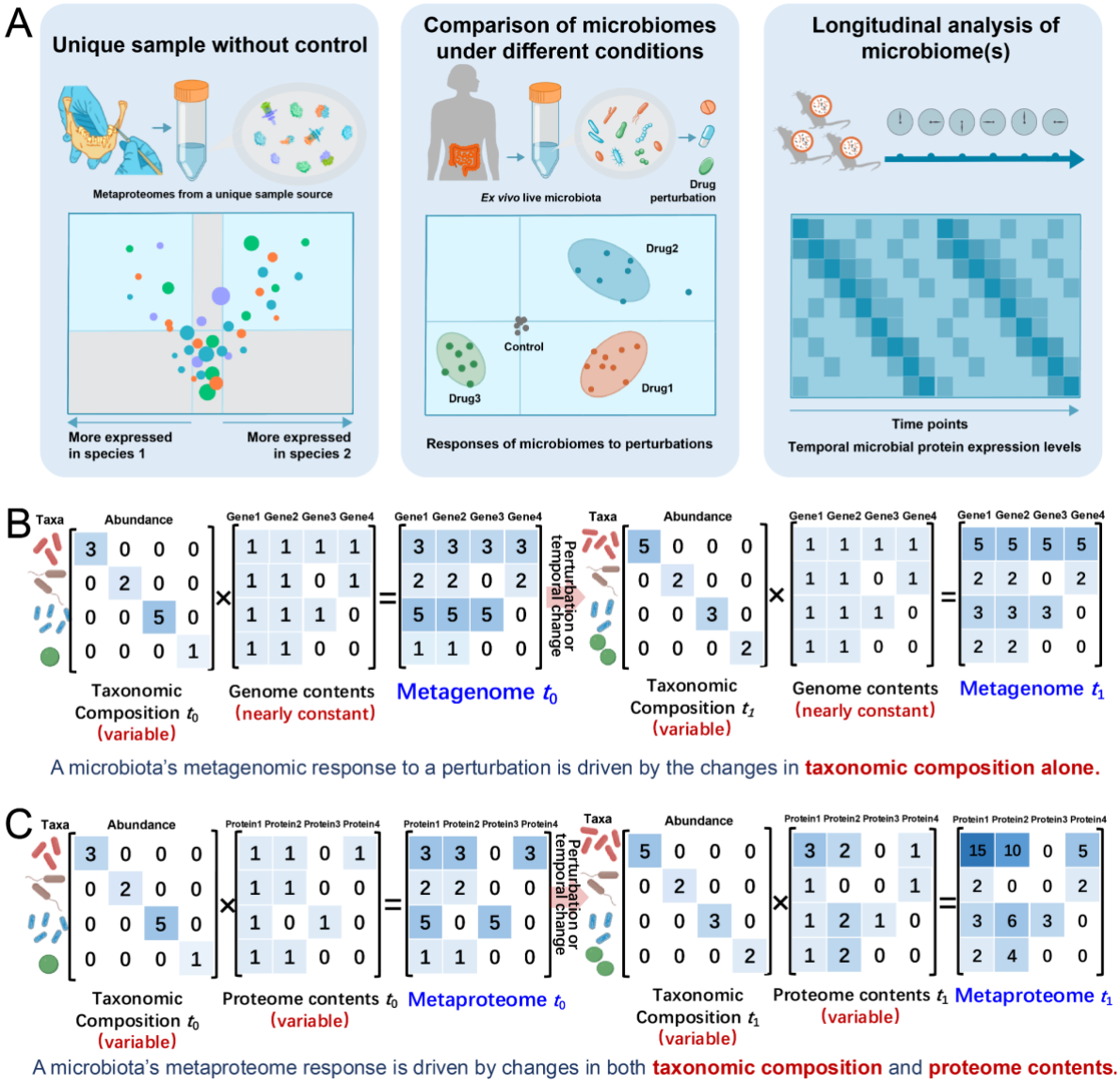


Figure 3. Metaproteomic experimental designs and their comparison with metagenomics in studying microbiome dynamics. (A) Overview of common metaproteomic experimental designs. The left panel illustrates the comparison of microbial protein expression between species within a unique sample source, lacking a control. The middle panel compares microbiomes under varying conditions, such as drug treatments, using *ex vivo* microbiomes to assess microbial responses. The right panel shows longitudinal studies that monitor temporal changes in microbial protein expression over time. (B) Metagenomic responses to perturbations, showing shifts in taxonomic composition while assuming genome content remains relatively constant. (C) Metaproteomic responses to perturbations, showing changes in both taxonomic composition and proteome content. This approach captures microbial abundances and their functional contributions, providing deeper insights into microbiome dynamics.

Some readers may already have experience designing experiments for metagenomics and understand its principles. In contrast, metaproteomics offers a different perspective on microbiome changes (**Figure 3B**). Metagenomics captures shifts driven by changes in taxonomic composition, as genomic content within a sample is relatively constant. This approach reveals species abundance and diversity but does not provide functional insights. Metaproteomics, on the other hand, measures not only taxonomic changes through taxon-specific peptide intensities but also dynamic functional responses through proteome variations across taxa. This makes metaproteomics particularly well-suited for comparing microbiomes under different conditions or for longitudinal studies.

When selecting conditions or time points for a kinetic analysis, careful consideration is essential. Comparisons between vastly different samples, such as a soil microbiome versus a human gut microbiome, are in general uninformative, while overly similar samples may show no significant differences. Selection should be guided by a clear rationale and preliminary observations. The reference condition or time point depends on the scientific question but may involve using a mixture of all samples as a reference. While this approach increases peptide diversity in the reference sample, it can complicate analysis if the full diversity is not captured by the analytical workflow (Armengaud 2023) as further detailed in **Sections 3.5 and 3.7**.

Potential confounding factors must also be accounted for during experimental design. Comprehensive metadata collection is critical, including information on sampling location, timing, storage, processing conditions, and data acquisition. Additional metadata, such as weather conditions on sampling days, patient medication, or health status, may also be essential for interpreting results. Additionally, researchers should also consider using additional material to create appropriate databases for matching spectra to peptides and for testing methodologies before processing all samples. More details on proteomics software and database creation are provided in **Section 4.1.1 and 4.1.2**, respectively.

Finally, while a limited number of metaproteomics studies have used metabolic labeling (e.g., to study host-microbiome or plant rhizosphere interactions (Z. Li et al. 2019; Smyth et al. 2020; Sachsenberg et al. 2015), this approach is often impractical for environmental or human microbiome samples. Metabolic labeling, as briefly mentioned in **Section 2**, involves incorporating heavy isotopes like 15N or 13C into proteins through labeled substrates, enabling the study of metabolic crosstalk and protein production rates. However, its limited applicability means that it is not further discussed in this review.

3.1.2. Reproducibility & statistics

The high complexity and heterogeneity of metaproteomics samples necessitate careful consideration of statistical power and steps to ensure reproducibility during experimental design. Biological, technical, and analytical replicates are key to producing reliable data and accurate interpretations. Increasing the number of biological replicates improves the ability to detect smaller differences, even in the presence of high variability. When only slight differences between conditions are expected, the use of pooled samples may also be considered. Technical and analytical replicates are necessary to account for noise introduced during measurement. It is advisable to first evaluate the variability of sample preparation and the analytical workflow using a representative sample. Additionally, randomizing the order of samples before LC-MS/MS analysis reduces the risk of bias due to the sequence in which they are processed (Nakayasu et al. 2021). For cases where specific sources of variability, such as batch effects, are known, blocked randomization is preferable to further minimize bias (Oberg and Vitek, 2009). Rigorous quality control (QC) is essential during the LC-MS/MS phase of the metaproteomics workflow to ensure data reliability and consistency. **Section 3.7.4** provides further details on these QC procedures.

Determining the appropriate number of biological replicates is essential to detect meaningful biological differences, such as variations in taxon biomasses, protein abundances, or metabolic pathways. Power analysis is typically used to calculate the required sample size, but it can be challenging in metaproteomics due to the complexity of experimental designs and the inherent variability of samples. When precise endpoints are unavailable, rough estimates from similar studies can serve as a guide. Power analysis considers several key factors: the effect size, which reflects the expected magnitude of differences between groups and helps determine the necessary sample size; the significance level (α), usually set at 0.05 to allow a 5% risk of false positives; statistical power (1 - β), often set at 0.8 or higher to reduce the likelihood of failing to detect a true effect; and the variability in the data, which can be estimated from pilot studies or previous literature on comparable experiments. In studies involving complex microbial communities, deriving precise sample size estimates may be impractical, but approximate estimates remain a valuable approach (Ferdous et al. 2022). Conducting power analysis is critical for avoiding underpowered studies and ensuring efficient use of resources (Levin 2011; Ferdous et al. 2022).

3.2. Sample collection, preservation and storage prior to preprocessing

3.2.1 Sample collection and preservation

Metaproteomics has been applied to a variety of samples, including microbial communities from environmental niches such as water, soil, sewage, aerosols, and rocks (Starke, Jehmlich, and Bastida 2019; Nebauer, Pearson, and Neilan 2024). It has also been used to analyze microbiomes in fermented foods and beverages (L. Yang, Fan, and Xu 2020; Okeke et al. 2021) and in associations with various higher eukaryotes, including arachnids, insects, worms, mollusks, fish, plants, birds, and mammals (Ezzeldin et al. 2019; Andersen et al. 2021). In mammals and other vertebrates, metaproteomics has been applied to numerous body sites across the digestive, respiratory, and urogenital systems (Y. Wang et al. 2020; Wolf et al. 2023). However, many microbiomes remain unexplored by metaproteomics.

The choice of collection method significantly influences the resulting metaproteomic profile by altering the ratios of microbial to non-microbial components and the relative abundances of microbial taxa. Collection strategies also introduce operator-dependent variability, making user-friendly devices especially valuable for self-sampling of clinical specimens. Microbiome samples are often collected directly into sterile tubes or containers. This method is common for non-invasive clinical samples, such as feces, saliva, sputum, and urine, which can often be self-collected by study participants (Long et al. 2020; Arıkan et al. 2023; Graf et al. 2021; XiaoLian Xiao et al. 2022). For clinical specimens requiring surface sampling, swabs, spatulas, or syringes are often used for oral, nasal, and cervicovaginal samples (Chen et al. 2024; Bihani et al. 2023; Berard et al. 2023), while periodontal curettes or paper strips are used for tooth- and gingiva-associated microbiomes (Rabe et al. 2022; Xiaolian Xiao et al. 2023). Invasive procedures, such as bronchoalveolar lavage, endotracheal aspiration (Pathak et al. 2020), intestinal biopsies (Jabbar et al. 2021), colonic luminal aspirates (X. Zhang et al. 2020), and surgical collection of colonic contents (Tanca et al. 2022), are necessary for some specimens. Similarly, gastrointestinal fistulation (Deusch et al. 2017) and post-mortem dissection (Haange et al. 2019) are used for collecting samples from laboratory or field animals. For environmental samples, specialized devices such as quartz filters for bioaerosols (Meyer et al. 2023) and large-volume water transfer/filter systems for aquatic environments (L.-F. Kong et al. 2021; S. Wang et al. 2024) are commonly employed. More complex ecosystems may require multi-step collection protocols (Aylward et al. 2012).

Microbiome sampling inherently involves the translocation of microbial communities from their native environment to laboratory conditions. During this transition, microbial communities are highly sensitive to environmental changes such as temperature, humidity, and exposure to chemical or biological agents. These factors can induce substantial alterations in the metaproteome profile. To minimize artifacts, protein extraction should ideally occur immediately after sampling. However, immediate processing is often impractical, particularly in large-scale studies or field collections. In such cases, proper transport and storage procedures are crucial to preserving the microbiome's original biological functions. This is especially important for low-biomass or low-diversity microbiomes, which are more vulnerable to rapid shifts in their composition and activity due to external stimuli.

3.2.2. Storage conditions to maintain sample integrity

Proper storage is critical to preserving the integrity of microbial proteins and ensuring reliable downstream analyses. Exposure to environmental changes, such as air exposure, temperature fluctuations, or nutrient depletion, can significantly alter protein profiles, leading to misleading results. For instance, air exposure can introduce oxidative stress and enrich bacterial superoxide dismutase enzymes, which may bias colorectal cancer studies by mimicking disease-specific characteristics (Long et al. 2020). Therefore, appropriate storage immediately after sample collection is essential to maintain the microbiome's original state.

The standard practice for preserving metaproteomic samples involves flash-freezing in liquid nitrogen, followed by storage at −80°C. This approach minimizes molecular degradation and prevents alterations in protein abundances. While this method is highly effective, some experimental setups do not allow for immediate freezing. In such cases, alternative preservation methods may be employed. Solutions like PBS (Delgado-Diaz et al. 2022), Amies liquid medium (Bankvall et al. 2023), NAP buffer (Mordant and Kleiner 2021), and other commercially available liquid reagents (Birse et al. 2020) have been tested for their ability to enhance storage conditions or enable room-temperature preservation in metaproteomics. Protease inhibitors are often added to biological fluids such as saliva to prevent uncontrolled proteolysis (Ruan et al. 2021). RNAlater or RNAlater-like treatments have shown potential for preserving protein profiles in intestinal and marine samples, although with conflicting results (Mordant and Kleiner 2021; Jensen, Wippler, and Kleiner 2021; Saito et al. 2011). Regardless of the method used, compatibility with downstream protein extraction, digestion, and analysis steps is crucial. Common pitfalls, including

polyethylene glycol (PEG) contamination from plasticware, keratin contamination from handling, and interference from detergents or salts, must be carefully managed.

Alternative long-term storage strategies, such as freeze-drying or storing samples at −20°C, in liquid nitrogen tanks, or as lyophilized powders, also require careful evaluation. These approaches may be suitable for some sample types but may not consistently maintain protein integrity. For example, frozen intact stool material has been shown to be more stable than extracted proteins when stored at −80°C, underscoring the importance of selecting storage strategies tailored to the specific sample type (Morris and Marchesi 2016).

It is important to note that the stability of proteins during storage is highly dependent on the sample type and storage conditions. For example, the activity and stability of soil proteins are influenced by temperature, duration of storage, and soil organic matter content (Bandick and Dick 1999; Keiblinger et al. 2016). For studies involving prolonged transport or storage, incorporating a straightforward mock community can provide valuable controls to assess sample stability and detect potential storage-induced changes (Nebauer, Pearson, and Neilan 2024).

3.3. Sample preprocessing

Sample preprocessing ensures the removal of contaminants and debris, which can hinder protein extraction, degrade analytical quality (Heyer et al. 2019), and dilute biologically relevant signals. This step, as in other gene expression measurement workflows, ensures the enrichment of microbial fractions and improves the quality of downstream analysis. Ideally, preprocessing should involve minimal, rapid, and reproducible steps. Since no standardized protocols for metaproteomics (or metagenomics) currently exist, methods must be tailored to the specific sample type and evaluated based on the study's objectives (Tanca et al. 2015; Salvato, Hettich, and Kleiner 2021; Pettersen et al. 2022). While the breadth of samples processed for metaproteomics remains limited, this field is rapidly evolving, and many more methods are expected to emerge.

For soil samples, humic substances derived from decomposed organic material often co-extract with proteins, interfering with MS measurements (Benndorf et al. 2007; Waibel et al. 2023). To address this, several methods have been developed to remove humic compounds while preserving protein integrity before digestion (Keiblinger et al. 2012; Giagnoni et al. 2011; Chourey et al. 2010; Bastida, Hernández, and García 2014). Alternatively, filter-aided sample preparation (FASP) can directly digest proteins within humic complexes. This method uses acidification to precipitate humic compounds and

undigested proteins while peptides are extracted via centrifugation through molecular weight cut-off filters (Qian and Hettich 2017).

For human gut microbiome samples, non-microbial proteins from host cells and food debris are often much more abundant than microbial proteins, reducing the efficiency of microbial metaproteome identification (X. Zhang and Figeys 2019). Techniques such as double filtering (Xiong et al. 2015) and differential centrifugation (Tanca et al. 2014) can enrich microbial cells to improve identification. However, these methods may introduce biases and depend on the study's goals (Tanca et al. 2015). For example, double filtration can remove host cells and exoproteins, while differential centrifugation may non-specifically remove microbial cells and proteins (Speda et al. 2017; Armengaud et al. 2012; A. Wang et al. 2024). Moreover, these methods are time-consuming and may be influenced by fecal variability, such as texture, fiber, and water content. Automation technologies, including solid-phase extraction clean-up, have been proposed to streamline processing for large longitudinal studies, reducing variability and improving reliability (Gonzalez et al. 2020).

In studies analyzing heterogeneous samples with high host protein content, such as viscous sputum of cystic fibrosis patients, certain plant tissues or environmental samples, a homogenization step can improve sample consistency. This step should be performed under conditions (temperature and duration) that minimize alterations to the *in vivo* metaproteome. Various mechanical strategies can achieve homogenization, including laboratory mills (Graf et al. 2021) and glass homogenizers (Salvato et al. 2022). The addition of protease inhibitors and DNase I to prevent protein degradation and disrupt DNA-based aggregates may also be beneficial, yet should be carefully evaluated based on the sample type and study objective.

For clinical samples containing bacterial or viral pathogens, inactivation is required before further processing outside appropriate biosafety level (BSL) containment. Since no standardized pipeline exists for this step, protocols must be tailored to the specific pathogen and sample type. Methods such as heat inactivation in lithium dodecyl sulfate buffers (Grenga et al. 2022) and MPLEx extraction, which uses chloroform, methanol, and water (8:4:3) for simultaneous pathogen inactivation and fractionation into metabolite, protein, and lipid phases, are commonly used (Burnum-Johnson et al. 2017). These approaches ensure both safety and compatibility with downstream metaproteomics workflows.

3.4 Protein sample preparation: from extraction to digestion

Preparing protein samples from biological material involves a series of interconnected steps, each essential for obtaining high-quality metaproteomic data. The term "protein extraction" is often used broadly to describe the entire workflow of isolating proteins from a biological sample. This process typically begins with cell lysis using extraction buffers and may also include subsequent protein clean-up steps, such as precipitation, filtration, or other methods. In some workflows, however, protein clean-up is treated as a distinct step, especially in protocols where extraction, clean-up, and digestion are streamlined into a single process. This section provides an overview of the key stages in protein sample preparation: cell lysis and extraction (**Section 3.4.1**), protein clean-up (**Section 3.4.2**), protein concentration (**Section 3.4.3**), and protein digestion (**Section 3.4.4**).

3.4.1 Cell lysis and protein extraction

Cell lysis releases the proteome from microbial cells, with a variety of methods available, each with distinct advantages (Hansmeier, Sharma, and Chao 2022). Mechanical disruption methods, such as direct ultrasonication, non-contact ultrasonication, and bead beating, are commonly used. Ultrasonication usually involves direct ultrasonication, where the probe is directly inserted into the sample, or non-contact ultrasonication, where the sample in a tube receives sonication energy from a cup horn through a coupling fluid. An advanced non-contact method termed Adaptive Focused Acoustic (AFA) technique provides precise control over parameters like amplitude and duration, achieving efficient lysis while minimizing protein denaturation (Dhabaria et al. 2015). Bead beating, which uses zirconia or silica beads, is effective for cell disruption, with bead size modulating efficiency (X. Zhang et al. 2018).

Chemical lysis methods use detergents such as urea buffers containing Triton X-100 or sodium dodecyl sulfate (SDS) to disrupt microbial cell membranes, often in combination with mechanical disruption/ultrasonication (X. Zhang et al. 2018). Notably, when combining urea-containing buffers with mechanical disruption or ultrasonication, one should be aware of the risk of urea-induced carbamylation caused by sample overheating (Kollipara and Zahedi 2013). Physical methods, including freeze-thaw cycles or high-pressure homogenization, are also effective, with pressure settings tailored to specific sample types (Cai et al. 2022). Since microbial cell structures vary significantly, for example between Gram-positive bacteria, Gram-negative bacteria, and fungi, optimizing lysis conditions is

crucial to preserve protein integrity, maximize yield, and ensure unbiased protein extraction (Starke et al. 2019; J. Wang et al. 2020).

Recently, some of the above approaches have been compared and found that a urea- and SDS-containing lysis buffer coupled to ultrasonication yielded higher protein recovery than bead beating in microbiome samples, with minimal sample loss, though both methods achieved similar peptide and protein identifications (X. Zhang et al. 2018). Careful selection of lysis buffers is also critical to avoid interference with downstream MS analysis. For example, ion suppression-inducing detergents like Tween-20 should be avoided unless they are removed during cleanup, as in methods like suspension trapping (S-trap) or FASP.

Table 1 compares commonly used protein sample preparation methods, summarizing their key advantages and disadvantages. The choice of lysis method depends on factors such as sample type, desired protein yield, and sensitivity of proteins to denaturation or degradation. The listed lysis methods can also be combined, for example, detergent-containing urea lysis buffers are often coupled with ultrasonication to achieve fast and unbiased bacterial cell lysis in complex microbiome samples.

Table 1. Comparison of standard protein sample preparation methods. This table summarizes commonly used protein sample preparation techniques, outlining their key advantages and potential disadvantages.

Method	Description	Advantages	Disadvantages
Chemical Lysis	Disrupts cell membranes with chemicals like urea or guanidine hydrochloride.	Can unfold complex proteins.	If not removed or sufficiently diluted, it can interfere with protease activity. Risk of urea-induced carbamylation.
Detergent Lysis	Uses detergents (e.g., SDS, Triton X-100) to solubilize cell membranes.	Mild, preserves protein function, ideal for membrane proteins.	If a detergent is not removed or sufficiently diluted, it can interfere with protease activity.
Freeze-Thaw	Repeatedly freezes	Simple, no special	Time-consuming,

Cycles	and thaws the sample to rupture cell membranes.	equipment needed.	may not fully lyse cells, risk of protein degradation.
Bead beating	Physical force such as using bead beating to break cell walls.	Effective for bacterial cell lysis.	Requires specific instrument, sample loss due to contact with beads, can generate heat, risk of protein degradation.
Ultrasonication	Uses ultrasound waves to break cell membranes/walls and release proteins.	Fast, effective and can be non-contact for small samples, no need for harsh chemicals.	Can denature proteins if overused, heat generation requires sample cooling.

3.4.2 Protein clean-up: precipitation and alternative methods

Protein precipitation addresses the challenges of complex environmental and fecal samples by removing contaminants such as lipids, nucleic acids, and polysaccharides that can interfere with downstream MS analysis. Following microbial cell lysis, effective separation of proteins from cellular debris and contaminants is essential to ensure high protein yield and purity. Removing contaminants not only improves protein recovery but also enhances MS sensitivity, enabling more accurate and reliable protein identification.

The trichloroacetic acid (TCA)/acetone precipitation method is widely employed for this purpose. This method involves adding cold (-20°C) TCA or acetone, or both, to the protein lysate to precipitate proteins, followed by centrifugation to pellet the proteins. The pellets are then washed with cold acetone (-20°C) to remove residual contaminants and insoluble particles (Nickerson and Doucette 2020). This approach has proven effective for high-yield protein precipitation in diverse sample types, including marine sediment and forest soil samples, which contain complex organic matrices (Niu et al. 2018). Similarly, acidified acetone/ethanol buffer has also been used in metaproteomics (X. Zhang et al. 2016).

An alternative method, phenol extraction, separates proteins into the organic phase while partitioning nucleic acids into the aqueous phase. This approach is particularly beneficial for "dirty" samples, such as soil and wastewater sludge, which are rich in organic and inorganic contaminants. Phenol extraction can reduce the interference caused by contaminants, thus improving the downstream analysis of target proteins (Benndorf et al. 2009). Phenol extraction also enables the simultaneous extraction of nucleic acids from the same sample, making it highly suitable for integrated omics studies, especially in microbiome research (Baldrian 2017).

For samples with low microbial load, such as fecal samples, river sediment, or air filters, maximizing protein recovery is critical. Organic solvent systems, such as chloroform/methanol or chloroform/methanol/water mixtures, have proven effective for enhancing protein recovery and minimizing the loss of low-abundance proteins by optimizing solvent ratios and conditions (Vertommen et al. 2010). Biphasic systems, such as phenol/chloroform or Triton X-114, can also be used to selectively partition proteins and facilitate the removal of contaminants (Wessel and Flügge 1984).

Traditional protein precipitation methods, while effective, can be labor-intensive and may not always completely eliminate contaminants that interfere with downstream analyses. To address these limitations, alternative methods have been developed to improve protein clean-up and digestion efficiency. Techniques such as FASP, SP3, and suspension trapping (S-Trap) have shown promise for processing challenging samples like human fecal protein extracts (Tanca et al. 2024). Solid-phase alkylation, a novel strategy designed for low-loss and anti-interference sample preparation, utilizing covalent binding and purification of proteins, has also been proved effective for marine microbiome samples (S. Wang et al. 2024). These approaches integrate clean-up and digestion steps into a single workflow, facilitating high-throughput applications.

3.4.3 Measuring protein concentration

Accurate protein concentration measurement ensures uniform loading in downstream LC-MS/MS analyses and for facilitating reliable data interpretation (Sapan and Lundblad 2015). Consistent peptide loading in LC-MS/MS is essential for accurate peptide quantification, as it maintains signal intensity and ensures reliable peptide detection across samples. Uniform loading also optimizes column performance, reducing variability in peak shapes and retention times. This consistency minimizes technical artifacts, enabling clearer biological insights when comparing samples.

Various methods are commonly used to determine protein concentration. The Bradford Assay, which utilizes Coomassie Brilliant Blue dye, measures protein concentration through a colorimetric change, requiring a standard curve prepared with known protein concentrations to ensure precision. The bicinchoninic acid (BCA) assay forms a purple-colored complex for protein quantification, with sensitivity optimized by adjusting reagent ratios and incubation conditions. Fluorescence-based assays, such as the Qubit Protein Assay, use dye-binding technology for highly sensitive quantification with minimal interference, making them suitable for samples with low protein concentrations.

The 2-D Quant Kit is another option, which quantitatively precipitates proteins while leaving interfering substances in solution. This method produces a color density inversely related to protein concentration, with a linear response in the range of 0–50 µg and a volume range of 1–50 µL. When selecting a protein concentration method, it is important to consider the required sensitivity, dynamic range, and compatibility with buffer components, as some assays show varying tolerance to substances like SDS or protease inhibitors, including PMSF.

If no suitable quantification assay is available, running SDS-PAGE gels can provide a rough estimate of protein abundance. While less precise, this approach can offer a practical alternative for assessing protein concentrations in certain scenarios.

This systematic approach to protein concentration measurement ensures consistency and reliability in downstream analyses, particularly when dealing with complex microbial samples containing proteins spanning a wide range of abundances.

3.4.4 Protein digestion

Bottom-up (shotgun) metaproteomic studies involve the enzymatic digestion of proteins into peptides, a process known as proteolysis, for untargeted protein identification. This method requires several preparatory steps to ensure efficient proteolysis. Initially, proteins are denatured using agents such as urea or guanidine hydrochloride to expose cleavage sites. Disulfide bonds are then reduced using reducing agents like dithiothreitol (DTT) or tris(2-carboxyethyl)phosphine (TCEP). To prevent the re-formation of disulfide bonds, cysteine residues are alkylated with agents like iodoacetamide, which react with sulfhydryl groups to form stable thioether adducts (Sechi and Chait 1998). This alkylation introduces mass changes that must be accounted for during peptide identification, as discussed in **Section 4.1.1**.

Following these preparatory steps, proteins are enzymatically cleaved into peptides suitable for downstream LC-MS/MS analysis (Hustoft et al. 2012). The most commonly used protease is trypsin due to its high specificity and efficiency. It cleaves proteins at the C-terminal side of lysine and arginine residues, producing peptides ideal for shotgun MS analysis. Lys-C, another commonly used protease, complements trypsin digestion by cleaving at the C-terminal side of lysine residues, particularly in high urea concentrations (8 M), enhancing peptide coverage. Alternative proteases such as chymotrypsin, Glu-C, and Asp-N may also be used to increase peptide diversity or for specific applications. However, the combination of trypsin and Lys-C is often the most practical and widely applied choice.

The enzyme-to-substrate ratio is another important factor, with typical ratios ranging from 1:50 to 1:100 (w/w). Digestion time is also critical and usually involves incubating the proteome mixture at an appropriate temperature (e.g., 37°C) for several hours to overnight, depending on sample complexity and enzyme properties. Digestion is quenched by acidification, commonly using formic acid or trifluoroacetic acid to achieve a pH of 2–3. In methods such as S-trap or FASP, peptides may also be eluted without an acidification step.

Peptide lysates are subsequently desalted or purified to remove salts and contaminants. Solid-phase extraction (SPE), C18 ZipTips (Millipore), or ultrafiltration are commonly used for this purpose. In some cases, the desalting step can be omitted if peptides are desalted on a trap column in the LC system.

Direct in-solution protein digestion methods have been developed to streamline the workflow, offering efficient and high-throughput options. Notable examples include SP3 (Hughes et al. 2014), FASP (Wiśniewski et al. 2009), S-trap (HaileMariam et al. 2018) and a commercial kit based on the in-StageTip (iST) (Kulak et al. 2014). These methods are designed to ensure high protein recovery and compatibility with downstream MS analysis, even when working with low protein amounts.

3.5 Separation and fractionation techniques

Separation and fractionation enable researchers to reduce sample complexity and enhance the depth and sensitivity of protein identification and quantification. These processes can be performed at multiple levels, including the peptide, protein, and cellular stages, depending on the specific goals of the analysis (Cheng et al. 2018). Techniques such as peptide fractionation are frequently used to enhance LC-MS/MS performance (**Section 3.5.1**), while enrichment approaches allow for the targeted analysis of PTMs (**Section 3.5.2**).

At the protein or cellular level, fractionation strategies can further refine sample complexity or enrich specific components of interest (**Section 3.5.3**).

3.5.1 On-line and off-line peptide fractionation

Peptide separation workflows can generally be categorized into one-dimensional (1D) and two-dimensional (2D) or multi-dimensional approaches. In 1D liquid chromatography (LC), which is widely used in metaproteomics, reverse-phase (RP) nano-high-performance liquid chromatography (nanoHPLC, mostly just abbreviated as LC or HPLC) employs C18 columns to separate peptides based on their hydrophobicity and is coupled directly with mass spectrometry for peptide analysis. 2D-LC, often based on multidimensional protein identification technology (MudPIT) (Washburn, Wolters, and Yates 2001), combines strong cation exchange (SCX) with RP-HPLC. Peptides are first fractionated on the SCX column based on their charge using salt or pH gradients for elution, and then further separated based on hydrophobicity on an RP-HPLC column (Verberkmoes et al. 2009). The 2D-LC strategy has been applied in metaproteomic analyses to improve identification depth, with online 2D LC-MS setups used for shotgun proteomics in studies of human gut and environmental microbiomes (Verberkmoes et al. 2009).

Off-line pre-fractionation, although less commonly used in metaproteomics due to its labor-intensive nature and the increased MS time required, offers potential for deeper peptide and protein identification (X. Zhang et al. 2017). High-pH RP chromatography is one such method and is orthogonal to low-pH RP-LC-MS gradients. This fractionation can be achieved using either stage-tip methods or HPLC systems. Stage-tip-based fractionation is straightforward to implement and is supported by commercially available kits (e.g., Pierce ™ High pH Reversed-Phase Peptide Fractionation Kit). On the other hand, micro-flow HPLC systems enable higher-resolution fractionation through continuous collection of numerous fractions and stepwise concatenation.

While extensive fractionation can significantly enhance the depth of metaproteomic analysis, it also increases costs, sample requirements, and instrument time, making it less feasible for large cohort studies. The adoption of multiplexing techniques, such as tandem mass tags (TMT) (Creskey et al. 2023), has mitigated these limitations by reducing MS time and the required sample quantity per condition. The combination of off-line peptide fractionation and multiplexing presents a promising and accessible option for researchers, particularly beginners, aiming to conduct in-depth metaproteomic analyses to investigate microbiome functionality.

3.5.2 Enrichment of peptides with post-translational modifications

PTMs are critical regulators of protein activity and function, and their study is uniquely possible through metaproteomics. Unlike other omics approaches, metaproteomics provides the direct capability to identify and quantify PTMs in microbial proteins, offering unparalleled insights into microbiome functionality. While analyzing PTMs at the metaproteome level is particularly challenging, several studies have successfully performed metaPTMomics on environmental and human gut microbiomes (Z. Li et al. 2014; W. Zhang et al. 2016; X. Zhang et al. 2021; 2020). These studies identified various PTMs, including methylation, hydroxylation, acylations, citrullination, deamination, phosphorylation, and nitrosylation, among others, with abundances varying across different microbiome types. Understanding the diversity and distribution of PTMs is essential for uncovering microbiome functionality. Recent advancements in the field have been detailed in two comprehensive reviews (Duchovni, Shmunis, and Lobel 2024; Duan, Zhang, and Figeys 2023).

Microbiome PTMs can be analyzed using non-enriched samples combined with tailored bioinformatics workflows (Z. Li et al. 2014; W. Zhang et al. 2016) or quantitatively profiled using enrichment techniques at the peptide or protein level (X. Zhang et al. 2021; 2020). Depending on the type of PTM, specific enrichment strategies may be employed to facilitate detection during MS analysis.

Immuno-affinity enrichment is widely used for protein acylations, such as lysine acetylation, propionylation, and succinylation, and has recently been applied to human gut microbiomes (X. Zhang et al. 2021). This technique uses antibodies bound to agarose or magnetic beads to selectively enrich acylated peptides, improving MS sensitivity and specificity. However, this approach can be limited by the availability of motif-specific antibodies and the inability to capture the full spectrum of modified peptides.

Immobilized metal affinity chromatography (IMAC) is a commonly used strategy in proteomics to enrich phosphorylated peptides for phosphoproteomic studies. Ti-IMAC and Fe-IMAC are typical examples, offering robust enrichment prior to LC-MS/MS analysis (Low et al. 2021).

Hydrophilic interaction liquid chromatography (HILIC) is another effective technique, particularly for enriching glycopeptides. This method capitalizes on its high selectivity and specificity for hydrophilic glycan moieties (Mysling et al. 2010). These enrichment approaches have been extensively applied to mammalian cells, tissues, and single bacterial strains, and they show potential for broader applications in microbiome studies.

3.5.3 Protein, cell-level and functional fractionation techniques

The high complexity of microbiomes often necessitates cellular and protein-level separations to complement peptide-level fractionation, enhancing the depth and resolution of metaproteomic analysis. Although high-speed, high-resolution mass spectrometers have made peptide fractionation sufficient for many proteomics workflows, the added complexity of microbiomes can still benefit from upstream fractionation approaches.

Capillary zone electrophoresis (CZE), a technique used to separate charged particles, shows promise for separating intact proteins and even bacterial cells (Cheng et al. 2018). Another method for separating proteomes from different bacteria is differential lysis, which, despite its relatively low granularity, can distinguish between bacterial types based on cell wall structure (J. Wang et al. 2020). In this approach, sequential lysis is achieved using buffers of increasing strength, such as those containing urea or varying concentrations of SDS. This method can separate the proteomes of Gram-negative bacteria, which have thinner cell walls, from those of Gram-positive bacteria with thicker, multilayered cell walls (J. Wang et al. 2020).

For host-associated microbiomes, removing abundant host cells is often critical to improving microbial signal detection. Techniques such as differential centrifugation and density gradient centrifugation (Hinzke, Kleiner, and Markert 2018) are commonly used to enrich microbial cells. Following lysis, additional separation of cellular components can be achieved through methods like ultracentrifugation (Henry et al. 2022), further increasing protein identification coverage.

Functional fractionation techniques, such as Activity-Based Protein Probing (ABPP), can be used to study enzymatic functions at the proteome level (Cravatt, Wright, and Kozarich 2008). ABPP employs small-molecule probes that covalently bind to active sites of proteins with specific functions or residues. These labeled proteins can then be captured or enriched for LC-MS/MS analysis, enabling detailed profiling of protein functions and aiding in drug target discovery. ABPP is particularly useful for annotating proteins with unknown functions (Barglow and Cravatt 2007), making it a relevant approach in microbiome studies. Recent applications of ABPP in both host-associated and environmental microbiomes have uncovered diverse microbial enzymes, including thiol-containing proteases, bile salt hydrolases (BSHs), glycoside hydrolases (GHs), and β-glucuronidases (Han and Chang 2023).

3.6 Automation

High-throughput techniques have transformed sample preparation, simplifying labor-intensive steps and revolutionizing workflows in proteomics, especially as datasets continue to grow in scale and complexity (Fu et al. 2023; Burns et al. 2021). These advancements have facilitated applications such as chemical proteomics (Lin et al. 2023), biomarker detection (Paramasivan et al. 2023) and drug target discovery (Qiong Wu et al. 2024). Although automation in metaproteomics has not advanced as rapidly as in proteomics, its potential for transforming the field is immense.

Automating metaproteomics workflows offers multiple benefits, including reduced sample handling time, minimized operator-induced variability, and enhanced reproducibility. These improvements provide broader coverage of microbiome responses to environmental factors within limited experimental timeframes. Furthermore, high-throughput automated workflows allow researchers to scale up the discovery of microbiome-associated biomarkers and explore dynamic functional landscapes across diverse microbiomes. Automation also generates large datasets, enabling the application of artificial intelligence (AI) to uncover hidden patterns within metaproteomic profiles.

Automated sample processing in metaproteomics can be broadly divided into four key steps: microbial cell disruption and protein extraction (**Section 3.6.1**), protein digestion and peptide clean-up (**Section 3.6.2**), and multiplexing (**Section 3.6.3**).

3.6.1 Microbial cell disruption and protein extraction

In certain scenarios, such as working with complex clinical samples like human stool or saliva, microbial cell enrichment is often required but poses significant challenges. Sample properties can vary greatly within a dataset, complicating efforts to standardize technical parameters for automated microbial cell purification. As a result, current automated metaproteomics workflows often exclude fully automated raw sample handling steps. For example, the RapidAIM 2.0 pipeline (L. Li et al. 2024) includes manual bacterial enrichment and cell washing, with a 96-channel liquid handler accelerating pipetting steps. In contrast, the SHT-Pro protocol (Gonzalez et al. 2020), the first high-throughput pipeline specifically designed for large-scale stool sample processing, begins with the lysis of raw stool samples without prior microbial enrichment. This approach is particularly beneficial when both host and microbial proteins are of interest.

Microbial cell disruption for protein extraction can be effectively automated in a 96-well format using ultra-sonication devices designed for high-throughput workflows. These instruments facilitate efficient protein extraction, enabling downstream high-throughput protein clean-up. Several methods, including FASP, SP3, and S-Trap, have been successfully adapted to microplate-based formats, with studies showing that the combination of FASP and SP3 with iST yields the most robust results for high-throughput protein processing (Tanca et al. 2024).

3.6.2 Protein digestion and peptide clean-up

Similar to manual metaproteomics workflows, automated protein preparation typically involves protein denaturation, reduction, alkylation, and protease digestion. These steps are relatively straightforward to automate and can be performed using liquid handling platforms equipped with low-volume pipetting accuracy and heater-shaker capabilities. Therefore, protein digestion is often considered one of the least complex steps to automate in metaproteomic workflows.

Peptide clean-up, however, presents greater challenges. Typically, this step is carried out manually by skilled personnel using solid-phase extraction (SPE), C18 ZipTips, or ultrafiltration, as described in **Section 3.4.4**. During automation, sample heterogeneity at this stage can introduce variability, complicating experimental parameter control. A promising solution involves replacing centrifugation through reverse-phase columns with pipette-based mixing of reverse-phase resins. This approach has been incorporated into workflows like RapidAIM 2.0 (L. Li et al. 2024) and is supported by established proteomics automation protocols. For example, the autoSISPROT system offers all-in-tip sample preparation capabilities, demonstrating compatibility with automated platforms (Qiong Wu et al. 2024).

3.6.3 Multiplexing

The integration of automated sample handling with techniques like TMT labeling significantly enhances throughput and accelerates the discovery process in metaproteomics. However, the high cost of TMT reagents might be a challenge for broader application. One solution involves pre-aliquoting and drying TMT reagents in a 96-well plate format, a strategy that reduces reagent waste and preparation time. This approach is compatible with automated workflows, such as those used in the RapidAIM 2.0 platform, and facilitates more efficient reagent utilization (L. Li et al. 2024).

While advancements in automation have enabled notable progress in metaproteomics, most current systems are semi-automated rather than fully automated. Continued development of automation technologies is essential to further streamline workflows, enhance sample processing speed, and achieve higher throughput.

3.7 Mass spectrometry data acquisition methods

Mass spectrometry analysis of (meta)proteomes is predominantly carried out using (HP)LC-MS/MS. A fundamental limitation of mass spectrometers, even when combined with multidimensional separations, is their inability to generate fragmentation spectra (or MS/MS spectra) for all peptides in a sample within a single run. This constraint has led to the widespread adoption of data-dependent acquisition (DDA) as the dominant approach in proteomics over the past 25 years.

DDA, as discussed in **Section 3.7.1**, involves selecting the most abundant precursor ions from the MS1 spectra for fragmentation in the MS2 (or MS/MS) stage, dynamically excluding previously fragmented ions to prioritize unfragmented targets. This strategy increases the diversity of identified peptides and proteins. In metaproteomics, however, the complexity of the samples presents significant challenges for DDA, particularly in achieving comprehensive sequencing depth and coverage. Even with the latest high-resolution and highly sensitive mass spectrometers, DDA is inherently biased toward the most abundant ions, leaving many lower-abundance peptides uncharacterized. Nevertheless, DDA remains the most widely used method due to its extensive validation, established workflows, and compatibility with a broad range of analytical tools.

Data-independent acquisition (DIA), as discussed in **Section 3.7.2**, is a more recent advancement that offers an alternative approach by fragmenting all peptide ions within predefined mass-to-charge (m/z) windows, rather than selectively targeting the most abundant ones. DIA addresses some of the limitations of DDA, particularly in terms of peptide coverage and reproducibility, making it increasingly attractive for metaproteomics. However, the broader data capture in DIA results in significantly more complex datasets that require advanced computational tools for processing and analysis. While progress has been made in developing such tools, further validation and optimization are needed before DIA can become a routine method for metaproteomics.

Both DDA and DIA have distinct advantages and limitations, and their choice depends on the specific goals of the experiment, the complexity of the sample, and the available computational resources.

3.7.1 DDA

DDA is the most widely used method in proteomics, particularly in shotgun proteomics, for identifying peptides in biological samples. In DDA mode, the mass spectrometer dynamically selects a specified number of the most abundant precursor ions (commonly referred to as the "topN") for fragmentation. This prioritization ensures that the most intense ions within each acquisition cycle are fragmented into smaller ions, generating MS/MS spectra that serve as unique fingerprints for peptide identification. To enhance the detection of lower-abundance peptides, DDA incorporates a process known as dynamic exclusion. Previously selected precursor ions are temporarily excluded from subsequent fragmentation, increasing the diversity of peptides analyzed within a single run. These MS/MS spectra are then analyzed using proteomics software packages (**Section 4.1.1**).

DDA has several advantages, making it a popular choice for metaproteomics workflows. It is relatively simple to configure and analyze compared to more complex approaches like DIA, making it accessible for both beginners and experienced researchers. The one-to-one relationship between spectra and peptides reduces computational demands during data analysis, particularly when a well-curated protein database is available. More information on creating a protein database is provided in **Section 4.1.2**. Furthermore, DDA supports relative quantification of proteins using both label-free quantification (LFQ) and labeling approaches, offering flexibility for various experimental designs (**Section 4.1.5**). Its longstanding use in proteomics has also led to the development of numerous software tools and well-established workflows, enhancing its reliability and versatility.

Despite its strengths, DDA has notable limitations. Its reliance on selecting the most intense precursor ions means that low-abundance proteins may go undetected, especially in complex samples. Additionally, DDA often fails to identify the same peptides consistently across multiple runs, resulting in missing values for low-abundance proteins and complicating large-scale quantitative studies.

Overall, while DDA is not without its limitations, it remains the most widely used and versatile technique in metaproteomics (Van Den Bossche, Kunath, et al. 2021). For studies requiring deeper proteome coverage or greater reproducibility, alternative methods like DIA may offer complementary advantages.

3.7.2 DIA

DIA mass spectrometry has emerged as a powerful approach in proteomics, providing broad protein coverage, high reproducibility, and quantitative accuracy. Unlike DDA, which focuses on fragmenting a limited number of the most intense precursor ions, DIA fragments all ions within predefined m/z windows. These windows are repeatedly scanned across the entire m/z range, generating complex MS/MS spectra that provide a more comprehensive view of the proteome. This inclusivity is particularly advantageous in metaproteomics, where samples contain an overwhelming diversity of peptides and low-abundance proteins that might be missed by DDA (E. Wu et al. 2024).

DIA has demonstrated significant potential in metaproteomics applications. Its application in metaproteomics was first evaluated in gut microbiome studies (Aakko et al. 2020) and

has since expanded to various contexts, including Chinese liquor fermenter starters (Zhao, Yang, Chen, et al. 2023), and multicenter diagnostic research on tongue coating samples for gastric cancer (Chen et al. 2024). Recent advances in MS instrumentation, such as DIA-PASEF (Gómez-Varela et al. 2023) and the Orbitrap Astral (Dumas et al. 2024), have significantly improved DIA's sensitivity and resolution, enabling deeper proteome coverage in highly complex microbial communities.

One of DIA's key advantages lies in its ability to capture a broader range of peptides compared to DDA, enabling deeper proteome coverage and improved detection of low-abundance proteins (Chen et al. 2024; Gómez-Varela et al. 2023; Pietilä, Suomi, and Elo 2022; Zhao, Yang, Xu, et al. 2023; Zhao, Yang, Chen, et al. 2023; Aakko et al. 2020; Zhao, Yang, Teng, et al. 2023). Another significant advantage is its reproducibility across samples, as it is less susceptible to variations in ionization efficiency (Fernández-Costa et al. 2020). This consistency makes DIA particularly well-suited for large-scale quantitative studies.

Despite its advantages, DIA also comes with challenges, particularly in data analysis. Indeed, analyzing the complex MS/MS spectra generated by DIA requires advanced computational tools and specialized expertise which is further discussed in **Section 4.1.1**. Additionally, because DIA fragments all ions within a given m/z window simultaneously, the resulting spectra are more complex and less specific to individual peptides compared to DDA. This reduced specificity can make it challenging to confidently resolve detailed structural or sequence-level information for single peptides, limiting DIA's utility for applications that require precise characterization, such as studying PTMs or differentiating highly similar peptide sequences. These inherent trade-offs highlight the importance of carefully tailoring DIA workflows to specific research objectives.

Nevertheless, DIA's rapid advancements make it a promising tool for metaproteomics, providing the depth and reproducibility required to explore the functional landscape of microbial communities comprehensively.

3.7.3 Critical parameters to optimize the HPLC and MS methods

Optimization of HPLC and MS methods is crucial for obtaining high-quality data in metaproteomics workflows. Each parameter below plays a significant role in ensuring accurate peptide separation, identification, and quantification. Metaproteomics, with its added complexity compared to standard proteomics workflows, requires specific adjustments to many of these parameters.

i) Analytical column quality, gradient and flow rates

Peptides are commonly separated using HPLC, which is directly coupled to the MS, using either commercial or in-house analytical HPLC columns. These separations are achieved with a mobile phase composed of increasing concentrations of acetonitrile (ACN). For laboratories using in-house columns, stringent QC checks are crucial to ensure consistent column performance, as explained in **Section 3.7.4**.

Metaproteomics samples present significantly greater chromatographic challenges than single-species proteomics due to their inherent complexity (Duan et al. 2022). To address this, typical mobile phase gradients of 5–35% of 80% ACN or 5–30% of 100% ACN over 1–2 hours are generally sufficient for tryptic peptide elution. However, adjustments may be required for specific experimental setups. For example, chemically labeled digests with increased hydrophobicity often require a steeper gradient with a higher final concentration of ACN for complete peptide elution.

Efficient gradient design is essential to optimize runtime and achieve an even distribution of peptide elution across the gradient. Since fewer peptides elute at the beginning and end of the gradient, tailoring the gradient can improve separation and detection (Xu, Duong, and Peng 2009). Accurate peptide quantification requires sufficient sampling points per LC peak, making short gradients (e.g., 10-minute gradients) generally unsuitable for metaproteomics in data-dependent acquisition (DDA) mode. Comprehensive tutorials on gradient optimization are available for general proteomics (Lenčo et al. 2022), and metaproteomics specifically (Hinzke et al. 2019).

LC flow rates typically range from 200–300 nL/min. Recently, higher flow rates have gained popularity to accelerate sample duty cycles. However, these higher flow rates compromise sensitivity. Strategies to offset this limitation include increasing the sample loading amount

or using dimethyl sulfoxide (DMSO) to boost signal intensity, making higher flow rates more viable for metaproteomics workflows.

ii) MS Settings in DDA workflows

Optimizing MS parameters plays a key role in obtaining high-quality data in metaproteomics. While those new to the field are generally not expected to configure MS settings, understanding key optimization steps can provide valuable context for interpreting data and troubleshooting issues.

Accurate mass measurements require regular calibration of the mass spectrometer, which is crucial for reliable peptide identification and quantification. Additionally, source parameters such as source temperature, flow rates, and nebulizer gas pressure must be optimized to enhance ionization efficiency and maximize signal intensity. The specific optimization steps vary depending on the type of mass analyzer used, such as time-of-flight (TOF) or Orbitrap instruments. Key parameters for these analyzers include scan range, resolution, and scan speed, which must be fine-tuned to ensure precise mass measurements and resolve closely spaced peptide ions. Similarly, collision energy settings for peptide fragmentation need careful adjustment to generate high-quality fragment spectra for peptide identification.

Dynamic exclusion is a critical parameter in DDA workflows, requiring careful calibration to align with the chromatographic gradient and peak width. This setting prevents repeated fragmentation of the same peptide by excluding it temporarily after its initial fragmentation, thereby increasing peptide diversity. However, this approach poses challenges, particularly in metaproteomics. Many researchers rely on spectral counting for relative quantification, as it has been shown robust for metaproteomic datasets with significant differences in cell numbers and total protein amounts between community members (Kleiner et al. 2017). Nonetheless, dynamic exclusion can limit the number of spectra acquired for abundant peptides, leading to fewer spectral counts than expected and potentially skewing quantification accuracy. This issue is exacerbated with modern high-resolution instruments, where the correlation between peptide abundance and peptide-spectrum matches (PSMs) becomes less relevant due to faster scan rates and increased resolving power. Dynamic exclusion times must therefore strike a balance, ensuring high-quality fragmentation spectra while maximizing the diversity of peptides analyzed. The choice between spectral counting and MS1-based quantification methods like area under the curve (AUC) remains a topic of debate in metaproteomics.

In DDA, selecting the isolation window width for precursor ions is a critical optimization step. A wider isolation window, up to 2 Da, allows the collection of more ions, resulting in higher-quality MS spectra. However, this increases the risk of generating chimeric spectra, where fragments from multiple precursor ions are combined, complicating peptide identification. Conversely, narrower isolation windows, down to 0.7 Da, reduce the likelihood of chimeric spectra but limit the number of ions isolated, potentially impacting signal intensity. In metaproteomics, the high density and diversity of precursor ions in certain mass ranges complicates this balance, as even narrow windows can capture multiple ions. Advances in mass spectrometers, such as faster scan speeds, now enable higher topN settings in DDA workflows, helping to address this challenge by acquiring more fragmentation spectra within a given run.

iii) MS Settings in DIA workflows

Optimizing data-independent acquisition (DIA) workflows requires careful calibration of several key parameters to achieve accurate and comprehensive peptide identification. The width of mass isolation windows is particularly critical, as narrower windows, such as 2 m/z, provide higher resolution and more precise fragmentation spectra, which are essential for resolving complex peptide mixtures. However, narrower windows can reduce proteome coverage, as fewer ions are isolated in each cycle. Balancing resolution with proteome coverage is thus a central challenge in DIA optimization. Recent advancements, such as the Orbitrap Astral mass spectrometer, support exceptionally narrow isolation windows while maintaining high scanning speeds, effectively bridging the gap between DDA and DIA methodologies.

In addition to tuning isolation windows, optimizing collision energy is required for generating high-quality fragment ions, while chromatographic conditions, including gradient length and flow rate, must be carefully calibrated to align with the DIA cycle time. Ensuring sufficient acquisition points across peptide elution peaks is essential for accurate quantification and peptide identification. DIA workflows in metaproteomics are advancing rapidly, providing enhanced resolution and deeper proteome coverage in complex microbial samples (E. Wu et al. 2024; Dumas et al. 2024). Detailed guidelines for these optimization strategies can be found in recent studies exploring advancements in DIA methodologies (Ishikawa et al. 2022; Demichev et al. 2022; Gu et al. 2024).

3.7.4 Quality control of LC-MS/MS

A comprehensive QC workflow begins with a blank injection of solvent without any sample to check for background contamination. Ideally, a blank run should produce minimal

identifications, which can be verified visually or through database searches. Contamination sources can include transport solvents used in HPLC systems, so these should be carefully monitored. Next, a standard injection of a known peptide mixture, such as cytochrome C or BSA digest, is performed to confirm instrument calibration and performance. Simple mixtures like these are useful for testing HPLC performance, while more complex peptide mixtures, such as HeLa digest, assess the mass spectrometer's ability to analyze complex samples. A representative microbiome sample digest can also be injected to refine the LC gradient profile, and such standards should be injected regularly throughout the run. Additionally, using reference microbiome material as a positive control can help verify the efficiency of protein extraction protocols. This ensures that the extraction method reliably captures a representative set of proteins from the sample, which is particularly important for metaproteomic studies. Database searches on complex standards should be used to monitor metrics like number of PSMs, peptide and protein identifications. Consistently tracking these values over time helps detect performance declines, signaling when the instrument requires cleaning or recalibration.

During the LC-MS/MS run, retention times for known peaks should be monitored closely, as significant shifts compared to previous runs may indicate issues such as column blockage, connector leakage, or valve wear. Similarly, column back pressure should be monitored as a potential indicator of problems. Peak shape should also be evaluated for symmetry and sharpness; tailing or broadening peaks may suggest problems with chromatography or ionization efficiency. Signal intensity is another important parameter, and any significant drop compared to expected values may point to reduced instrument sensitivity or ionization issues.

After the run, each raw file must be carefully reviewed to identify potential issues. Failed runs should be rerun immediately to avoid batch effects caused by delayed reanalysis. The total ion current (TIC) chromatogram provides valuable information on instrument performance, and it should be examined for unexpected peaks or a noisy baseline, both of which may point to contamination or hardware issues. The base peak chromatogram provides additional insights into LC resolution. Comparing the TIC-to-base peak intensity ratio is also informative, as higher values often reflect increased sample complexity or poor chromatographic performance. Retention times and peak intensities across samples should be consistent, indicating good repeatability. Additional QC checks, such as PCA clustering or heatmaps, can help pinpoint variations between runs and ensure data quality.

Metrics collected after protein identification and quantification are also essential for evaluating QC (Bielow, Mastrobuoni, and Kempa 2016). For example, the number of

identified PSMs to the total number of MS2 spectra, the PSM identification rate, serves as a key indicator of data quality. Using a 1-hour gradient on a Q-Exactive mass spectrometer with optimized conditions and high-quality sample preparation, metaproteomic samples can achieve an ID rate of approximately 50%, meaning that 50% of spectra yield identified peptide sequences after 1% FDR filtering. Note that for samples in less trivial environments, such as soil, the PSM identification rate will be much lower. It is crucial to analyze high-quality QC samples using the same LC-MS/MS methods, as the identification rate depends heavily on both the instrument's performance and sample preparation.

In large-scale projects lasting several weeks, retention time drift and signal drops are common. Blocking and randomizing samples during analysis is recommended to reduce systematic biases caused by these performance variations (Oberg and Vitek 2009). Implementing rigorous QC procedures at each step of LC-MS/MS is essential to maintain data reliability and consistency, with standardized QC samples serving as valuable benchmarks for long-term performance evaluation.

Several dedicated QC tools, such as MaCProQC (Rozanova et al. 2023), QCloud2 (Olivella et al. 2021), Rawtools (Cortay et al. 1988) are available to evaluate the quality of LC-MS/MS data. These tools provide a range of functionalities, from tracking performance metrics to generating clustering analyses for data quality evaluation. However, more recently, the HUPO-PSI Quality Control working group has introduced the mzQC file format, a JSON-based standard designed to streamline the reporting and exchange of mass spectrometry (MS) quality control metrics. To facilitate adoption, they have also developed open-source software libraries in Python (pymzqc), R (rmzqc), and Java (jmzqc), which provide functionalities for creating, validating, and analyzing mzQC files. These libraries enable researchers to integrate mzQC into diverse workflows for proteomics, metabolomics, and other MS applications, ensuring consistent data quality assessment and fostering interoperability across different analytical platforms (Bielow et al. 2024).

3.7.5 Data management and data sharing

Effective data management and sharing are essential to advancing metaproteomics research, ensuring data integrity, reproducibility, and collaboration. A robust data management plan should include secure, redundant storage solutions to protect against data loss, particularly for large-scale studies conducted over extended periods. Implementing version control for raw and processed data facilitates systematic tracking of updates and reanalyses, improving reproducibility and transparency.

Adhering to community standards, such as those established by the Human Proteome Organization Proteomics Standards Initiative (HUPO-PSI) (Deutsch, Vizcaíno, et al. 2023), is crucial for consistency and interoperability. The HUPO-PSI defines data representation standards in proteomics to facilitate data comparison, exchange, and verification. Using standardized formats like mzML for mass spectrometry data (Martens et al. 2011), mzIdentML for identification results (Combe et al. 2024), and the Universal Spectrum Identifier (USI) for referring to any mass spectrum in publicly deposited proteomics datasets (Deutsch et al. 2021), ensures compatibility across platforms and tools, thereby streamlining collaborative efforts and enabling more efficient data use.

Metadata plays a critical role in making datasets interpretable, reusable, and comparable across studies. Comprehensive metadata should capture sample origins, preparation protocols, instrument settings, and data processing workflows, ideally using standardized ontologies like PSI-MS Ontology. In proteomics, this information is collected in the Sample and Data Relationship Format for Proteomics (SDRF-Proteomics) format, which provides a structured, tab-delimited format for describing the relationships between samples and data files, mirroring the experimental workflow in proteomics (Dai et al. 2021). Tools like lesSDRF offer user-friendly interfaces to annotate metadata in SDRF format, facilitating standardization (Claeys et al. 2023). Recognizing the added complexity of microbial environments, the Metaproteomics Initiative is developing SDRF-Proteomics templates tailored for metaproteomics, as current formats for single-species proteomics do not fully address the nuances of microbial data. Standardized metadata not only supports computational analyses but also ensures structured inputs for machine learning models, advancing reproducibility and consistency across the field.

Depositing both data and metadata in recognized international ProteomeXchange repositories (Deutsch, Bandeira, et al. 2023), such as PRIDE (Perez-Riverol et al. 2024), aligns with the FAIR (Findable, Accessible, Interoperable, and Reusable) principles, promoting open science and innovation. These repositories make data accessible to the broader research community, enabling others to validate findings, conduct systematic reviews, and perform large-scale analyses. Sharing practices in metaproteomics help with benchmarking studies, development of new interpretation tools, and the ability to draw broader conclusions, significantly improving the field's collaborative potential and impact.

4. Computational analysis of metaproteomics data

4.1 Peptide identification, protein inference and quantification

After acquiring MS/MS spectra from mass spectrometry, the next step is to identify the peptides present in the sample. This involves analyzing the fragmentation patterns in the MS/MS spectra to determine the specific amino acid sequences of the peptides. This process is performed using search engines, often integrated into comprehensive proteomics software packages (**Section 4.1.1**). Typically, these algorithms match the experimental MS/MS spectra to a theoretical protein sequence database, and the success of this step depends heavily on the selection or construction of an appropriate database, as outlined in **Section 4.1.2**. The search engine then applies a false discovery rate (FDR) threshold to filter out potential false positives (**Section 4.1.3**). Peptides passing this filter are subsequently used for protein inference (**Section 4.1.4**) and quantification (**Section 4.1.5**). All these sections focus on DDA MS, while **Section 4.1.6** is dedicated to tools specifically designed for analyzing DIA MS data.

4.1.1 Peptide identification with proteomics search engines

Shotgun metaproteomics experiments generate large datasets of MS1 and MS2 spectra, which form the basis for downstream analysis. With advancements in high-throughput MS, these datasets now range from thousands to millions of spectra, making manual interpretation impractical. To address this challenge, search engines are essential for interpreting the data and identifying peptides. Peptide identification relies on three main strategies: (i) sequence database searching, where experimental spectra are matched to theoretical spectra derived from protein or peptide sequences in a database; (ii) *de novo* sequencing, which directly infers peptide sequences from spectra without a reference database; and (iii) spectral library searching, where experimental spectra are compared to curated libraries of previously validated spectra. These methods are often complemented by post-processing steps to enhance accuracy and confidence in peptide identification, as outlined in **Section 4.1.3**. Additionally, most proteomics software packages integrate peptide identification with protein inference and quantification, a topic discussed in **Section 4.1.4** and **Section 4.1.5**. Some specific metaproteomics software also integrates taxonomic and functional analyses, as outlined in **Section 4.2**.

i) **Protein sequence database searching**

Database search algorithms are fundamental for interpreting mass spectrometry data, particularly in metaproteomics, where the complexity of microbial communities poses significant analytical challenges. These algorithms match experimental MS/MS spectra to theoretical spectra generated from protein sequence databases. The success of this process depends on the choice of search engine, the search parameters used, and the composition of the database, all of which influence the number and type of peptides and proteins detected.

Database search engines start by using a selected reference protein sequence database, which is *in silico* digested to emulate the cleavage rules of the enzyme used during protein digestion, most commonly trypsin. From these digested sequences, theoretical MS/MS spectra are generated and compared to the experimental MS/MS spectra obtained during mass spectrometry. Each combination of theoretical peptide and spectrum (peptide-spectrum match, PSM) is assigned a similarity score, with the search engine ranking and filtering potential PSMs based on the score and peptide properties. The exact method of score calculation varies between search engines, and these differences can affect both sensitivity and specificity. An in-depth explanation of the various scoring algorithms used in database search engines can be found in this comprehensive review (Verheggen et al. 2020).

Each database search engine offers unique advantages and limitations, including variations in processing speed, compatibility with input and output formats, support for post-processing tools, and overall user-friendliness. These factors significantly influence their performance in metaproteomics workflows, where the complexity and scale of datasets demand highly efficient and reliable analysis tools. A detailed discussion of these tools and their applications is available in a comprehensive review (Schiebenhoefer et al. 2019). A selection of database search engines and proteomics software commonly used in metaproteomics research is highlighted below:

- SearchGUI (Vaudel et al. 2011) provides simultaneous access to multiple complementary search algorithms, including X!Tandem (Craig and Beavis 2004), Comet (Eng, Jahan, and Hoopmann 2013), Andromeda (Cox et al. 2011), OMSSA (Geer et al. 2004), Sage (Lazear 2023), and others. Its companion tool, PeptideShaker (Vaudel et al. 2015), seamlessly imports SearchGUI output and offers a comprehensive, user-friendly interface for interpreting and visualizing results. Additionally, PeptideShaker includes a direct export feature to Unipept,

enabling streamlined downstream taxonomic and functional analysis (Vande Moortele, Devlaminck, et al. 2024; Van Den Bossche et al. 2020). A detailed tutorial is available on the CompOmics web page to guide users through these workflows (Vaudel et al. 2014).

- Andromeda (Cox et al. 2011), used in MaxQuant (Cox and Mann 2008), is widely used for its ease of use and MS1 quantitative capabilities. Users benefit from a well-established community, including annual user meetings and a dedicated forum for support.
- Mascot (Matrix Science) and Proteome Discoverer (Thermo Fisher Scientific) are popular commercial tools with extensive user bases.
- FragPipe, using MSFragger (A. T. Kong et al. 2017), and pFind (Le-heng Wang et al. 2007) incorporate open search strategies, which improve sensitivity by enabling the identification of PTMs.
- Sipros (Guo et al. 2018), ProteoStorm (Beyter et al. 2018) and COMPIL 2.0 (Park et al. 2019) are tailored specifically for metaproteomics but are perceived less user-friendly than mainstream software.
- Tools such as Sage (Lazear 2023) and MSFragger (A. T. Kong et al. 2017) leverage advanced spectral and sequence indexing strategies to significantly accelerate database searches, making them highly promising for improving the speed of metaproteomics analysis.

For researchers that want more integrated solutions, several software suites can simplify metaproteomics workflows by consolidating multiple steps and managing the high density of information inherent to the field.

- Galaxy for Proteomics (Galaxy-P) is another versatile platform offering numerous tools and workflows tailored to metaproteomics, including database generation, discovery analysis, verification, quantitation, and statistical analysis (Blank et al. 2018; P. D. Jagtap et al. 2015; Do et al. 2024). With public gateway availability (The Galaxy Community 2024) and access to training resources via the Galaxy Training Network (Hiltemann et al. 2023), Galaxy-P is a valuable resource for researchers seeking an open and user-friendly platform for users to access metaproteomic workflows.
- The MetaProteomeAnalyzer (MPA) software suite (Muth, Behne, et al. 2015) offers modules for protein database creation, database searching, protein grouping, annotation, and results visualization. Its user-oriented design makes it a suitable option for both beginners and experienced researchers.

- MetaLab (Cheng et al. 2017) is an integrated data processing pipeline that includes tools for sample-specific database generation, peptide determination, taxonomic and functional profiling, and abundance analysis. Its open search strategy enables comprehensive profiling of PTMs and improved sensitivity. Additionally, MetaLab offers workflows for taxonomic analysis based on metagenome-assembled genome (MAG) databases, allowing peptide-to-genome linkages for improved specificity compared to traditional lowest common ancestor (LCA) methods.

In these tools, selecting appropriate search parameters is essential for reliable and meaningful results. The choices regarding modifications, enzyme specificity, and mass tolerance significantly impact the identification of PSMs. Below are key considerations:

- Selection of modifications: It is important to distinguish between modifications introduced by the experimental workflow and biological modifications. Fixed modifications, like carbamidomethylation of cysteine, are commonly applied across all peptides to account for standard sample preparation artifacts, as discussed in **Section 3.4.4**. Variable modifications, such as methionine oxidation, are applied selectively to explore biologically relevant modifications. However, including too many variable modifications can expand the search space excessively, reducing identification rates. It is often best to limit variable modifications to the most biologically relevant ones.
- Enzyme specificity and number of missed cleavages: Choosing the correct enzyme and setting an appropriate number of allowed missed cleavages affects the range of detectable peptides. For instance, trypsin, the most commonly used enzyme in proteomics, may occasionally miss cleavages after lysine (K) or arginine (R). Allowing one or two missed cleavages is generally a good compromise in metaproteomics, as it accounts for incomplete digestion without excessively broadening the search. Semi-specific or non-specific cleavage settings might be useful in some cases but can lead to longer processing times and a lower identification rate due to the expanded search space.
- Mass tolerance: Mass tolerance settings should match the resolution capabilities of the mass spectrometer. For example, on a high-resolution Q Exactive instrument with HCD data, setting a precursor mass tolerance of 10 ppm (for MS1) and a fragment mass tolerance of 0.02 Da (for MS2) can balance accuracy and computational efficiency, restricting the search to relevant matches while taking advantage of the instrument's resolution.

Thoughtful parameter selection helps balance sensitivity and specificity, leading to high-quality data that accurately reflects the sample's biological characteristics. Parameter adjustments should consider the mass spectrometer type, sample complexity, and specific research objectives.

ii) *de novo* searching

De novo peptide sequencing assigns amino acid sequences to MS/MS spectra without requiring a protein sequence database for spectral matching. This approach provides an unbiased method for detecting peptides, independent of the quality and completeness of the protein sequence database. Several *de novo* sequencing algorithms have been introduced in recent years, including PEAKS, Casanovo (Yilmaz et al. 2024), PepNovo (Frank and Pevzner 2005), and the newly developed π-HelixNovo (T. Yang et al. 2024), metaSpectraST (Hao et al. 2023), and NovoBridge (Kleikamp et al. 2021).

When applied effectively, *de novo* sequencing can sensitively and accurately estimate the taxonomic composition and functional content of the microbiome without prior knowledge of the system under study. It also has the potential to identify unsequenced members of the microbial community. Furthermore, *de novo* sequencing can be used to evaluate the completeness and suitability of a protein sequence database for metaproteomics research (R. S. Johnson et al. 2020). Recently, the progress and opportunities in *de novo* sequencing for metaproteomics were reviewed, emphasizing its potential for unsequenced species detection and deeper functional insights into microbial communities (Van Den Bossche, Beslic, et al. 2024).

Despite its promise, there remains a need for systematic benchmarking of *de novo* sequencing tools to assess their applicability to metaproteomics. In particular, most tools and approaches for *de novo* metaproteomic analysis still require some input from databases either to help selecting peptides or to gain information from the identified peptides. Evaluating their performance in terms of sensitivity, accuracy, and throughput is essential to ensure their effectiveness in the complex and diverse datasets characteristic of microbiome studies.

iii) Spectral library searching

Spectral library search engines operate on principles similar to database searching but differ by directly comparing experimental MS/MS spectra to pre-existing libraries of validated spectra. These libraries consist of MS/MS spectra previously acquired through the analysis of complex peptide mixtures and conventional sequence database searches

or generated using predictive deep-learning algorithms. Unlike sequence database searching, spectral library searching can incorporate additional parameters, such as retention time on the LC column and the relative intensities of fragment peaks within the spectra, enhancing both accuracy and confidence in peptide identification.

The development of AI-based tools like MS²PIP (Degroeve, Maddelein, and Martens 2015) and Prosit (Gessulat et al. 2019) has made it possible to generate high-quality spectral libraries from protein sequence databases (Lautenbacher et al. 2024). These advancements have expanded the applicability of spectral library searches by enabling the generation of predictive libraries tailored to specific experiments. Newer spectral library search tools designed for data-dependent acquisition (DDA) data, such as Mistle (Nowatzky et al. 2023) and Scribe (Searle, Shannon, and Wilburn 2023), have also emerged for metaproteomics research.

Spectral library searching offers a fast and efficient approach to match peptide sequences to MS/MS data, often outperforming traditional database searching in terms of speed and precision for well-curated libraries. However, despite its potential, spectral library tools for metaproteomics require further evaluation, particularly regarding their usability and effectiveness for highly complex microbial datasets.

4.1.2 Database construction or selection

For single-organism proteomics, constructing a protein sequence database is relatively straightforward, as it can be derived directly from the organism's genome. In metaproteomics, however, the complexity of microbial communities, the diversity of organisms, and the prevalence of unknown proteins present significant challenges. Selecting or generating an appropriate database is crucial, as the database must balance comprehensiveness and specificity. An incomplete database risks missing or falsely identifying proteins, while an excessively large database decreases the sensitivity of the analysis and inflates the FDR, as detailed in **Section 4.1.3** (Nesvizhskii and Aebersold 2005; Blakeley-Ruiz and Kleiner 2022).

An optimal database for metaproteomics should be both comprehensive and specific. Comprehensive, as it should include all proteins potentially present in the sample. Missing sequences lead to false negatives, reducing peptide and protein identification rates. Specific, because it should exclude sequences unexpected to be present in the sample. Including irrelevant sequences increases random matches, inflates the FDR, and therefore negatively affects peptide (and protein) identification (see also **Section 4.1.3**). Additionally,

metaproteomic analyses often include contaminants from sample processing, such as leftover trypsin, BSA carry-over, or keratin from handling. Incorporating these contaminants into the database, using resources like the common Repository of Adventitious Proteins (cRAP, https://www.thegpm.org/crap/), allows for their accurate identification and prevents misidentification with other proteins in the sample.

To create a suitable database, prior knowledge of the community composition is essential. This information can be derived from various sources, including prior literature, 16S rRNA amplicon sequencing, or metagenomic and/or metatranscriptomic sequencing, each offering different levels of resolution and success. Literature reviews provide only limited insights, whereas meta-omics approaches offer the most comprehensive and detailed characterization of the community (Kleiner et al. 2012; Blakeley-Ruiz et al. 2022; Minniti et al. 2019). Additionally, depending on the sample's environment, host or dietary proteins may need to be included in the database. While adding these proteins can improve identification rates, it also increases database size and complexity, potentially complicating the analysis. The inclusion of nearly identical sequences, often inevitable in large databases, can further exacerbate protein inference issues (see **Section 4.1.5**). Sequence clustering algorithms (W. Li, Jaroszewski, and Godzik 2001) or protein grouping tools (Audain et al. 2017; The et al. 2016) can address these challenges by consolidating redundant entries while retaining essential taxonomic and functional annotations.

The choice of database type depends on the sample type, the level of understanding of the microbial community, and the available resources. Based on these factors, different types of databases can be used, each with its own set of advantages and limitations (see **Table 2**). These include public repositories, reference catalogs, and meta-omic databases, as detailed below.

i) Public repositories

Public repositories like UniProtKB (The UniProt Consortium 2023) and NCBI RefSeq (O'Leary et al. 2016) provide extensive reference collections of protein sequences. However, these untailored (or unrestricted) databases often lack specificity and contain many unrelated sequences, leading to reduced identification rates and increased FDR (**Section 4.1.3**). Furthermore, public repositories are biased toward well-characterized microbes, such as model organisms or pathogens, and heavily studied environments or systems, such as clinical and human samples. This bias results in significant gaps for less-studied environmental microbial communities, making these repositories incomplete for many metaproteomics applications. Filtering (or restricting) these repositories based on

16S rRNA analysis results can improve specificity, but the resolution of 16S rRNA sequencing is limited. Entire genera or sets of species often need to be included, preventing strain-level specificity (Odom et al. 2023; J. S. Johnson et al. 2019).

ii) Reference catalogs

Reference catalogs are curated collections of protein sequences tailored to specific environments or systems. They are available for well-studied ecosystems such as the human gut (J. Li et al. 2014; Almeida et al. 2021), the cow rumen (Stewart et al. 2019; Xie et al. 2021), and the mouse gut (Kieser, Zdobnov, and Trajkovski 2022; Beresford-Jones et al. 2022; Lesker et al. 2020). These catalogues are typically constructed by combining data from isolated microbes and metagenomic studies (Gurbich et al. 2023). Although smaller and more targeted than public repositories, reference catalogs can still be relatively large for metaproteomic analyses and often aggregate data from many samples, including different individuals and studies - yet, not from the study itself, therefore also called unmatched meta-omics databases. This composite nature introduces challenges, as even samples from similar environments can exhibit substantial variation in species composition and strain diversity. Consequently, reference catalogs can suffer from inaccuracies, incompleteness, and overrepresentation of certain subsamples (Van Den Bossche, Kunath, et al. 2021; Abdill, Adamowicz, and Blekhman 2022). Like repositories, the specificity of reference catalogs can be improved by incorporating prior knowledge of the microbial community, such as results from 16S rRNA analysis, to narrow down the included sequences to those most relevant to the sample.

Alternatively, to address the challenges posed by large and composite catalogs, database-reduction methods have been developed. These methods include the two-step search approach (P. Jagtap et al. 2013), iterative workflows such as MetaPro-IQ (X. Zhang et al. 2016) and MetaLab (Cheng et al. 2017), next to others. While these methods are often used in the field and increase the number of identified PSMs and peptides, some have been shown to significantly raise the number of false positives at both levels, exceeding the FDR estimate (Muth, Kolmeder, et al. 2015). These methods should therefore be treated with caution, and additional validation might be appropriate prior to drawing biological conclusions.

iii) (matched) meta-omics databases

Meta-omic databases are constructed using metagenomic and/or metatranscriptomic data collected from the same sample as the metaproteomic analysis, making them the most specific databases available. These databases accurately reflect the species composition and strain diversity of the sample (Blakeley-Ruiz and Kleiner 2022; Heintz-Buschart and Wilmes 2018; B. J. Kunath et al. 2022). However, generating a high-quality meta-omic database requires significant sequencing effort, cost, computational resources, and technical expertise. Although the specific details of this process are beyond the scope of this manuscript, they have been extensively covered elsewhere (Blakeley-Ruiz and Kleiner 2022; Benoit J. Kunath et al. 2019). Briefly, constructing a meta-omic database involves four key steps: sequencing, assembly, binning, and annotation.

To create a comprehensive database suitable for metaproteomic analysis, the sequencing effort must be sufficiently deep to capture the complexity of the community. One major advantage of meta-omic databases is their ability to provide precise insights into the species and strain diversity of the sample, enabling direct linkage between genomes and identified proteins. This requires genome reconstruction through binning, where contigs are grouped into MAGs based on shared features. However, due to the complexity of microbial communities and limitations in sequencing depth, some MAGs may remain incomplete. Therefore, a robust meta-omic database should include both binned and unbinned sequences to retain as much information as possible (Benoit J. Kunath et al. 2017; Narayanasamy et al. 2016).

Once reconstructed, MAGs and contigs are taxonomically annotated, and protein sequences or open-reading frames (ORFs) are predicted and functionally annotated. The choice of tools and resources for these steps depends on the study's objectives (Queirós et al. 2021). Despite their specificity, meta-omic databases can still be incomplete due to insufficient sequencing depth or the inability to recover all relevant MAGs from the sample. This issue can be partially addressed by performing exploratory 16S rRNA gene sequencing to assess the required sequencing depth for optimal metagenomic analysis (Blakeley-Ruiz et al. 2022).

Combining metagenomic data with metatranscriptomic data further improves the quality and specificity of the database (Narayanasamy et al. 2016; F. Delogu et al. 2020). Since metatranscriptomics focuses on mRNA, it captures the active portion of the community, providing a gene-centric view that aligns closely with the functional content of interest for metaproteomics.

Table 2: Comparison of database types for metaproteomics: public repositories, reference catalogs, and meta-omic databases. The color indicates our preference: green represents favorable choices, yellow indicates intermediate choices, and red highlights unfavorable choices.

	Public repositories (*)	Reference catalogs	Meta-omic databases
Monetary cost	Free	Free	Sample type dependent $100-$2,000/sample or pooled samples
Time cost (labor & computation)	Days	Days	Genome-resolved month-year, otherwise weeks
Comprehensiveness	Low to Medium depending on the sample representation in the repository	Medium to High depending on sequencing effort and multi-omics integration	Medium to High depending on sequencing effort and multi-omics integration
Identification probability	Low	Medium	High
Specificity	Low due to high diversity of the repository	Medium due to lack of strains resolution	High due to sample specificity
Misidentification probability	High	Medium	Low
Sequence Redundancy and Impact	High and difficult to resolve due to high diversity of the repository	Medium but can be resolved depending the curation level	Low and can be resolved as part of the metagenomic processing
Taxonomic Annotation and Resolution	Taxonomy not curated and potentially outdated	Depends on curation level (potential for misidentification due to closely related taxa)	Possibility of *de novo* annotation and species resolution based on metagenomic processing
Certainty/Applicability	Easily available but lacks the guarantee of appropriate sequences	Available for few sample types only and lacks of accuracy	High accuracy but requires particular expertise and extra time/cost

(*) Restricted repositories have similar characteristics to reference catalogs in terms of specificity and sequence redundancy.

4.1.3 PSM FDR control

A critical step in the process of peptide identification is acquiring a set of reliable PSMs. After PSMs are acquired, they are evaluated based on the scoring function of the search engine, retaining the highest-ranked PSM for each spectrum — that is, the peptide sequence whose theoretical spectrum most closely matches the experimental MS/MS spectrum. However, regardless of the scoring algorithm used, some PSMs will inevitably represent false matches, making robust control of false positives essential.

The most commonly used strategy to manage false positives in (meta)proteomics is the target-decoy approach (Elias and Gygi 2007). In this approach, the protein sequences in the target database are processed *in silico* to emulate enzymatic digestion, generating theoretical peptides. The same procedure is applied to the reversed or shuffled sequences of a decoy database, ensuring that these decoy peptides are biologically implausible and not present in the sample. During the search, the experimental spectra are matched to both the target and decoy sequences in a concatenated target-decoy database. This process results in PSMs being labeled as either target or decoy. The proportion of decoy PSMs in the final result serves as an estimate of the FDR, calculated as the number of decoy PSMs divided by the total number of accepted PSMs (**Figure 4**). The FDR is typically controlled at 1% in proteomics and metaproteomics experiments, but for highly complex samples such as soil microbiomes, the FDR threshold can be increased to 5% to retain a sufficient number of identifications for biological interpretation.

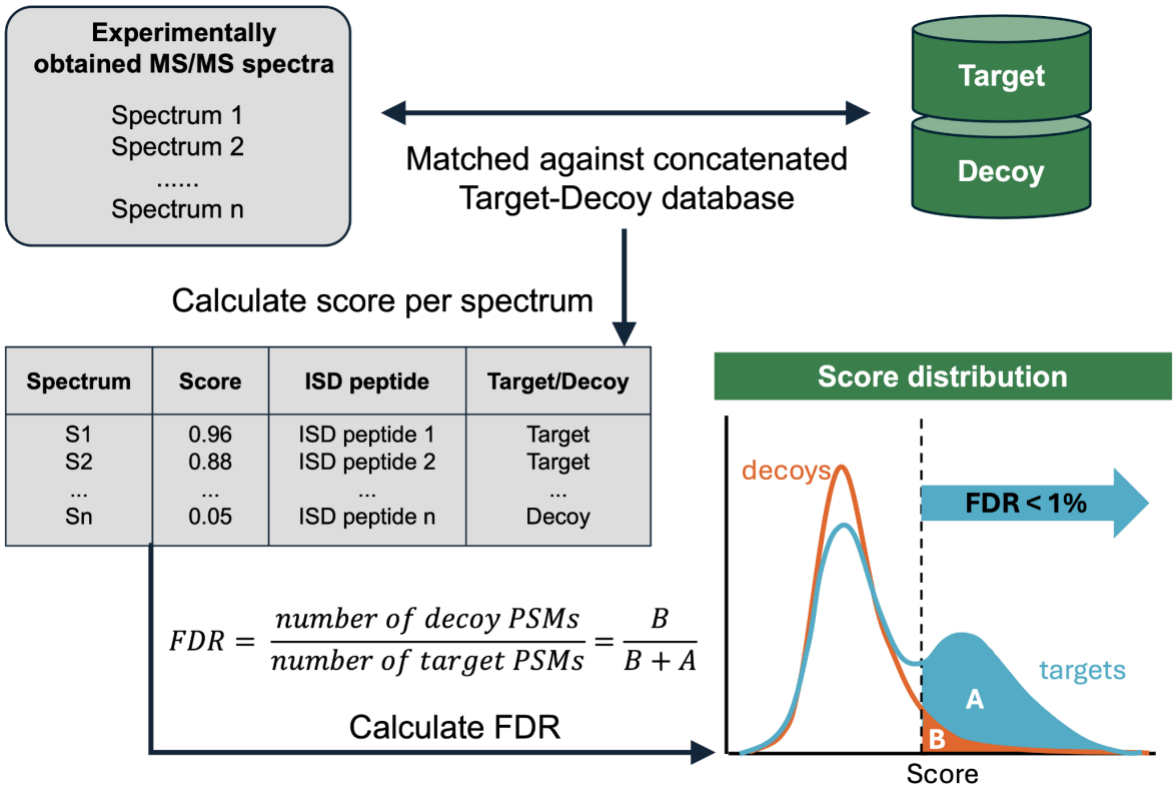


Figure 4. **Principle of target-decoy analysis and False Discovery Rate (FDR) calculation.** (Top) The experimentally obtained MS/MS spectra are matched to *in silico* generated spectra of the concatenated target/decoy protein sequence database. (Middle) For each obtained spectrum, the match with the highest score is retained, together with the assigned *in silico* digested peptide sequence and its target or decoy label. (Bottom) The score distribution is used to select which PSMs will be considered as true matches. The metric to control the false positives is the FDR, and is calculated as the number of decoy PSMs divided by the number of target PSMs (in the Figure depicted as area B divided by the sum of areas B and A). Figure of (schematic) target/decoy distribution adjusted from (Käll et al. 2008)**.**

The specific challenges of metaproteomics add complexity to FDR control. The larger, more diverse protein sequence databases required for metaproteomics often increase the search space significantly, leading to a greater overlap between the score distributions of target and decoy PSMs. This overlap reduces the resolution of FDR estimation and necessitates careful database construction to limit irrelevant sequences, as discussed in **Section 4.1.2**. Overly large but unspecific databases inflate the FDR by increasing random matches to both target and decoy sequences, resulting in fewer confident peptide identifications (Schiebenhoefer et al. 2019; Tanca et al. 2016). Conversely, overly restrictive databases risk excluding true target sequences, resulting in missed matches, false negatives, and reduced proteome coverage. Therefore, achieving an optimal balance between database

specificity and comprehensiveness is crucial to minimize false positives from decoy matches while maximizing target identifications, ensuring effective FDR control.

Metaproteomics workflows often rely on advanced post-processing tools to improve the accuracy and confidence of peptide identifications. Tools such as Percolator (Käll et al. 2007) and MS²Rescore (C. Silva et al. 2019) refine PSM scores using machine learning algorithms that consider additional features beyond the initial search engine score, such as precursor intensity, fragmentation patterns, and retention time. These tools can substantially improve the separation between target and decoy PSMs, enabling more accurate FDR estimation even for complex datasets.

In metaproteomics, where samples often contain thousands of species, the challenge of FDR control is even larger by the inherent complexity and diversity of the microbial communities under study. Careful database construction (**Section 4.1.2**), combined with robust FDR control during the search and advanced post-processing techniques, is critical to ensure reliable peptide and protein identifications, thereby enabling meaningful biological insights from metaproteomics datasets.

4.1.4 Protein inference

Protein inference is a fundamental challenge in shotgun proteomics where the goal is to determine the proteins present in a sample based on the peptides identified through tandem mass spectrometry (Nesvizhskii and Aebersold 2005). This process is complicated by the fact that peptides can often be mapped to multiple proteins or protein isoforms present in the commonly large protein database. This is especially the case in complex samples such as microbial communities where multiple species may contribute homologous proteins, making it difficult to conclusively infer which proteins are actually present (Schallert et al. 2022).

To address this complexity, protein grouping is commonly used to generate a more manageable list of identified protein (sub)groups for downstream analysis. However, different methods for protein grouping exist, as depicted in **Figure 5**, and these are typically performed by the search engine. It is essential to verify the default settings of the search engine to understand which grouping approach it applies, and if needed, adjust it to align with your research hypothesis. The two main approaches are Occam's razor and anti-Occam's razor.

Occam’s razor is based on the principle of maximum parsimony, providing the smallest set of proteins that can explain all observed peptides. However, this approach discards proteins

not matched by a unique peptide, potentially losing their associated taxonomy and functions that might be present in the sample. Occam's razor is particularly suited for simpler, single-species samples or targeted proteomics experiments, where reducing complexity is key.

In contrast, anti-Occam's razor adopts a more inclusive strategy, retaining all proteins that can be mapped to at least one peptide, regardless of whether those peptides are shared with other proteins. This approach is beneficial for complex metaproteomic samples, where the goal is to capture as much protein diversity as possible. By being more inclusive, anti-Occam's razor ensures that proteins from different species with minimal unique peptides are not overlooked, providing a more comprehensive picture of the microbial community. However, this inclusivity comes at the cost of increased complexity in the resulting protein list.

After choosing between Occam's and anti-Occam's razor principles, proteins can then be grouped into protein groups or protein subgroups. Protein groups cluster proteins that share at least one peptide, offering a broader overview of potential protein identifications. Protein subgroups, on the other hand, are more specific and include proteins that share the exact same set of peptides. For example, the anti-Occam's razor approach often benefits from subgrouping to prevent excessively large and uninformative protein groups. In metaproteomics, this approach helps disentangle the contributions of individual species, even when closely related proteins share substantial sequence similarity (Schallert et al. 2022).

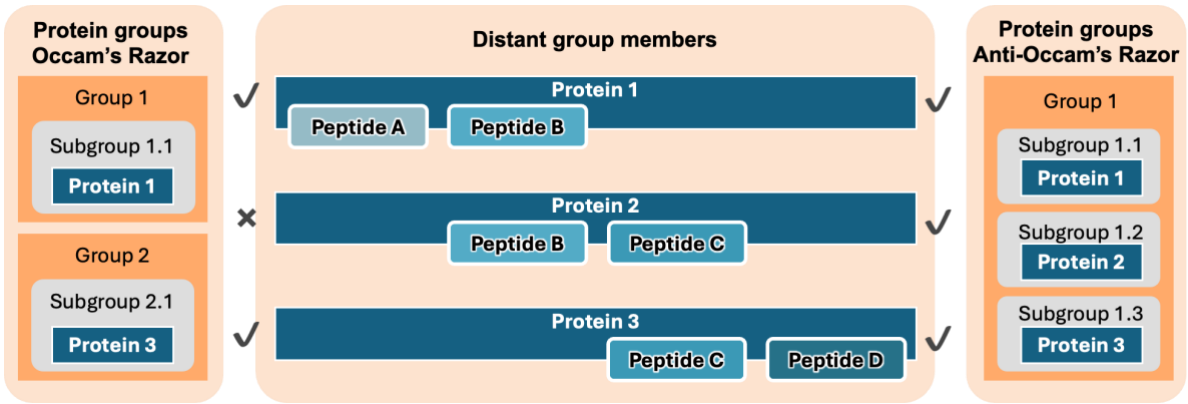


Figure 5: Practical example of (sub)grouping approaches. This grouping case deals with distant group members, meaning that certain proteins in the group don't share a single peptide, in this case protein 1 and 3. Applying the rule of parsimony separates the group in this specific case. In the anti-Occam case, protein 2 remains in a separate subgroup.

The choice of protein inference approach should align with the complexity of the sample and the research objectives. For single-species or targeted studies, Occam's razor

combined with protein grouping is advantageous for reducing false positives and simplifying downstream analyses. This strategy was used, for example, in analyzing the SIHUMIx mock community (Schäpe et al. 2019) as part of the CAMPI study (Van Den Bossche, Kunath, et al. 2021). For complex, multi-species metaproteomic samples, anti-Occam's razor combined with protein subgrouping is often preferred, as it maximizes protein diversity while maintaining manageable group sizes. This inclusive approach was used for fecal sample analysis in the CAMPI study (Van Den Bossche, Kunath, et al. 2021). Ultimately, the selection of a protein inference method depends on the specific characteristics of the sample and the research objectives. Researchers must balance the need for comprehensive protein identification with the practical considerations of data complexity and interpretability (Schallert et al. 2022).

4.1.5 Protein quantification

Protein quantification is a central component of metaproteomics, offering valuable insights into the functional dynamics of microbial communities. By quantifying proteins, researchers can assess how microbes respond to environmental changes, revealing shifts in physiology and metabolic processes. For example, changes in nutrient availability can trigger significant alterations in protein expression within individual microbes (Caglar et al. 2017) or entire microbial populations (Patnode et al. 2019). This section outlines the key concepts, strategies, and challenges in metaproteomic quantification, focusing on label-free and labeling-based approaches, as well as methods for downstream data analysis.

Metaproteomics workflows typically rely on two main quantification strategies: label-free quantification (LFQ) and labeling-based quantification. LFQ methods are widely used because they do not require stable isotope labels, making them more suitable for diverse and complex samples. Two common LFQ approaches are MS1 intensity-based quantification and MS2 spectral counting. MS1 quantification measures precursor ion intensities by calculating the area under the curve or apex intensity for each identified peptide, with tools such as MaxQuant (Cox et al. 2014) or standalone alternatives like moFF (Argentini et al. 2019) or FlashLFQ (Millikin et al. 2018). MS2 spectral counting, in contrast, quantifies peptides based on the number of matched MS2 spectra. Although simpler to implement, spectral counting typically has a narrower dynamic range and slightly lower precision. Currently, there is limited validation to determine which of the two primary quantification approaches—MS1 intensity-based quantification or MS2 spectral counting—is more accurate for metaproteomics, or under which conditions one might outperform the other. One study demonstrated that spectral counting provided a more accurate measure of the proteinaceous biomass of members within a synthetic community compared to MS1

intensities (Kleiner et al. 2017). Nonetheless, the prevailing consensus in the field suggests that both methods are generally suitable for metaproteomic quantification, with their applicability depending on the specific context and experimental goals.

Labeling-based quantification approaches, while valuable in proteomics, are less commonly used in metaproteomics due to the complexity of microbial communities. These methods, including TMT and stable isotope labeling by amino acids in cell culture (SILAC), enable absolute quantification and are particularly effective for controlled experimental designs requiring precise comparisons across samples. However, applying these methods to metaproteomics presents significant challenges. The diverse microbial populations and high sample complexity of environmental or clinical samples make labeling-based approaches less practical, favoring label-free strategies for most metaproteomics workflows. Nevertheless, labeling remains a viable option for targeted studies with well-defined microbial communities.

Quantification in metaproteomics faces several challenges, particularly in aggregating peptide-level data to infer protein abundances. This aggregation process is influenced by the protein inference problem (Nesvizhskii and Aebersold 2005), which determines how peptides are assigned to proteins or protein (sub)groups (see also **Section 4.1.4)**. Most software tools automatically assign peptides to proteins or protein groups, facilitating the quantification process. Once protein abundance data is obtained, normalization and transformation steps are crucial for meaningful statistical analysis. While various normalization methods have been proposed for proteomic data (Bubis et al. 2017; Pavelka et al. 2008; Välikangas, Suomi, and Elo 2018), the optimal approach for metaproteomics remains an area of active research.

One widely used normalization method, particularly for spectral count data, is the normalized spectral abundance factor (NSAF) (Florens et al. 2006). This approach compensates for biases introduced by protein length and sample variability. It involves dividing a protein's PSM count by its amino acid length to account for protein size, followed by normalizing against the total PSM count within the sample to reduce between-run batch effects. NSAF is relatively simple to calculate, robust to missing values, and particularly suited to the sparse data often encountered in metaproteomics. Further transformation, such as log or square root normalization, is typically applied to meet the assumptions of statistical tests.

A key distinction between standard proteomics and metaproteomics is the need to account for the diverse and complex nature of microbial communities. In metaproteomics, it may be

advantageous to normalize protein abundances specifically for organisms or groups of organisms within the community. This targeted normalization allows researchers to focus on changes in gene expression and function within specific taxa, providing more granular insights into microbial activity. The normalized spectral abundance factor per organism (orgNSAF) normalization method has been proposed as a solution for this purpose, as it enables normalization of protein abundances within defined taxonomic groups (Hinzke et al. 2021; Mueller et al. 2010; Ponnudurai et al. 2020).

A unique advantage of metaproteomic data is its ability to generate multiple datasets based on the research question. These datasets generally involve summing the abundance of constituent proteins into relevant categories. Broadly, there are three main categories: (1) individual proteins or groups of proteins with similar sequences, which can offer insights into the specific functionalities of individual organisms within the community; (2) categories of biological functions assigned to proteins associated with the measured peptides, enabling researchers to investigate shifts in overall community functions; and (3) taxonomic categories, where protein abundances can be used to estimate the relative contributions of different organisms within a microbial community.

The accuracy of both functional and taxonomic quantification is heavily dependent on the quality and completeness of protein annotations in the databases used. Functional categories can range from highly specific annotations, such as biochemical reactions, to broader descriptions of cellular processes like metabolism, gene expression, transport, or replication. Similarly, taxonomic quantification can achieve high resolution, down to the strain or species level (Brooks et al. 2015; Xiong et al. 2017), but this depends on the depth and accuracy of protein annotations. In some cases, it is limited to higher taxonomic ranks when annotations are incomplete or ambiguous (Blakeley-Ruiz et al. 2019). Metaproteomic measurements, when processed correctly, can provide an accurate representation of the relative proteinaceous biomass of microbial species within a community (Kleiner et al. 2017). However, the specificity and accuracy of these measurements are closely tied to the reliability of the annotations used for protein classification (Blakeley-Ruiz and Kleiner 2022; Tanca et al. 2016).

While these approaches enable the generation of robust datasets for understanding microbial abundance and function, further validation is necessary to refine these methodologies. Current quantification strategies in metaproteomics require additional benchmarking to identify optimal or equivalent approaches for various types of studies. Future research using mock communities with defined compositions and spike-in proteins

will be crucial for systematically evaluating the accuracy, reproducibility, and reliability of protein quantification methods in metaproteomics.

4.1.6 DIA data analysis

The application of DIA-MS in metaproteomics, as discussed in **Section 3.7.2**, demands tailored analytical workflows to manage the unique challenges posed by the complexity and scale of microbial communities. Unlike DDA, which prioritizes peptide selection, DIA generates complex spectra by fragmenting all ions within a predefined m/z range simultaneously. This comprehensive approach requires advanced computational tools and strategies to handle the resulting data.

Extracting quantitative and identification data from DIA-MS involves specialized software, such as Spectronaut (Bruderer et al. 2017), DIA-NN (Demichev et al. 2020), and EncyclopeDIA (Searle et al. 2018). These tools rely heavily on pre-existing spectral libraries to match experimental spectra to theoretical peptides. Such libraries are often generated through prior DDA experiments or predicted from protein sequence databases. While promising, library-free approaches that predict spectra directly from protein sequences remain computationally intensive and impractical for complex metaproteomics samples without additional data reduction strategies. One effective approach is using genome sequencing to limit the database search space or performing a preliminary DDA step to construct a targeted spectral library. These steps, although resource-intensive, are essential for reducing ambiguity in protein and peptide identifications.

Metaproteomics datasets amplify the inherent analytical challenges of DIA-MS due to their immense scale, which frequently involves millions of proteins and peptides. This complexity can lead to significant computational demands and requires extensive data processing pipelines. Direct library-free DIA analysis for such datasets is virtually impossible with current technology unless supplemental genome sequencing or DDA-based library construction is performed. These preparatory steps add complexity but are critical for optimizing DIA's utility in resolving the intricate dynamics of microbial communities.

Recent advancements in MS, including DIA-PASEF (Gómez-Varela et al. 2023) and the Orbitrap Astral analyzer (Dumas et al. 2024), have shown potential for enhancing the application of DIA-MS in metaproteomics. These technologies allow for deeper proteome coverage, improved sensitivity, and more accurate quantification. However, their integration into workflows must be carefully aligned with the computational tools and spectral library strategies mentioned above to fully exploit their capabilities.

A recent benchmarking study has demonstrated the reproducibility and accuracy of DIA-MS for metaproteomic workflows in comparison to DDA-MS methods (Rajczewski et al. 2024). Using mock communities of known taxonomic composition, DIA-MS consistently identified and quantified more peptides and proteins across laboratories. Additionally, the reproducibility of protein and peptide identifications was higher in DIA-MS workflows, which also provided accurate quantification of both protein abundances and taxonomic groups. These findings underscore the advantages of DIA-MS for metaproteomics, including its capacity for deep sequencing, robust quantitation, and reproducibility across samples. However, current studies also highlight the limitations of existing DIA tools when applied to metaproteomic datasets, emphasizing the need for improvements in software capabilities to handle the unique complexities of microbiome samples. These insights stress the importance of optimizing library generation, computational tools, and workflows to fully leverage the potential of DIA-MS for microbial community analysis.

Although DIA-MS presents substantial benefits for reproducible and quantitative analysis, its application in metaproteomics is still evolving and faces several technical and computational challenges. Advances in mass spectrometry and bioinformatics hold promise for addressing these hurdles, enabling deeper insights into microbial community dynamics. Ongoing research is needed to refine workflows, optimize computational methods, and explore the potential of library-free approaches to broaden its applicability in metaproteomics.

4.2 Taxonomic and functional analysis

In metaproteomics, researchers aim to characterize microbial communities by determining the organisms present (taxonomic analysis) and elucidating their physiological roles (functional analysis). These analyses provide critical insights into the composition, diversity, and ecological functions of microbial communities across diverse environments. The accuracy of these assignments depends on the quality of peptide and protein identifications (see **Section 4.1.1**) and is significantly influenced by the choice of database (see **Section 4.1.2**). Below, we describe the methodologies and tools available for taxonomic and functional annotation in metaproteomics, emphasizing the importance of robust annotation strategies and computational resources.

4.2.1 Taxonomic analysis

Taxonomic analysis in metaproteomics identifies the organisms present in a sample based on their expressed proteins. This analysis provides insights into microbial community

composition and diversity, linking proteins to their taxonomic origins. Taxonomic assignment can be achieved using exact matching or homology-based searches against comprehensive databases such as UniProtKB (The UniProt Consortium 2023) or NCBI NR (O'Leary et al. 2016).

While numerous metaproteomics-specific tools are available (described in **Section 4.2.4**), researchers can also use tools originally developed for metagenomics, such as Centrifuge (Kim et al. 2016) and Kraken 2 (Wood, Lu, and Langmead 2019). These tools match peptides or proteins to known taxa, but their accuracy depends on the completeness of publicly available genome databases. If organisms in the sample have not been previously sequenced and deposited, taxonomic assignments may be incomplete or inaccurate.

Alternatively, taxonomic assignments can leverage meta-omics databases derived from metagenomic assemblies. Proteins are inherently tied to genomes, and clustering metagenomic sequences into MAGs enables genome-centric taxonomy assignment. Tools like GTDB-Tk (Chaumeil et al. 2020) use MAG taxonomy to assign taxa to proteins. For proteins not linked to MAGs, tools such as CAT (von Meijenfeldt et al. 2019) can infer taxonomy based on the context of all the genes in an assembled contig. Advances in long-read sequencing are revolutionizing genome assembly from metagenomes, further improving taxonomic assignments (Liu et al. 2022).

4.2.2 Functional analysis

Functional analysis of metaproteomes reveals how microbial communities contribute to environmental processes, human health, and disease. By measuring the abundance of proteins involved in processes such as metabolism, transport, replication, and defense, functional analysis provides a window into microbial community dynamics and their roles in ecosystems.

To describe microbial functions, various functional ontologies are used:

- **Gene Ontology (GO)**: Organizes annotations into three categories: molecular functions, biological processes, and cellular components. GO terms are used to describe what a gene product does (molecular function), the biological goals it helps achieve (biological process), and where in the cell it acts (cellular component) (The Gene Ontology Consortium 2019)
- **Enzyme Commission (EC) numbers**: Categorizes enzymes by the chemical reactions they catalyze, particularly useful in studies of enzymatic activity and the role these enzymes play in metabolic pathways.

- **Kyoto Encyclopedia of Genes and Genomes (KEGG)**: Maps proteins to metabolic and signaling pathways, illustrating their interactions within larger biological systems (Kanehisa and Goto 2000)

There are also more specialized ontologies such as MEROPS (Rawlings et al. 2018) for proteases and CAZy (Drula et al. 2022) for carbohydrate-active enzymes, including glycoside hydrolases, offer enhanced specificity for analyzing distinct functional categories within microbial communities.

Functional annotations can rely on computational tools commonly used in metagenomics, such as KoFamKOALA (Aramaki et al. 2020), InterProScan (Quevillon et al. 2005), and eggNOG-mapper (Cantalapiedra et al. 2021). However, while these tools provide robust frameworks for mapping protein functions, more tailored tools specifically designed for the unique requirements of metaproteomics are available and discussed in **Section 4.2.4**.

4.2.3 Peptide-centric vs protein-centric approach

In metaproteomics, taxonomic and functional analyses can be performed using either a peptide-centric or protein-centric approach:

- **Peptide-centric approach**: Peptides identified through MS are directly annotated with taxa and functions based on their matches to *in silico* tryptic digests of known protein sequences. This approach ensures that all potential protein matches are retained during annotation, providing a broader view of possible taxa and functions.
- **Protein-centric approach**: Peptides are first mapped to their corresponding proteins or protein (sub)groups, aggregating peptides that share common proteins. This step addresses the protein inference problem, a challenge in assigning peptides to proteins due to shared sequences among multiple proteins (see **Section 4.1.4**).

The peptide-centric approach typically considers all proteins that a peptide could originate from, whereas protein-centric tools may discard information deemed redundant based on the chosen protein (sub)grouping strategy. These different approaches may lead to variations in the resulting annotations, and the debate over which method provides the most accurate results remains an active topic in metaproteomics research (Van Den Bossche, Kunath, et al. 2021).

4.2.4 Metaproteomics tools for taxonomic and functional analysis

Various tools have been developed for taxonomic and functional analysis in metaproteomics, each with distinct features and applications (Sajulga et al. 2020).

Unipept is a powerful ecosystem of tools for the taxonomic and functional analysis of metaproteomics samples, offering a command-line interface (CLI), a desktop application, a web application, and an application programming interface (API) to accommodate diverse user preferences and workflows. (Vande Moortele, Devlaminck, et al. 2024; Verschaffelt et al. 2023; 2020). It follows a peptide-centric approach, assigning taxa and functions directly to peptides by mapping them to the UniProtKB database. For taxonomic classification, Unipept calculates the LCA by identifying the most specific, or lowest, shared taxonomic rank among all taxa associated with a peptide's matched proteins (**Figure 6**). More details on how the LCA is calculated can be found in a recent comprehensive tutorial (Van Den Bossche, Verschaffelt, et al. 2024). Unipept also supports extensive functional analysis by reporting functions based on the GO, EC, and InterPro classifications. For each peptide, it aggregates all annotations associated with proteins matching the input peptide and counts their occurrences. This information is displayed in a table within the web application. Detailed tutorials and examples for using Unipept have been published (Mesuere et al. 2018; Van Den Bossche, Verschaffelt, et al. 2024), and the documentation available on the website (https://unipept.ugent.be/) offers additional guidance to help users navigate the tool.

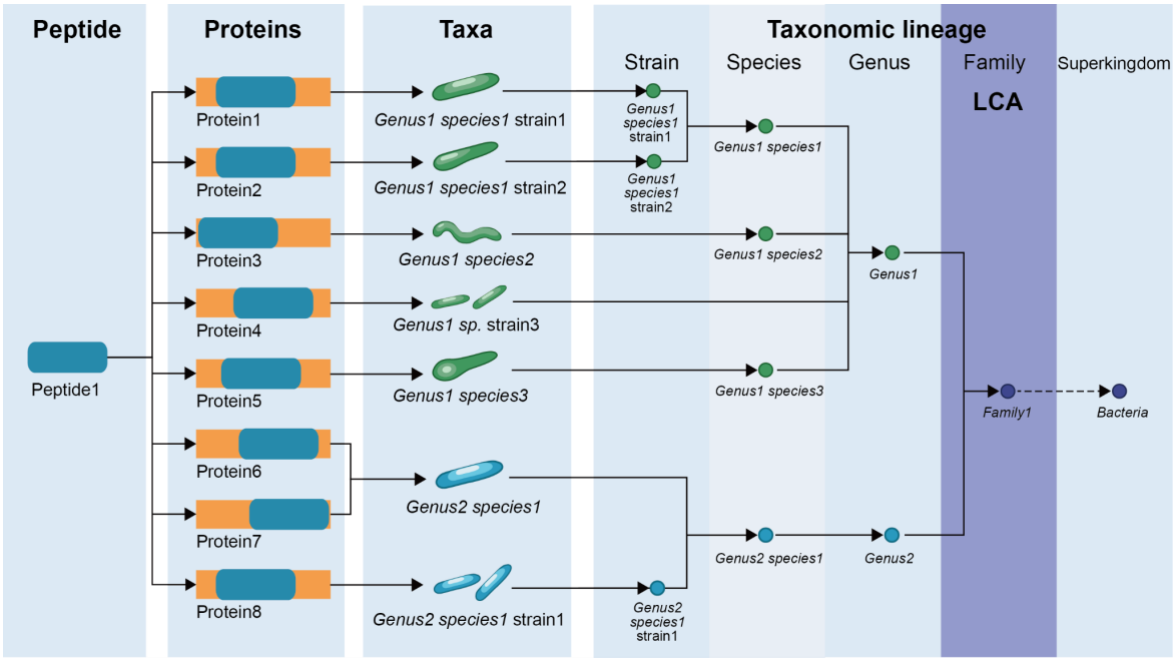


Figure 6. Calculation of the Lowest Common Ancestor (LCA) for a tryptic peptide. In this figure, the hypothetical Peptide 1 is present in eight different proteins, which are

associated with seven distinct organisms. The LCA for these organisms is identified as the hypothetical Family 1. Figure adjusted from (Van Den Bossche, Verschaffelt, et al. 2024)

The Peptonizer2000 is a novel metaproteomics pipeline for taxonomic inference that models the errors and uncertainties introduced by a typical metaproteomics analysis pipeline (Holstein et al. 2024). Indeed, the analysis of mass spectra is inherently challenging: researchers need to match observed data to databases of protein sequences, where factors such as database bias, ambiguous spectra, degenerate peptide sequences, and interspecies sequence homology come into play. The Peptonizer2000 pipeline uses Bayesian statistics to model peptide sequences, associated taxa, and the possible errors and uncertainties introduced earlier as a graph. Then, subsequently, the Belief Propagation algorithm is utilized on this graph to compute probability scores that indicate the potential presence of a taxon in a sample under study.

MetaLab (Cheng et al. 2017; 2023; Liao et al. 2018; L. Li et al. 2022) is an integrated software platform that provides a streamlined pipeline for microbial identification, quantification, and taxonomic profiling using mass spectrometry raw data. Employing a hybrid approach, MetaLab combines information derived from both peptide-centric and protein-centric metaproteomics analyses. MetaLab utilizes a precomputed index of the UniProtKB for taxonomic classification of identified peptides and retrieves functional annotations from the eggNOG database (Hernández-Plaza et al. 2023). The latest version supports DDA and DIA workflows across various mass spectrometry platforms (Cheng et al. 2024). Comprehensive resources on iMetaLab (L. Li et al. 2022) can be found on their dedicated Wiki-page (https://wiki.imetalab.ca/).

Prophane (Schiebenhoefer et al. 2020) is a software tool designed for taxonomic and functional annotation of metaproteomes, offering interactive result visualization and an intuitive web-based interface. It integrates data from various annotation databases, including NCBI (Schoch et al. 2020), UniProtKB (The UniProt Consortium 2023), eggNOG (Hernández-Plaza et al. 2023) or Pfam (Mistry et al. 2021). Unlike tools such as Unipept and MetaLab, Prophane adopts a purely protein-centric approach for its analyses. The software is accessible both as a Conda package (https://anaconda.org/bioconda/prophane) and via a web service (https://prophane.de/login). Tutorials and example datasets are available on the tool's website (https://prophane.de/about/tutorial).

The MetaProteomeAnalyzer (MPA) (Muth, Behne, et al. 2015) is an open-source Java tool designed for the taxonomic and functional analysis of metaproteomics data. MPA employs both sequence-based and spectral-based approaches to identify organisms and functional

pathways in a sample, enabling researchers to explore the metabolic activities of microbial communities and their environmental interactions. The software supports multiple search engines and incorporates features to reduce data redundancy by grouping protein hits into so-called meta-proteins. MPA is available as a desktop application, and extensive tutorials, documentation, and other resources are provided on its homepage (www.mpa.ovgu.de).

4.3 Downstream statistics

A common question among researchers is how to determine the optimal approach for downstream processing of metaproteomic data. Unfortunately, there is no universal workflow that fits every scenario. This section aims to guide readers in constructing a tailored decision tree for analyzing metaproteomic datasets. In earlier sections, we detailed the generation of various metaproteomic data tables, including peptides, proteins, taxonomy, and functional attributes. The next step involves uncovering the underlying patterns and biological insights within these datasets through statistical analysis. Designing a robust statistical analysis pipeline for metaproteomics requires researchers to make several informed decisions, which are summarized in a “cheat sheet” in **Figure 7**.

4.3.1 Identifying relevant scientific questions

The foundation of any metaproteomics analysis begins with defining the key scientific question(s) of the study. Metaproteomics allows us to address a variety of research objectives. Below are some common examples of questions that can be explored (**Figure 7A**):

i. **Cohort studies:** What differential features distinguish healthy individuals from those with a disease? Are there potential biomarkers for specific conditions?

ii. **Microbiome dynamics:** How does the microbiome vary over space and time? Can beta diversity be observed at the functional ecological level? What is the impact of specific environmental factors on the microbiome?

iii. **Perturbation study**: How do microbial communities respond to external perturbations at the taxonomic, functional, and ecological levels?

iv. **Multi-omics study:** What (holistic) insights can be gained by integrating metaproteomics with other omics approaches?

4.3.2 Selecting appropriate levels of analytical insights

Once the primary research questions are defined, the next step is to determine the level of insights required to address these questions (**Figure 7B**). This involves selecting between different analytical approaches tailored to the objectives of the study:

i. Feature-centric analysis:

Feature-based methods are the most commonly applied in metaproteomics. These analyses focus on identifying differential features, which are quantifiable variables that exhibit statistically significant differences between groups or conditions. Examples include specific peptides, proteins, taxonomic groups, or annotated functions that vary significantly under different experimental conditions.

There are two key considerations that underpin feature-centric analysis: (i) the assumption of standard statistical distributions, such as normality, to validate analytical methods, and (ii) the treatment of features as independent variables, enabling the use of widely applied statistical approaches like parametric or non-parametric tests.

By adhering to these principles, feature-centric analyses enable robust identification of biologically meaningful differences across datasets.

ii. Community-centric analysis:

Unlike feature-centric analysis, community-centric analysis considers the dataset as a reflection of a living ecological community. Here, proteins are viewed not as isolated features but as components of interconnected networks, with functions linked through evolutionary relationships and taxonomic origins. For example, proteins from different taxa may exhibit functional redundancy, while ecological dynamics may influence functional and taxonomic interactions.

Due to these complex interactions, traditional statistical methods that assume feature independence may not be suitable. To address these challenges, novel ecological approaches have been developed in metaproteomics, inspired by advancements in metagenomics.

For example, metrics for functional redundancy utilize bipartite networks to link taxonomic and functional attributes, serving as indicators of community health and stability (Blakeley-Ruiz et al. 2019; L. Li et al. 2023). Similarly, PhyloFunc, integrates phylogenetic

composition into functional beta diversity analysis by incorporating functional distances at nodes of phylogenetic trees and applying a unifrac-like weighting scheme (Luman Wang et al. 2024). This approach distinguishes whether functional changes result from compensation among closely related species or shifts between distantly related taxa, offering valuable insights into ecological dynamics.

iii. Cross-omics analysis

The metaproteome is not independent of other meta-omes; therefore, the integration of multiple omics datasets is crucial for a deeper understanding of the systems ecology of microbiomes. Different meta-omics approaches possess complementary strengths as they collectively capture variations along the central dogma of molecular biology (DNA → RNA → Protein), favoring a comprehensive understanding of biological processes and ecological interactions within microbiomes.

Despite the complementary nature of these datasets, most studies have traditionally analyzed meta-omics using separate, stand-alone workflows. However, recent advances in bioinformatics tools and platforms, such as Galaxy (Schiml et al. 2023) and MOSCA (Sequeira et al. 2024), have facilitated the integration of these datasets, enabling more seamless and coherent analysis. Cross-omics analysis can also provide an in-depth view of the functional dynamics of community ecology.

In a recent study, metagenomics and metaproteomics were paired to assess whether certain proteins serve as niche proteins (proteins that contribute to the ecological role or niche that a microbial community occupies within its environment) or play essential metabolic roles within a community (T. Wang et al. 2024). To achieve this, genome- and proteome-level functional redundancy within the community were compared simultaneously. A larger discrepancy might indicate that certain genes are present but not expressed as proteins, suggesting a more specialized or niche role. Smaller discrepancies might indicate that the genes are actively translated into proteins, suggesting essential metabolic functions.

4.3.3 Data preprocessing strategies

After making the relevant decisions outlined in **Sections 4.3.1 and 4.3.2**, the first step in downstream analysis is data preprocessing. Common preprocessing steps include data

filtering, data transformation, data imputation, and data scaling (**Figure 7C**). However, there is no universal approach for data preprocessing; the best strategy depends on the specific research questions under investigation.

i) Data transformation

Common data transformations used in proteomics and metaproteomics include logarithmic transformations (e.g., log2 or log10) and square root transformations. However, not all scenarios are suitable for data transformation.

When to use data transformation: Transformation is recommended when achieving near-normality in the data is necessary. For feature-level analyses, log transformation of peak intensities can make the data approximate a normal distribution. Normal distributions are crucial for many commonly applied metaproteomic feature selection methods, such as linear models, empirical Bayes, univariate t-tests, partial least squares discriminant analysis (PLS-DA), and orthogonal partial least squares discriminant analysis (OPLS-DA). If the data are not normally distributed, alternative non-parametric methods may be considered to meet the assumptions of the chosen analysis.

When not to use data transformation: Transformation should be avoided when reflecting protein abundance. For example, volcano plots, often used for identifying differential features, plot statistical significance (-log10(p-value)) against fold change (log2 fold change). While fold change values are log-transformed for visualization purposes, the original fold change data should remain untransformed during statistical analyses or comparisons. Additionally, in community-level analyses, log transformation can obscure protein biomass information, which is essential for estimating taxonomic and functional compositions. Protein intensities or PSM counts can serve as reliable measures of protein biomass contributions by taxa (Kleiner et al. 2017). Therefore, composition-based analyses, such as alpha and beta diversity or functional redundancy assessments, should use untransformed data.

ii) Data centering and scaling

In standard metaproteomics workflows, an equal amount of protein is typically extracted from each sample, digested, and loaded into the mass spectrometer to ensure consistency and comparability. However, in specific cases, metaproteomics may quantify overall protein biomass responses based on the total protein biomass in a given system volume rather than standardizing based on protein content (L. Li et al. 2020). In such cases, centering and

scaling are not recommended. Alternative normalization techniques, such as total spectral count normalization or median normalization, may be more appropriate for these scenarios.

iii) Data filtering

Filtering the dataset typically helps remove noise, irrelevant features, or outliers. The application of data filtering should be tailored to the specific context of the study:

When to use stringent data filtering? For feature-centric analysis. When identifying biomarkers, it is essential to apply stringent data filtering. This involves setting a higher threshold for the presence of proteins across samples to ensure that the identified biomarkers are consistently found in the majority of subjects. By requiring proteins to be present in a large percentage of samples (e.g., 70-90%), researchers can improve the reliability and relevance of the identified biomarkers. This consistency is critical for validating potential biomarkers, as it reduces the likelihood of identifying false positives. Data filtering is also typically stringent for other types of feature-centric analysis to ensure the validity of statistical hypotheses. However, the threshold and method of filtering (e.g., by the whole dataset or by group) must be properly applied to prevent over-filtering, which could remove features that are truly missing in specific subgroups.

When is data filtering optional? For community-centric analysis. While some level of filtering is still beneficial to remove obvious noise, the thresholds can be less stringent compared to feature-centric analysis. This allows for a more comprehensive view of community dynamics. For example, unfiltered taxon-specific functional data can provide a better review of the degree distribution of functions in a microbiome (L. Li et al. 2023).

iv) Data imputation

In a metaproteomic dataset, missingness often arises from two simultaneous mechanisms. First, the diversity and sparse nature of the metaproteome lead to a significant proportion of true missing proteins (missing not at random) (Plancade et al., 2022). Second, the inherent depth limitation of current common metaproteomic techniques results in highly sparse detection of low-abundance proteins across samples (missing at random) (Plancade et al. 2022).

Data imputation is the step that requires the most caution. Improper selection of the data imputation approach can induce false positives. When a large proportion (e.g., >50%) of a feature is missing, excessive imputation can lead to the creation of artificial values that do not reflect the true biological scenario and, in some cases, can further lead to false positives.

If the imputation method does not accurately reflect the nature of the missing data, it can introduce bias, particularly if the data contains a mixture of both missingness mechanisms. If features have been selected through a statistical test following data imputation, it is recommended to always revisit the un-imputed data to double-check if the feature-level difference is true before drawing solid conclusions.

Alternatively, a univariate selection method has been which combines a test of association between missingness and classes with a test for the difference in observed intensities between classes. This method provides a robust alternative for handling missing data without relying on imputation (Plancade et al. 2022).

Notably, data imputation is essential for feature selection analysis, whereas for community-level approaches, it is typically unnecessary, for reasons similar to those explained above.

4.3.4 Choosing data analysis methods

After a thorough understanding and careful selection of preprocessing steps, the final step in downstream data analysis is the selection of appropriate methods. This stage presents significant opportunities for deriving diverse insights from the dataset and is often the most engaging and time-consuming phase, allowing researchers to explore the data and uncover meaningful biological or ecological patterns and conclusions. These strategies typically include, but are not limited to:

- **Dimensionality reduction**: Dimensionality reduction methods are commonly used to uncover underlying patterns or structures within the dataset and to assess similarities between samples. Unsupervised methods such as Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), hierarchical clustering, and k-means clustering are frequently applied. Supervised methods, such as Partial Least Squares Discriminant Analysis (PLS-DA), are also widely utilized. Dimensionality reduction is applicable not only to peptide, protein, taxonomic, and functional tables but also at the MS1 level, especially when the primary goal is to reveal patterns between samples (Simopoulos et al. 2022).
- **Enrichment analysis:** Enrichment analysis determines whether a subset of selected features is significantly overrepresented compared to a background database. While enrichment analysis can be implemented using programming languages such as R, iMetaShiny (L. Li et al. 2022) offers interactive functionality for taxonomic and functional enrichment analysis of protein IDs or COG IDs.

However, protein ID-based enrichment analysis is currently restricted to human gut metaproteome analysis using the IGC database.

- **Feature Selection:** Several online tools, such as MetaFS (Tang et al. 2021), MetaQuantome (Easterly et al. 2019), MetaX (Qing Wu et al. 2024), iMetaShiny (L. Li et al. 2022), and stand-alone tools, such as Meta4P (Porcheddu et al. 2023) have been developed to facilitate feature-based metaproteomic data analysis without requiring extensive programming expertise.
- **Pathway analysis:** Pathway analysis is typically employed to gain an overview of detected functions or to compare differentially expressed or enriched pathways across groups. The most commonly used tools for pathway analysis include KEGG mapper (Kanehisa and Goto 2000) and iPath (Letunic et al. 2008). More recently, PathwayPilot was developed to easily compare functions at the KEGG pathway level, either between selected taxa within a single sample or across different samples, by leveraging EC numbers to identify active enzymes as proxies for metabolites linked to KEGG maps, thereby facilitating investigations into functions associated with specific conditions while allowing targeted analysis of selected species (Vande Moortele, Verschaffelt, et al. 2024).
- **Community analysis:** Beyond feature-driven analysis, community-level analysis focuses on viewing the entire metaproteome as a dynamic system. Such analyses may include inferring community composition, alpha diversity, beta diversity, and functional redundancy using metaproteomic data.

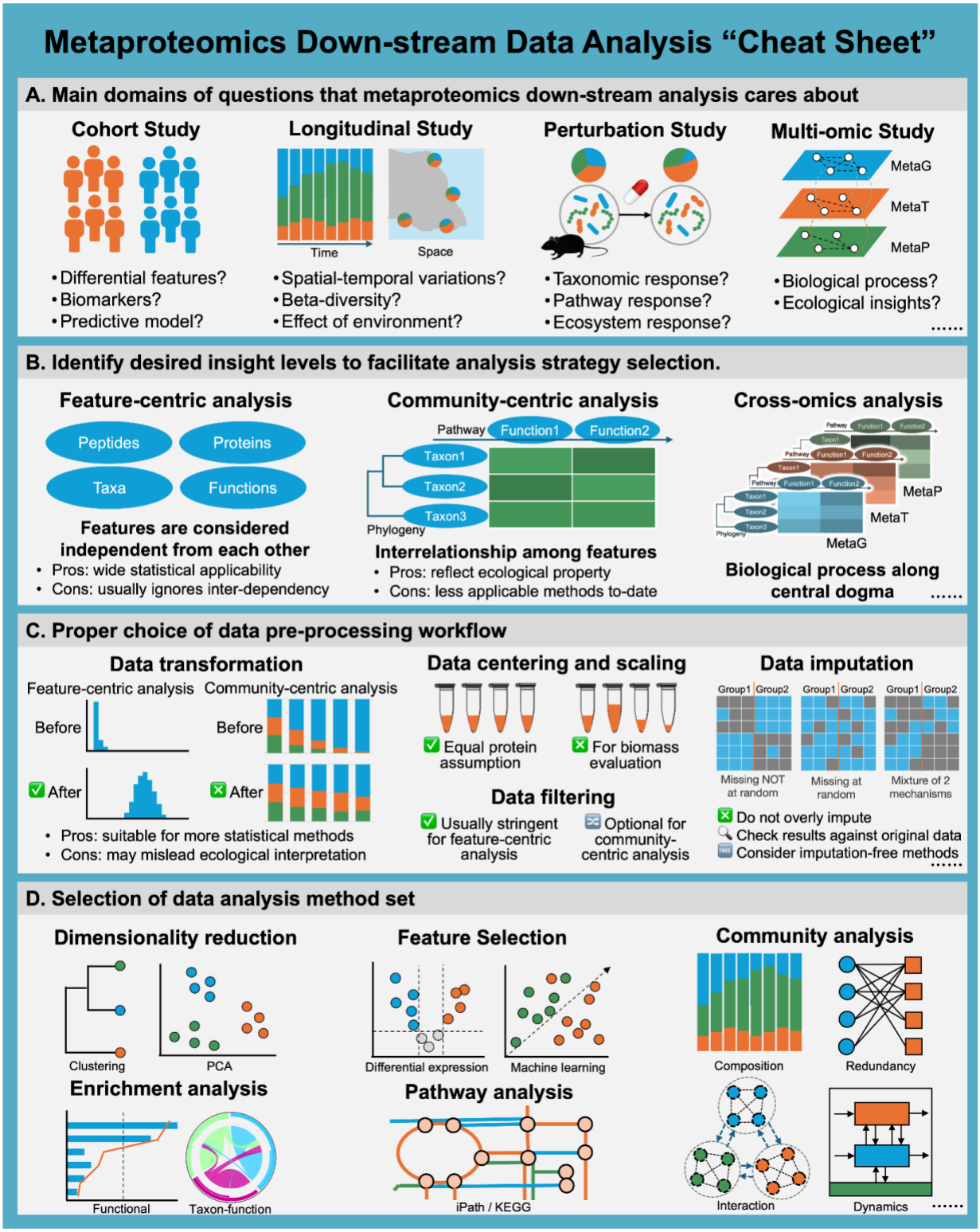


Figure 7. Metaproteomics down-stream data analysis "cheat sheet"

5. A collaborative effort: writing a comprehensive review with members of the Metaproteomics Initiative

The Metaproteomics Initiative is an international community dedicated to advancing the field of metaproteomics within microbiome research. Supported by HUPO and EuPA and in collaboration with ELIXIR, this initiative serves as a central hub for researchers to disseminate advancements, share methodologies, and establish standards across the metaproteomics community.

This Initiative aims to facilitate communication between experts and newcomers, standardize practices, and accelerate developments in metaproteomic methodologies. Its primary mission is to be the go-to resource for metaproteomics fundamentals, advancements, and applications, fostering a collaborative network to drive forward experimental and bioinformatic methodologies.

The Metaproteomics Initiative supports on three pillars:

1. **Communication & Collaboration**: This pillar focuses on sharing field advancements, organizing benchmark studies like CAMPI, and hosting the International Metaproteomics Symposium (IMS).
2. **Education & Outreach**: The initiative educates the broader microbiome community through accessible resources, including webinars and workshops, and facilitates expert interactions.
3. **Standardization**: Efforts are directed toward developing robust (meta)data standards, promoting FAIR data principles to ensure accessible and reusable research outputs.

As part of our commitment to Education & Outreach, we created this review to make metaproteomics accessible to a broad audience. To ensure a thorough and well-rounded perspective, we first invited experts in various areas to draft individual sections. These drafts were then reviewed internally, where initial feedback helped refine each section. Once authors made adjustments, the document went through additional rounds, allowing all contributors to share insights and address any remaining comments.

In the next step, we brought in microbiome researchers who were new to metaproteomics to review the manuscript, helping us ensure it was clear and approachable for those outside the field. With their feedback integrated, all co-authors—including section authors and both

expert and novice reviewers—had a final opportunity to review the work. This collaborative approach allowed us to prepare a comprehensive, accessible resource, which we shared as a preprint before journal submission.

6. Conclusion

This *Microbiologist's Guide to Metaproteomics* is designed for microbiome researchers starting in metaproteomics, offering a practical introduction to reduce barriers to entry. It covers the essentials of metaproteomics, including experimental design, sample preparation, mass spectrometry data acquisition, peptide identification, protein inference, taxonomic and functional analysis, and basic statistical methods. The guide provides the foundational knowledge needed to apply metaproteomic technologies in microbiology and microbiome studies.

Metaproteomics is a rapidly evolving field with unresolved technical challenges and unexplored areas. This guide focuses on foundational concepts rather than providing exhaustive coverage. To address these challenges, the Metaproteomics Initiative launched the "Critical Assessment of Metaproteome Investigations" (CAMPI) series, which facilitates multi-laboratory collaborations to compare and improve workflows, including sample preparation, mass spectrometry methods, and bioinformatics.

Looking ahead, the next decade promises remarkable advancements in mass spectrometry, with continually improving performance deepening the coverage of metaproteomic analysis. These advancements, coupled with ongoing and future enhancements in wet-lab protocols, strategies, and bioinformatic tools, will further propel the field. Collaborative efforts, such as the CAMPI series of the Metaproteomics Initiative, underscores the power of cooperation in driving metaproteomic progress. These developments, supported by input from microbiome researchers, will help deepen our understanding of microbiomes and their functions in diverse ecosystems.

Author contributions

This review is a collaborative effort led by Tim Van Den Bossche and Leyuan Li, and overseen by the Scientific Committee of the Metaproteomics Initiative, who provided overall guidance. Each section was contributed by nominated authors and internal reviewers as follows: **Section 1: Why metaproteomics?** was written by Robert Hettich, Jean Armengaud, Dirk Benndorf, and Paul Wilmes, and reviewed by Daniel Figeys.

Section 2: Basics of proteomics was written by Zhibin Ning and Daniel Figeys, and reviewed by Leyuan Li. **Section 3: Experimental methods in metaproteoimcs** includes several subsections: **3.1 Experiment Design** was written by Lucia Grenga and Jean Armengaud, reviewed by Céline Henry and Leyuan Li; **3.2 Sample collection, preservation, and storage prior to preprocessing**, where **3.2.1 Sample collection and preservation** was written by Sergio Uzzau and Alessandro Tanca, and **3.2.2 Storage conditions to maintain sample integrity** was written by Lucia Grenga and Jean Armengaud, both reviewed by Céline Henry and Leyuan Li; **3.3 Sample preprocessing** was written by Lucia Grenga and Jean Armengaud, reviewed by Céline Henry; **3.4 Protein sample preparation: from extraction to digestion** was written by Nico Jehmlich, reviewed by Xu Zhang and Céline Henry; **3.5 Separation and fractionation techniques** was written by Xu Zhang and Marybeth Creskey, reviewed by Céline Henry; **3.6 Automation** was written by Leyuan Li, reviewed by Sergio Uzzau and Alessandro Tanca; **3.7 Mass spectrometry data acquisition methods** was written by Zhibin Ning and Daniel Figeys, reviewed by Jean Armengaud and Céline Henry. **Section 4: Computational analysis of metaproteomics data** includes several subsections: **4.1.1 Peptide identification with proteomics search engines** was written by Pratik Jagtap, Subina Mehta, and Timothy Griffin, reviewed by Tanja Holstein and Kai Cheng; **4.1.2 Database construction or selection** was written by Paul Wilmes and Benoit Kunath, reviewed by Jose Alfredo Blakely-Ruiz; **4.1.3 PSM FDR control**, by Tim Van Den Bossche and Lennart Martens, reviewed by Tanja Holstein; **4.1.4 Protein inference** was written by Tim Van Den Bossche, reviewed by Tanja Holstein; **4.1.5 Protein quantification** was written by Jose Alfredo Blakely-Ruiz and Manuel Kleiner, reviewed by Tanja Holstein and Kai Cheng; **4.1.6 DIA data analysis** was written by Pratik Jagtap, reviewed by Tanja Holstein and Kai Cheng; **4.2: Taxonomic and functional Analysis** was written by Pieter Verschaffelt and Bart Mesuere, reviewed by Tanja Holstein and Tim Van Den Bossche; **4.3 Downstream statistics** was written by Leyuan Li, reviewed by Tanja Holstein and Lucia Grenga. **Section 5: A collaborative effort: writing a comprehensive review with members of the Metaproteomics Initiative** was written by Tim Van Den Bossche, and reviewed by Leyuan Li. We invited Madita Brauer, Xuxa Malliet, Jing Wang, Xin Zhang, Jong Kim to review the manuscript to ensure its accessibility. All figures were artistically designed by Leyuan Li based on author drafts. To homogenize the text, ensure consistency and avoid redundancy across sections, all sections were rewritten by Tim Van Den Bossche. All authors commented and approved the final version of the manuscript.

Abbreviations

ABP	activity-based protein probing
ACN	acetonitrile
AI	artificial intelligence
API	application programming interface
AUC	area under the curve
BSHs	bile salt hydrolases
BSL	biosafety level
CLI	command-line interface
cRAP	common repository of adventitious proteins
CZE	capillary zone electrophoresis
DDA	data-dependent acquisition
DIA	data-independent acquisition
FASP	filter-aided sample preparation
FDR	false discovery rate
Galaxy-P	Galaxy for proteomics
GHs	glycoside hydrolases
HILIC	hydrophilic interaction liquid chromatography
HPLC	high-performance liquid chromatography
IMAC	Immobilized metal affinity chromatography
iST	in-stage tips
LC	liquid chromatography
LCA	lowest common ancestor
LFQ	label-free quantification
MAG	metagenome-assembled genome
MPA	MetaProteome Analyzer
MS	mass spectrometry

MuDPIT	multidimensional protein identification technology
NSAF	normalized spectral abundance factor
ORF	open-reading frame
PSMs	peptide-spectrum matches
PTM	post-translational modification
QC	quality control
RP	reverse phase
SCX	strong cation exchange
SILAC	stable isotope labeling by amino acids in cell culture
SP3	single-pot solid-phase-enhanced sample preparation
SPE	solid-phase extraction
TMT	tandem mass tags

Acknowledgements

TV acknowledges funding from the Research Foundation Flanders (FWO) [grant 1286824N]. JA acknowledges funding from the French National Agency for Research (Agence Nationale de la Recherche, grant ANR-20-CE34-0012) and Occitanie Région (grant 21023526-DeepMicro). JB acknowledges funding from the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number R35GM138362. RH acknowledges funding from the United States Department of Energy, Biological and Environmental Research Program. PJ and TG acknowledge funding from the Minnesota Ovarian Cancer Alliance, the National Institutes of Health/National Cancer Institute [grants 5R01CA262153 (A.P.N.S.), 1R21CA267707], and The National Institutes of Health/National Cancer Institute [grant P30CA077598]. MK acknowledges funding from the National Institute Of General Medical Sciences of the National Institutes of Health [grant R35GM138362]. BK acknowledges funding from the FNR INTERMOBILITY/2022/BM/16965254. LM acknowledges funding from the Research Foundation Flanders (FWO) [grants G028821N, G010023N]. AT acknowledges funding from the Next Generation EU [grant PNRR-MAD-2022-12376416]. SU acknowledges funding from the Next Generation EU [grant PNRR-MAD-2022-12376416]. PV acknowledges funding from the Ghent University (BOF) [grant BOF/01P10623]. JW, XZ, LL

acknowledges funding from the State Key Laboratory of Medical Proteomics, National Center for Protein Sciences (Beijing), China. PW acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program [grant 863664].

Conflicts of Interest

DF is a Co-founder of MedBiome inc.

References

Aakko, Juhani, Sami Pietilä, Tomi Suomi, Mehrad Mahmoudian, Raine Toivonen, Petri Kouvonen, Anne Rokka, Arno Hänninen, and Laura L. Elo. 2020. 'Data-Independent Acquisition Mass Spectrometry in Metaproteomics of Gut Microbiota—Implementation and Computational Analysis'. *Journal of Proteome Research* 19 (1): 432–36. https://doi.org/10.1021/acs.jproteome.9b00606.

Abdill, Richard J., Elizabeth M. Adamowicz, and Ran Blekhman. 2022. 'Public Human Microbiome Data Are Dominated by Highly Developed Countries'. *PLoS Biology* 20 (2): e3001536. https://doi.org/10.1371/journal.pbio.3001536.

Almeida, Alexandre, Stephen Nayfach, Miguel Boland, Francesco Strozzi, Martin Beracochea, Zhou Jason Shi, Katherine S. Pollard, et al. 2021. 'A Unified Catalog of 204,938 Reference Genomes from the Human Gut Microbiome'. *Nature Biotechnology* 39 (1): 105–14. https://doi.org/10.1038/s41587-020-0603-3.

Andersen, Thea Os, Benoit J. Kunath, Live H. Hagen, Magnus Ø. Arntzen, and Phillip B. Pope. 2021. 'Rumen Metaproteomics: Closer to Linking Rumen Microbial Function to Animal Productivity Traits'. *Methods*, Methods to face the challenges of ruminant phenotyping, 186 (February):42–51. https://doi.org/10.1016/j.ymeth.2020.07.011.

Aramaki, Takuya, Romain Blanc-Mathieu, Hisashi Endo, Koichi Ohkubo, Minoru Kanehisa, Susumu Goto, and Hiroyuki Ogata. 2020. 'KofamKOALA: KEGG Ortholog Assignment Based on Profile HMM and Adaptive Score Threshold'. *Bioinformatics* 36 (7): 2251–52. https://doi.org/10.1093/bioinformatics/btz859.

Argentini, Andrea, An Staes, Björn Grüning, Subina Mehta, Caleb Easterly, Timothy J. Griffin, Pratik Jagtap, Francis Impens, and Lennart Martens. 2019. 'Update on the moFF Algorithm for Label-Free Quantitative Proteomics'. *Journal of Proteome Research* 18 (2): 728–31. https://doi.org/10.1021/acs.jproteome.8b00708.

Arıkan, Muzaffer, Tuğçe Kahraman Demir, Zeynep Yıldız, Özkan Ufuk Nalbantoğlu, Nur Damla Korkmaz, Nesrin H. Yılmaz, Aysu Şen, et al. 2023. 'Metaproteogenomic Analysis of Saliva Samples from Parkinson's Disease Patients with Cognitive Impairment'. *Npj Biofilms and Microbiomes* 9 (1): 1–10. https://doi.org/10.1038/s41522-023-00452-x.

Armengaud, Jean. 2023. 'Metaproteomics to Understand How Microbiota Function: The Crystal Ball Predicts a Promising Future'. *Environmental Microbiology* 25 (1): 115–25. https://doi.org/10.1111/1462-2920.16238.

Armengaud, Jean, Joseph A Christie-Oleza, Gérémy Clair, Véronique Malard, and Catherine Duport. 2012. 'Exoproteomics: Exploring the World around Biological Systems'. *Expert Review of Proteomics* 9 (5): 561–75. https://doi.org/10.1586/epr.12.52.

Audain, Enrique, Julian Uszkoreit, Timo Sachsenberg, Julianus Pfeuffer, Xiao Liang,

Henning Hermjakob, Aniel Sanchez, et al. 2017. 'In-Depth Analysis of Protein Inference Algorithms Using Multiple Search Engines and Well-Defined Metrics'. *Journal of Proteomics* 150 (January):170–82. https://doi.org/10.1016/j.jprot.2016.08.002.

Aylward, Frank O., Kristin E. Burnum, Jarrod J. Scott, Garret Suen, Susannah G. Tringe, Sandra M. Adams, Kerrie W. Barry, et al. 2012. 'Metagenomic and Metaproteomic Insights into Bacterial Communities in Leaf-Cutter Ant Fungus Gardens'. *The ISME Journal* 6 (9): 1688–1701. https://doi.org/10.1038/ismej.2012.10.

Baldrian, Petr. 2017. 'Microbial Activity and the Dynamics of Ecosystem Processes in Forest Soils'. *Current Opinion in Microbiology* 37 (June):128–34. https://doi.org/10.1016/j.mib.2017.06.008.

Bandick, Anna K., and Richard P. Dick. 1999. 'Field Management Effects on Soil Enzyme Activities'. *Soil Biology and Biochemistry* 31 (11): 1471–79. https://doi.org/10.1016/S0038-0717(99)00051-6.

Bankvall, Maria, Miguel Carda-Diéguez, Alex Mira, Anders Karlsson, Bengt Hasséus, Roger Karlsson, and Jairo Robledo-Sierra. 2023. 'Metataxonomic and Metaproteomic Profiling of the Oral Microbiome in Oral Lichen Planus - a Pilot Study'. *Journal of Oral Microbiology* 15 (1): 2161726. https://doi.org/10.1080/20002297.2022.2161726.

Barglow, Katherine T., and Benjamin F. Cravatt. 2007. 'Activity-Based Protein Profiling for the Functional Annotation of Enzymes'. *Nature Methods* 4 (10): 822–27. https://doi.org/10.1038/nmeth1092.

Bastida, F., T. Hernández, and C. García. 2014. 'Metaproteomics of Soils from Semiarid Environment: Functional and Phylogenetic Information Obtained with Different Protein Extraction Methods'. *Journal of Proteomics* 101 (April):31–42. https://doi.org/10.1016/j.jprot.2014.02.006.

Benndorf, Dirk, Gerd U. Balcke, Hauke Harms, and Martin von Bergen. 2007. 'Functional Metaproteome Analysis of Protein Extracts from Contaminated Soil and Groundwater'. *The ISME Journal* 1 (3): 224–34. https://doi.org/10.1038/ismej.2007.39.

Benndorf, Dirk, Carsten Vogt, Nico Jehmlich, Yvonne Schmidt, Henrik Thomas, Gary Woffendin, Andrej Shevchenko, Hans-Hermann Richnow, and Martin von Bergen. 2009. 'Improving Protein Extraction and Separation Methods for Investigating the Metaproteome of Anaerobic Benzene Communities within Sediments'. *Biodegradation* 20 (6): 737–50. https://doi.org/10.1007/s10532-009-9261-3.

Berard, Alicia R., Douglas K. Brubaker, Kenzie Birse, Alana Lamont, Romel D. Mackelprang, Laura Noël-Romas, Michelle Perner, et al. 2023. 'Vaginal Epithelial Dysfunction Is Mediated by the Microbiome, Metabolome, and mTOR Signaling'. *Cell Reports* 42 (5): 112474. https://doi.org/10.1016/j.celrep.2023.112474.

Beresford-Jones, Benjamin S., Samuel C. Forster, Mark D. Stares, George Notley, Elisa Viciani, Hilary P. Browne, Daniel J. Boehmler, et al. 2022. 'The Mouse Gastrointestinal Bacteria Catalogue Enables Translation between the Mouse and Human Gut Microbiotas via Functional Mapping'. *Cell Host & Microbe* 30 (1): 124-138.e8. https://doi.org/10.1016/j.chom.2021.12.003.

Berg, Gabriele, Daria Rybakova, Doreen Fischer, Tomislav Cernava, Marie-Christine Champomier Vergès, Trevor Charles, Xiaoyulong Chen, et al. 2020. 'Microbiome Definition Re-Visited: Old Concepts and New Challenges'. *Microbiome* 8 (1): 103. https://doi.org/10.1186/s40168-020-00875-0.

Beyter, Doruk, Miin S. Lin, Yanbao Yu, Rembert Pieper, and Vineet Bafna. 2018. 'ProteoStorm: An Ultrafast Metaproteomics Database Search Framework'. *Cell Systems* 7 (4): 463-467.e6. https://doi.org/10.1016/j.cels.2018.08.009.

Bielow, Chris, Nils Hoffmann, David Jimenez-Morales, Tim Van Den Bossche, Juan Antonio Vizcaíno, David L. Tabb, Wout Bittremieux, and Mathias Walzer. 2024. 'Communicating Mass Spectrometry Quality Information in mzQC with Python, R, and Java'. *Journal of the American Society for Mass Spectrometry* 35 (8): 1875–

82. https://doi.org/10.1021/jasms.4c00174.
Bielow, Chris, Guido Mastrobuoni, and Stefan Kempa. 2016. 'Proteomics Quality Control: Quality Control Software for MaxQuant Results'. *Journal of Proteome Research* 15 (3): 777–87. https://doi.org/10.1021/acs.jproteome.5b00780.
Bihani, Surbhi, Aryan Gupta, Subina Mehta, Andrew T. Rajczewski, James Johnson, Dhanush Borishetty, Timothy J. Griffin, Sanjeeva Srivastava, and Pratik D. Jagtap. 2023. 'Metaproteomic Analysis of Nasopharyngeal Swab Samples to Identify Microbial Peptides in COVID-19 Patients'. *Journal of Proteome Research* 22 (8): 2608–19. https://doi.org/10.1021/acs.jproteome.3c00040.
Birse, Kenzie D., Kateryna Kratzer, Christina Farr Zuend, Sarah Mutch, Laura Noël-Romas, Alana Lamont, Max Abou, et al. 2020. 'The Neovaginal Microbiome of Transgender Women Post-Gender Reassignment Surgery'. *Microbiome* 8 (1): 61. https://doi.org/10.1186/s40168-020-00804-1.
Blakeley-Ruiz, J. Alfredo, Alison R. Erickson, Brandi L. Cantarel, Weili Xiong, Rachel Adams, Janet K. Jansson, Claire M. Fraser, and Robert L. Hettich. 2019. 'Metaproteomics Reveals Persistent and Phylum-Redundant Metabolic Functional Stability in Adult Human Gut Microbiomes of Crohn's Remission Patients despite Temporal Variations in Microbial Taxa, Genomes, and Proteomes'. *Microbiome* 7 (1): 18. https://doi.org/10.1186/s40168-019-0631-8.
Blakeley-Ruiz, J. Alfredo, and Manuel Kleiner. 2022. 'Considerations for Constructing a Protein Sequence Database for Metaproteomics'. *Computational and Structural Biotechnology Journal* 20 (January):937–52. https://doi.org/10.1016/j.csbj.2022.01.018.
Blakeley-Ruiz, J Alfredo, Carlee S McClintock, Him K Shrestha, Suresh Poudel, Zamin K Yang, Richard J Giannone, James J Choo, et al. 2022. 'Morphine and High-Fat Diet Differentially Alter the Gut Microbiota Composition and Metabolic Function in Lean versus Obese Mice'. *ISME Communications* 2 (1): 66. https://doi.org/10.1038/s43705-022-00131-6.
Blank, Clemens, Caleb Easterly, Bjoern Gruening, James Johnson, Carolin A. Kolmeder, Praveen Kumar, Damon May, et al. 2018. 'Disseminating Metaproteomic Informatics Capabilities and Knowledge Using the Galaxy-P Framework'. *Proteomes* 6 (1): 7. https://doi.org/10.3390/proteomes6010007.
Brooks, Brandon, Ryan S. Mueller, Jacque C. Young, Michael J. Morowitz, Robert L. Hettich, and Jillian F. Banfield. 2015. 'Strain-Resolved Microbial Community Proteomics Reveals Simultaneous Aerobic and Anaerobic Function during Gastrointestinal Tract Colonization of a Preterm Infant'. *Frontiers in Microbiology* 6 (July). https://doi.org/10.3389/fmicb.2015.00654.
Bruderer, Roland, Oliver M. Bernhardt, Tejas Gandhi, Yue Xuan, Julia Sondermann, Manuela Schmidt, David Gomez-Varela, and Lukas Reiter. 2017. 'Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results *'. *Molecular & Cellular Proteomics* 16 (12): 2296–2309. https://doi.org/10.1074/mcp.RA117.000314.
Bubis, Julia A., Lev I. Levitsky, Mark V. Ivanov, Irina A. Tarasova, and Mikhail V. Gorshkov. 2017. 'Comparative Evaluation of Label-Free Quantification Methods for Shotgun Proteomics'. *Rapid Communications in Mass Spectrometry* 31 (7): 606–12. https://doi.org/10.1002/rcm.7829.
Burns, Andrew P., Ya-Qin Zhang, Tuan Xu, Zhengxi Wei, Qin Yao, Yuhong Fang, Valeriu Cebotaru, et al. 2021. 'A Universal and High-Throughput Proteomics Sample Preparation Platform'. *Analytical Chemistry* 93 (24): 8423–31. https://doi.org/10.1021/acs.analchem.1c00265.
Burnum-Johnson, Kristin E., Jennifer E. Kyle, Amie J. Eisfeld, Cameron P. Casey, Kelly G. Stratton, Juan F. Gonzalez, Fabien Habyarimana, et al. 2017. 'MPLEx: A Method for Simultaneous Pathogen Inactivation and Extraction of Samples for Multi-Omics Profiling'. *Analyst* 142 (3): 442–48. https://doi.org/10.1039/C6AN02486F.

C. Silva, Ana S, Robbin Bouwmeester, Lennart Martens, and Sven Degroeve. 2019. 'Accurate Peptide Fragmentation Predictions Allow Data Driven Approaches to Replace and Improve upon Proteomics Search Engine Scoring Functions'. *Bioinformatics* 35 (24): 5243–48. https://doi.org/10.1093/bioinformatics/btz383.

Caglar, Mehmet U., John R. Houser, Craig S. Barnhart, Daniel R. Boutz, Sean M. Carroll, Aurko Dasgupta, Walter F. Lenoir, et al. 2017. 'The E. Coli Molecular Phenotype under Different Growth Conditions'. *Scientific Reports* 7 (1): 45303. https://doi.org/10.1038/srep45303.

Cai, Xue, Zhangzhi Xue, Chunlong Wu, Rui Sun, Liujia Qian, Liang Yue, Weigang Ge, et al. 2022. 'High-Throughput Proteomic Sample Preparation Using Pressure Cycling Technology'. *Nature Protocols* 17 (10): 2307–25. https://doi.org/10.1038/s41596-022-00727-1.

Cantalapiedra, Carlos P, Ana Hernández-Plaza, Ivica Letunic, Peer Bork, and Jaime Huerta-Cepas. 2021. 'eggNOG-Mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale'. *Molecular Biology and Evolution* 38 (12): 5825–29. https://doi.org/10.1093/molbev/msab293.

Charlier, Philippe, Virginie Bourdin, Didier N'Dah, Mélodie Kielbasa, Olivier Pible, and Jean Armengaud. 2024. 'Metaproteomic Analysis of King Ghezo Tomb Wall (Abomey, Benin) Confirms 19th Century Voodoo Sacrifices'. *Proteomics* 24 (16): e2400048. https://doi.org/10.1002/pmic.202400048.

Chaumeil, Pierre-Alain, Aaron J Mussig, Philip Hugenholtz, and Donovan H Parks. 2020. 'GTDB-Tk: A Toolkit to Classify Genomes with the Genome Taxonomy Database'. *Bioinformatics* 36 (6): 1925–27. https://doi.org/10.1093/bioinformatics/btz848.

Chen, Jiahui, Yingying Sun, Jie Li, Mengge Lyu, Li Yuan, Jiancheng Sun, Shangqi Chen, et al. 2024. 'In-Depth Metaproteomics Analysis of Tongue Coating for Gastric Cancer: A Multicenter Diagnostic Research Study'. *Microbiome* 12 (1): 6. https://doi.org/10.1186/s40168-023-01730-8.

Cheng, Kai, Zhibin Ning, Leyuan Li, Xu Zhang, Joeselle M. Serrana, Janice Mayne, and Daniel Figeys. 2023. 'MetaLab-MAG: A Metaproteomic Data Analysis Platform for Genome-Level Characterization of Microbiomes from the Metagenome-Assembled Genomes Database'. *Journal of Proteome Research* 22 (2): 387–98. https://doi.org/10.1021/acs.jproteome.2c00554.

Cheng, Kai, Zhibin Ning, Xu Zhang, Haonan Duan, Janice Mayne, and Daniel Figeys. 2024. 'MetaLab Platform Enables Comprehensive DDA and DIA Metaproteomics Analysis'. *bioRxiv*, January, 2024.09.27.615406. https://doi.org/10.1101/2024.09.27.615406.

Cheng, Kai, Zhibin Ning, Xu Zhang, Leyuan Li, Bo Liao, Janice Mayne, Alain Stintzi, and Daniel Figeys. 2017. 'MetaLab: An Automated Pipeline for Metaproteomic Data Analysis'. *Microbiome* 5 (1): 157. https://doi.org/10.1186/s40168-017-0375-2.

Cheng, Kai, Zhibin Ning, Xu Zhang, Janice Mayne, and Daniel Figeys. 2018. 'Separation and Characterization of Human Microbiomes by Metaproteomics'. *TrAC Trends in Analytical Chemistry* 108 (November):221–30. https://doi.org/10.1016/j.trac.2018.09.006.

Chourey, Karuna, Janet Jansson, Nathan VerBerkmoes, Manesh Shah, Krystle L. Chavarria, Lauren M. Tom, Eoin L. Brodie, and Robert L. Hettich. 2010. 'Direct Cellular Lysis/Protein Extraction Protocol for Soil Metaproteomics'. *Journal of Proteome Research* 9 (12): 6615–22. https://doi.org/10.1021/pr100787q.

Claeys, Tine, Tim Van Den Bossche, Yasset Perez-Riverol, Kris Gevaert, Juan Antonio Vizcaíno, and Lennart Martens. 2023. 'lesSDRF Is More: Maximizing the Value of Proteomics Data through Streamlined Metadata Annotation'. *Nature Communications* 14 (1): 6743. https://doi.org/10.1038/s41467-023-42543-5.

Combe, Colin W., Lars Kolbowski, Lutz Fischer, Ville Koskinen, Joshua Klein, Alexander Leitner, Andrew R. Jones, Juan Antonio Vizcaíno, and Juri Rappsilber. 2024. 'mzIdentML 1.3.0 – Essential Progress on the Support of Crosslinking and Other Identifications Based on Multiple Spectra'. *PROTEOMICS* 24 (17): 2300385.

https://doi.org/10.1002/pmic.202300385.
Cortay, Jean-Claude, Corinne Rieul, Françoise Bleicher, Mustapha Dadssi, and Alain J. Cozzone. 1988. 'Evidence of Protein Kinase Activity and Characterization of Substrate Proteins in Escherichia Coli'. In *Advances in Post-Translational Modifications of Proteins and Aging*, edited by Vincenzo Zappia, Patrizia Galletti, Raffaele Porta, and Finn Wold, 467–74. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4684-9042-8_39.
Cox, Jürgen, Marco Y. Hein, Christian A. Luber, Igor Paron, Nagarjuna Nagaraj, and Matthias Mann. 2014. 'Accurate Proteome-Wide Label-Free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ *'. *Molecular & Cellular Proteomics* 13 (9): 2513–26. https://doi.org/10.1074/mcp.M113.031591.
Cox, Jürgen, and Matthias Mann. 2008. 'MaxQuant Enables High Peptide Identification Rates, Individualized p.p.b.-Range Mass Accuracies and Proteome-Wide Protein Quantification'. *Nature Biotechnology* 26 (12): 1367–72. https://doi.org/10.1038/nbt.1511.
Cox, Jürgen, Nadin Neuhauser, Annette Michalski, Richard A. Scheltema, Jesper V. Olsen, and Matthias Mann. 2011. 'Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment'. *Journal of Proteome Research* 10 (4): 1794–1805. https://doi.org/10.1021/pr101065j.
Craig, Robertson, and Ronald C. Beavis. 2004. 'TANDEM: Matching Proteins with Tandem Mass Spectra'. *Bioinformatics* 20 (9): 1466–67. https://doi.org/10.1093/bioinformatics/bth092.
Cravatt, Benjamin F., Aaron T. Wright, and John W. Kozarich. 2008. 'Activity-Based Protein Profiling: From Enzyme Chemistry to Proteomic Chemistry'. *Annual Review of Biochemistry* 77 (Volume 77, 2008): 383–414. https://doi.org/10.1146/annurev.biochem.75.101304.124125.
Creskey, Marybeth, Leyuan Li, Zhibin Ning, Emily EF Fekete, Janice Mayne, Krystal Walker, Anna Ampaw, Robert Ben, Xu Zhang, and Daniel Figeys. 2023. 'An Economic and Robust TMT Labeling Approach for High Throughput Proteomic and Metaproteomic Analysis'. *PROTEOMICS* 23 (21–22): 2200116. https://doi.org/10.1002/pmic.202200116.
Dai, Chengxin, Anja Füllgrabe, Julianus Pfeuffer, Elizaveta M. Solovyeva, Jingwen Deng, Pablo Moreno, Selvakumar Kamatchinathan, et al. 2021. 'A Proteomics Sample Metadata Representation for Multiomics Integration and Big Data Analysis'. *Nature Communications* 12 (1): 5854. https://doi.org/10.1038/s41467-021-26111-3.
Degroeve, Sven, Davy Maddelein, and Lennart Martens. 2015. 'MS2PIP Prediction Server: Compute and Visualize MS2 Peak Intensity Predictions for CID and HCD Fragmentation'. *Nucleic Acids Research* 43 (W1): W326–30. https://doi.org/10.1093/nar/gkv542.
Delgado-Diaz, David Jose, Brianna Jesaveluk, Joshua A. Hayward, David Tyssen, Arghavan Alisoltani, Matthys Potgieter, Liam Bell, et al. 2022. 'Lactic Acid from Vaginal Microbiota Enhances Cervicovaginal Epithelial Barrier Integrity by Promoting Tight Junction Protein Expression'. *Microbiome* 10 (1): 141. https://doi.org/10.1186/s40168-022-01337-5.
Delogu, F., B. J. Kunath, P. N. Evans, M. Ø. Arntzen, T. R. Hvidsten, and P. B. Pope. 2020. 'Integration of Absolute Multi-Omics Reveals Dynamic Protein-to-RNA Ratios and Metabolic Interplay within Mixed-Domain Microbiomes'. *Nature Communications* 11 (1): 4708. https://doi.org/10.1038/s41467-020-18543-0.
Delogu, Francesco, Benoit J. Kunath, Pedro M. Queirós, Rashi Halder, Laura A. Lebrun, Phillip B. Pope, Patrick May, Stefanie Widder, Emilie E. L. Muller, and Paul Wilmes. 2024. 'Forecasting the Dynamics of a Complex Microbial Community Using Integrated Meta-Omics'. *Nature Ecology & Evolution* 8 (1): 32–44. https://doi.org/10.1038/s41559-023-02241-3.
Demichev, Vadim, Christoph B. Messner, Spyros I. Vernardis, Kathryn S. Lilley, and

Markus Ralser. 2020. ‘DIA-NN: Neural Networks and Interference Correction Enable Deep Proteome Coverage in High Throughput’. *Nature Methods* 17 (1): 41–44. https://doi.org/10.1038/s41592-019-0638-x.

Demichev, Vadim, Lukasz Szyrwiel, Fengchao Yu, Guo Ci Teo, George Rosenberger, Agathe Niewienda, Daniela Ludwig, et al. 2022. ‘Dia-PASEF Data Analysis Using FragPipe and DIA-NN for Deep Proteomics of Low Sample Amounts’. *Nature Communications* 13 (1): 3944. https://doi.org/10.1038/s41467-022-31492-0.

Deusch, Simon, Amélia Camarinha-Silva, Jürgen Conrad, Uwe Beifuss, Markus Rodehutscord, and Jana Seifert. 2017. ‘A Structural and Functional Elucidation of the Rumen Microbiome Influenced by Various Diets and Microenvironments’. *Frontiers in Microbiology* 8 (August). https://doi.org/10.3389/fmicb.2017.01605.

Deutsch, Eric W, Nuno Bandeira, Yasset Perez-Riverol, Vagisha Sharma, Jeremy J Carver, Luis Mendoza, Deepti J Kundu, et al. 2023. ‘The ProteomeXchange Consortium at 10 Years: 2023 Update’. *Nucleic Acids Research* 51 (D1): D1539–48. https://doi.org/10.1093/nar/gkac1040.

Deutsch, Eric W., Yasset Perez-Riverol, Jeremy Carver, Shin Kawano, Luis Mendoza, Tim Van Den Bossche, Ralf Gabriels, et al. 2021. ‘Universal Spectrum Identifier for Mass Spectra’. *Nature Methods* 18 (7): 768–70. https://doi.org/10.1038/s41592-021-01184-6.

Deutsch, Eric W., Juan Antonio Vizcaíno, Andrew R. Jones, Pierre-Alain Binz, Henry Lam, Joshua Klein, Wout Bittremieux, et al. 2023. ‘Proteomics Standards Initiative at Twenty Years: Current Activities and Future Work’. *Journal of Proteome Research* 22 (2): 287–301. https://doi.org/10.1021/acs.jproteome.2c00637.

Dhabaria, Avantika, Paolo Cifani, Casie Reed, Hanno Steen, and Alex Kentsis. 2015. ‘A High-Efficiency Cellular Extraction System for Biological Proteomics’. *Journal of Proteome Research* 14 (8): 3403–8. https://doi.org/10.1021/acs.jproteome.5b00547.

Diz, Angel P., and Paula Sánchez-Marín. 2021. ‘A Primer and Guidelines for Shotgun ProteomicShotgun Proteomics Analysis in Non-Model Organisms’. In *Shotgun Proteomics: Methods and Protocols*, edited by Mónica Carrera and Jesús Mateos, 77–102. New York, NY: Springer US. https://doi.org/10.1007/978-1-0716-1178-4_6.

Do, Katherine, Subina Mehta, Reid Wagner, Dechen Bhuming, Andrew T. Rajczewski, Amy P. N. Skubitz, James E. Johnson, Timothy J. Griffin, and Pratik D. Jagtap. 2024. ‘A Novel Clinical Metaproteomics Workflow Enables Bioinformatic Analysis of Host-Microbe Dynamics in Disease’. *mSphere* 9 (6): e00793-23. https://doi.org/10.1128/msphere.00793-23.

Drula, Elodie, Marie-Line Garron, Suzan Dogan, Vincent Lombard, Bernard Henrissat, and Nicolas Terrapon. 2022. ‘The Carbohydrate-Active Enzyme Database: Functions and Literature’. *Nucleic Acids Research* 50 (D1): D571–77. https://doi.org/10.1093/nar/gkab1045.

Duan, Haonan, Kai Cheng, Zhibin Ning, Leyuan Li, Janice Mayne, Zhongzhi Sun, and Daniel Figeys. 2022. ‘Assessing the Dark Field of Metaproteome’. *Analytical Chemistry* 94 (45): 15648–54. https://doi.org/10.1021/acs.analchem.2c02452.

Duan, Haonan, Xu Zhang, and Daniel Figeys. 2023. ‘An Emerging Field: Post-Translational Modification in Microbiome’. *PROTEOMICS* 23 (3–4): 2100389. https://doi.org/10.1002/pmic.202100389.

Duchovni, Lirit, Genrieta Shmunis, and Lior Lobel. 2024. ‘Posttranslational Modifications: An Emerging Functional Layer of Diet-Host-Microbe Interactions’. *mBio* 0 (0): e02387-24. https://doi.org/10.1128/mbio.02387-24.

Dumas, Thibaut, Roxana Martinez Pinna, Clément Lozano, Sonja Radau, Olivier Pible, Lucia Grenga, and Jean Armengaud. 2024. ‘The Astounding Exhaustiveness and Speed of the Astral Mass Analyzer for Highly Complex Samples Is a Quantum Leap in the Functional Analysis of Microbiomes’. *Microbiome* 12 (1): 46. https://doi.org/10.1186/s40168-024-01766-4.

Easterly, Caleb W., Ray Sajulga, Subina Mehta, James Johnson, Praveen Kumar, Shane Hubler, Bart Mesuere, Joel Rudney, Timothy J. Griffin, and Pratik D. Jagtap. 2019. 'metaQuantome: An Integrated, Quantitative Metaproteomics Approach Reveals Connections Between Taxonomy and Protein Function in Complex Microbiomes *'. *Molecular & Cellular Proteomics* 18 (8): S82–91. https://doi.org/10.1074/mcp.RA118.001240.

Elias, Joshua E., and Steven P. Gygi. 2007. 'Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry'. *Nature Methods* 4 (3): 207–14. https://doi.org/10.1038/nmeth1019.

Eng, Jimmy K., Tahmina A. Jahan, and Michael R. Hoopmann. 2013. 'Comet: An Open-Source MS/MS Sequence Database Search Tool'. *PROTEOMICS* 13 (1): 22–24. https://doi.org/10.1002/pmic.201200439.

Ezzeldin, Shahd, Aya El-Wazir, Shymaa Enany, Abdelrahman Muhammad, Dina Johar, Aya Osama, Eman Ahmed, Hassan Shikshaky, and Sameh Magdeldin. 2019. 'Current Understanding of Human Metaproteome Association and Modulation'. *Journal of Proteome Research* 18 (10): 3539–54. https://doi.org/10.1021/acs.jproteome.9b00301.

Ferdous, Tahsin, Lai Jiang, Irina Dinu, Julie Groizeleau, Anita L. Kozyrskyj, Celia M. T. Greenwood, and Marie-Claire Arrieta. 2022. 'The Rise to Power of the Microbiome: Power and Sample Size Calculation for Microbiome Studies'. *Mucosal Immunology* 15 (6): 1060–70. https://doi.org/10.1038/s41385-022-00548-1.

Fernández-Costa, Carolina, Salvador Martínez-Bartolomé, Daniel B. McClatchy, Anthony J. Saviola, Nam-Kyung Yu, and John R. III Yates. 2020. 'Impact of the Identification Strategy on the Reproducibility of the DDA and DIA Results'. *Journal of Proteome Research* 19 (8): 3153–61. https://doi.org/10.1021/acs.jproteome.0c00153.

Florens, Laurence, Michael J. Carozza, Selene K. Swanson, Marjorie Fournier, Michael K. Coleman, Jerry L. Workman, and Michael P. Washburn. 2006. 'Analyzing Chromatin Remodeling Complexes Using Shotgun Proteomics and Normalized Spectral Abundance Factors'. *Methods*, Chromatin and Transcriptional Regulation, 40 (4): 303–11. https://doi.org/10.1016/j.ymeth.2006.07.028.

Frank, Ari, and Pavel Pevzner. 2005. 'PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling'. *Analytical Chemistry* 77 (4): 964–73. https://doi.org/10.1021/ac048788h.

Fu, Qin, Christopher I. Murray, Oleg A. Karpov, and Jennifer E. Van Eyk. 2023. 'Automated Proteomic Sample Preparation: The Key Component for High Throughput and Quantitative Mass Spectrometry Analysis'. *Mass Spectrometry Reviews* 42 (2): e21750. https://doi.org/10.1002/mas.21750.

Geer, Lewis Y., Sanford P. Markey, Jeffrey A. Kowalak, Lukas Wagner, Ming Xu, Dawn M. Maynard, Xiaoyu Yang, Wenyao Shi, and Stephen H. Bryant. 2004. 'Open Mass Spectrometry Search Algorithm'. *Journal of Proteome Research* 3 (5): 958–64. https://doi.org/10.1021/pr0499491.

The Gene Ontology Consortium. 2019. 'The Gene Ontology Resource: 20 Years and Still GOing Strong'. *Nucleic Acids Research* 47 (D1): D330–38. https://doi.org/10.1093/nar/gky1055.

Gessulat, Siegfried, Tobias Schmidt, Daniel Paul Zolg, Patroklos Samaras, Karsten Schnatbaum, Johannes Zerweck, Tobias Knaute, et al. 2019. 'Prosit: Proteome-Wide Prediction of Peptide Tandem Mass Spectra by Deep Learning'. *Nature Methods* 16 (6): 509–18. https://doi.org/10.1038/s41592-019-0426-7.

Giagnoni, L., F. Magherini, L. Landi, S. Taghavi, A. Modesti, L. Bini, P. Nannipieri, D. Van der lelie, and G. Renella. 2011. 'Extraction of Microbial Proteome from Soil: Potential and Limitations Assessed through a Model Study'. *European Journal of Soil Science* 62 (1): 74–81. https://doi.org/10.1111/j.1365-2389.2010.01322.x.

Gómez-Varela, David, Feng Xian, Sabrina Grundtner, Julia Regina Sondermann,

Giacomo Carta, and Manuela Schmidt. 2023. ‘Increasing Taxonomic and Functional Characterization of Host-Microbiome Interactions by DIA-PASEF Metaproteomics’. *Frontiers in Microbiology* 14 (October). https://doi.org/10.3389/fmicb.2023.1258703.

Gonzalez, Carlos G., Hannah C. Wastyk, Madeline Topf, Christopher D. Gardner, Justin L. Sonnenburg, and Joshua E. Elias. 2020. ‘High-Throughput Stool Metaproteomics: Method and Application to Human Specimens’. *mSystems* 5 (3): 10.1128/msystems.00200-20. https://doi.org/10.1128/msystems.00200-20.

Graf, Alexander C., Johanna Striesow, Jan Pané-Farré, Thomas Sura, Martina Wurster, Michael Lalk, Dietmar H. Pieper, Dörte Becher, Barbara C. Kahl, and Katharina Riedel. 2021. ‘An Innovative Protocol for Metaproteomic Analyses of Microbial Pathogens in Cystic Fibrosis Sputum’. *Frontiers in Cellular and Infection Microbiology* 11 (August). https://doi.org/10.3389/fcimb.2021.724569.

Grenga, Lucia, Olivier Pible, Guylaine Miotello, Karen Culotta, Sylvie Ruat, Marie-Anne Roncato, Fabienne Gas, et al. 2022. ‘Taxonomical and Functional Changes in COVID-19 Faecal Microbiome Could Be Related to SARS-CoV-2 Faecal Load’. *Environmental Microbiology* 24 (9): 4299–4316. https://doi.org/10.1111/1462-2920.16028.

Gu, Kongxin, Haruka Kumabe, Takumi Yamamoto, Naoto Tashiro, Takeshi Masuda, Shingo Ito, and Sumio Ohtsuki. 2024. ‘Improving Proteomic Identification Using Narrow Isolation Windows with Zeno SWATH Data-Independent Acquisition’. *Journal of Proteome Research* 23 (8): 3484–95. https://doi.org/10.1021/acs.jproteome.4c00149.

Guo, Xuan, Zhou Li, Qiuming Yao, Ryan S Mueller, Jimmy K Eng, David L Tabb, William Judson Hervey IV, and Chongle Pan. 2018. ‘Sipros Ensemble Improves Database Searching and Filtering for Complex Metaproteomics’. *Bioinformatics* 34 (5): 795–802. https://doi.org/10.1093/bioinformatics/btx601.

Gurbich, Tatiana A., Alexandre Almeida, Martin Beracochea, Tony Burdett, Josephine Burgin, Guy Cochrane, Shriya Raj, et al. 2023. ‘MGnify Genomes: A Resource for Biome-Specific Microbial Genome Catalogues’. *Journal of Molecular Biology* 435 (14): 168016. https://doi.org/10.1016/j.jmb.2023.168016.

Haange, Sven-Bastiaan, Nico Jehmlich, Maximilian Hoffmann, Klaus Weber, Jörg Lehmann, Martin von Bergen, and Ulla Slanina. 2019. ‘Disease Development Is Accompanied by Changes in Bacterial Protein Abundance and Functions in a Refined Model of Dextran Sulfate Sodium (DSS)-Induced Colitis’. *Journal of Proteome Research* 18 (4): 1774–86. https://doi.org/10.1021/acs.jproteome.8b00974.

Habeck, Tanja, Kyle A. Brown, Benjamin Des Soye, Carter Lantz, Mowei Zhou, Novera Alam, Md Amin Hossain, et al. 2024. ‘Top-down Mass Spectrometry of Native Proteoforms and Their Complexes: A Community Study’. *Nature Methods*, May, 1–9. https://doi.org/10.1038/s41592-024-02279-6.

Han, Lin, and Pamela V. Chang. 2023. ‘Activity-Based Protein Profiling in Microbes and the Gut Microbiome’. *Current Opinion in Chemical Biology* 76 (October):102351. https://doi.org/10.1016/j.cbpa.2023.102351.

Hansmeier, N., S Sharma, and TC Chao. 2022. ‘Protein Purification and Digestion Methods for Bacterial Proteomic Analyses.’ In *Geddes-McAlister, J (Eds). Proteomics in Systems Biology. Methods in Molecular Biology.* Vol. 2456. New York, NY: Humana.

Hao, Chunlin, Joshua E. Elias, Patrick K. H. Lee, and Henry Lam. 2023. ‘metaSpectraST: An Unsupervised and Database-Independent Analysis Workflow for Metaproteomic MS/MS Data Using Spectrum Clustering’. *Microbiome* 11 (1): 176. https://doi.org/10.1186/s40168-023-01602-1.

Heintz-Buschart, Anna, and Paul Wilmes. 2018. ‘Human Gut Microbiome: Function Matters’. *Trends in Microbiology* 26 (7): 563–74. https://doi.org/10.1016/j.tim.2017.11.002.

Henry, Céline, Ariane Bassignani, Magali Berland, Olivier Langella, Harry Sokol, and Catherine Juste. 2022. 'Modern Metaproteomics: A Unique Tool to Characterize the Active Microbiome in Health and Diseases, and Pave the Road towards New Biomarkers—Example of Crohn's Disease and Ulcerative Colitis Flare-Ups'. *Cells* 11 (8): 1340. https://doi.org/10.3390/cells11081340.

Herbst, Florian-Alexander, Vanessa Lünsmann, Henrik Kjeldal, Nico Jehmlich, Andreas Tholey, Martin von Bergen, Jeppe Lund Nielsen, Robert L. Hettich, Jana Seifert, and Per Halkjær Nielsen. 2016. 'Enhancing Metaproteomics—The Value of Models and Defined Environmental Microbial Systems'. *PROTEOMICS* 16 (5): 783–98. https://doi.org/10.1002/pmic.201500305.

Hernández-Plaza, Ana, Damian Szklarczyk, Jorge Botas, Carlos P Cantalapiedra, Joaquín Giner-Lamia, Daniel R Mende, Rebecca Kirsch, et al. 2023. 'eggNOG 6.0: Enabling Comparative Genomics across 12 535 Organisms'. *Nucleic Acids Research* 51 (D1): D389–94. https://doi.org/10.1093/nar/gkac1022.

Hettich, Robert L., Chongle Pan, Karuna Chourey, and Richard J. Giannone. 2013. 'Metaproteomics: Harnessing the Power of High Performance Mass Spectrometry to Identify the Suite of Proteins That Control Metabolic Activities in Microbial Communities'. *Analytical Chemistry* 85 (9): 4203–14. https://doi.org/10.1021/ac303053e.

Heyer, Robert, Patrick Hellwig, Irena Maus, Daniel Walke, Andreas Schlüter, Julia Hassa, Alexander Sczyrba, et al. 2024. 'Breakdown of Hardly Degradable Carbohydrates (Lignocellulose) in a Two-Stage Anaerobic Digestion Plant Is Favored in the Main Fermenter'. *Water Research* 250 (February):121020. https://doi.org/10.1016/j.watres.2023.121020.

Heyer, Robert, Kay Schallert, Anja Büdel, Roman Zoun, Sebastian Dorl, Alexander Behne, Fabian Kohrs, et al. 2019. 'A Robust and Universal Metaproteomics Workflow for Research Studies and Routine Diagnostics Within 24 h Using Phenol Extraction, FASP Digest, and the MetaProteomeAnalyzer'. *Frontiers in Microbiology* 10 (August). https://doi.org/10.3389/fmicb.2019.01883.

Hiltemann, Saskia, Helena Rasche, Simon Gladman, Hans-Rudolf Hotz, Delphine Larivière, Daniel Blankenberg, Pratik D. Jagtap, et al. 2023. 'Galaxy Training: A Powerful Framework for Teaching!' *PLOS Computational Biology* 19 (1): e1010752. https://doi.org/10.1371/journal.pcbi.1010752.

Hinzke, Tjorven, Manuel Kleiner, and Stephanie Markert. 2018. 'Centrifugation-Based Enrichment of Bacterial Cell Populations for Metaproteomic Studies on Bacteria–Invertebrate Symbioses'. In *Microbial Proteomics: Methods and Protocols*, edited by Dörte Becher, 319–34. New York, NY: Springer. https://doi.org/10.1007/978-1-4939-8695-8_22.

Hinzke, Tjorven, Manuel Kleiner, Mareike Meister, Rabea Schlüter, Christian Hentschker, Jan Pané-Farré, Petra Hildebrandt, et al. 2021. 'Bacterial Symbiont Subpopulations Have Different Roles in a Deep-Sea Symbiosis'. Edited by Gisela Storz and Victoria Orphan. *eLife* 10 (January):e58371. https://doi.org/10.7554/eLife.58371.

Hinzke, Tjorven, Angela Kouris, Rebecca-Ayme Hughes, Marc Strous, and Manuel Kleiner. 2019. 'More Is Not Always Better: Evaluation of 1D and 2D-LC-MS/MS Methods for Metaproteomics'. *Frontiers in Microbiology* 10 (February). https://doi.org/10.3389/fmicb.2019.00238.

Holstein, Tanja, Pieter Verschaffelt, Tim Van den Bossche, Lennart Martens, and Thilo Muth. 2024. 'The Peptonizer2000: Graphical Model Based Taxonomic Identifications of Metaproteomic Samples'. bioRxiv. https://doi.org/10.1101/2024.05.20.594958.

Hughes, Christopher S, Sophia Foehr, David A Garfield, Eileen E Furlong, Lars M Steinmetz, and Jeroen Krijgsveld. 2014. 'Ultrasensitive Proteome Analysis Using Paramagnetic Bead Technology'. *Molecular Systems Biology* 10 (10): 757. https://doi.org/10.15252/msb.20145625.

Hustoft, Hanne, Helle Malerod, Steven Wilson, Léon Reubsaet, Elsa Lundanes, and Tyge Greibrokk. 2012. 'A Critical Review of Trypsin Digestion for LC-MS Based Proteomics'. In . https://doi.org/10.13140/2.1.2226.7846.

Ishikawa, Masaki, Ryo Konno, Daisuke Nakajima, Mari Gotoh, Keiko Fukasawa, Hironori Sato, Ren Nakamura, Osamu Ohara, and Yusuke Kawashima. 2022. 'Optimization of Ultrafast Proteomics Using an LC-Quadrupole-Orbitrap Mass Spectrometer with Data-Independent Acquisition'. *Journal of Proteome Research* 21 (9): 2085–93. https://doi.org/10.1021/acs.jproteome.2c00121.

Jabbar, Karolina S., Brendan Dolan, Lisbeth Eklund, Catharina Wising, Anna Ermund, Åsa Johansson, Hans Törnblom, Magnus Simren, and Gunnar C. Hansson. 2021. 'Association between Brachyspira and Irritable Bowel Syndrome with Diarrhoea'. *Gut* 70 (6): 1117–29. https://doi.org/10.1136/gutjnl-2020-321466.

Jagtap, Pratik D., Alan Blakely, Kevin Murray, Shaun Stewart, Joel Kooren, James E. Johnson, Nelson L. Rhodus, Joel Rudney, and Timothy J. Griffin. 2015. 'Metaproteomic Analysis Using the Galaxy Framework'. *PROTEOMICS* 15 (20): 3553–65. https://doi.org/10.1002/pmic.201500074.

Jagtap, Pratik, Jill Goslinga, Joel A. Kooren, Thomas McGowan, Matthew S. Wroblewski, Sean L. Seymour, and Timothy J. Griffin. 2013. 'A Two-Step Database Search Method Improves Sensitivity in Peptide Sequence Matches for Metaproteomics and Proteogenomics Studies'. *PROTEOMICS* 13 (8): 1352–57. https://doi.org/10.1002/pmic.201200352.

James, P., M. Quadroni, E. Carafoli, and G. Gonnet. 1993. 'Protein Identification by Mass Profile Fingerprinting'. *Biochemical and Biophysical Research Communications* 195 (1): 58–64. https://doi.org/10.1006/bbrc.1993.2009.

Jarman, Kristin H., Natalie C. Heller, Sarah C. Jenson, Janine R. Hutchison, Brooke L. Deatherage Kaiser, Samuel H. Payne, David S. Wunschel, and Eric D. Merkley. 2018. 'Proteomics Goes to Court: A Statistical Foundation for Forensic Toxin/Organism Identification Using Bottom-Up Proteomics'. *Journal of Proteome Research* 17 (9): 3075–85. https://doi.org/10.1021/acs.jproteome.8b00212.

Jensen, Marlene, Juliane Wippler, and Manuel Kleiner. 2021. 'Evaluation of RNAlater as a Field-Compatible Preservation Method for Metaproteomic Analyses of Bacterium-Animal Symbioses'. *Microbiology Spectrum* 9 (2): e01429-21. https://doi.org/10.1128/Spectrum.01429-21.

Jersie-Christensen, Rosa R., Liam T. Lanigan, David Lyon, Meaghan Mackie, Daniel Belstrøm, Christian D. Kelstrup, Anna K. Fotakis, et al. 2018. 'Quantitative Metaproteomics of Medieval Dental Calculus Reveals Individual Oral Health Status'. *Nature Communications* 9 (1): 4744. https://doi.org/10.1038/s41467-018-07148-3.

Jiang, Yuming, Devasahayam Arokia Balaya Rex, Dina Schuster, Benjamin A. Neely, Germán L. Rosano, Norbert Volkmar, Amanda Momenzadeh, et al. 2024. 'Comprehensive Overview of Bottom-Up Proteomics Using Mass Spectrometry'. *ACS Measurement Science Au* 4 (4): 338–417. https://doi.org/10.1021/acsmeasuresciau.3c00068.

Johnson, Jethro S., Daniel J. Spakowicz, Bo-Young Hong, Lauren M. Petersen, Patrick Demkowicz, Lei Chen, Shana R. Leopold, et al. 2019. 'Evaluation of 16S rRNA Gene Sequencing for Species and Strain-Level Microbiome Analysis'. *Nature Communications* 10 (1): 5029. https://doi.org/10.1038/s41467-019-13036-1.

Johnson, Richard S., Brian C. Searle, Brook L. Nunn, Jason M. Gilmore, Molly Phillips, Chris T. Amemiya, Michelle Heck, and Michael J. MacCoss. 2020. 'Assessing Protein Sequence Database Suitability Using De Novo Sequencing *'. *Molecular & Cellular Proteomics* 19 (1): 198–208. https://doi.org/10.1074/mcp.TIR119.001752.

Justice, Nicholas B., Zhou Li, Yingfeng Wang, Susan E. Spaudling, Annika C. Mosier, Robert L. Hettich, Chongle Pan, and Jillian F. Banfield. 2014. '(15)N- and (2)H Proteomic Stable Isotope Probing Links Nitrogen Flow to Archaeal Heterotrophic Activity'. *Environmental Microbiology* 16 (10): 3224–37.

https://doi.org/10.1111/1462-2920.12488.
Käll, Lukas, Jesse D. Canterbury, Jason Weston, William Stafford Noble, and Michael J. MacCoss. 2007. 'Semi-Supervised Learning for Peptide Identification from Shotgun Proteomics Datasets'. *Nature Methods* 4 (11): 923–25. https://doi.org/10.1038/nmeth1113.
Käll, Lukas, John D. Storey, Michael J. MacCoss, and William Stafford Noble. 2008. 'Posterior Error Probabilities and False Discovery Rates: Two Sides of the Same Coin'. *Journal of Proteome Research* 7 (1): 40–44. https://doi.org/10.1021/pr700739d.
Kanehisa, Minoru, and Susumu Goto. 2000. 'KEGG: Kyoto Encyclopedia of Genes and Genomes'. *Nucleic Acids Research* 28 (1): 27–30. https://doi.org/10.1093/nar/28.1.27.
Keiblinger, Katharina M., Stephan Fuchs, Sophie Zechmeister-Boltenstern, and Katharina Riedel. 2016. 'Soil and Leaf Litter Metaproteomics—a Brief Guideline from Sampling to Understanding'. *FEMS Microbiology Ecology* 92 (11): fiw180. https://doi.org/10.1093/femsec/fiw180.
Keiblinger, Katharina M., Inés C. Wilhartitz, Thomas Schneider, Bernd Roschitzki, Emanuel Schmid, Leo Eberl, Kathrin Riedel, and Sophie Zechmeister-Boltenstern. 2012. 'Soil Metaproteomics – Comparative Evaluation of Protein Extraction Protocols'. *Soil Biology and Biochemistry* 54 (November):14–24. https://doi.org/10.1016/j.soilbio.2012.05.014.
Kieser, Silas, Evgeny M. Zdobnov, and Mirko Trajkovski. 2022. 'Comprehensive Mouse Microbiota Genome Catalog Reveals Major Difference to Its Human Counterpart'. *PLoS Computational Biology* 18 (3): e1009947. https://doi.org/10.1371/journal.pcbi.1009947.
Kim, Daehwan, Li Song, Florian P. Breitwieser, and Steven L. Salzberg. 2016. 'Centrifuge: Rapid and Sensitive Classification of Metagenomic Sequences'. *Genome Research* 26 (12): 1721–29. https://doi.org/10.1101/gr.210641.116.
Kleikamp, Hugo B. C., Denis Grouzdev, Pim Schaasberg, Ramon van Valderen, Ramon van der Zwaan, Roel van de Wijgaart, Yuemei Lin, et al. 2023. 'Metaproteomics, Metagenomics and 16S rRNA Sequencing Provide Different Perspectives on the Aerobic Granular Sludge Microbiome'. *Water Research* 246 (November):120700. https://doi.org/10.1016/j.watres.2023.120700.
Kleikamp, Hugo B. C., Mario Pronk, Claudia Tugui, Leonor Guedes da Silva, Ben Abbas, Yue Mei Lin, Mark C. M. van Loosdrecht, and Martin Pabst. 2021. 'Database-Independent de Novo Metaproteomics of Complex Microbial Communities'. *Cell Systems* 12 (5): 375-383.e5. https://doi.org/10.1016/j.cels.2021.04.003.
Kleiner, Manuel. 2019. 'Metaproteomics: Much More than Measuring Gene Expression in Microbial Communities'. *mSystems* 4 (3): 10.1128/msystems.00115-19. https://doi.org/10.1128/msystems.00115-19.
Kleiner, Manuel, Angela Kouris, Marlene Violette, Grace D'Angelo, Yihua Liu, Abigail Korenek, Nikola Tolić, et al. 2023. 'Ultra-Sensitive Isotope Probing to Quantify Activity and Substrate Assimilation in Microbiomes'. *Microbiome* 11 (1): 24. https://doi.org/10.1186/s40168-022-01454-1.
Kleiner, Manuel, Erin Thorson, Christine E. Sharp, Xiaoli Dong, Dan Liu, Carmen Li, and Marc Strous. 2017. 'Assessing Species Biomass Contributions in Microbial Communities via Metaproteomics'. *Nature Communications* 8 (1): 1558. https://doi.org/10.1038/s41467-017-01544-x.
Kleiner, Manuel, Cecilia Wentrup, Christian Lott, Hanno Teeling, Silke Wetzel, Jacque Young, Yun-Juan Chang, et al. 2012. 'Metaproteomics of a Gutless Marine Worm and Its Symbiotic Microbial Community Reveal Unusual Pathways for Carbon and Energy Use'. *Proceedings of the National Academy of Sciences* 109 (19). https://doi.org/10.1073/pnas.1121198109.
Kollipara, Laxmikanth, and René P. Zahedi. 2013. 'Protein Carbamylation: In Vivo Modification or in Vitro Artefact?' *PROTEOMICS* 13 (6): 941–44.

https://doi.org/10.1002/pmic.201200452.
Kong, Andy T., Felipe V. Leprevost, Dmitry M. Avtonomov, Dattatreya Mellacheruvu, and Alexey I. Nesvizhskii. 2017. 'MSFragger: Ultrafast and Comprehensive Peptide Identification in Mass Spectrometry–Based Proteomics'. *Nature Methods* 14 (5): 513–20. https://doi.org/10.1038/nmeth.4256.
Kong, Ling-Fen, Yan-Bin He, Zhang-Xian Xie, Xing Luo, Hao Zhang, Sheng-Hui Yi, Zhi-Long Lin, et al. 2021. 'Illuminating Key Microbial Players and Metabolic Processes Involved in the Remineralization of Particulate Organic Carbon in the Ocean's Twilight Zone by Metaproteomics'. *Applied and Environmental Microbiology* 87 (20): e00986-21. https://doi.org/10.1128/AEM.00986-21.
Kulak, Nils A., Garwin Pichler, Igor Paron, Nagarjuna Nagaraj, and Matthias Mann. 2014. 'Minimal, Encapsulated Proteomic-Sample Processing Applied to Copy-Number Estimation in Eukaryotic Cells'. *Nature Methods* 11 (3): 319–24. https://doi.org/10.1038/nmeth.2834.
Kunath, B. J., O. Hickl, P. Queirós, C. Martin-Gallausiaux, L. A. Lebrun, R. Halder, C. C. Laczny, et al. 2022. 'Alterations of Oral Microbiota and Impact on the Gut Microbiome in Type 1 Diabetes Mellitus Revealed by Integrated Multi-Omic Analyses'. *Microbiome* 10 (1): 243. https://doi.org/10.1186/s40168-022-01435-4.
Kunath, Benoit J., Andreas Bremges, Aaron Weimann, Alice C. McHardy, and Phillip B. Pope. 2017. 'Metagenomics and CAZyme Discovery'. *Methods in Molecular Biology (Clifton, N.J.)* 1588:255–77. https://doi.org/10.1007/978-1-4939-6899-2_20.
Kunath, Benoit J., Giusi Minniti, Morten Skaugen, Live H. Hagen, Gustav Vaaje-Kolstad, Vincent G. H. Eijsink, Phillip B. Pope, and Magnus Ø Arntzen. 2019. 'Metaproteomics: Sample Preparation and Methodological Considerations.' In *Emerging Sample Treatments in Proteomics (Ed. Capelo-Martínez, J.-L.)*, 187–215. Cham: Springer International Publishing. https://dokumen.pub/emerging-sample-treatments-in-proteomics-1st-ed-978-3-030-12297-3978-3-030-12298-0.html.
Lautenbacher, Ludwig, Kevin L. Yang, Tobias Kockmann, Christian Panse, Matthew Chambers, Elias Kahl, Fengchao Yu, et al. 2024. 'Koina: Democratizing Machine Learning for Proteomics Research'. bioRxiv. https://doi.org/10.1101/2024.06.01.596953.
Lazear, Michael R. 2023. 'Sage: An Open-Source Tool for Fast Proteomics Searching and Quantification at Scale'. *Journal of Proteome Research* 22 (11): 3652–59. https://doi.org/10.1021/acs.jproteome.3c00486.
Lenčo, Juraj, Siddharth Jadeja, Denis K. Naplekov, Oleg V. Krokhin, Maria A. Khalikova, Petr Chocholouš, Jiří Urban, Ken Broeckhoven, Lucie Nováková, and František Švec. 2022. 'Reversed-Phase Liquid Chromatography of Peptides for Bottom-Up Proteomics: A Tutorial'. *Journal of Proteome Research* 21 (12): 2846–92. https://doi.org/10.1021/acs.jproteome.2c00407.
Lesker, Till R., Abilash C. Durairaj, Eric J. C. Gálvez, Ilias Lagkouvardos, John F. Baines, Thomas Clavel, Alexander Sczyrba, Alice C. McHardy, and Till Strowig. 2020. 'An Integrated Metagenome Catalog Reveals New Insights into the Murine Gut Microbiome'. *Cell Reports* 30 (9): 2909-2922.e6. https://doi.org/10.1016/j.celrep.2020.02.036.
Letunic, Ivica, Takuji Yamada, Minoru Kanehisa, and Peer Bork. 2008. 'iPath: Interactive Exploration of Biochemical Pathways and Networks'. *Trends in Biochemical Sciences* 33 (3): 101–3. https://doi.org/10.1016/j.tibs.2008.01.001.
Levin, Yishai. 2011. 'The Role of Statistical Power Analysis in Quantitative Proteomics'. *PROTEOMICS* 11 (12): 2565–67. https://doi.org/10.1002/pmic.201100033.
Lezcano, María Ángeles, Laura Sánchez-García, Antonio Quesada, Daniel Carrizo, Miguel Ángel Fernández-Martínez, Erika Cavalcante-Silva, and Víctor Parro. 2022. 'Comprehensive Metabolic and Taxonomic Reconstruction of an Ancient Microbial Mat From the McMurdo Ice Shelf (Antarctica) by Integrating Genetic,

Metaproteomic and Lipid Biomarker Analyses'. *Frontiers in Microbiology* 13 (April). https://doi.org/10.3389/fmicb.2022.799360.
Li, Junhua, Huijue Jia, Xianghang Cai, Huanzi Zhong, Qiang Feng, Shinichi Sunagawa, Manimozhiyan Arumugam, et al. 2014. 'An Integrated Catalog of Reference Genes in the Human Gut Microbiome'. *Nature Biotechnology* 32 (8): 834–41. https://doi.org/10.1038/nbt.2942.
Li, Leyuan, Janice Mayne, Adrian Beltran, Xu Zhang, Zhibin Ning, and Daniel Figeys. 2024. 'RapidAIM 2.0: A High-Throughput Assay to Study Functional Response of Human Gut Microbiome to Xenobiotics'. *Microbiome Research Reports* 3 (2): N/A-N/A. https://doi.org/10.20517/mrr.2023.57.
Li, Leyuan, Zhibin Ning, Kai Cheng, Xu Zhang, Caitlin M. A. Simopoulos, and Daniel Figeys. 2022. 'iMetaLab Suite: A One-Stop Toolset for Metaproteomics'. *iMeta* 1 (2): e25. https://doi.org/10.1002/imt2.25.
Li, Leyuan, Zhibin Ning, Xu Zhang, Janice Mayne, Kai Cheng, Alain Stintzi, and Daniel Figeys. 2020. 'RapidAIM: A Culture- and Metaproteomics-Based Rapid Assay of Individual Microbiome Responses to Drugs'. *Microbiome* 8 (1): 33. https://doi.org/10.1186/s40168-020-00806-z.
Li, Leyuan, Tong Wang, Zhibin Ning, Xu Zhang, James Butcher, Joeselle M. Serrana, Caitlin M. A. Simopoulos, et al. 2023. 'Revealing Proteome-Level Functional Redundancy in the Human Gut Microbiome Using Ultra-Deep Metaproteomics'. *Nature Communications* 14 (1): 3428. https://doi.org/10.1038/s41467-023-39149-2.
Li, W., L. Jaroszewski, and A. Godzik. 2001. 'Clustering of Highly Homologous Sequences to Reduce the Size of Large Protein Databases'. *Bioinformatics (Oxford, England)* 17 (3): 282–83. https://doi.org/10.1093/bioinformatics/17.3.282.
Li, Zhou, Yingfeng Wang, Qiuming Yao, Nicholas B. Justice, Tae-Hyuk Ahn, Dong Xu, Robert L. Hettich, Jillian F. Banfield, and Chongle Pan. 2014. 'Diverse and Divergent Protein Post-Translational Modifications in Two Growth Stages of a Natural Microbial Community'. *Nature Communications* 5 (1): 4405. https://doi.org/10.1038/ncomms5405.
Li, Zhou, Qiuming Yao, Xuan Guo, Alexander Crits-Christoph, Melanie A. Mayes, William Judson Hervey Iv, Sarah L. Lebeis, et al. 2019. 'Genome-Resolved Proteomic Stable Isotope Probing of Soil Microbial Communities Using 13CO2 and 13C-Methanol'. *Frontiers in Microbiology* 10 (December):2706. https://doi.org/10.3389/fmicb.2019.02706.
Liao, Bo, Zhibin Ning, Kai Cheng, Xu Zhang, Leyuan Li, Janice Mayne, and Daniel Figeys. 2018. 'iMetaLab 1.0: A Web Platform for Metaproteomics Data Analysis'. *Bioinformatics* 34 (22): 3954–56. https://doi.org/10.1093/bioinformatics/bty466.
Lin, Zongtao, Joanna Gongora, Xingyu Liu, Yixuan Xie, Chenfeng Zhao, Dongwen Lv, and Benjamin A. Garcia. 2023. 'Automation to Enable High-Throughput Chemical Proteomics'. *Journal of Proteome Research* 22 (12): 3676–82. https://doi.org/10.1021/acs.jproteome.3c00467.
Liu, Lei, Yu Yang, Yu Deng, and Tong Zhang. 2022. 'Nanopore Long-Read-Only Metagenomics Enables Complete and High-Quality Genome Reconstruction from Mock and Complex Metagenomes'. *Microbiome* 10 (1): 209. https://doi.org/10.1186/s40168-022-01415-8.
Long, Shuping, Yi Yang, Chengpin Shen, Yiwen Wang, Anmei Deng, Qin Qin, and Liang Qiao. 2020. 'Metaproteomics Characterizes Human Gut Microbiome Function in Colorectal Cancer'. *Npj Biofilms and Microbiomes* 6 (1): 1–10. https://doi.org/10.1038/s41522-020-0123-4.
Low, Teck Yew, M. Aiman Mohtar, Pey Yee Lee, Nursyazwani Omar, Houjiang Zhou, and Mingliang Ye. 2021. 'Widening the Bottleneck of Phosphoproteomics: Evolving Strategies for Phosphopeptide Enrichment'. *Mass Spectrometry Reviews* 40 (4): 309–33. https://doi.org/10.1002/mas.21636.
Martens, Lennart, Matthew Chambers, Marc Sturm, Darren Kessner, Fredrik Levander,

Jim Shofstahl, Wilfred H. Tang, et al. 2011. ‘mzML—a Community Standard for Mass Spectrometry Data *’. *Molecular & Cellular Proteomics* 10 (1). https://doi.org/10.1074/mcp.R110.000133.

Matthiesen, Rune, and Jakob Bunkenborg. 2013. ‘Introduction to Mass Spectrometry-Based Proteomics’. In *Mass Spectrometry Data Analysis in Proteomics*, edited by Rune Matthiesen, 1–45. Totowa, NJ: Humana Press. https://doi.org/10.1007/978-1-62703-392-3_1.

Meijenfeldt, F. A. Bastiaan von, Ksenia Arkhipova, Diego D. Cambuy, Felipe H. Coutinho, and Bas E. Dutilh. 2019. ‘Robust Taxonomic Classification of Uncharted Microbial Sequences and Bins with CAT and BAT’. *Genome Biology* 20 (1): 217. https://doi.org/10.1186/s13059-019-1817-x.

Mesuere, Bart, Felix Van der Jeugt, Toon Willems, Tom Naessens, Bart Devreese, Lennart Martens, and Peter Dawyndt. 2018. ‘High-Throughput Metaproteomics Data Analysis with Unipept: A Tutorial’. *Journal of Proteomics*, Tutorials in Bioinformatics for Biological Science, 171 (January):11–22. https://doi.org/10.1016/j.jprot.2017.05.022.

Meyer, Susann, Nicole Hüttig, Marianne Zenk, Udo Jäckel, and Dierk-Christoph Pöther. 2023. ‘Bioaerosols in Swine Confinement Buildings: A Metaproteomic View’. *Environmental Microbiology Reports* 15 (6): 684–97. https://doi.org/10.1111/1758-2229.13208.

Millikin, Robert J., Stefan K. Solntsev, Michael R. Shortreed, and Lloyd M. Smith. 2018. ‘Ultrafast Peptide Label-Free Quantification with FlashLFQ’. *Journal of Proteome Research* 17 (1): 386–91. https://doi.org/10.1021/acs.jproteome.7b00608.

Minniti, Giusi, Simen Rød Sandve, János Tamás Padra, Live Heldal Hagen, Sara Lindén, Phillip B. Pope, Magnus Ø Arntzen, and Gustav Vaaje-Kolstad. 2019. ‘The Farmed Atlantic Salmon (Salmo Salar) Skin-Mucus Proteome and Its Nutrient Potential for the Resident Bacterial Community’. *Genes* 10 (7): 515. https://doi.org/10.3390/genes10070515.

Mistry, Jaina, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik L L Sonnhammer, Silvio C E Tosatto, et al. 2021. ‘Pfam: The Protein Families Database in 2021’. *Nucleic Acids Research* 49 (D1): D412–19. https://doi.org/10.1093/nar/gkaa913.

Mordant, Angie, and Manuel Kleiner. 2021. ‘Evaluation of Sample Preservation and Storage Methods for Metaproteomics Analysis of Intestinal Microbiomes’. *Microbiology Spectrum* 9 (3): e01877-21. https://doi.org/10.1128/Spectrum.01877-21.

Morris, Laura S., and Julian R. Marchesi. 2016. ‘Assessing the Impact of Long Term Frozen Storage of Faecal Samples on Protein Concentration and Protease Activity’. *Journal of Microbiological Methods* 123 (April):31–38. https://doi.org/10.1016/j.mimet.2016.02.001.

Mueller, Ryan S, Vincent J Denef, Linda H Kalnejais, K Blake Suttle, Brian C Thomas, Paul Wilmes, Richard L Smith, et al. 2010. ‘Ecological Distribution and Population Physiology Defined by Proteomics in a Natural Microbial Community’. *Molecular Systems Biology* 6 (1): 374. https://doi.org/10.1038/msb.2010.30.

Muth, Thilo, Alexander Behne, Robert Heyer, Fabian Kohrs, Dirk Benndorf, Marcus Hoffmann, Miro Lehtevä, Udo Reichl, Lennart Martens, and Erdmann Rapp. 2015. ‘The MetaProteomeAnalyzer: A Powerful Open-Source Software Suite for Metaproteomics Data Analysis and Interpretation’. *Journal of Proteome Research* 14 (3): 1557–65. https://doi.org/10.1021/pr501246w.

Muth, Thilo, Carolin A. Kolmeder, Jarkko Salojärvi, Salla Keskitalo, Markku Varjosalo, Froukje J. Verdam, Sander S. Rensen, et al. 2015. ‘Navigating through Metaproteomics Data: A Logbook of Database Searching’. *PROTEOMICS* 15 (20): 3439–53. https://doi.org/10.1002/pmic.201400560.

Mysling, Simon, Giuseppe Palmisano, Peter Højrup, and Morten Thaysen-Andersen. 2010. ‘Utilizing Ion-Pairing Hydrophilic Interaction Chromatography Solid Phase

Extraction for Efficient Glycopeptide Enrichment in Glycoproteomics'. *Analytical Chemistry* 82 (13): 5598–5609. https://doi.org/10.1021/ac100530w.

Nakayasu, Ernesto S., Marina Gritsenko, Paul D. Piehowski, Yuqian Gao, Daniel J. Orton, Athena A. Schepmoes, Thomas L. Fillmore, et al. 2021. 'Tutorial: Best Practices and Considerations for Mass-Spectrometry-Based Protein Biomarker Discovery and Validation'. *Nature Protocols* 16 (8): 3737–60. https://doi.org/10.1038/s41596-021-00566-6.

Narayanasamy, Shaman, Yohan Jarosz, Emilie E. L. Muller, Anna Heintz-Buschart, Malte Herold, Anne Kaysen, Cédric C. Laczny, Nicolás Pinel, Patrick May, and Paul Wilmes. 2016. 'IMP: A Pipeline for Reproducible Reference-Independent Integrated Metagenomic and Metatranscriptomic Analyses'. *Genome Biology* 17 (1): 260. https://doi.org/10.1186/s13059-016-1116-8.

Nebauer, Daniel J., Leanne A. Pearson, and Brett A. Neilan. 2024. 'Critical Steps in an Environmental Metaproteomics Workflow'. *Environmental Microbiology* 26 (5): e16637. https://doi.org/10.1111/1462-2920.16637.

Nesvizhskii, Alexey I., and Ruedi Aebersold. 2005. 'Interpretation of Shotgun Proteomic Data'. *Molecular & Cellular Proteomics* 4 (10): 1419–40. https://doi.org/10.1074/mcp.R500012-MCP200.

Nickerson, Jessica L., and Alan A. Doucette. 2020. 'Rapid and Quantitative Protein Precipitation for Proteome Analysis by Mass Spectrometry'. *Journal of Proteome Research* 19 (5): 2035–42. https://doi.org/10.1021/acs.jproteome.9b00867.

Niu, Liangjie, Hang Zhang, Zhaokun Wu, Yibo Wang, Hui Liu, Xiaolin Wu, and Wei Wang. 2018. 'Modified TCA/Acetone Precipitation of Plant Proteins for Proteomic Analysis'. *PLOS ONE* 13 (12): e0202238. https://doi.org/10.1371/journal.pone.0202238.

Nowatzky, Yannek, Philipp Benner, Knut Reinert, and Thilo Muth. 2023. 'Mistle: Bringing Spectral Library Predictions to Metaproteomics with an Efficient Search Index'. *Bioinformatics* 39 (6): btad376. https://doi.org/10.1093/bioinformatics/btad376.

Oberg, Ann L., and Olga Vitek. 2009. 'Statistical Design of Quantitative Mass Spectrometry-Based Proteomic Experiments'. *Journal of Proteome Research* 8 (5): 2144–56. https://doi.org/10.1021/pr8010099.

Odom, Aubrey R., Tyler Faits, Eduardo Castro-Nallar, Keith A. Crandall, and W. Evan Johnson. 2023. 'Metagenomic Profiling Pipelines Improve Taxonomic Classification for 16S Amplicon Sequencing Data'. *Scientific Reports* 13 (1): 13957. https://doi.org/10.1038/s41598-023-40799-x.

Okeke, Emmanuel Sunday, Richard Ekeng Ita, Egong John Egong, Lydia Etuk Udofia, Chiamaka Linda Mgbechidinma, and Otobong Donald Akan. 2021. 'Metaproteomics Insights into Fermented Fish and Vegetable Products and Associated Microbes'. *Food Chemistry: Molecular Sciences* 3 (December):100045. https://doi.org/10.1016/j.fochms.2021.100045.

O'Leary, Nuala A., Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, et al. 2016. 'Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation'. *Nucleic Acids Research* 44 (D1): D733-745. https://doi.org/10.1093/nar/gkv1189.

Olivella, Roger, Cristina Chiva, Marc Serret, Daniel Mancera, Luca Cozzuto, Antoni Hermoso, Eva Borràs, et al. 2021. 'QCloud2: An Improved Cloud-Based Quality-Control System for Mass-Spectrometry-Based Proteomics Laboratories'. *Journal of Proteome Research* 20 (4): 2010–13. https://doi.org/10.1021/acs.jproteome.0c00853.

Pan, Haixia, Ruddy Wattiez, and David Gillan. 2024. 'Soil Metaproteomics for Microbial Community Profiling: Methodologies and Challenges'. *Current Microbiology* 81 (8): 257. https://doi.org/10.1007/s00284-024-03781-y.

Paramasivan, Selvam, Janna L. Morrison, Mitchell C. Lock, Jack R. T. Darby, Roberto A. Barrero, Paul C. Mills, and Pawel Sadowski. 2023. 'Automated Proteomics

Workflows for High-Throughput Library Generation and Biomarker Detection Using Data-Independent Acquisition'. *Journal of Proteome Research* 22 (6): 2018–29. https://doi.org/10.1021/acs.jproteome.3c00074.
Park, Sung Kyu Robin, Titus Jung, Peter S. Thuy-Boun, Ana Y. Wang, John R. III Yates, and Dennis W. Wolan. 2019. 'ComPIL 2.0: An Updated Comprehensive Metaproteomics Database'. *Journal of Proteome Research* 18 (2): 616–22. https://doi.org/10.1021/acs.jproteome.8b00722.
Pathak, Khyatiben V., Marissa I. McGilvrey, Charles K. Hu, Krystine Garcia-Mansfield, Karen Lewandoski, Zahra Eftekhari, Yate-Ching Yuan, Frederic Zenhausern, Emmanuel Menashi, and Patrick Pirrotte. 2020. 'Molecular Profiling of Innate Immune Response Mechanisms in Ventilator-Associated Pneumonia'. *Molecular & Cellular Proteomics* 19 (10): 1688–1705. https://doi.org/10.1074/mcp.RA120.002207.
Patnode, Michael L., Zachary W. Beller, Nathan D. Han, Jiye Cheng, Samantha L. Peters, Nicolas Terrapon, Bernard Henrissat, et al. 2019. 'Interspecies Competition Impacts Targeted Manipulation of Human Gut Bacteria by Fiber-Derived Glycans'. *Cell* 179 (1): 59-73.e13. https://doi.org/10.1016/j.cell.2019.08.011.
Pavelka, Norman, Marjorie L. Fournier, Selene K. Swanson, Mattia Pelizzola, Paola Ricciardi-Castagnoli, Laurence Florens, and Michael P. Washburn. 2008. 'Statistical Similarities between Transcriptomics and Quantitative Shotgun Proteomics Data *'. *Molecular & Cellular Proteomics* 7 (4): 631–44. https://doi.org/10.1074/mcp.M700240-MCP200.
Perez-Riverol, Yasset, Chakradhar Bandla, Deepti J Kundu, Selvakumar Kamatchinathan, Jingwen Bai, Suresh Hewapathirana, Nithu Sara John, et al. 2024. 'The PRIDE Database at 20 Years: 2025 Update'. *Nucleic Acids Research*, November, gkae1011. https://doi.org/10.1093/nar/gkae1011.
Pettersen, Veronika Kuchařová, Luis Caetano Martha Antunes, Antoine Dufour, and Marie-Claire Arrieta. 2022. 'Inferring Early-Life Host and Microbiome Functions by Mass Spectrometry-Based Metaproteomics and Metabolomics'. *Computational and Structural Biotechnology Journal* 20 (January):274–86. https://doi.org/10.1016/j.csbj.2021.12.012.
Pible, Olivier, Pauline Petit, Gérard Steinmetz, Corinne Rivasseau, and Jean Armengaud. 2023. 'Taxonomical Composition and Functional Analysis of Biofilms Sampled from a Nuclear Storage Pool'. *Frontiers in Microbiology* 14 (April). https://doi.org/10.3389/fmicb.2023.1148976.
Pietilä, Sami, Tomi Suomi, and Laura L Elo. 2022. 'Introducing Untargeted Data-Independent Acquisition for Metaproteomics of Complex Microbial Samples'. *ISME Communications* 2 (1): 51. https://doi.org/10.1038/s43705-022-00137-0.
Plancade, Sandra, Magali Berland, Mélisande Blein-Nicolas, Olivier Langella, Ariane Bassignani, and Catherine Juste. 2022. 'A Combined Test for Feature Selection on Sparse Metaproteomics Data—an Alternative to Missing Value Imputation'. *PeerJ* 10 (June):e13525. https://doi.org/10.7717/peerj.13525.
Ponnudurai, Ruby, Stefan E Heiden, Lizbeth Sayavedra, Tjorven Hinzke, Manuel Kleiner, Christian Hentschker, Horst Felbeck, et al. 2020. 'Comparative Proteomics of Related Symbiotic Mussel Species Reveals High Variability of Host–Symbiont Interactions'. *The ISME Journal* 14 (2): 649–56. https://doi.org/10.1038/s41396-019-0517-6.
Porcheddu, Massimo, Marcello Abbondio, Laura De Diego, Sergio Uzzau, and Alessandro Tanca. 2023. 'Meta4P: A User-Friendly Tool to Parse Label-Free Quantitative Metaproteomic Data and Taxonomic/Functional Annotations'. *Journal of Proteome Research* 22 (6): 2109–13. https://doi.org/10.1021/acs.jproteome.2c00803.
Qian, Chen, and Robert L. Hettich. 2017. 'Optimized Extraction Method To Remove Humic Acid Interferences from Soil Samples Prior to Microbial Proteome Measurements'. *Journal of Proteome Research* 16 (7): 2537–46. https://doi.org/10.1021/acs.jproteome.7b00103.

Queirós, Pedro, Francesco Delogu, Oskar Hickl, Patrick May, and Paul Wilmes. 2021. 'Mantis: Flexible and Consensus-Driven Genome Annotation'. *GigaScience* 10 (6): giab042. https://doi.org/10.1093/gigascience/giab042.

Quevillon, E., V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, and R. Lopez. 2005. 'InterProScan: Protein Domains Identifier'. *Nucleic Acids Research* 33 (suppl_2): W116–20. https://doi.org/10.1093/nar/gki442.

Rabe, Alexander, Manuela Gesell Salazar, Stephan Michalik, Thomas Kocher, Harald Below, Uwe Völker, and Alexander Welk. 2022. 'Impact of Different Oral Treatments on the Composition of the Supragingival Plaque Microbiome'. *Journal of Oral Microbiology* 14 (1): 2138251. https://doi.org/10.1080/20002297.2022.2138251.

Rajczewski, Andrew T., J. Alfredo Blakeley-Ruiz, Annaliese Meyer, Simina Vintila, Matthew R. McIlvin, Tim Van Den Bossche, Brian C. Searle, et al. 2024. 'Data-Independent Acquisition Mass Spectrometry as a Tool for Metaproteomics: Interlaboratory Comparison Using a Model Microbiome'. bioRxiv. https://doi.org/10.1101/2024.09.18.613707.

Rawlings, Neil D, Alan J Barrett, Paul D Thomas, Xiaosong Huang, Alex Bateman, and Robert D Finn. 2018. 'The MEROPS Database of Proteolytic Enzymes, Their Substrates and Inhibitors in 2017 and a Comparison with Peptidases in the PANTHER Database'. *Nucleic Acids Research* 46 (D1): D624–32. https://doi.org/10.1093/nar/gkx1134.

Rozanova, Svitlana, Julian Uszkoreit, Karin Schork, Bettina Serschnitzki, Martin Eisenacher, Lars Tönges, Katalin Barkovits-Boeddinghaus, and Katrin Marcus. 2023. 'Quality Control—A Stepchild in Quantitative Proteomics: A Case Study for the Human CSF Proteome'. *Biomolecules* 13 (3): 491. https://doi.org/10.3390/biom13030491.

Ruan, Wenhua, Chao Sun, Qikang Gao, and Neeraj Shrivastava. 2021. 'Metaproteomics Associated with Severe Early Childhood Caries Highlights the Differences in Salivary Proteins'. *Archives of Oral Biology* 131 (November):105220. https://doi.org/10.1016/j.archoralbio.2021.105220.

Sachsenberg, Timo, Florian-Alexander Herbst, Martin Taubert, René Kermer, Nico Jehmlich, Martin von Bergen, Jana Seifert, and Oliver Kohlbacher. 2015. 'MetaProSIP: Automated Inference of Stable Isotope Incorporation Rates in Proteins for Functional Metaproteomics'. *Journal of Proteome Research* 14 (2): 619–27. https://doi.org/10.1021/pr500245w.

Saito, Mak A., Vladimir V. Bulygin, Dawn M. Moran, Craig Taylor, and Chris Scholin. 2011. 'Examination of Microbial Proteome Preservation Techniques Applicable to Autonomous Environmental Sample Collection'. *Frontiers in Microbiology* 2 (November). https://doi.org/10.3389/fmicb.2011.00215.

Sajulga, Ray, Caleb Easterly, Michael Riffle, Bart Mesuere, Thilo Muth, Subina Mehta, Praveen Kumar, et al. 2020. 'Survey of Metaproteomics Software Tools for Functional Microbiome Analysis'. *PLOS ONE* 15 (11): e0241503. https://doi.org/10.1371/journal.pone.0241503.

Salvato, Fernanda, Robert L. Hettich, and Manuel Kleiner. 2021. 'Five Key Aspects of Metaproteomics as a Tool to Understand Functional Interactions in Host-Associated Microbiomes'. *PLOS Pathogens* 17 (2): e1009245. https://doi.org/10.1371/journal.ppat.1009245.

Salvato, Fernanda, Simina Vintila, Omri M. Finkel, Jeffery L. Dangl, and Manuel Kleiner. 2022. 'Evaluation of Protein Extraction Methods for Metaproteomic Analyses of Root-Associated Microbes'. *Molecular Plant-Microbe Interactions*® 35 (11): 977–88. https://doi.org/10.1094/MPMI-05-22-0116-TA.

Sapan, Christine V., and Roger L. Lundblad. 2015. 'Review of Methods for Determination of Total Protein and Peptide Concentration in Biological Samples'. *PROTEOMICS – Clinical Applications* 9 (3–4): 268–76. https://doi.org/10.1002/prca.201400088.

Saunders, Jaclyn K., Matthew R. McIlvin, Chris L. Dupont, Drishti Kaul, Dawn M. Moran,

Tristan Horner, Sarah M. Laperriere, et al. 2022. ‘Microbial Functional Diversity across Biogeochemical Provinces in the Central Pacific Ocean’. *Proceedings of the National Academy of Sciences* 119 (37): e2200014119. https://doi.org/10.1073/pnas.2200014119.

Schallert, Kay, Pieter Verschaffelt, Bart Mesuere, Dirk Benndorf, Lennart Martens, and Tim Van Den Bossche. 2022. ‘Pout2Prot: An Efficient Tool to Create Protein (Sub)Groups from Percolator Output Files’. *Journal of Proteome Research* 21 (4): 1175–80. https://doi.org/10.1021/acs.jproteome.1c00685.

Schäpe, Stephanie Serena, Jannike Lea Krause, Beatrice Engelmann, Katarina Fritz-Wallace, Florian Schattenberg, Zishu Liu, Susann Müller, et al. 2019. ‘The Simplified Human Intestinal Microbiota (SIHUMIx) Shows High Structural and Functional Resistance against Changing Transit Times in In Vitro Bioreactors’. *Microorganisms* 7 (12): 641. https://doi.org/10.3390/microorganisms7120641.

Schiebenhoefer, Henning, Kay Schallert, Bernhard Y. Renard, Kathrin Trappe, Emanuel Schmid, Dirk Benndorf, Katharina Riedel, Thilo Muth, and Stephan Fuchs. 2020. ‘A Complete and Flexible Workflow for Metaproteomics Data Analysis Based on MetaProteomeAnalyzer and Prophane’. *Nature Protocols* 15 (10): 3212–39. https://doi.org/10.1038/s41596-020-0368-7.

Schiebenhoefer, Henning, Tim Van Den Bossche, Stephan Fuchs, Bernhard Y. Renard, Thilo Muth, and Lennart Martens. 2019. ‘Challenges and Promise at the Interface of Metaproteomics and Genomics: An Overview of Recent Progress in Metaproteogenomic Data Analysis’. *Expert Review of Proteomics* 16 (5): 375–90. https://doi.org/10.1080/14789450.2019.1609944.

Schiml, Valerie C., Francesco Delogu, Praveen Kumar, Benoit Kunath, Bérénice Batut, Subina Mehta, James E. Johnson, et al. 2023. ‘Integrative Meta-Omics in Galaxy and Beyond’. *Environmental Microbiome* 18 (1): 56. https://doi.org/10.1186/s40793-023-00514-9.

Schoch, Conrad L, Stacy Ciufo, Mikhail Domrachev, Carol L Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe, et al. 2020. ‘NCBI Taxonomy: A Comprehensive Update on Curation, Resources and Tools’. *Database* 2020 (January):baaa062. https://doi.org/10.1093/database/baaa062.

Searle, Brian C., Lindsay K. Pino, Jarrett D. Egertson, Ying S. Ting, Robert T. Lawrence, Brendan X. MacLean, Judit Villén, and Michael J. MacCoss. 2018. ‘Chromatogram Libraries Improve Peptide Detection and Quantification by Data Independent Acquisition Mass Spectrometry’. *Nature Communications* 9 (1): 5128. https://doi.org/10.1038/s41467-018-07454-w.

Searle, Brian C., Ariana E. Shannon, and Damien Beau Wilburn. 2023. ‘Scribe: Next Generation Library Searching for DDA Experiments’. *Journal of Proteome Research* 22 (2): 482–90. https://doi.org/10.1021/acs.jproteome.2c00672.

Sechi, Salvatore, and Brian T. Chait. 1998. ‘Modification of Cysteine Residues by Alkylation. A Tool in Peptide Mapping and Protein Identification’. *Analytical Chemistry* 70 (24): 5150–58. https://doi.org/10.1021/ac9806005.

Sequeira, João C., Vítor Pereira, M. Madalena Alves, M. Alcina Pereira, Miguel Rocha, and Andreia F. Salvador. 2024. ‘MOSCA 2.0: A Bioinformatics Framework for Metagenomics, Metatranscriptomics and Metaproteomics Data Analysis and Visualization’. *Molecular Ecology Resources* 24 (7): e13996. https://doi.org/10.1111/1755-0998.13996.

Shuken, Steven R. 2023. ‘An Introduction to Mass Spectrometry-Based Proteomics’. *Journal of Proteome Research* 22 (7): 2151–71. https://doi.org/10.1021/acs.jproteome.2c00838.

Simopoulos, Caitlin M. A., Zhibin Ning, Leyuan Li, Mona M. Khamis, Xu Zhang, Mathieu Lavallée-Adam, and Daniel Figeys. 2022. ‘MetaProClust-MS1: An MS1 Profiling Approach for Large-Scale Microbiome Screening’. *mSystems* 7 (4): e00381-22. https://doi.org/10.1128/msystems.00381-22.

Sinitcyn, Pavel, Jan Daniel Rudolph, and Jürgen Cox. 2018. ‘Computational Methods for

Understanding Mass Spectrometry–Based Shotgun Proteomics Data'. *Annual Review of Biomedical Data Science* 1 (1): 207–34. https://doi.org/10.1146/annurev-biodatasci-080917-013516.

Smyth, Patrick, Xu Zhang, Zhibin Ning, Janice Mayne, Jasmine Isabelle Moore, Krystal Walker, Mathieu Lavallée-Adam, and Daniel Figeys. 2020. 'Studying the Temporal Dynamics of the Gut Microbiota Using Metabolic Stable Isotope Labeling and Metaproteomics'. *Analytical Chemistry* 92 (24): 15711–18. https://doi.org/10.1021/acs.analchem.0c02070.

Speda, Jutta, Mikaela A. Johansson, Uno Carlsson, and Martin Karlsson. 2017. 'Assessment of Sample Preparation Methods for Metaproteomics of Extracellular Proteins'. *Analytical Biochemistry* 516 (January):23–36. https://doi.org/10.1016/j.ab.2016.10.008.

Starke, Robert, Nico Jehmlich, Trinidad Alfaro, Alice Dohnalkova, Petr Capek, Sheryl L. Bell, and Kirsten S. Hofmockel. 2019. 'Incomplete Cell Disruption of Resistant Microbes'. *Scientific Reports* 9 (1): 5618. https://doi.org/10.1038/s41598-019-42188-9.

Starke, Robert, Nico Jehmlich, and Felipe Bastida. 2019. 'Using Proteins to Study How Microbes Contribute to Soil Ecosystem Services: The Current State and Future Perspectives of Soil Metaproteomics'. *Journal of Proteomics*, 10 Year Anniversary of Proteomics, 198 (April):50–58. https://doi.org/10.1016/j.jprot.2018.11.011.

Stewart, Robert D., Marc D. Auffret, Amanda Warr, Alan W. Walker, Rainer Roehe, and Mick Watson. 2019. 'Compendium of 4,941 Rumen Metagenome-Assembled Genomes for Rumen Microbiome Biology and Enzyme Discovery'. *Nature Biotechnology* 37 (8): 953–61. https://doi.org/10.1038/s41587-019-0202-3.

Sun, Zhongzhi, Zhibin Ning, and Daniel Figeys. 2024. 'The Landscape and Perspectives of the Human Gut Metaproteomics'. *Molecular & Cellular Proteomics* 23 (5): 100763. https://doi.org/10.1016/j.mcpro.2024.100763.

Tanca, Alessandro, Marcello Abbondio, Giovanni Fiorito, Giovanna Pira, Rosangela Sau, Alessandra Manca, Maria Rosaria Muroni, et al. 2022. 'Metaproteomic Profile of the Colonic Luminal Microbiota From Patients With Colon Cancer'. *Frontiers in Microbiology* 13 (April). https://doi.org/10.3389/fmicb.2022.869523.

Tanca, Alessandro, Maria Antonietta Deledda, Laura De Diego, Marcello Abbondio, and Sergio Uzzau. 2024. 'Benchmarking Low- and High-Throughput Protein Cleanup and Digestion Methods for Human Fecal Metaproteomics'. *mSystems* 9 (7): e00661-24. https://doi.org/10.1128/msystems.00661-24.

Tanca, Alessandro, Antonio Palomba, Cristina Fraumene, Daniela Pagnozzi, Valeria Manghina, Massimo Deligios, Thilo Muth, et al. 2016. 'The Impact of Sequence Database Choice on Metaproteomic Results in Gut Microbiota Studies'. *Microbiome* 4 (1): 51. https://doi.org/10.1186/s40168-016-0196-8.

Tanca, Alessandro, Antonio Palomba, Salvatore Pisanu, Maria Filippa Addis, and Sergio Uzzau. 2015. 'Enrichment or Depletion? The Impact of Stool Pretreatment on Metaproteomic Characterization of the Human Gut Microbiota'. *PROTEOMICS* 15 (20): 3474–85. https://doi.org/10.1002/pmic.201400573.

Tanca, Alessandro, Antonio Palomba, Salvatore Pisanu, Massimo Deligios, Cristina Fraumene, Valeria Manghina, Daniela Pagnozzi, Maria Filippa Addis, and Sergio Uzzau. 2014. 'A Straightforward and Efficient Analytical Pipeline for Metaproteome Characterization'. *Microbiome* 2 (1): 49. https://doi.org/10.1186/s40168-014-0049-2.

Tang, Jing, Minjie Mou, Yunxia Wang, Yongchao Luo, and Feng Zhu. 2021. 'MetaFS: Performance Assessment of Biomarker Discovery in Metaproteomics'. *Briefings in Bioinformatics* 22 (3): bbaa105. https://doi.org/10.1093/bib/bbaa105.

The Galaxy Community. 2024. 'The Galaxy Platform for Accessible, Reproducible, and Collaborative Data Analyses: 2024 Update'. *Nucleic Acids Research* 52 (W1): W83–94. https://doi.org/10.1093/nar/gkae410.

The, Matthew, Michael J. MacCoss, William S. Noble, and Lukas Käll. 2016. 'Fast and

Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0'. *Journal of the American Society for Mass Spectrometry* 27 (11): 1719–27. https://doi.org/10.1007/s13361-016-1460-7.

The UniProt Consortium. 2023. 'UniProt: The Universal Protein Knowledgebase in 2023'. *Nucleic Acids Research* 51 (D1): D523–31. https://doi.org/10.1093/nar/gkac1052.

Välikangas, Tommi, Tomi Suomi, and Laura L Elo. 2018. 'A Systematic Evaluation of Normalization Methods in Quantitative Label-Free Proteomics'. *Briefings in Bioinformatics* 19 (1): 1–11. https://doi.org/10.1093/bib/bbw095.

Van Den Bossche, Tim, Magnus Ø. Arntzen, Dörte Becher, Dirk Benndorf, Vincent G. H. Eijsink, Céline Henry, Pratik D. Jagtap, et al. 2021. 'The Metaproteomics Initiative: A Coordinated Approach for Propelling the Functional Characterization of Microbiomes'. *Microbiome* 9 (1): 243. https://doi.org/10.1186/s40168-021-01176-w.

Van Den Bossche, Tim, Denis Beslic, Sam van Puyenbroeck, Tomi Suomi, Tanja Holstein, Lennart Martens, Laura L. Elo, and Thilo Muth. 2024. 'Metaproteomics beyond Databases: Addressing the Challenges and Potentials of de Novo Sequencing'. ChemRxiv. https://doi.org/10.26434/chemrxiv-2024-4v6q0.

Van Den Bossche, Tim, Benoit J. Kunath, Kay Schallert, Stephanie S. Schäpe, Paul E. Abraham, Jean Armengaud, Magnus Ø Arntzen, et al. 2021. 'Critical Assessment of MetaProteome Investigation (CAMPI): A Multi-Laboratory Comparison of Established Workflows'. *Nature Communications* 12 (1): 7305. https://doi.org/10.1038/s41467-021-27542-8.

Van Den Bossche, Tim, Pieter Verschaffelt, Kay Schallert, Harald Barsnes, Peter Dawyndt, Dirk Benndorf, Bernhard Y. Renard, Bart Mesuere, Lennart Martens, and Thilo Muth. 2020. 'Connecting MetaProteomeAnalyzer and PeptideShaker to Unipept for Seamless End-to-End Metaproteomics Data Analysis'. *Journal of Proteome Research* 19 (8): 3562–66. https://doi.org/10.1021/acs.jproteome.0c00136.

Van Den Bossche, Tim, Pieter Verschaffelt, Tibo Vande Moortele, Peter Dawyndt, Lennart Martens, and Bart Mesuere. 2024. 'Biodiversity Analysis of Metaproteomics Samples with Unipept: A Comprehensive Tutorial'. In *Protein Bioinformatics*, edited by Frédérique Lisacek, 183–215. New York, NY: Springer US. https://doi.org/10.1007/978-1-0716-4007-4_11.

Vande Moortele, Tibo, Bram Devlaminck, Simon Van de Vyver, Tim Van Den Bossche, Lennart Martens, Peter Dawyndt, Bart Mesuere, and Pieter Verschaffelt. 2024. 'Unipept 6.0: Expanding Metaproteomics Analysis with Support for Missed Cleavages, Semi-Tryptic and Non-Tryptic Peptides'. bioRxiv. https://doi.org/10.1101/2024.09.26.615136.

Vande Moortele, Tibo, Pieter Verschaffelt, Qingyao Huang, Nadezhda T. Doncheva, Tanja Holstein, Caroline Jachmann, Peter Dawyndt, Lennart Martens, Bart Mesuere, and Tim Van Den Bossche. 2024. 'PathwayPilot: A User-Friendly Tool for Visualizing and Navigating Metabolic Pathways'. bioRxiv. https://doi.org/10.1101/2024.06.21.599989.

Vaudel, Marc, Harald Barsnes, Frode S. Berven, Albert Sickmann, and Lennart Martens. 2011. 'SearchGUI: An Open-Source Graphical User Interface for Simultaneous OMSSA and X!Tandem Searches'. *PROTEOMICS* 11 (5): 996–99. https://doi.org/10.1002/pmic.201000595.

Vaudel, Marc, Julia M. Burkhart, René P. Zahedi, Eystein Oveland, Frode S. Berven, Albert Sickmann, Lennart Martens, and Harald Barsnes. 2015. 'PeptideShaker Enables Reanalysis of MS-Derived Proteomics Data Sets'. *Nature Biotechnology* 33 (1): 22–24. https://doi.org/10.1038/nbt.3109.

Vaudel, Marc, A. Saskia Venne, Frode S. Berven, René P. Zahedi, Lennart Martens, and Harald Barsnes. 2014. 'Shedding Light on Black Boxes in Protein Identification'. *PROTEOMICS* 14 (9): 1001–5. https://doi.org/10.1002/pmic.201300488.

Verberkmoes, Nathan C., Alison L. Russell, Manesh Shah, Adam Godzik, Magnus

Rosenquist, Jonas Halfvarson, Mark G. Lefsrud, et al. 2009. ‘Shotgun Metaproteomics of the Human Distal Gut Microbiota’. *The ISME Journal* 3 (2): 179–89. https://doi.org/10.1038/ismej.2008.108.

Verheggen, Kenneth, Helge Ræder, Frode S. Berven, Lennart Martens, Harald Barsnes, and Marc Vaudel. 2020. ‘Anatomy and Evolution of Database Search Engines—a Central Component of Mass Spectrometry Based Proteomic Workflows’. *Mass Spectrometry Reviews* 39 (3): 292–306. https://doi.org/10.1002/mas.21543.

Verschaffelt, Pieter, Alessandro Tanca, Marcello Abbondio, Tim Van Den Bossche, Tibo Vande Moortele, Peter Dawyndt, Lennart Martens, and Bart Mesuere. 2023. ‘Unipept Desktop 2.0: Construction of Targeted Reference Protein Databases for Metaproteogenomics Analyses’. *Journal of Proteome Research* 22 (8): 2620–28. https://doi.org/10.1021/acs.jproteome.3c00091.

Verschaffelt, Pieter, Philippe Van Thienen, Tim Van Den Bossche, Felix Van der Jeugt, Caroline De Tender, Lennart Martens, Peter Dawyndt, and Bart Mesuere. 2020. ‘Unipept CLI 2.0: Adding Support for Visualizations and Functional Annotations’. *Bioinformatics (Oxford, England)* 36 (14): 4220–21. https://doi.org/10.1093/bioinformatics/btaa553.

Vertommen, Annelies, Bart Panis, Rony Swennen, and Sebastien Christian Carpentier. 2010. ‘Evaluation of Chloroform/Methanol Extraction to Facilitate the Study of Membrane Proteins of Non-Model Plants’. *Planta* 231 (5): 1113–25. https://doi.org/10.1007/s00425-010-1121-1.

Waibel, Matthias, Kevin McDonnell, Maria Tuohy, Sally Shirran, Sylvia Synowsky, Barry Thornton, Eric Paterson, Fiona Brennan, and Florence Abram. 2023. ‘Assessing the Impact of Interfering Organic Matter on Soil Metaproteomic Workflow’. *European Journal of Soil Science* 74 (3): e13392. https://doi.org/10.1111/ejss.13392.

Wang, Angela, Emily E. F. Fekete, Marybeth Creskey, Kai Cheng, Zhibin Ning, Annabelle Pfeifle, Xuguang Li, Daniel Figeys, and Xu Zhang. 2024. ‘Assessing Fecal Metaproteomics Workflow and Small Protein Recovery Using DDA and DIA PASEF Mass Spectrometry’. *Microbiome Research Reports* 3 (3): N/A-N/A. https://doi.org/10.20517/mrr.2024.21.

Wang, Jiaqin, Xu Zhang, Leyuan Li, Zhibin Ning, Janice Mayne, Cian Schmitt-Ulms, Krystal Walker, Kai Cheng, and Daniel Figeys. 2020. ‘Differential Lysis Approach Enables Selective Extraction of Taxon-Specific Proteins for Gut Metaproteomics’. *Analytical Chemistry* 92 (7): 5379–86. https://doi.org/10.1021/acs.analchem.0c00062.

Wang, Le-heng, De-Quan Li, Yan Fu, Hai-Peng Wang, Jing-Fen Zhang, Zuo-Fei Yuan, Rui-Xiang Sun, Rong Zeng, Si-Min He, and Wen Gao. 2007. ‘pFind 2.0: A Software Package for Peptide and Protein Identification via Tandem Mass Spectrometry’. *Rapid Communications in Mass Spectrometry* 21 (18): 2985–91. https://doi.org/10.1002/rcm.3173.

Wang, Luman, Caitlin M. A. Simopoulos, Joeselle M. Serrana, Zhibin Ning, Boyan Sun, Jinhui Yuan, Daniel Figeys, and Leyuan Li. 2024. ‘PhyloFunc: Phylogeny-Informed Functional Distance as a New Ecological Metric for Metaproteomic Data Analysis’. bioRxiv. https://doi.org/10.1101/2024.05.28.596184.

Wang, Songduo, Zenghu Zhang, Kaiguang Yang, Jiulong Zhao, Weijie Zhang, Zhiting Wang, Zhen Liang, et al. 2024. ‘SMMP: A Deep-Coverage Marine Metaproteome Method for Microbial Community Analysis throughout the Water Column Using 1 L of Seawater’. *Analytical Chemistry* 96 (29): 12030–39. https://doi.org/10.1021/acs.analchem.4c02079.

Wang, Tong, Leyuan Li, Daniel Figeys, and Yang-Yu Liu. 2024. ‘Pairing Metagenomics and Metaproteomics to Characterize Ecological Niches and Metabolic Essentiality of Gut Microbiomes’. *ISME Communications* 4 (1): ycae063. https://doi.org/10.1093/ismeco/ycae063.

Wang, Yuqiu, Yanting Zhou, Xiao Xiao, Jing Zheng, and Hu Zhou. 2020. ‘Metaproteomics:

A Strategy to Study the Taxonomy and Functionality of the Gut Microbiota'. *Journal of Proteomics* 219 (May):103737. https://doi.org/10.1016/j.jprot.2020.103737.

Washburn, Michael P., Dirk Wolters, and John R. Yates. 2001. 'Large-Scale Analysis of the Yeast Proteome by Multidimensional Protein Identification Technology'. *Nature Biotechnology* 19 (3): 242–47. https://doi.org/10.1038/85686.

Wessel, D., and U. I. Flügge. 1984. 'A Method for the Quantitative Recovery of Protein in Dilute Solution in the Presence of Detergents and Lipids'. *Analytical Biochemistry* 138 (1): 141–43. https://doi.org/10.1016/0003-2697(84)90782-6.

Wilkins, Marc R., Christian Pasquali, Ron D. Appel, Keli Ou, Olivier Golaz, Jean-Charles Sanchez, Jun X. Yan, et al. 1996. 'From Proteins to Proteomes: Large Scale Protein Identification by Two-Dimensional Electrophoresis and Arnino Acid Analysis'. *Bio/Technology* 14 (1): 61–65. https://doi.org/10.1038/nbt0196-61.

Wilmes, Paul, Anna Heintz-Buschart, and Philip L. Bond. 2015. 'A Decade of Metaproteomics: Where We Stand and What the Future Holds'. *PROTEOMICS* 15 (20): 3409–17. https://doi.org/10.1002/pmic.201500183.

Wiśniewski, Jacek R., Alexandre Zougman, Nagarjuna Nagaraj, and Matthias Mann. 2009. 'Universal Sample Preparation Method for Proteome Analysis'. *Nature Methods* 6 (5): 359–62. https://doi.org/10.1038/nmeth.1322.

Wolf, Maximilian, Kay Schallert, Luca Knipper, Albert Sickmann, Alexander Sczyrba, Dirk Benndorf, and Robert Heyer. 2023. 'Advances in the Clinical Use of Metaproteomics'. *Expert Review of Proteomics* 20 (4–6): 71–86. https://doi.org/10.1080/14789450.2023.2215440.

Wood, Derrick E., Jennifer Lu, and Ben Langmead. 2019. 'Improved Metagenomic Analysis with Kraken 2'. *Genome Biology* 20 (1): 257. https://doi.org/10.1186/s13059-019-1891-0.

Wu, Enhui, Guanyang Xu, Dong Xie, and Liang Qiao. 2024. 'Data-Independent Acquisition in Metaproteomics'. *Expert Review of Proteomics* 21 (7–8): 271–80. https://doi.org/10.1080/14789450.2024.2394190.

Wu, Qing, Zhibin Ning, Ailing Zhang, Xu Zhang, Zhongzhi Sun, and Daniel Figeys. 2024. 'MetaX: A Peptide Centric Metaproteomic Data Analysis Platform Using Operational Taxa-Functions (OTF)'. bioRxiv. https://doi.org/10.1101/2024.04.19.590315.

Wu, Qiong, Jiangnan Zheng, Xintong Sui, Changying Fu, Xiaozhen Cui, Bin Liao, Hongchao Ji, et al. 2024. 'High-Throughput Drug Target Discovery Using a Fully Automated Proteomics Sample Preparation Platform'. *Chemical Science* 15 (8): 2833–47. https://doi.org/10.1039/D3SC05937E.

Xiao, XiaoLian, Haidan Sun, Xiaoyan Liu, Zhengguang Guo, Shuxin Zheng, Jiyu Xu, Jiameng Sun, Ying Lan, Chen Shao, and Wei Sun. 2022. 'Qualitative and Quantitative Proteomic and Metaproteomic Analyses of Healthy Human Urine Sediment'. *PROTEOMICS – Clinical Applications* 16 (2): 2100007. https://doi.org/10.1002/prca.202100007.

Xiao, Xiaolian, Xiaoping Xiao, Yaoran Liu, Haidan Sun, Xiaoyan Liu, Zhengguang Guo, Qian Li, and Wei Sun. 2023. 'Metaproteomics Characterizes the Human Gingival Crevicular Fluid Microbiome Function in Periodontitis'. *Journal of Proteome Research* 22 (7): 2411–20. https://doi.org/10.1021/acs.jproteome.3c00143.

Xie, Fei, Wei Jin, Huazhe Si, Yuan Yuan, Ye Tao, Junhua Liu, Xiaoxu Wang, et al. 2021. 'An Integrated Gene Catalog and over 10,000 Metagenome-Assembled Genomes from the Gastrointestinal Microbiome of Ruminants'. *Microbiome* 9 (1): 137. https://doi.org/10.1186/s40168-021-01078-x.

Xiong, Weili, Christopher T. Brown, Michael J. Morowitz, Jillian F. Banfield, and Robert L. Hettich. 2017. 'Genome-Resolved Metaproteomic Characterization of Preterm Infant Gut Microbiota Development Reveals Species-Specific Metabolic Shifts and Variabilities during Early Life'. *Microbiome* 5 (1): 72. https://doi.org/10.1186/s40168-017-0290-6.

Xiong, Weili, Richard J. Giannone, Michael J. Morowitz, Jillian F. Banfield, and Robert L. Hettich. 2015. 'Development of an Enhanced Metaproteomic Approach for Deepening the Microbiome Characterization of the Human Infant Gut'. *Journal of Proteome Research* 14 (1): 133–41. https://doi.org/10.1021/pr500936p.

Xu, Ping, Duc M. Duong, and Junmin Peng. 2009. 'Systematical Optimization of Reverse-Phase Chromatography for Shotgun Proteomics'. *Journal of Proteome Research* 8 (8): 3944–50. https://doi.org/10.1021/pr900251d.

Xue, Ming-Yuan, Yun-Yi Xie, Xin-Wei Zang, Yi-Fan Zhong, Xiao-Jiao Ma, Hui-Zeng Sun, and Jian-Xin Liu. 2024. 'Deciphering Functional Groups of Rumen Microbiome and Their Underlying Potentially Causal Relationships in Shaping Host Traits'. *iMeta* 3 (4): e225. https://doi.org/10.1002/imt2.225.

Yang, Liang, Wenlai Fan, and Yan Xu. 2020. 'Metaproteomics Insights into Traditional Fermented Foods and Beverages'. *Comprehensive Reviews in Food Science and Food Safety* 19 (5): 2506–29. https://doi.org/10.1111/1541-4337.12601.

Yang, Tingpeng, Tianze Ling, Boyan Sun, Zhendong Liang, Fan Xu, Xiansong Huang, Linhai Xie, et al. 2024. 'Introducing π-HelixNovo for Practical Large-Scale de Novo Peptide Sequencing'. *Briefings in Bioinformatics* 25 (2): bbae021. https://doi.org/10.1093/bib/bbae021.

Yilmaz, Melih, William E. Fondrie, Wout Bittremieux, Carlo F. Melendez, Rowan Nelson, Varun Ananth, Sewoong Oh, and William Stafford Noble. 2024. 'Sequence-to-Sequence Translation from Mass Spectra to Peptides with a Transformer Model'. bioRxiv. https://doi.org/10.1101/2023.01.03.522621.

Zhang, Weipeng, Jin Sun, Huiluo Cao, Renmao Tian, Lin Cai, Wei Ding, and Pei-Yuan Qian. 2016. 'Post-Translational Modifications Are Enriched within Protein Functional Groups Important to Bacterial Adaptation within a Deep-Sea Hydrothermal Vent Environment'. *Microbiome* 4 (1): 49. https://doi.org/10.1186/s40168-016-0194-x.

Zhang, Xu, Wendong Chen, Zhibin Ning, Janice Mayne, David Mack, Alain Stintzi, Ruijun Tian, and Daniel Figeys. 2017. 'Deep Metaproteomics Approach for the Study of Human Microbiomes'. *Analytical Chemistry* 89 (17): 9407–15. https://doi.org/10.1021/acs.analchem.7b02224.

Zhang, Xu, Kai Cheng, Zhibin Ning, Janice Mayne, Krystal Walker, Hao Chi, Charles L. Farnsworth, Kimberly Lee, and Daniel Figeys. 2021. 'Exploring the Microbiome-Wide Lysine Acetylation, Succinylation, and Propionylation in Human Gut Microbiota'. *Analytical Chemistry* 93 (17): 6594–98. https://doi.org/10.1021/acs.analchem.1c00962.

Zhang, Xu, and Daniel Figeys. 2019. 'Perspective and Guidelines for Metaproteomics in Microbiome Studies'. *Journal of Proteome Research* 18 (6): 2370–80. https://doi.org/10.1021/acs.jproteome.9b00054.

Zhang, Xu, Leyuan Li, Janice Mayne, Zhibin Ning, Alain Stintzi, and Daniel Figeys. 2018. 'Assessing the Impact of Protein Extraction Methods for Human Gut Metaproteomics'. *Journal of Proteomics*, Proteomics in Infectious Diseases, 180 (May):120–27. https://doi.org/10.1016/j.jprot.2017.07.001.

Zhang, Xu, Zhibin Ning, Janice Mayne, Jasmine I. Moore, Jennifer Li, James Butcher, Shelley Ann Deeke, et al. 2016. 'MetaPro-IQ: A Universal Metaproteomic Approach to Studying Human and Mouse Gut Microbiota'. *Microbiome* 4 (1): 31. https://doi.org/10.1186/s40168-016-0176-z.

Zhang, Xu, Zhibin Ning, Janice Mayne, Yidai Yang, Shelley A. Deeke, Krystal Walker, Charles L. Farnsworth, et al. 2020. 'Widespread Protein Lysine Acetylation in Gut Microbiome and Its Alterations in Patients with Crohn's Disease'. *Nature Communications* 11 (1): 4120. https://doi.org/10.1038/s41467-020-17916-9.

Zhao, Jinzhi, Yi Yang, Liangqiang Chen, Jianxujie Zheng, Xibin Lv, Dandan Li, Ziyu Fang, et al. 2023. 'Quantitative Metaproteomics Reveals Composition and Metabolism Characteristics of Microbial Communities in Chinese Liquor Fermentation Starters'. *Frontiers in Microbiology* 13 (January).

https://doi.org/10.3389/fmicb.2022.1098268.
Zhao, Jinzhi, Yi Yang, Mengjing Teng, Jianxujie Zheng, Bing Wang, Vijini Mallawaarachchi, Yu Lin, et al. 2023. 'Metaproteomics Profiling of the Microbial Communities in Fermentation Starters (Daqu) during Multi-Round Production of Chinese Liquor'. *Frontiers in Nutrition* 10 (June). https://doi.org/10.3389/fnut.2023.1139836.
Zhao, Jinzhi, Yi Yang, Hua Xu, Jianxujie Zheng, Chengpin Shen, Tian Chen, Tao Wang, et al. 2023. 'Data-Independent Acquisition Boosts Quantitative Metaproteomics for Deep Characterization of Gut Microbiota'. *Npj Biofilms and Microbiomes* 9 (1): 1–14. https://doi.org/10.1038/s41522-023-00373-9.