

UNIVERSITY OF OKLAHOMA
GRADUATE COLLEGE

ADVANCED TOOLS FOR ISOTOPE PROBING AND MASS SPECTROMETRY IN
MICROBIAL PROTEOMICS WITH AERITH AND SIPROS

A DISSERTATION
SUBMITTED TO THE GRADUATE FACULTY
in partial fulfillment of the requirements for the
Degree of
DOCTOR OF PHILOSOPHY

By YI XIONG
Norman, Oklahoma

2026

ADVANCED TOOLS FOR ISOTOPE PROBING AND MASS SPECTROMETRY IN
MICROBIAL PROTEOMICS WITH AERITH AND SIPROS

A DISSERTATION APPROVED FOR THE
SCHOOL OF BIOLOGICAL SCIENCES

BY THE COMMITTEE CONSISTING OF

Dr. Chongle Pan, Chair

Dr. Qinggong Tang

Dr. Jizhong Zhou

Dr. Kara Deleon

Dr. Wei Qin

© Copyright by YI XIONG 2026

All Rights Reserved.

Acknowledgments

“The important thing is not to stop questioning. Curiosity has its own reason for existing.”

— *Albert Einstein*

This dissertation would not have been possible without the support and guidance of many remarkable individuals.

First and foremost, I thank my advisor, **Dr. Chongle Pan**, for his unwavering mentorship, scientific vision, and patience throughout my doctoral training. His guidance has been instrumental to every chapter of this dissertation, from the original conception of the Sipros isotope-resolved search engine to the design of the Aerith visualization framework and the extension of Sipros to data-independent acquisition on the Orbitrap Astral.

I also thank the members of my dissertation committee, **Dr. Qinggong Tang**, **Dr. Jizhong Zhou**, **Dr. Kara Deleon**, and **Dr. Wei Qin**, for their thoughtful feedback across mass spectrometry, biomedical engineering, environmental microbiology, and microbial biogeochemistry, and for the constructive critiques that improved both the science and its presentation. I am also grateful to **Dr. Zhibo Yang** for his guidance and support during my doctoral training.

I gratefully acknowledge my collaborators and co-authors on the manuscripts that became chapters of this dissertation: **Dr. Ryan S. Mueller** (Oregon State University) for collaboration on the Sipros 4 and 5 benchmarks and **Dr. Zhou Li** (Pfizer) for sharing the ^{13}C -methanol and $^{13}\text{CO}_2$ soil SIP datasets that anchored those benchmarks; **Dr. Shichao Feng** (University of North Texas) for her close partnership on Sipros 4 and Sipros 5 development, benchmarking, and writing; **Bailu Zhang** (University of North Texas) for collaborative work on SEMQuant and WinnowNet that sharpened the metaproteomics tooling environment in which Aerith and Sipros matured; and **Dr. Xuan Guo** (University of North Texas) for his contributions to the Sipros Ensemble framework.

I am grateful to the members of **thepanlab** for the camaraderie, code reviews, and server time. The mass spectrometry datasets that made this work possible originated in collaborations with

proteomic SIP groups at University of North Texas and Oregon State University; I thank the **ProteomeXchange** community for the public RAW files under accessions PXD011738, PXD011739, PXD011740, PXD041958, PXD041414, PXD017451, PXD075964, PXD075589, PXD075720, and PXD076077. I thank the **IDeA National Resource for Quantitative Proteomics** at the **University of Arkansas for Medical Sciences (UAMS)** and the **Proteomics and Metabolomics Core Facility at Duke University School of Medicine** for performing LC–MS-based proteomic analyses on the standards and labeled samples that benchmarked Sipros 5.

I thank the **University of Oklahoma School of Biological Sciences** for the doctoral training environment, the **OU Supercomputing Center for Education & Research (OSCER)** for sustained access to the Schooner cluster on which the enrichment-resolved database searches for Chapters 2 and 3 were performed, and the staff of the **OU Mass Spectrometry, Proteomics, and Metabolomics Core** for technical support. I am particularly grateful to **Matt Caldwell** (Supervisor) and **Duncan Ziegler** at the OU School of Biological Sciences MPC for providing the bacterial strains used in this work, and to **Mike Todd** for equipment installation and technical support. I gratefully acknowledge the **National Institutes of Health**, specifically the **National Center for Complementary & Integrative Health** and the **National Institute of General Medical Sciences** (Grant R01AT011618), for funding support.

Finally, I thank my family and friends for their patient encouragement throughout this journey. This work is as much theirs as it is mine.

Table of Contents

Acknowledgments	iv
Table of Contents	vi
List of Tables	viii
List of Figures	ix
Abstract	xi
1 Introduction	1
Stable isotope probing in microbial ecology	1
Protein-based stable isotope probing	1
Algorithms for proteomic SIP and their limitations	3
Modern mass spectrometry	5
Goals and structure of this dissertation	6
2 Proteomic Stable Isotope Probing with an Upgraded Sipros Algorithm for Improved Identification and Quantification of Isotopically Labeled Proteins	7
Abstract	7
Introduction	8
Results	9
Discussion	17
Conclusions	18
Materials and Methods	19
Author contributions and funding	25
Supplementary Material	26
3 Sipros 5 Significantly Enhances Proteomic Isotope Probing in DDA Orbitrap and Narrow-Window DIA Astral Mass Spectrometry Analyses	27
Abstract	27
Introduction	27
Results	29
Discussion	43
Methods	48
Acknowledgments and funding	53

	Supplementary Material	54
4	Aerith: Visualization and Annotation of Isotopic Enrichment Patterns of Peptides and Metabolites with Stable Isotope Labeling from Proteomics and Metabolomics	55
	Abstract	55
	Introduction	57
	Feature and Implementation	58
	Results and Discussion	63
	Conclusions	64
	Software and data availability	64
	Acknowledgments and funding	65
	Supplementary Material	65
5	Summary and Output	66
	Synthesis	66
	Future directions	66
	Output of this dissertation	67
	References	69
	Supplementary Material for Chapter 2	74
	Supplementary Material for Chapter 3	90
	Supplementary Material for Chapter 4	117

List of Tables

Table 2.1	Comparison of Sipros 3, Sipros 4, Calisp, and MetaProSIP on <i>E. coli</i> standard samples	12
Table 2.2	Number of ¹³ C-labeled PSMs, peptides, and proteins identified at ≥5 atom% ¹³ C in soil SIP experiments.....	13
Table 3.1	Performance of filtering tools on Orbitrap DDA ¹³ C-labeled <i>E. coli</i> spike-ins	37
Table A.1	Isolation-window-size comparison at 50% ¹³ C-labeled <i>E. coli</i>	74
Table A.2	Sipros 4 vs. Sipros 3 on <i>E. coli</i> standards	74
Table A.3	Unlabeled regular-search results across tools.....	75
Table A.4	Sipros 4 vs. Sipros 3 on ¹⁵ N-labeled acid mine drainage samples	75
Table A.5	Unlabeled soil PSM, peptide, and protein totals	75
Table A.6	M9 media recipes	76
Table A.7	Genome-resolved proteomic SIP results	77
Table A.8	Abstract of large Appendix A supplementary workbooks	79
Table B.1	Sipros 5 vs. Sipros 4 benchmark on ¹³ C-glucose <i>E. coli</i>	90
Table B.2	Astral DIA spike-in filtering summary.....	91
Table B.3	Gifu anaerobic broth and working medium	91
Table B.4	GO biological-process enrichment in the ¹³ C-glucose incubation	92
Table B.5	GO molecular-function enrichment in the ¹³ C-glucose incubation.....	92
Table B.6	KEGG orthology enrichment in the ¹³ C-glucose incubation	92
Table B.7	GO biological-process enrichment in the ¹³ C-inulin incubation	93
Table B.8	GO molecular-function enrichment in the ¹³ C-inulin incubation	93
Table B.9	KEGG orthology enrichment in the ¹³ C-inulin incubation	94
Table B.10	Inulin-utilization enzyme dynamics	94
Table B.11	Functional-enrichment counts by annotation class.....	95
Table B.12	GO biological-process enrichment among ¹⁵ N-labeled microbiome proteins .	95
Table B.13	KEGG orthology enrichment among ¹⁵ N-labeled microbiome proteins	96
Table B.14	GO biological-process enrichment among ¹⁵ N-labeled host proteins	96
Table B.15	Reactome enrichment among ¹⁵ N-labeled host proteins	97
Table B.16	Abstract of large Sipros 5 supplementary workbook	97

List of Figures

Figure 1.1	Overview of Sipros-based proteomic SIP study designs	4
Figure 2.1	MS/MS measurement of ^{13}C -labeled vs. unlabeled isotopologues of an illustrative peptide	10
Figure 2.2	Accuracy and precision of ^{13}C atom% quantification on PSM level	10
Figure 2.3	^{13}C enrichment levels and label abundances of proteins labeled by ^{13}C -methanol SIP	14
Figure 2.4	^{13}C enrichment levels and label abundances of proteins labeled by $^{13}\text{CO}_2$ SIP	15
Figure 3.1	Overview of the Sipros 5 workflow and SIP-aware scoring modules	30
Figure 3.2	Sipros 5 benchmarking: PSM density and identification yields across enrichment levels	31
Figure 3.3	UpSet peptide-overlap analysis at 50 atom% ^{13}C between Astral DIA and Orbitrap DDA	32
Figure 3.4	Visualization of PSMs at 1% and 50% ^{13}C labeling on Orbitrap and Astral ..	33
Figure 3.5	Sipros 5 validation using ^{13}C -labeled <i>E. coli</i> peptide spike-ins (part 1)	35
Figure 3.6	Sipros 5 validation: q-value threshold analysis and density plots	36
Figure 3.7	Sipros 5 performance in mouse cecal ^{13}C -glucose SIP	39
Figure 3.8	Time-resolved ^{13}C -inulin assimilation by the mouse cecal microbiome	41
Figure 3.9	^{15}N -SILAM diet labeling experiment in mouse stool	44
Figure 4.1	Simulated isotopic fine structures and envelopes of SIP-labeled metabolites .	56
Figure 4.2	Visualization of an LC-MS/MS run in Aerith	58
Figure 4.3	Simulated isotopic fine structures and envelopes of SIP-labeled metabolites .	60
Figure 4.4	Simulated isotopic envelopes of SIP-labeled peptide precursors and fragments	61
Figure 4.5	Visualization of PSMs with observed and theoretical spectra	62
Figure 4.6	Comparison of scoring functions for ^{13}C enrichment estimation	62
Figure A.1	Venn diagrams of protein identifications across algorithms	80
Figure A.2	Taxonomic tree of soil microbial proteins	81
Figure A.3	Functional analysis of the ^{13}C -methanol SIP results	82
Figure A.4	Functional analysis of the $^{13}\text{CO}_2$ SIP results	83
Figure B.1	Astral DIA validation of Sipros 5 on ^{13}C -labeled spike-ins	98
Figure B.2	Genus-level annotation in ^{13}C -glucose cecum incubation	99
Figure B.3	GO enrichment among glucose-incorporating proteins	100
Figure B.4	Genus-level annotation in ^{13}C -inulin time course	101
Figure B.5	GO enrichment and enzyme-level shift in ^{13}C -inulin assimilation	102
Figure B.6	Hexbin density of precursor intensity vs. ^{15}N abundance	103

Figure B.7	GO enrichment landscape of ^{15}N -labeled microbiome and host proteins	104
Figure B.8	^{15}N protein-turnover half-life across enriched KO terms	105
Figure C.1	Reference spectrum of glucose from MZcloud	120
Figure C.2	Reference spectrum of penicillin from MZcloud	121

Abstract

Stable isotope probing (SIP) is a foundational technique in microbial ecology for linking microbial identity and activity to substrate utilization in complex communities. Proteomic SIP extends this idea to the peptide and protein level, where labeled amino acids carry both taxonomic information from their sequence and functional information from the annotation of their parent protein. However, proteomic SIP has historically been limited by three bottlenecks: the high computational cost of searching mass spectra against a protein database over a continuous range of isotopic enrichment levels, the difficulty of separating weakly labeled peptides from unlabeled background in complex metaproteomes, and the lack of accessible tools to visualize and manually validate isotopically labeled peptide-spectrum matches.

In this dissertation, I develop a coordinated suite of computational tools that addresses these bottlenecks and extends proteomic SIP from data-dependent acquisition (DDA) on Orbitrap mass spectrometers into the high-throughput narrow-window data-independent acquisition (DIA) regime of the Orbitrap Astral platform. Chapter 2 presents Sipros 4, an upgraded enrichment-resolved database search engine that is more than twenty-fold faster than Sipros 3, accurately recovers the median ^{13}C atom% expected in *E. coli* standards across 1.07–99 atom% enrichment, and identifies 153 ^{13}C -labeled proteins/protein groups in a ^{13}C -methanol soil SIP experiment versus 81, 15, and 13 by Sipros 3, Calisp, and MetaProSIP respectively. Chapter 3 introduces Sipros 5, an end-to-end SIP proteomics platform with universal percentage-labeling support that models precursor and fragment isotope envelopes in DDA and DIA data, uses unlabeled negative controls to control the label-specific false discovery rate (FDR), and is paired with a convolutional neural network (CNN)-based WinnowNet rescorer. Sipros 5 produced $2\text{--}3\times$ more peptide-spectrum matches (PSMs) than Sipros 4 in DDA and reached approximately fivefold higher PSM yields in 35-minute Astral DIA runs (130,000–330,000 PSMs) compared with 90-minute Orbitrap DDA runs ($\sim 30,000\text{--}50,000$ PSMs); applied to a 43-day mouse experiment

using ^{15}N stable isotope labeling of mammals (SILAM), Sipros 5 quantified protein-turnover kinetics that distinguish microbial proteins (half-life ~ 3 days; 90% conversion by ~ 10 days) from host proteins (half-life ~ 5 days; 90% conversion by ~ 17 days). Chapter 4 describes **Aerith**, an R package that simulates SIP-labeled isotopic fine structures via Monte Carlo and isotopic envelopes via fast Fourier transform (FFT) for arbitrary chemical formulas, simulates fragment-ion envelopes of peptides via residue-level convolution and binomial distributions, scores PSMs with multivariate hypergeometric (MVH), cross-correlation (Xcorr), weighted dot product (WDP), and spectral-entropy scores, and produces publication-quality visualizations through ggplot2 for manual validation across the full 0–100% enrichment range. Together, these tools provide a comprehensive analytical pipeline that improves the sensitivity, accuracy, and interpretability of proteomic and metaproteomic SIP and enables the routine quantitative analysis of substrate assimilation and protein turnover at single-organism and microbiome-community scales.

Keywords: stable isotope probing, mass spectrometry, proteomics, microbial proteomics, metaproteomics, Sipros, Aerith, Orbitrap, Astral, data-dependent acquisition, data-independent acquisition, protein turnover, gut microbiome.

Introduction

Stable isotope probing in microbial ecology

Microbial communities drive the global cycling of carbon, nitrogen, sulfur, and many other elements, yet most of the organisms involved are uncultivated and their in situ activities are not directly observable. Stable isotope probing (SIP) traces rare, non-radioactive isotopes such as ^{13}C , ^{15}N , ^2H , or ^{18}O from a labeled substrate into the biomolecules of microorganisms that have assimilated that substrate. Because the isotopic label is inherited only by organisms that were metabolically active during the labeling period, SIP provides a direct link between identity and activity in complex communities (Radajewski et al., 2000; Dumont and Murrell, 2005; Neufeld et al., 2007). The first SIP experiments tracked ^{13}C from labeled methanol into the DNA of methylotrophs and, shortly thereafter, into ribosomal RNA, establishing DNA-SIP and RNA-SIP as workhorses for identifying active community members (Radajewski et al., 2000; Manefield et al., 2002). Applications have spanned soil and freshwater bioremediation (Jeon et al., 2003), methane and methanol cycling in lakes and soils, microbial interactions in the rhizosphere (Li et al., 2019), and the dynamics of host-associated microbiomes such as the human and animal gut (Berry et al., 2013; Mordant et al., 2025). Despite their power, nucleic-acid SIP methods reveal only the identity of active organisms; they do not directly report which proteins or biochemical pathways are responsible for substrate assimilation (Coyotzi et al., 2016). This functional gap motivated the development of proteomic SIP.

Protein-based stable isotope probing

Protein-based SIP, or protein-SIP, was introduced as a method to couple the phylogenetic resolution of peptide sequencing with the quantitative isotope incorporation measurable by mass spectrometry (Jehmlich et al., 2008, 2010; Seifert et al., 2012). In a typical workflow, a microbial community

is incubated with a substrate enriched in a stable isotope; proteins are extracted, digested, and analyzed by liquid chromatography coupled to tandem mass spectrometry (LC–MS/MS); and a database search identifies peptides while estimating the isotopic enrichment of each labeled peptide. Because the unlabeled peptide sequence carries species-level information through its parent protein, while the enrichment level reports on the time-integrated metabolic activity of the source organism, protein-SIP simultaneously recovers *who* is active and *what* they are doing (von Bergen et al., 2013).

Two distinct chemistries support protein-SIP. In a fully labeled experiment, the substrate is supplied at near-100 atom% so that proteins synthesized during labeling carry an unambiguous mass shift relative to their unlabeled counterparts (Jehmlich et al., 2010). In partial-labeling experiments—common in environmental and gut microbiome settings, where complete substitution of endogenous substrates is impractical—proteins acquire intermediate enrichment levels that broaden their isotopic envelopes across many isotopologues. The latter regime is biologically more realistic but analytically far more challenging, because the labeling shifts both precursor and fragment-ion isotopic patterns in a way that confounds conventional database search engines (Pan et al., 2011; von Bergen et al., 2013).

Protein-SIP has revealed taxon-resolved substrate use in acid mine drainage biofilms with $^{15}\text{NH}_4\text{Cl}$ labeling (Justice et al., 2014), in ^{13}C -methanol- and $^{13}\text{CO}_2$ -amended soils (Li et al., 2019), in dietary fiber utilization by the gut microbiome (Berry et al., 2013; Mordant et al., 2025), and in the assimilation of host-derived substrates such as mucin glycans (Berry et al., 2013).

The Sipros-based proteomic SIP studies summarized in Figure 1.1 illustrate how this strategy matured from controlled algorithm-development settings into ecological and host-associated microbiome applications. The original Sipros framework enabled enrichment-resolved peptide identification and isotope-flow quantification in microbial community proteomes (Pan et al., 2011); subsequent work applied the same core strategy to ^{15}N - and ^2H -labeled acid mine drainage biofilms (Justice et al., 2014), genome-resolved ^{13}C transfer from $^{13}\text{CO}_2$ and ^{13}C -methanol into rhizosphere microorganisms (Li et al., 2019), and time-resolved isotope incorporation in gut micro-

biome metaproteomes (Smyth et al., 2020; Xiong et al., 2026). An in-house, unpublished benchmark uses $^{15}\text{NH}_4\text{Cl}$ spanning 1–99 atom% ^{15}N to label ammonia-oxidizing archaea (AOA) and ammonia-oxidizing bacteria (AOB), providing defined nitrogen-enrichment samples for evaluating enrichment-resolved peptide identification. In the marine proteomic-SIP experiment, ^{13}C -labeled exudates and lysates prepared from the diatom *Thalassiosira pseudonana* and the cyanobacterium *Synechococcus* sp. WH8102 were added separately to coastal-seawater microcosms; isotope incorporation into newly synthesized community proteins then identified lineage-specific assimilation of phytoplankton-derived dissolved organic matter (Kieft et al., 2021). Together with the defined *E. coli* enrichment standards used to benchmark Sipros 4, these examples motivate my focus on both accurate isotope-enrichment estimation and scalable database searching across heterogeneous SIP designs (Xiong et al., 2024).

Algorithms for proteomic SIP and their limitations

Three open-source algorithms have dominated proteomic SIP analysis over the last decade. **Sipros**, originally developed by Pan et al. (2011) and extended through Sipros Ensemble (Guo et al., 2018), performs enrichment-resolved database searches by scoring every candidate peptide-spectrum match at 101 isotopic enrichment levels (0–100 atom% in 1% increments). **MetaProSIP** (Sachsenberg et al., 2015) re-fits the precursor isotopic envelopes of peptides already identified by a label-free search to infer their atomic-percent enrichment. **Calisp** (Kleiner et al., 2018) similarly estimates atom% from precursor envelopes of pre-identified peptides and adds species-resolved biomass assignment.

In practice, three problems have limited the broad adoption of proteomic SIP. First, the *computational cost*: Sipros 3 performed the enrichment-resolved search by polynomial expansion of the isotopic distribution at each atom% level, which scaled poorly to deep metaproteomes and could require more than 12 hours of compute time per sample on a 24-core server (Pan et al., 2011). Second, *sensitivity at intermediate enrichment*: near 50 atom% ^{13}C , the precursor isotopic envelope is at its broadest; signal is distributed across many isotopologues and effective intensity drops, de-

SIPROS-BASED PROTEOMIC SIP STUDY

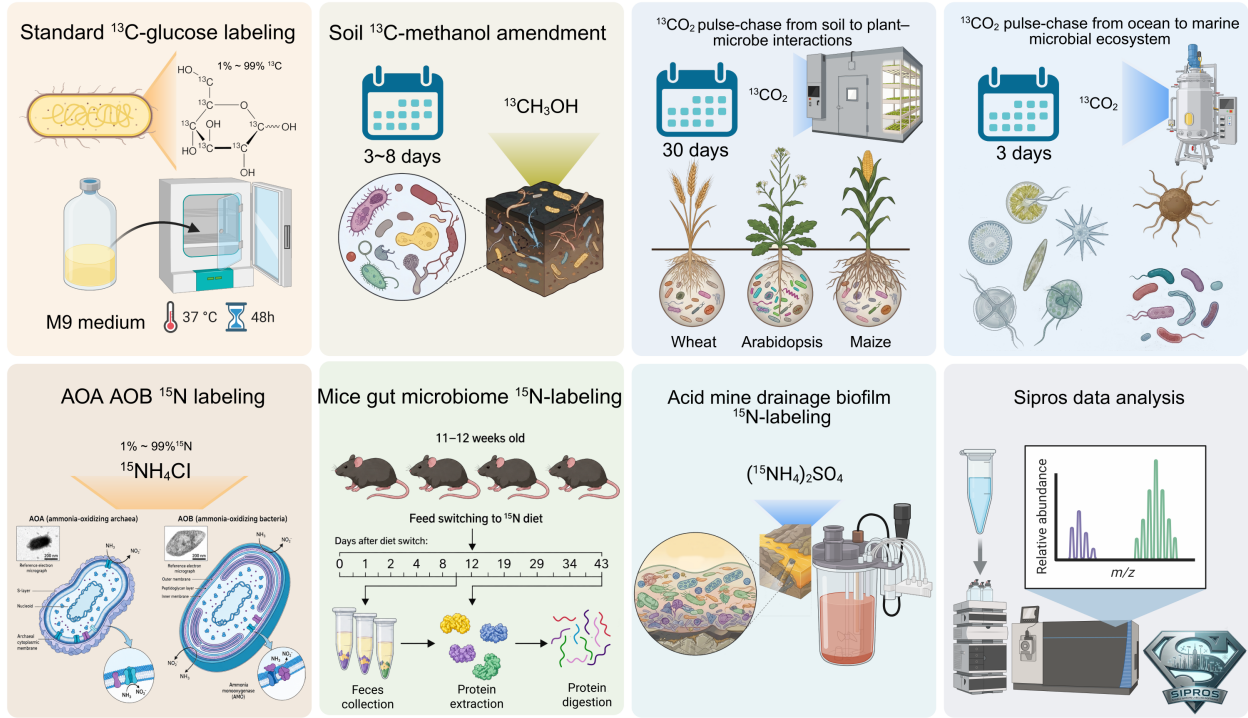


Figure 1.1: Overview of Sipros-based proteomic SIP study designs. From left to right, the top row shows: (1) defined *E. coli* cultures grown for 48 h in M9 medium with ^{13}C -glucose spanning 1.07–99 atom% ^{13}C , providing standards for enrichment-resolved peptide identification (Xiong et al., 2024); (2) soil communities amended with ^{13}C -methanol and sampled after 3 and 8 days to identify active methylotrophs and labeled metabolic pathways (Li et al., 2019; Xiong et al., 2024); (3) a 30-day plant–soil pulse–chase in which wheat, *Arabidopsis*, and maize fixed $^{13}\text{CO}_2$ and transferred labeled photosynthate to their rhizosphere microorganisms (Li et al., 2019); and (4) a marine experiment in which ^{13}C -labeled exudates and lysates from the diatom *Thalassiosira pseudonana* and the cyanobacterium *Synechococcus* sp. WH8102 were supplied separately to coastal-seawater microcosms, allowing proteomic SIP to resolve assimilation by heterotrophic taxa (Kieft et al., 2021). From left to right, the bottom row shows: (5) an in-house, unpublished benchmark in which ammonia-oxidizing archaea (AOA) and ammonia-oxidizing bacteria (AOB) were labeled with $^{15}\text{NH}_4\text{Cl}$ across 1–99 atom% ^{15}N ; (6) a 43-day ^{15}N diet-switch experiment in mice, with serial fecal sampling used to measure microbial and host protein labeling and turnover (Smyth et al., 2020; Xiong et al., 2026); (7) an acid mine drainage biofilm bioreactor supplied with $(^{15}\text{NH}_4)_2\text{SO}_4$ to trace nitrogen incorporation into community proteins (Justice et al., 2014); and (8) the shared analytical workflow of protein extraction and digestion, LC–MS/MS acquisition, and enrichment-resolved Sipros analysis of labeled peptide spectra (Pan et al., 2011; Xiong et al., 2024, 2026).

pressing identification rates for exactly the enrichment levels that are most biologically informative (von Bergen et al., 2013). Third, *limited acquisition support*: all three legacy tools were designed for data-dependent acquisition (DDA), in which stochastic precursor selection limits reproducibility and proteome coverage. Modern data-independent acquisition (DIA) on instruments such as the Thermo Orbitrap Astral provides systematic precursor sampling and deeper coverage (Heil et al., 2023; Guzman et al., 2024; Stewart et al., 2023), but SIP-specific algorithms are required because fixed isolation windows can clip parts of broadened isotopic envelopes and distort the apparent enrichment of MS2 fragments.

A fourth, often overlooked, gap is the absence of accessible visualization and manual-annotation tools for labeled spectra. Existing proteomic visualizers such as PDV do not allow user-defined chemical formulas or arbitrary isotopic enrichments and therefore cannot render the isotopic fine structure that is diagnostic of correct SIP assignments (Xiong et al., 2025).

Modern mass spectrometry: from DDA Orbitrap to narrow-window DIA Astral

The instrumental landscape underlying proteomic SIP has changed substantially during the timeline of this dissertation. The Orbitrap mass analyzer remains the workhorse of high-resolution shotgun proteomics, and its DDA mode is the basis for most existing SIP datasets. The introduction of the Thermo Orbitrap Astral hybrid in 2023 paired the high resolving power of the Orbitrap with the ultra-fast acquisition rate of the Astral mass analyzer, supporting narrow-window DIA at acquisition rates that quantify roughly five times more peptides per unit time than state-of-the-art legacy Orbitraps (Stewart et al., 2023; Heil et al., 2023; Guzman et al., 2024). For proteomic SIP, the move from DDA to narrow-window DIA is doubly attractive: DIA’s increased coverage partially compensates for the sensitivity loss that isotopic broadening imposes on intermediate-enrichment peptides, and the systematic precursor sampling reduces the stochastic dropout that has plagued comparisons between labeled and unlabeled controls. However, capturing these benefits requires new algorithms that can reason about precursor isolation windows and chimeric MS2 spectra in the presence of broadened SIP envelopes.

Goals and structure of this dissertation

The work described in this dissertation advances proteomic and metaproteomic SIP through a coordinated suite of tools that address the bottlenecks summarized above. The three data chapters can be read independently, but together they form an integrated analytical pipeline for isotope-resolved microbial proteomics. Chapter 2 (Xiong et al., 2024, *Microbiome*, 2024) presents Sipros 4. Chapter 3 extends the Sipros framework to a unified DDA/DIA platform, Sipros 5, and demonstrates its use on ^{13}C -glucose and time-resolved ^{13}C -inulin gut microbiome experiments together with a 43-day ^{15}N -SILAM mouse host–microbiome study. Chapter 4 (Xiong et al., 2025, *Analytical Chemistry*, 2025) introduces Aerith, an R package that simulates SIP-labeled isotopic fine structures and envelopes for arbitrary peptide sequences and chemical formulas. Chapter 5 synthesizes conclusions and outlines future directions.

Proteomic Stable Isotope Probing with an Upgraded Sipros Algorithm for Improved Identification and Quantification of Isotopically Labeled Proteins

Published as Xiong et al. (2024) in Microbiome 12, 148 (2024).

Reproduced with permission under the journal's open-access policy.

Abstract

Proteomic stable isotope probing (SIP) is used in microbial ecology to trace a non-radioactive isotope from a labeled substrate into de novo synthesized proteins in specific populations that are actively assimilating and metabolizing the substrate in a complex microbial community. The Sipros algorithm is used in proteomic SIP to identify variably labeled proteins and quantify their isotopic enrichment levels by performing enrichment-resolved database searching. In this study, we upgraded the Sipros algorithm to improve labeled protein identification, isotopic enrichment quantification, and database searching speed. The new Sipros 4 was compared with existing Sipros 3, Calisp, and MetaProSIP in terms of identifications and accuracy/precision of atom% quantification on both peptide and protein levels using standard *E. coli* cultures with 1.07, 2, 5, 25, 50, and 99 atom% ^{13}C enrichment. Sipros 4 outperformed Calisp and MetaProSIP across all samples, especially in samples with ≥ 5 atom% ^{13}C labeling. Additionally, the computational speed of Sipros 4 was over twenty times higher than Sipros 3 and was on par with the overall speed of Calisp and MetaProSIP-based pipelines. Sipros 4 also demonstrated higher sensitivity for the detection of labeled proteins in two ^{13}C -SIP experiments on real-world soil communities. The taxonomic and functional information of the labeled proteins provided rich biological insights into the soil community metabolism.

Introduction

Stable isotope probing (SIP) is a molecular method to identify which microorganisms within a complex community are actively assimilating a specific substrate labeled with a stable isotope such as ^{13}C , ^{15}N , or ^2H . It has been used to study biomass decomposition processes in a variety of ecosystems, such as digestion of dietary nutrients by the mouse gut microbiome (Berry et al., 2013), degradation of lignocellulose in soils, turnover of plant exudates in the rhizosphere (Li et al., 2019), and substrate switching captured by metaproteomics-based stable isotope fingerprinting (Mordant et al., 2025).

A variety of SIP methods have been developed to trace a stable isotope from a labeled substrate into the biomass of microorganisms that have assimilated the substrate. Nucleic-acid SIP (DNA- and RNA-SIP) involves the isolation of isotopically enriched DNAs or RNAs from SIP-labeled microorganisms using density-gradient ultracentrifugation (Radajewski et al., 2000; Manefield et al., 2002); sequencing of these nucleic acids can reveal the taxonomy of active microorganisms and reconstruct their genomes (Coyotzi et al., 2016). The isotopic labeling of phospholipid-derived fatty acids (PLFAs) in microbial communities can be measured using gas chromatography–mass spectrometry (GC–MS) in a PLFA-SIP approach. Unlike DNA- or RNA-SIP, the enrichment levels of PLFAs may be accurately quantified by mass spectrometry; however, labeled PLFAs can only be linked to very broad taxonomic categories of organisms, such as Gram-negative vs. Gram-positive bacteria.

Labeled proteins in a microbial community provide an alternative target for SIP analysis. The first protein-based SIP approach relied on two-dimensional gel electrophoresis of a SIP-labeled proteome and a parallel unlabeled proteome (Jehmlich et al., 2008): labeled protein spots were identified by matching them to corresponding unlabeled protein spots at the same gel positions, and the enrichment levels of the labeled proteins were quantified from their mass spectra. Subsequent developments shifted to shotgun proteomics with LC–MS/MS, enabling community-scale measurements (Pan et al., 2011; Sachsenberg et al., 2015; Kleiner et al., 2018).

Our recent ^{13}C SIP study of soil communities (Li et al., 2019) highlighted the high computational cost of SIP searches by Sipros 3 and the difficulty of finding labeled peptides in extremely complex communities. In this study, we upgraded the Sipros algorithm to overcome these two challenges. Sipros 4 was more than twenty-fold faster than Sipros 3 and identified more labeled proteins from SIP samples. The performance of Sipros 4 was compared with Sipros 3, Calisp, and MetaProSIP using *E. coli* cultures with predefined ^{13}C atom% levels and ^{13}C -methanol- and $^{13}\text{CO}_2$ -amended soil communities. The biological significance of the Sipros 4 SIP results was illustrated by the detection of ^{13}C -labeled metabolic enzymes in the soil community samples.

Results

Validation of the proteomic SIP performance using standard E. coli proteomes

Peptides collected from triplicate *E. coli* cultures grown under ^{13}C -SIP labeling conditions at six pre-defined atom% levels (1.07, 2, 5, 25, 50, and 99 atom% ^{13}C) were analyzed with LC-MS/MS, which produced an average of 133,698 MS2 scans for each pre-defined atom% level. Sipros was used to compare each observed MS2 spectrum with the theoretical spectrum of every peptide candidate at every enrichment level (0–100 atom% in 1 atom% increments) and to report the peptide-spectrum match (PSM) with the best score.

The 50-atom% ^{13}C -labeled *E. coli* proteome posed the greatest challenge: a single peptide candidate is represented by many isotopologues whose individual abundances are small. Sipros 4 explicitly models the broadened envelopes at the MS2 level via a vectorized convolution algorithm and rewards the matching of multiple isotopic peaks rather than only the monoisotopic peak. Sipros 4 identified and enrichment-quantified 64,113 PSMs in the 1.07-atom% samples, 65,131 PSMs in the 2-atom% samples, 46,665 PSMs in the 5-atom% samples, 43,526 PSMs in the 25-atom% samples, 37,659 PSMs in the 50-atom% samples, and 26,583 PSMs in the 99-atom% samples, indicating deep coverage of the labeled proteomes across the entire atom% range.

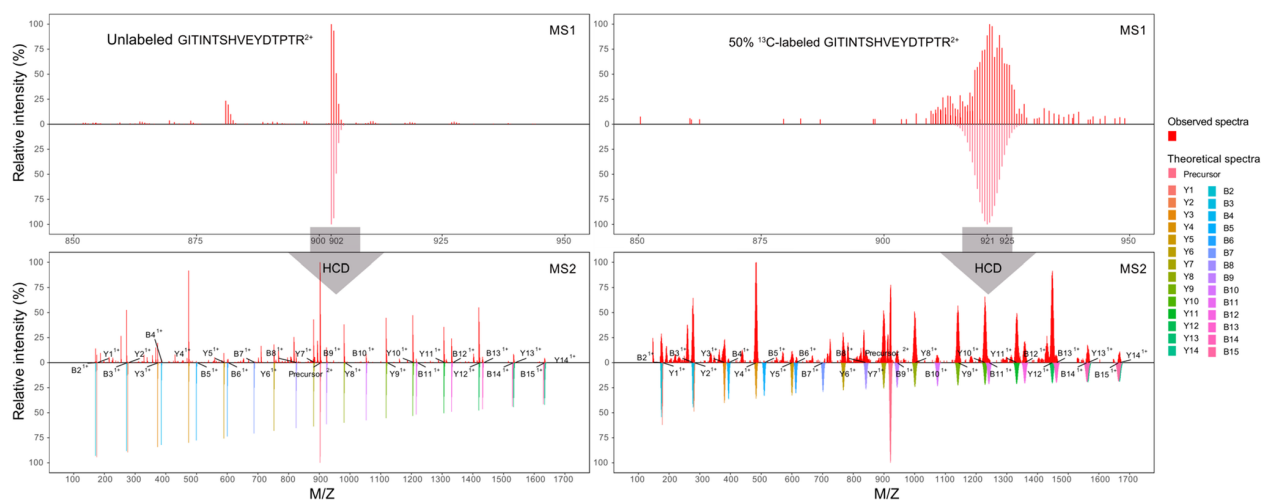


Figure 2.1: MS/MS measurement of a peptide's unlabeled isotopologue with 1.07 atom% ^{13}C and its labeled isotopologue with 50 atom% ^{13}C . The MS1 scans (upper panels) and MS2 scans (lower panels) are shown for the unlabeled isotopologue (left panels) and the 50 atom% ^{13}C -labeled isotopologue (right panels). Higher ^{13}C enrichment shifts and broadens the isotopic envelopes of both precursor and fragment ions.

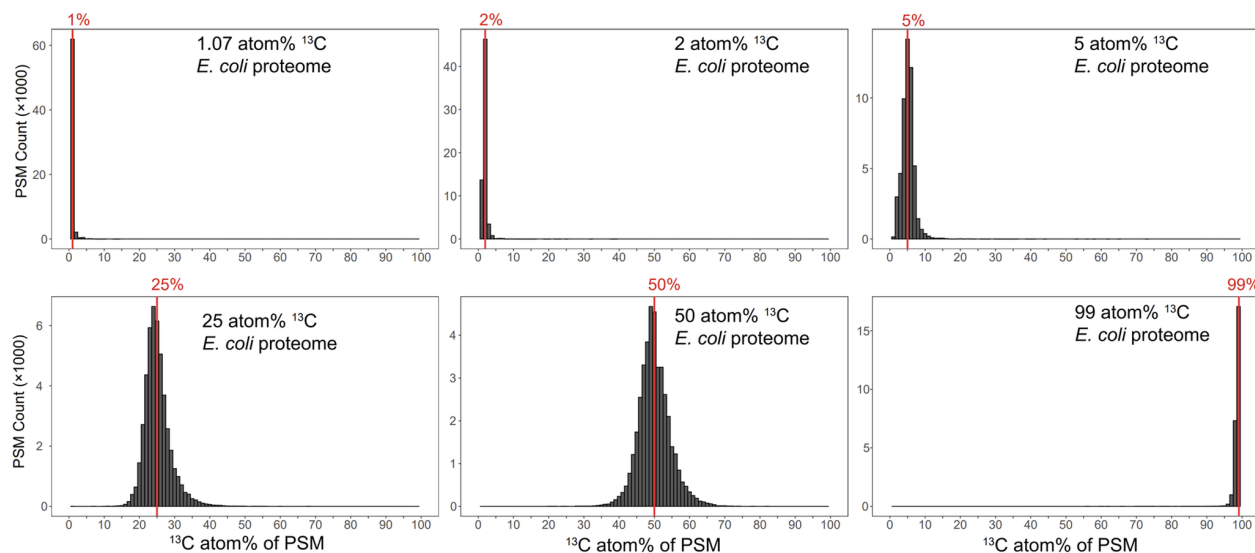


Figure 2.2: Accuracy and precision of ^{13}C atom% quantification by Sipros 4 on the PSM level for standard *E. coli* samples. The atom% estimates for all PSMs identified by Sipros 4 in each *E. coli* proteome are shown in its corresponding histogram with 1-atom% bin width. The medians of the atom% estimates were 1%, 2%, 5%, 25%, 50%, and 99% in the six samples, matching the expected enrichment levels exactly.

Comparison of the proteomic SIP performance using standard *E. coli* samples

Sipros 4 was compared with Sipros 3, Calisp, and MetaProSIP using the standard *E. coli* samples (Table 2.1). Because Calisp and MetaProSIP do not provide results at the PSM level, the performances of the four algorithms were compared at the peptide and protein levels. Calisp produced accurate atom% medians in the three samples for which it ran successfully, at 1.07, 2, and 5 atom% ^{13}C , but failed at 25 atom% or higher because Proteome Discoverer was unable to identify proteins in those samples. MetaProSIP severely under-estimated the atom% of proteins in the four samples other than the unlabeled and 99% ^{13}C -labeled samples. The median atom% values estimated by Sipros 4 aligned with the expected values in all six samples, and the median absolute deviation (MAD) of atom% estimates by Sipros 4 was lower than Calisp and MetaProSIP across all samples.

Table 2.1: Comparison of Sipros 3, Sipros 4, Calisp, and MetaProSIP on *E. coli* standard samples (peptide level). For each tool, median ^{13}C atom%, the median absolute deviation (MAD) of atom%, and the count of identifications are shown at each expected enrichment. NA = not available because database searching failed to produce any identification.

Metric	Tool	1.07%	2%	5%	25%	50%	99%
Median atom%	Sipros 4	1.1%	2.0%	5.0%	25.0%	50.0%	99.0%
	Sipros 3	1.0%	1.0%	1.0%	25.0%	51.0%	99.0%
	Calisp	1.1%	2.0%	4.9%	NA	NA	NA
	MetaProSIP	0.8%	1.2%	5.3%	10.2%	11.8%	97.7%
MAD atom%	Sipros 4	0.0%	0.0%	1.5%	1.5%	3.0%	0.0%
	Sipros 3	0.0%	0.0%	0.0%	4.4%	5.2%	0.0%
	Calisp	1.2%	1.3%	1.5%	NA	NA	NA
	MetaProSIP	0.3%	0.1%	2.7%	5.7%	7.6%	1.0%
Peptide count	Sipros 4	18,147	14,523	11,297	8,188	7,275	9,901
	Sipros 3	18,003	14,824	12,420	5,478	2,218	10,042
	Calisp	5,259	6,936	6,137	NA	NA	NA
	MetaProSIP	11,695	10,131	5,741	1,003	730	85
Protein/group count	Sipros 4	1,815	1,458	1,210	973	893	1,490
	Sipros 3	1,834	1,546	1,428	871	476	1,491
	Calisp	723	1,105	973	NA	NA	NA
	MetaProSIP	1,592	1,429	1,202	549	466	59

Sipros 4 outperformed Sipros 3 in atom% quantification accuracy for the 2-atom% and 5-atom% ^{13}C -labeled samples, as well as in the atom% precision and the proteome coverage for the 50-atom% sample. Notably, Sipros 4 identified approximately 2 $\times$ and 4 $\times$ more PSMs than Sipros 3 at 25-atom% and 50-atom% ^{13}C , respectively. Code optimization for Sipros 4 increased the computational efficiency of the enrichment-resolved database searching. Sipros 4 used ~ 0.5 hours and Sipros 3 used more than 12 hours to search the standard *E. coli* datasets on a 24-core

compute server.

Comparison using ^{13}C SIP soil communities

The performance of the four algorithms was benchmarked using a set of soil community samples analyzed in a previous SIP study (Li et al., 2019) where either ^{13}C -methanol or $^{13}\text{CO}_2$ was used as the labeling substrate (Table 2.2). These samples are substantially more complex and diverse than standard *E. coli* proteomes and are representative of real-world SIP applications.

Table 2.2: Number of ^{13}C -labeled PSMs, peptides, and proteins identified at ≥ 5 atom% ^{13}C in the initial soil (no labeling), ^{13}C -methanol SIP soil, and $^{13}\text{CO}_2$ SIP rhizosphere soil. FDRs at peptide and protein level were controlled at 1 %. NA = Calisp/MetaProSIP do not provide PSM identifications.

SIP	Origin	Algorithm	PSMs	Peptides	Proteins/groups
Initial soil	Microbes	Sipros 4	0	0	0
		Sipros 3	1	1	1
		Calisp	NA	0	0
		MetaProSIP	NA	0	0
^{13}C -methanol SIP	Microbes	Sipros 4	614	270	153
		Sipros 3	342	129	81
		Calisp	NA	18	15
		MetaProSIP	NA	13	13
$^{13}\text{CO}_2$ SIP	Microbes	Sipros 4	244	124	84
		Sipros 3	73	37	29
		Calisp	NA	19	19
		MetaProSIP	NA	1	1
	Plants	Sipros 4	161	36	26
		Sipros 3	69	20	15
		Calisp	NA	1	1
		MetaProSIP	NA	2	2

Biological analysis of the proteomic SIP results from ^{13}C SIP soil communities

For all the identified proteins in a proteomic SIP sample, Sipros 4 quantified both their isotopic enrichment levels in terms of atom% and their label abundances in terms of labeled spectral counts (Figure 2.3). In the ^{13}C -methanol SIP experiment, the soil community was sampled after 3 days and after 8 days of daily ^{13}C -methanol addition. 186 ^{13}C -labeled PSMs, 97 peptides, and 63 proteins/protein groups were identified in the day-3 samples. 428 ^{13}C -labeled PSMs, 255 peptides, and 135 proteins/protein groups were identified in the day-8 samples. The median ^{13}C enrichment levels of labeled proteins increased from 43.5 atom% on day 3 to 53.5 atom% on day 8.

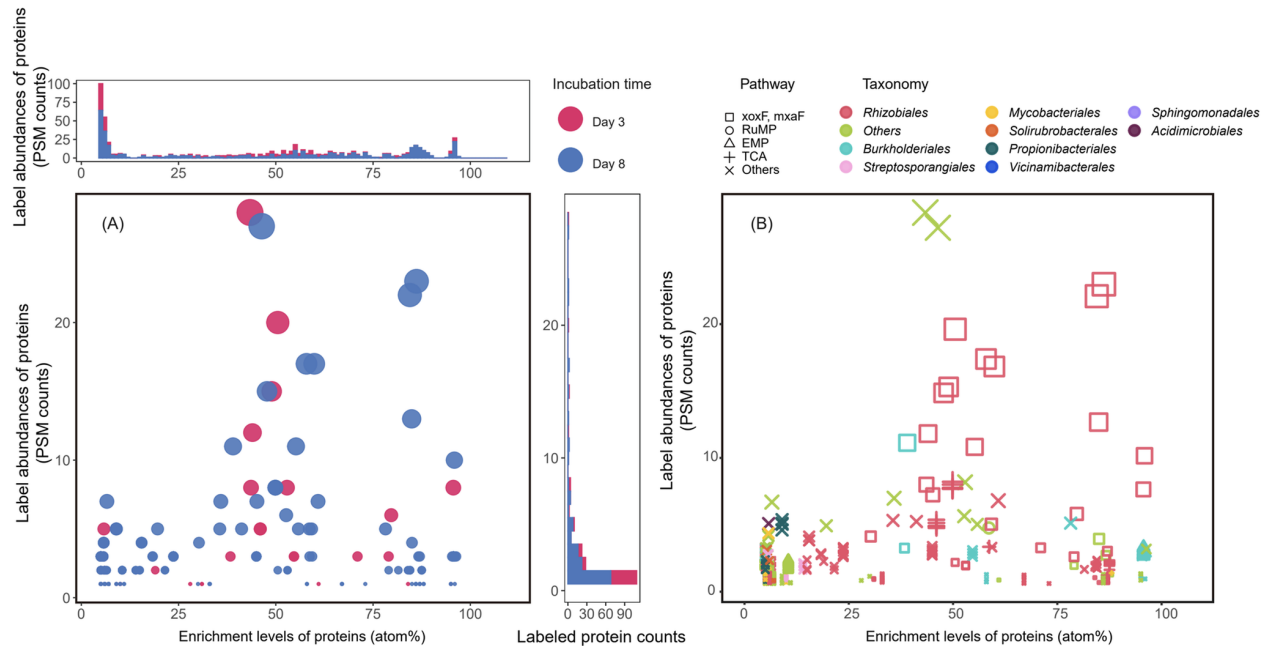


Figure 2.3: ^{13}C enrichment levels and label abundances of the proteins labeled by ^{13}C -methanol SIP. Both proteomic SIP scatterplots show all identified proteins, with their ^{13}C -labeled spectral counts on the x-axis and atom% ^{13}C on the y-axis. Coloring annotates phylum-level taxonomy and functional categories of the proteins. Day-3 and day-8 samples are shown; methanol-labeling identified methanol dehydrogenases and downstream C1-assimilation enzymes at high enrichment.

In the $^{13}\text{CO}_2$ SIP experiment, ^{13}C was expected to be fixed by the plants in their leaves, then transported to their roots, and finally transferred to the rhizosphere communities. Sipros 4 identified both rhizosphere microbial proteins and plant proteins from the rhizosphere soil samples (Figure 2.4). The rhizosphere communities yielded 244 ^{13}C -labeled PSMs, 124 labeled peptides,

and 84 labeled microbial proteins/protein groups; plant proteins from the rhizosphere yielded 161 labeled PSMs, 36 labeled peptides, and 26 labeled proteins/protein groups. The median enrichment levels of ^{13}C were 11 % for the labeled microbial proteins and 54 % for the labeled plant proteins.

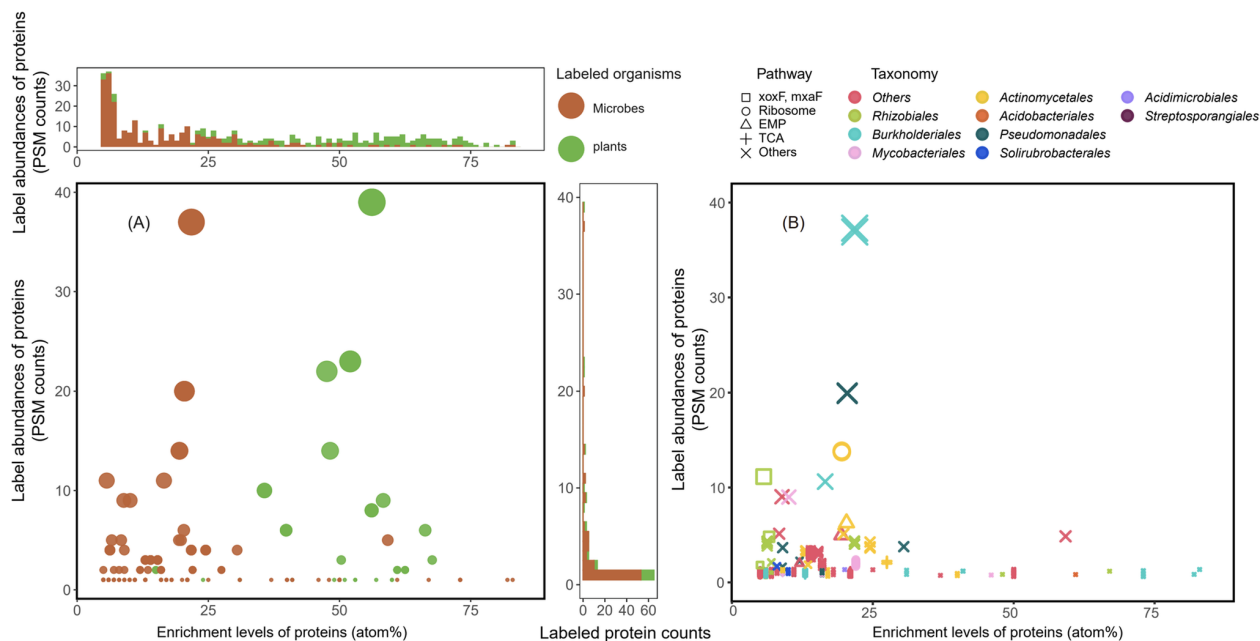


Figure 2.4: ^{13}C enrichment levels and label abundances of the proteins labeled by $^{13}\text{CO}_2$ SIP. Microbial proteins from the rhizosphere community and plant proteins are shown. Plant proteins carry higher enrichment levels (median 54%) than rhizosphere microbial proteins (median 11%), consistent with the expected fixation of $^{13}\text{CO}_2$ in plant leaves followed by transfer to rhizosphere communities.

The sequences of the labeled microbial proteins identified by Sipros 4 were used to infer their taxonomic origins and biological functions. In the ^{13}C -methanol SIP experiment, 136 labeled proteins/protein groups had phylum-level taxonomic assignments, including 77 Proteobacteria proteins/protein groups (391 total PSMs at 35% median ^{13}C enrichment), 47 Actinobacteriota proteins/protein groups (170 PSMs at 36% median ^{13}C enrichment), and 12 from other phyla. From genome-resolved binning—only 54 metagenome-assembled genomes (MAGs) were generated from the shallow soil metagenome sequencing (Li et al., 2019)—Sipros 4 identified ^{13}C -labeled unique proteins from 1 Rhizobiales MAG, 1 Sphingomonadales MAG, 1 Propionibacteriales MAG, and 3 additional MAGs from other Orders. In the $^{13}\text{CO}_2$ SIP experiment, Sipros 4 identified ^{13}C -labeled unique proteins from 2 Rhodospirillales MAGs and 4 MAGs from other

Orders.

The functional annotations of the large number of labeled proteins identified by Sipros 4 uncovered the de novo protein synthesis activities of the labeled microorganisms using the assimilated labeled substrates. In the ^{13}C -methanol SIP experiment, Sipros 4 labeled 22 methanol dehydrogenase (XoxF/MxaF) proteins/protein groups (258 PSMs at 59 % median ^{13}C enrichment) that convert methanol to formaldehyde (Macey et al., 2020; Keltjens et al., 2014; Picone and Op den Camp, 2019). Other labeled enzymes in formaldehyde assimilation included formaldehyde-activating enzymes (Fae), tetrahydromethanopterin/tetrahydrofolate-linked C1 transferases, and the downstream Embden–Meyerhof–Parnas (EMP), tricarboxylic acid (TCA), and serine/glycine cycle enzymes that consume the C1 carbon, along with ribosomal proteins of the labeled methylotrophs.

The $^{13}\text{CO}_2$ SIP results from Sipros 4 showed that $^{13}\text{CO}_2$ SIP labeled 6 XoxF/MxaF proteins/protein groups (21 PSMs at 6% median ^{13}C enrichment), 4 proteins/protein groups (14 PSMs at 20% median) in the EMP pathway, 7 proteins/protein groups (8 PSMs at 9% median) in the TCA, and 3 ribosomal proteins/protein groups (19 PSMs at 20% median). The low enrichment of microbial ^{13}C reflects the dilution of plant-fixed ^{13}C as it transits from leaves through roots to rhizosphere bacteria (Stewart and Metherell, 1999; Grayston et al., 1998; Haichar et al., 2008, 2014; Chen et al., 2019), while the high enrichment of plant proteins reflects direct photosynthetic incorporation. In a parallel reanalysis of the same rhizosphere data with the low-resolution ion-trap MS2 setting (Smyth et al., 2020), Sipros 4 retained the same labeled XoxF/MxaF identifications using the most intense isotopic peak per envelope.

The proteome of an organism reveals which biological processes it invests its scarce energy budget into (Lane and Martin, 2010; Lane, 2020). In a competitive community, an organism should make its energy investment decisions prudently based on the anticipated future return from its present investment in light of the perceived opportunities arising from its external environment. The ^{13}C -methanol SIP results showed that many methanol dehydrogenases (e.g., XoxF and MxaF) and other enzymes (e.g., *tal* and *fae*) involved in methanol utilization were labeled during their de novo synthesis after methanol amendments (Chen et al., 2020). Lanthanide-dependent methanol

dehydrogenases (XoxF type) from Rhizobiales or other Proteobacterial taxa have been reported to contribute to methanol degradation in soil (Keltjens et al., 2014; Picone and Op den Camp, 2019), demonstrating that the increased methanol availability prompted the organisms to invest in the biological processes to collect and consume methanol. The labeled proteome of the ^{13}C -methanol SIP and $^{13}\text{CO}_2$ SIP also included many housekeeping proteins, which can be attributed to the general growth of the source organisms.

Discussion

Our benchmarking results from the standard *E. coli* cultures and the natural soil samples showed that Sipros 4 identified more labeled PSMs, peptides, and proteins with greater atom% quantification precision and accuracy than the alternative algorithms Calisp and MetaProSIP. The benchmarks also demonstrated the unique capability of Sipros 4 to identify proteins with isotopic enrichment across the full range from 1 to 99 atom%, including the challenging intermediate enrichment levels where the isotopic envelopes are broadest.

The performance differences between algorithms originate in two distinct design choices. Calisp and MetaProSIP both use the isotopic envelope of the precursor ion in the MS1 scan to estimate atom%, whereas Sipros 3 and Sipros 4 estimate atom% from the isotopic envelopes of all observed fragment ions in the MS2 scan. Because MS2 spectra in DDA acquisitions typically include many fragment ions for a single peptide, the MS2 envelope provides richer evidence for quantification than the MS1 envelope alone. Sipros 4 additionally uses an updated scoring function that rewards both mass accuracy and isotopic envelope agreement, with explicit penalties for charge-state mismatch and missing complementary fragments.

A drawback of Sipros 3, compared with Calisp and MetaProSIP, had been its higher computational cost from enrichment-resolved searches at 101 atom% levels. To address this, we systematically profiled and optimized the Sipros codebase. In addition to the multi-node process-level parallelism and the multi-core thread-level parallelism that Sipros 3 already used, Sipros 4 adds instruction-level vectorization via single instruction, multiple data (SIMD) intrinsics and replaces

the polynomial-expansion algorithm with a convolution-based algorithm for computing theoretical isotopic envelopes of B and Y fragments. These optimizations together delivered the observed twenty-fold speedup.

Proteomic SIP quantifies both the label abundances and the enrichment levels of the labeled proteins and, by extension, their source organisms. The enrichment level reflects the percentage of the labeled substrate, relative to the unlabeled background substrates, that were assimilated and used by an organism for amino acid synthesis. The label abundances measure the relative abundances of labeled proteins and labeled organisms in terms of labeled PSM counts. Each of these axes carries distinct biological information: the enrichment level constrains the biochemistry of substrate assimilation, while the label abundance constrains the magnitude of de novo translational activity.

Conclusions

Our benchmarking tests demonstrated the high performance and computational efficiency of Sipros 4 for sensitive detection of labeled proteins and accurate quantification of their enrichment levels in SIP experiments. The label abundances and enrichment levels of the labeled proteins provided rich taxonomic and functional information about their source organisms. Analyses of real-world SIP experiments on soil communities labeled with ^{13}C -methanol and $^{13}\text{CO}_2$ showed that Sipros 4 enables proteomic SIP at the scale and complexity of natural metaproteomes. These real-world analyses demonstrated the use of labeled proteins to identify microbial consumers of labeled substrates, reconstruct their genomes, and define their de novo protein-synthesis activities during SIP labeling. Continued development of analytical tools such as Sipros 4 expands our capacity to understand metabolic activities in complex microbial communities by directly linking substrate assimilation with phylogeny and function.

Materials and Methods

Preparation of E. coli standards with known ^{13}C atom%

Each *E. coli* culture with a pre-defined ^{13}C incorporation level was grown in defined M9 medium. First, 5 μL of *E. coli* DH5 α (New England Biolabs) was inoculated into 10 mL of lysogeny broth (LB) and incubated for 1 day at 37 °C in an incubator (Robbins Scientific). Next, 700 μL of the *E. coli*–LB culture solution was mixed with 300 μL of 50 % glycerin in a 1-mL centrifuge tube and stored at -80 °C as a new inoculant. Then, the 1-mL *E. coli* inoculant was washed twice with 1 mL of phosphate-buffered saline (PBS), and 5 μL of the washed cells was inoculated into 10 mL of a ^{13}C -labeled M9 growth medium in a 50-mL centrifuge tube in a biological safety cabinet (Thermo Scientific). Appendix Table A.6 lists the recipes of the M9 growth media at different ^{13}C atom%, including the amounts of ^{12}C -glucose and ^{13}C -glucose (D-Glucose- $^{13}\text{C}_6$, ≥ 99 ^{13}C atom%, Sigma-Aldrich) as the sole carbon source for bacterial growth. Finally, the ^{13}C -labeled culture was incubated at 37 °C for 2 days, reaching an optical density at 600 nm (OD_{600}) greater than 0.5. Three replicate cultures were grown for each ^{13}C atom% level.

Protein extraction and LC–MS/MS

The protein extraction was performed as described previously (Li et al., 2019) with minor modifications. A cell pellet in 5 mL of a ^{13}C -labeled *E. coli* culture was harvested immediately after centrifugation at 4 °C and $10,000 \times g$ for 1 min. The pellet was washed twice with 1 mL of 10 mM Tris–HCl buffer at pH 7.0 and resuspended in 0.5 mL of lysis buffer (100 mM Tris–HCl, 4 % sodium dodecyl sulfate [SDS], and 0.1 M freshly added dithiothreitol [DTT]). The cells were sonicated on ice for five cycles of 30 s each with pulses. After centrifugation at 4 °C and $10,000 \times g$ for 1 min, the supernatant was collected and mixed with 0.5 mL of chilled (-20 °C) 50 % trichloroacetic acid to reach a final concentration of 25 %. The tubes were stored at -20 °C overnight to precipitate the proteins. The samples were centrifuged at $20,800 \times g$ for 20 min to collect protein pellets, which were then washed twice with 1 mL of chilled 80 % acetone and once

with 1 mL of chilled acetone, with centrifugation at $20,800 \times g$ for 20 min between washes. The protein pellets were dried in a centrifugal evaporator and resolubilized in 585 μL of urea-Tris-HCl solution (8 M urea, 0.1 M Tris-HCl, pH 8.0). An aliquot of 3 μL of fresh 1 M DTT was added to the urea-Tris-HCl solution to reach a final concentration of ~ 5 mM DTT. The mixture was vortexed for 20 min to dissolve the protein. Bubbles were removed by centrifugation at $10,000 \times g$ for 10 min. Then, an aliquot of 12 μL of 1 M iodoacetamide solution was added to achieve a final concentration of 20 mM iodoacetamide. After vortexing for 10 s, the solution was incubated in the dark at room temperature for 30 min. After centrifugation at $10,000 \times g$ for 5 min, the supernatant was divided into three 1-mL tubes (200 μL each), and protein concentration was measured using the Pierce bicinchoninic acid (BCA) Protein Assay Kit (Thermo Scientific).

Protein digestion was performed using filter-aided sample preparation (FASP) (Wiśniewski et al., 2009) in a 1-mL 30-kDa ultrafiltration unit (Vivacon 500, Sartorius). Each sample aliquot with 50 μg of protein was digested overnight at 37°C with 2 μg of sequencing-grade modified trypsin (V5113, Promega). The peptide digest was desalted by Pierce Peptide Desalting Spin Columns (Thermo Scientific) and quantified by NanoDrop 2000 Spectrophotometers (Thermo Scientific). The peptide separation was performed by reverse-phase XSelect CSH C18 2.5 μm resin (Waters) on a 150×0.075 mm column using an UltiMate 3000 RSLCnano system (Thermo Scientific) with 1 μg of peptides. The peptides were eluted with a 90-min gradient from 98 % solution A and 2 % solution B to 65 % solution A and 35 % solution B (solution A: 0.1 % formic acid, 0.5 % acetonitrile, and 99.4 % water; solution B: 0.1 % formic acid and 99.9 % acetonitrile). The eluted peptides were ionized by electrospray (2.4 kV) and analyzed by an Orbitrap Fusion Tribrid mass spectrometer (Thermo Scientific) in DDA mode. MS1 data were acquired using the Orbitrap analyzer in profile mode at 120,000 resolution over m/z 375–1500. MS2 data were acquired using the Orbitrap analyzer in centroid mode at 30,000 resolution after higher-energy collisional dissociation (HCD) activation. The precursor isolation window for MS2 was set to 5 Da. Dynamic exclusion time was 20 s, isotope exclusion was enabled, and the isolation mass tolerance was 10 ppm. The HCD energy was 28 % for precursors with charge states +3 to +7 and precursors in m/z 375–650,

and 31 % for precursors with charge state +2 and m/z 650–1500. Precursors with an unknown charge state or charge $< +2$ or $> +7$ were excluded.

MS/MS data extraction

RAW files from the labeled *E. coli* analyses were converted to FT1 and FT2 files using Raxport (<https://github.com/theapanlab/Raxport.net>). Raxport was upgraded to support both Linux and Windows via the RawFileReader library. The $^{15}\text{NH}_4\text{Cl}$ -labeled acid mine drainage (AMD) datasets were obtained from ProteomeXchange (PXD041958), and soil community datasets from PXD011738 (unlabeled), PXD011739 (^{13}C -methanol), and PXD011740 ($^{13}\text{CO}_2$).

Algorithmic improvements in Sipros 4 for SIP searches

By default, Sipros 4 performs database searching across 101 atom% levels, ranging from 0 to 100 % in 1 % increments, as specified in the configuration files. Subsequently, the PSMs at these pre-defined integer atom% levels are filtered based on their scores to reach a chosen FDR level. Users may customize the atom% increment (e.g., 0.5 %) and the search range (e.g., 0 to 10 %) in the configuration file based on their experimental requirements. A PSM identifies a peptide at an atom% level that best explains the corresponding MS/MS spectrum. The most abundant isotopic mass of a peptide candidate is approximated by the sum of the most abundant isotopic masses of all its residues. A series of precursor mass tolerance windows is opened for a range of unit mass offsets from the measured MS/MS precursor mass. In Sipros 4, the unit mass offset range is customized according to the enrichment level of the SIP searches based on the simulation results using the poly-Averagine peptides (Senko et al., 1995). The size of the mass tolerance windows is configurable by users with a default value of ± 0.01 Da for Orbitrap mass spectrometers.

To reconstruct the theoretical spectrum of a peptide candidate at a given atom%, Sipros computes the isotopic envelopes of all B and Y ions from this peptide. To speed up this computation task, the polynomial expansion algorithm used in Sipros 3 was replaced with the convolution algorithm (Rockwood and Haimi, 2006) in Sipros 4. The convolution algorithm was vectorized using

single instruction multiple data (SIMD) provided by the `omp simd` directive in OpenMP 4.0. The SIMD parallelism in Sipros 4 accelerated the convolution computation on individual CPU cores on top of the thread-level parallelism on multi-core CPUs and the process-level parallelism across compute nodes implemented in Sipros 3.

The scoring function was optimized in Sipros 4 to improve PSM identification performance. The score for a PSM, p , is a sum of the scores of the n B/Y ions found in an observed MS/MS spectrum:

$$p = \sum_{k=1}^n s_k c_k h_k g_k \quad (2.1)$$

where, for the k th matched B/Y ion, h_k is the mass-accuracy score defined in Eq. 2.2, s_k is the isotopic envelope score defined in Eq. 2.3, c_k is the charge-state penalty, and g_k is the complementary-fragment penalty. c_k takes a value of 1 when the expected charge state matches the observed charge state; otherwise, it assumes a value of 0.5. g_k is assigned a value of 2 in the presence of the complementary fragment ion; otherwise, it is set to 1.

The mass-accuracy score of the k th matched B/Y ion, h_k , is defined as

$$h_k = 2 [1 - \Phi(0, t/2, m_k)] \quad (2.2)$$

where $\Phi(\cdot)$ is the cumulative density function (CDF) at the threshold of m_k of a normal distribution with mean 0 and standard deviation $t/2$, t is the user-defined fragment mass tolerance, and m_k is the observed average mass error of the isotopic peaks of the matched B/Y ion.

The isotopic envelope score of the k th matched B/Y ion, s_k , is

$$s_k = 1 + \sum_i e_i - \sum_j u_j \quad (2.3)$$

where e_i is the reward for finding an expected isotopic peak i (Eq. 2.4) and u_j is the penalty for

missing an expected isotopic peak j (Eq. 2.5).

$$e_i = g \left[1 - \operatorname{erf} \left(\frac{|x_i - y_i|}{\sqrt{x_i^2 + y_i^2}} \right) \right] \quad (2.4)$$

$$u_j = x_j [a + b(q - 50\%)^c] \quad (2.5)$$

where x is the expected relative intensity, y is the observed relative intensity matched within the mass error tolerance, $\operatorname{erf}(\cdot)$ is the Gauss error function (Oldham et al., 2009), and q is the isotopic atom% being searched. The constants a , b , c , and g were set to 0.005, 4, 8, and 0.5, respectively, based on heuristic optimization and are user-configurable in the search configuration file. The code and user manual of Sipros 4 were released at <https://github.com/thePanlab/Sipros4>.

Database searching of SIP samples by Sipros 4

The FT2 files of the *E. coli* samples at different ^{13}C atom% levels were searched against a target–decoy protein sequence database consisting of the *Escherichia coli* (strain K12) proteome from UniProt and the non-*E. coli* contaminant proteins from <https://www.thegpm.org/crap/>. The reverse sequences of these target proteins were added to the database as decoys. The mass error tolerance was set to 0.01 Da for precursors and fragments. The false discovery rate (FDR) of peptide identifications was controlled to 1 % by adjusting the score thresholds of PSMs. Protein identifications were filtered to reach 1 % FDR based on the highest PSM score for protein identification. At least one unique peptide was required for each identified protein/protein group.

The $^{15}\text{NH}_4\text{Cl}$ -labeled AMD datasets were processed similarly using Sipros 4. The target–decoy protein sequence database was constructed from the AMD metagenome assemblies (Justice et al., 2014). The SIP isotope was changed to ^{15}N . The mass error tolerance was set to 0.05 for precursors and 0.02 for fragments. The FDRs of peptides and proteins were all controlled to 1 % as described above. For the analysis of the low-resolution ion-trap MS2 data from the ^{15}N -labeled spiked mouse gut microbiome sample (Smyth et al., 2020), the mass error tolerance was set to 0.02 for precursors and 0.11 for fragments. In this low-resolution MS2 setting, Sipros used the most

intense peak within each isotopic envelope to score PSMs, as opposed to all isotopic peaks in the high-resolution MS2 setting.

Regular label-free searches were performed on the soil ^{13}C -methanol and $^{13}\text{CO}_2$ SIP datasets using Sipros Ensemble (Guo et al., 2018) against a protein database containing all predicted proteins from the soil metagenome assemblies. All identified proteins were used to construct the protein database for ^{13}C SIP searches using Sipros 4. Protein sequences of *Arabidopsis*, wheat, and maize were also added to the protein database for the $^{13}\text{CO}_2$ SIP datasets. The mass error tolerance was set to 0.05 for precursors and 0.02 for fragments. The MS/MS spectra containing high-density clusters of noise peaks (i.e., >255 peaks within any 250-wide m/z window of a spectrum) were removed.

The RAW files for the standard *E. coli* samples were uploaded to the ProteomeXchange repository under the accession PXD041414. The RAW files for the $^{15}\text{NH}_4\text{Cl}$ -labeled AMD community were generated in a previous study (Pan et al., 2011) and were uploaded to ProteomeXchange under PXD041958. The RAW files for the ^{13}C -methanol-labeled soil communities and the $^{13}\text{CO}_2$ -labeled soil communities (Li et al., 2019) were downloaded from ProteomeXchange under the accessions PXD011738 (unlabeled initial soil), PXD011739 (^{13}C -methanol-labeled soil), PXD011737 ($^{13}\text{CO}_2$ -labeled *Arabidopsis* rhizosphere soil), PXD011891 ($^{13}\text{CO}_2$ -labeled maize rhizosphere soil), and PXD011892 ($^{13}\text{CO}_2$ -labeled wheat rhizosphere soil).

The database search for the *E. coli* samples and the AMD samples was performed on a compute node equipped with dual 22-core Intel Xeon CPUs (Gold 6152) and 376 GB of system memory on the OSCER Schooner supercomputer, and on a server with a 24-core AMD Ryzen CPU (5965WX) and 512 GB of memory. The database search for the soil samples was completed on compute nodes equipped with dual 10-core Intel Xeon Haswell CPUs and 32 GB of memory on Schooner.

SIP analysis by Calisp and MetaProSIP

Calisp (Kleiner et al., 2018) and MetaProSIP (Sachsenberg et al., 2015) quantify the atom% of peptides previously identified by label-free database searching using Proteome Discoverer (for

Calisp) or Comet (for MetaProSIP) (Eng et al., 1994). RAW files were converted to mzML using ProteoWizard. Regular database searching was performed with Proteome Discoverer at default parameters for Calisp and through the OpenMS TOPP pipeline at default parameters for MetaProSIP. FDRs of identified PSMs, peptides, and proteins were controlled at 1 %, and the labeled identifications were exported for downstream comparison.

Analysis and visualization of SIP search results

Protein functional annotation used eggNOG-mapper (Huerta-Cepas et al., 2017) with GO, EC, and KEGG terms. Taxonomic assignment used annoTree with DIAMOND (Buchfink et al., 2015) and MEGAN6. A protein group was assigned to a taxon if more than 70 % of its member proteins were assigned to that taxon. Similarly, a protein group was assigned a functional annotation only when more than 70 % of its member proteins carried that functional assignment. Identified proteins were grouped to representative sequences within each protein group, and labeled spectral counts were summed across the group. For genome-resolved proteomic SIP analysis, a MAG was marked as labeled if at least one unique labeled peptide was identified from that MAG. Functional enrichment analysis of labeled proteins was performed using clusterProfiler; the phylogenetic tree of labeled microorganisms was visualized using ggtree. Student's *t*-test and Wilcoxon tests were performed in R 4.2.1.

Author contributions and funding

C.P., X.G., R.S.M., and Y.X. designed the study; Y.X. upgraded the Sipros software with assistance from X.G. and C.P.; Y.X. prepared the ¹³C-labeled *E. coli* cultures and conducted the performance benchmarking; S.F. validated the analysis reproducibility; Y.X. and C.P. drafted the manuscript; and all authors read and approved the final version. The authors thank Dr. Zhou Li for technical assistance with Sipros and SIP data, Dr. Timo Sachsenberg and Dr. Michael Strous for MetaProSIP and Calisp respectively, the IDeA National Resource for Quantitative Proteomics for the LC-MS service, and the OU Supercomputer Center for Education and Research (OSCER) for the

computational resources. This work was supported by the National Center for Complementary & Integrative Health and the National Institute of General Medical Sciences at the National Institutes of Health (R01AT011618). The source article declared that no new datasets were generated or analyzed during that study, ethics approval and consent to participate were not applicable, all authors approved publication, and the authors had no competing interests.

Supplementary Material

The supplementary material accompanying this chapter (Xiong et al., 2024) is reproduced or summarized in Appendix A. Appendix Figure A.1 shows Venn diagrams of the proteins identified by Sipros 3, Sipros 4, Calisp, and MetaProSIP on the standard *E. coli* samples, demonstrating that Sipros 4 captures the most identifications at every enrichment level. Appendix Figure A.2 is a taxonomic tree of the microbial proteins identified in the initial soils, the ^{13}C -methanol SIP soils, and the $^{13}\text{CO}_2$ SIP soils, with bar charts showing the numbers of unlabeled and labeled proteins per Order and a heatmap showing the average enrichment levels. Appendix Figures A.3 and A.4 provide functional analyses of the methanol and CO_2 SIP results, including KO and GO enrichment of the labeled proteins. Appendix Tables A.1–A.8 contain the corresponding benchmark, search-throughput, media, soil metaproteome, genome-resolved, and workbook-summary data.

Sipros 5 Significantly Enhances Proteomic Isotope Probing in DDA Orbitrap and Narrow-Window DIA Astral Mass Spectrometry Analyses

Manuscript under review (Xiong et al., 2026).

Abstract

Proteomic stable-isotope probing (SIP) can connect substrate assimilation to taxa, functions, and protein renewal, but complex samples and weak isotope incorporation limit confident labeled-peptide recovery. To address this bottleneck, we developed Sipros 5, a SIP proteomics platform with universal percentage-labeling support that models precursor and fragment isotope envelopes in data-dependent acquisition (DDA) and data-independent acquisition (DIA) data and controls the label-specific false discovery rate (FDR) using unlabeled negative controls. In defined ^{13}C -labeled *E. coli* standards spanning 1–99 % enrichment, Sipros 5 increased identifications over Sipros 4 while preserving isotope-abundance accuracy, with narrow-window Astral DIA further improving depth. In fecal peptide spike-ins, it retained label-FDR-controlled *E. coli* peptide-spectrum match (PSM), peptide, and protein recovery at 2 % ^{13}C in a 1:100 background. Having established broad, FDR-controlled recovery, we applied Sipros 5 to ^{13}C -glucose, ^{13}C -inulin, and ^{15}N diet datasets, resolving substrate assimilation and protein-turnover kinetics across microbial taxa and host–microbiome compartments. This workflow enables quantitative protein-renewal analysis across organisms and microbiome ecosystems.

Introduction

Stable isotope probing (SIP) links isotope incorporation to biological activity and is widely used to trace elemental fluxes through natural and host-associated systems. Proteomic SIP is especially powerful because it measures isotope incorporation at the peptide and protein levels, thereby con-

necting substrate assimilation to taxa and functions. A central challenge, however, is no longer simply whether isotope incorporation can be detected, but whether enough confidently labeled peptides can be recovered from complex samples to support quantitative biological interpretation.

Existing SIP proteomics workflows have established the feasibility of isotope-aware peptide and protein identification. Sipros 4 (Xiong et al., 2024), MetaProSIP (Sachsenberg et al., 2015), and Calisp (Kleiner et al., 2018) have supported studies of ^{13}C -glucose-labeled *E. coli*, ^{15}N -labeled acid mine drainage biofilms (Justice et al., 2014), ^{13}C -methanol amendment in bulk soil, $^{13}\text{CO}_2$ -derived carbon flux into rhizosphere communities (Li et al., 2019), and assimilation of ^{13}C -labeled phytoplankton exudates by marine microbial consortia (Kieft et al., 2021). Yet broad application remains constrained by the chemistry and acquisition physics of SIP data. Isotope incorporation broadens precursor and fragment-ion envelopes, most severely near intermediate enrichment ($\sim 50\%$ elemental abundance), distributing signal across many isotopologues and reducing effective intensity. In complex metaproteomes, this effect compounds incomplete protein catalogs, co-fragmentation, and background interference, making weakly labeled peptides difficult to distinguish from unlabeled or spuriously enriched identifications.

The acquisition landscape creates both a limitation and a new opportunity. In DDA, the instrument surveys an MS1 spectrum and then selects a limited set of abundant precursor ions for fragmentation; this stochastic selection can reduce reproducibility, proteome coverage, and quantitative consistency. In DIA, the instrument systematically fragments all ions within a series of predefined isolation windows, providing more consistent sampling and deeper coverage (Stewart et al., 2023; Heil et al., 2023; Guzman et al., 2024). SIP-specific DIA algorithms are nevertheless needed because a fixed isolation window can capture only part of a broadened isotope envelope, and SIP workflows require explicit control of labeling FDR. This gap is particularly important for gut microbiome metaproteomics, where SIP can move beyond cataloging expressed proteins toward measuring active substrate use and protein renewal in situ (Berry et al., 2013; Mordant et al., 2025).

Here we present Sipros 5, an optimized SIP proteomics platform that improves labeled-peptide

recovery and provides principled label-FDR control. We benchmark Sipros 5 using ^{13}C -labeled *E. coli* acquired by Orbitrap DDA and Astral DIA, evaluate specificity in ^{13}C -labeled *E. coli* spike-ins into mouse fecal peptides, and apply the workflow to mouse gut microbiome SIP experiments. In ^{13}C -glucose and time-resolved ^{13}C -inulin incubations of mouse cecal contents and a 43-day ^{15}N -SILAM experiment in mouse stool, Sipros 5 recovers large sets of labeled peptides and proteins, resolves their taxonomic and functional origins, and enables quantitative analysis of substrate assimilation and protein turnover at community scale.

Results

Validation of Sipros 5 using Orbitrap DDA and Astral DIA for proteomic SIP analysis

Sipros 5 implements an end-to-end workflow for proteomic SIP (Figure 3.1), including theoretical isotopic peak generation and matching across arbitrary ^{13}C enrichment levels, enrichment-aware feature extraction, PSM rescoring and filtering, and protein inference. Downstream validation modules include semi-supervised support vector machine (SVM) filtering for target/SIP-labeled and decoy/control PSMs, a convolutional neural network (CNN) model for SIP-labeled PSM confidence and isotope-enrichment prediction, and Philosopher-based protein assembly.

We validated Sipros 5 using proteomic datasets from *E. coli* cultured in M9 medium supplemented with ^{13}C -glucose at six enrichment levels (1, 2, 5, 25, 50, and 99%), acquired by Orbitrap DDA and Astral DIA. Across labeling conditions, identified PSMs concentrated near expected ^{13}C abundances and extended across precursor intensities from 2^{15} to 2^{30} (Figure 3.2). In Orbitrap DDA, Sipros 5 consistently produced $\sim 2\text{--}3\times$ more PSMs than Sipros 4 while maintaining comparable isotopic-abundance accuracy. The largest gains were observed with Astral DIA: Sipros 5, together with the ultra-fast Astral analyzer and DIA-specific processing, increased PSM identifications approximately fivefold ($\approx 130,000\text{--}330,000$ in a 35-min run) relative to Orbitrap DDA ($\approx 30,000\text{--}50,000$ in a 90-min run). Across all labeling conditions, Sipros 5–Astral DIA yielded the highest identification totals, with $\sim 13,000$ to $>30,000$ peptides and $\sim 1,300$ to $>2,300$ proteins.

As ^{13}C enrichment approaches the intermediate range, isotopic envelopes are maximally broad-

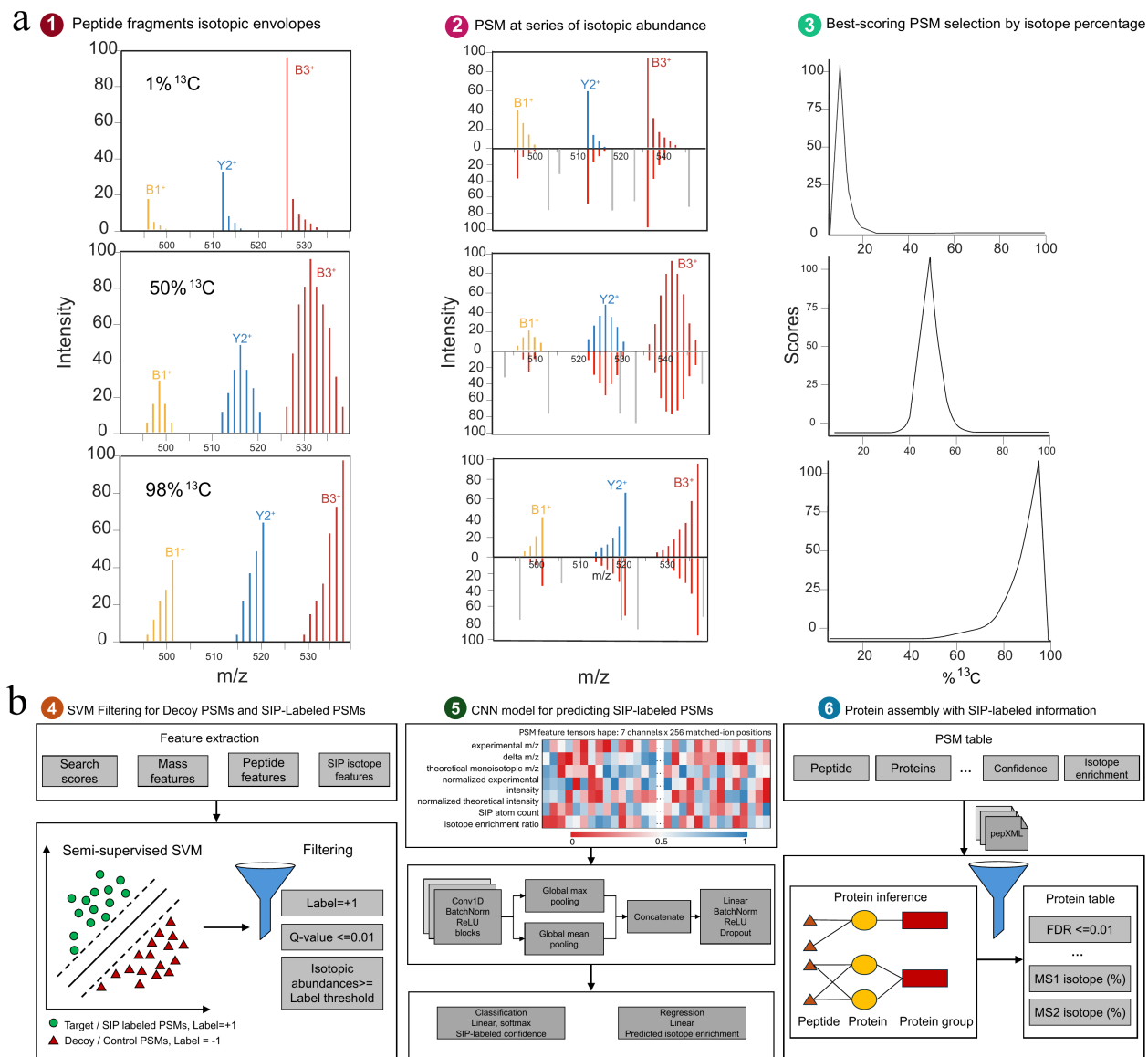


Figure 3.1: Overview of the Sipros 5 workflow and SIP-aware scoring modules. (a) Sipros 5 workflow for proteomic stable-isotope probing, including isotope-envelope modeling, feature extraction, PSM validation and protein-level inference. (b) SIP-aware validation modules, including semi-supervised SVM filtering for target/SIP-labeled and decoy/control PSMs, CNN-based prediction of SIP-labeled PSM confidence and isotope enrichment, and protein assembly with SIP-labeled peptide information.

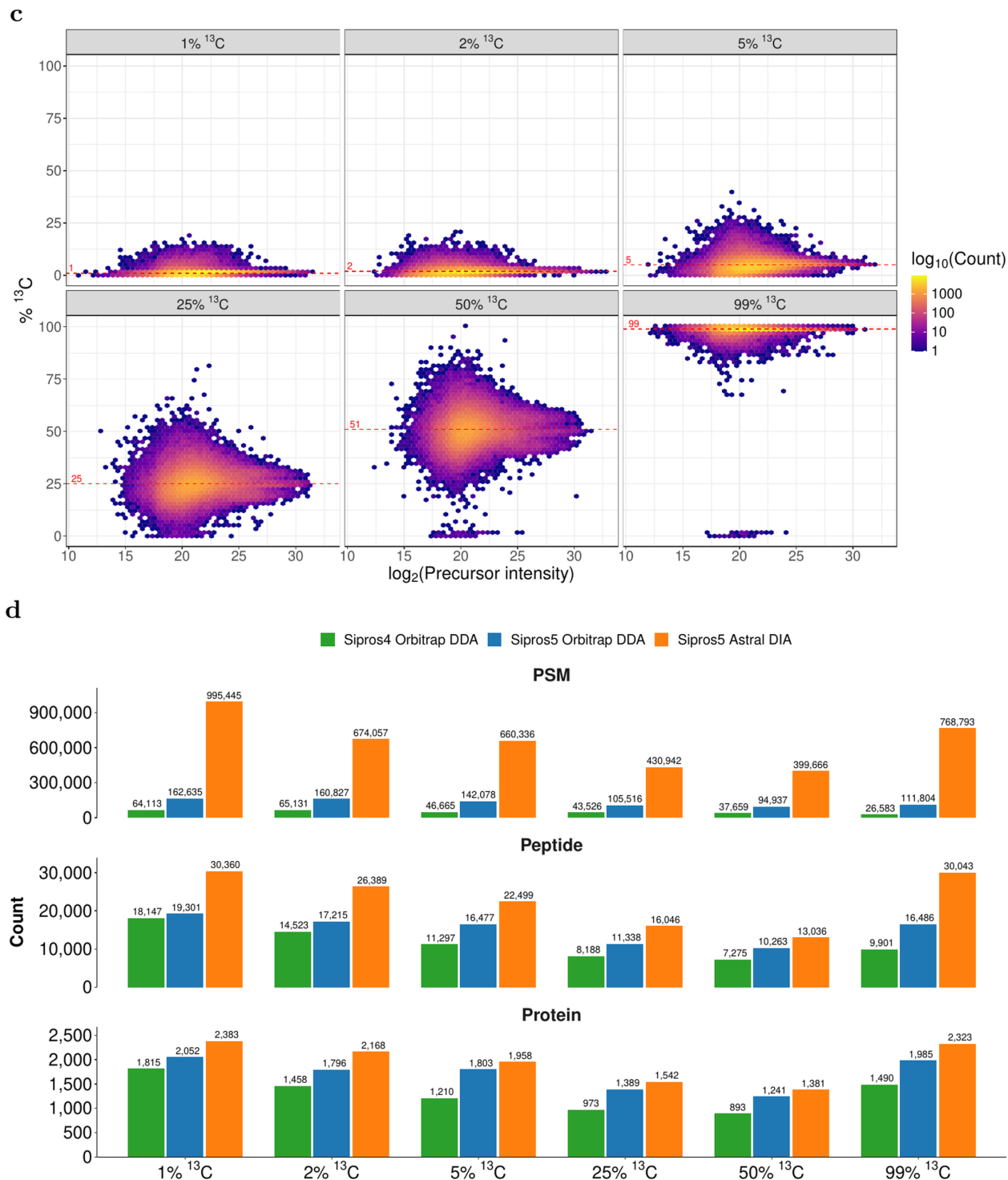


Figure 3.2: Sipros 5 benchmarking on ¹³C-glucose-labeled *E. coli*, continued. (c) Orbitrap DDA PSM density distributions across ¹³C enrichment levels, plotted by log₂ precursor intensity and inferred ¹³C abundance. (d) Identification yield summary for PSMs, peptides, and proteins across Sipros and acquisition workflows; Sipros 5 with Astral DIA consistently produces 2–3× more PSMs in Orbitrap DDA and approximately 5× more in Astral DIA than Sipros 4.

ened and peptide signal is distributed across the largest number of isotopologues, reducing effective spectrum intensity. The lowest identification yields appeared near 50% ^{13}C , marking this condition as the most difficult case. At 50% ^{13}C , replicate-level peptide-overlap analysis showed that Astral DIA yielded a larger reproducibly detected peptide core than Orbitrap DDA, with 7,679 peptides shared across all three Astral DIA replicates compared with 6,107 peptides shared across all three Orbitrap DDA replicates (Figure 3.3). The 4,420 peptides common to all six runs indicate substantial shared coverage, whereas the additional Astral-reproducible peptides highlight an advantage of DIA for repeatable peptide recovery under intermediate isotopic labeling.

e

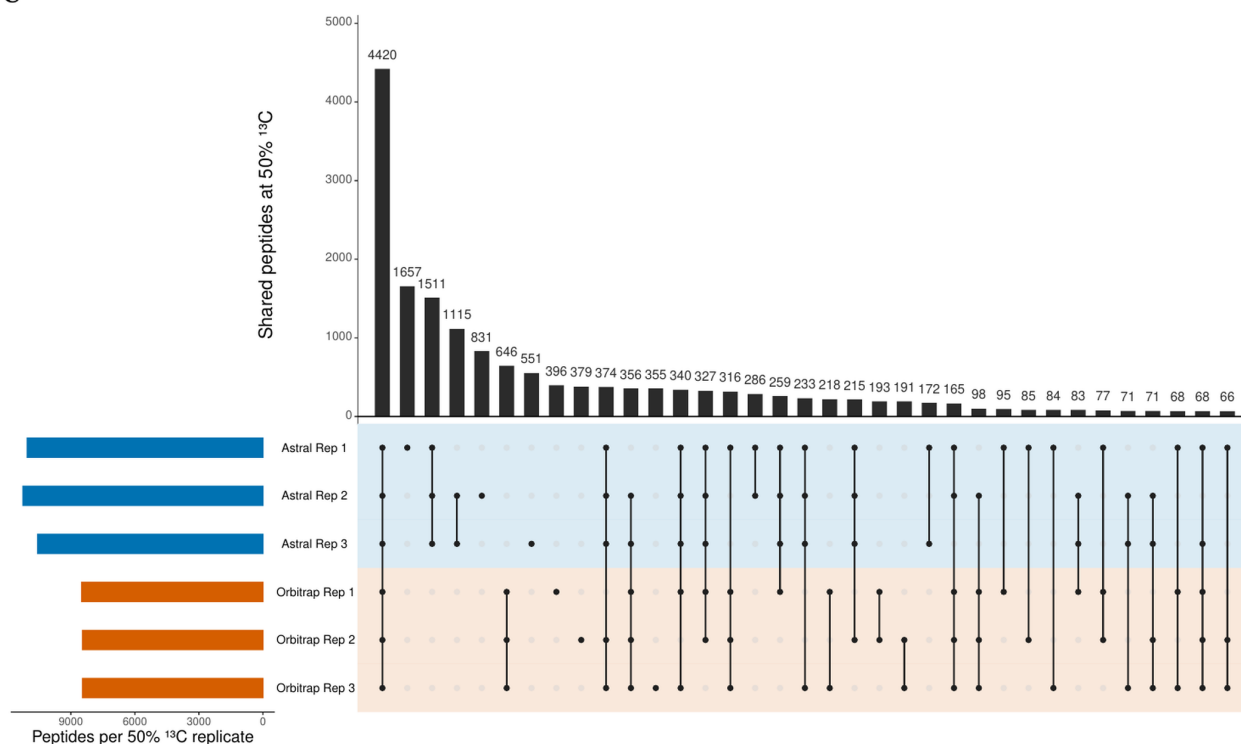


Figure 3.3: UpSet plot summarizing peptide-identification overlap among Astral DIA and Orbitrap DDA replicates at 50% ^{13}C labeling. 4,420 peptides were common to all six replicates; Astral DIA replicates contributed 6,000–9,000 peptides each, with 7,679 peptides shared across all three Astral runs vs. 6,107 across all three DDA runs.

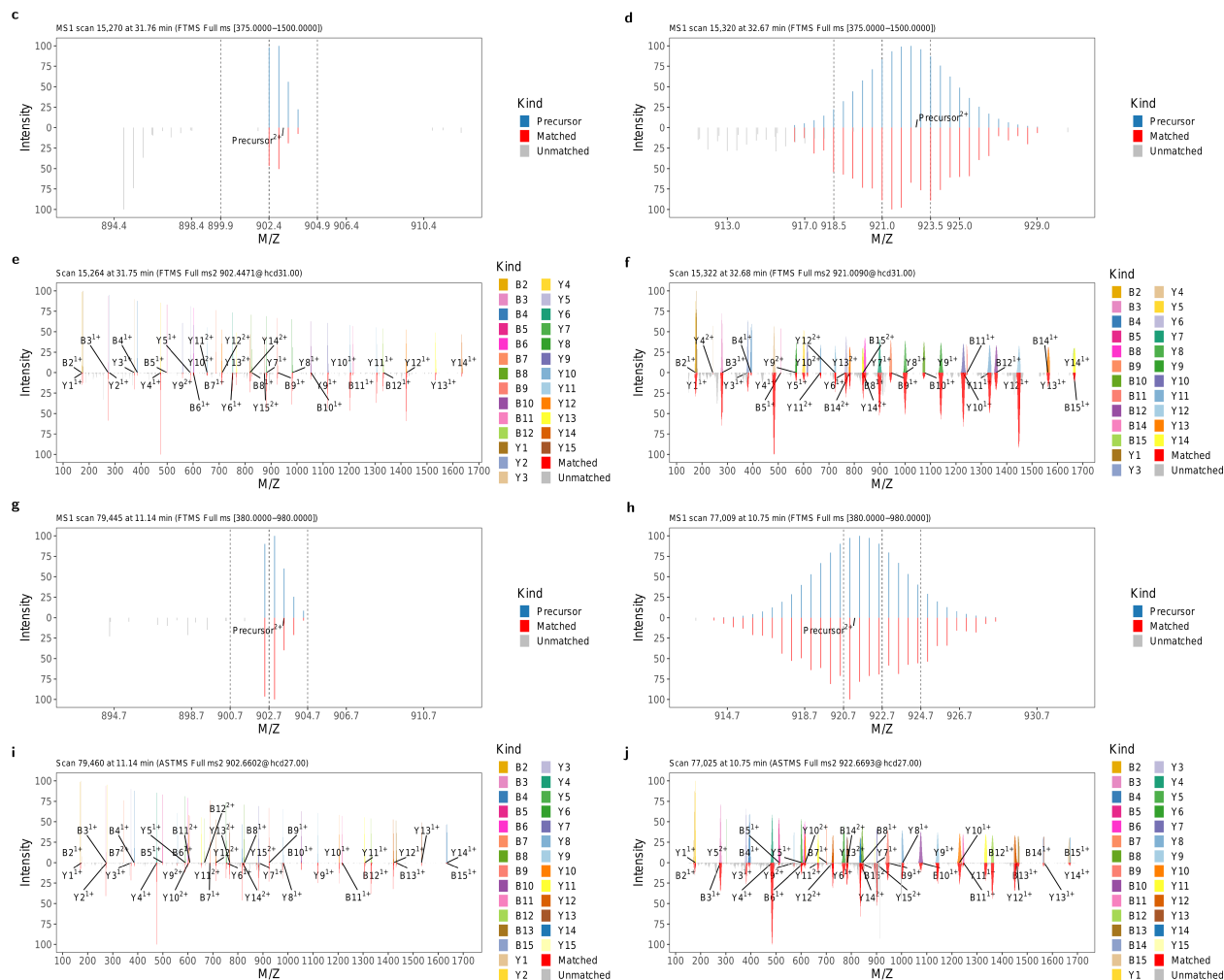


Figure 3.4: Visualization of Sipros 5 PSMs with 1% and 50% ^{13}C labeling acquired on Orbitrap DDA and Astral DIA. Representative MS1 spectra in panels (a), (b), (e), and (f) and MS2 spectra in panels (c), (d), (g), and (h) are shown for the identified peptide GITINTSHVEYDTPTR, acquired on Orbitrap (a–d) and Astral (e–h). Panels (a), (c), (e), and (g) correspond to 1% ^{13}C labeling, whereas panels (b), (d), (f), and (h) correspond to 50% ^{13}C labeling. In each panel, the upper trace shows the expected isotopic envelope for annotated b-, y-, and precursor ions, and the lower trace shows the measured isotopic envelope. In MS1 spectra, the isolation-window center and edges are indicated by dashed vertical lines. Fragment ions are labeled and color-coded by ion type; matched peaks are highlighted in red and unmatched peaks are shown in gray.

Spike-in validation in a complex fecal peptidome background

We tested the specificity of Sipros 5 on ^{13}C -labeled *E. coli* peptides spiked into a mouse fecal peptidome background at 1:10 and 1:100 (*E. coli*:fecal peptides, w/w). Three labeling levels were used (2, 5, and 50% ^{13}C), and SIP-aware q-value thresholding was applied for label-specific FDR control using unlabeled fecal peptides as negative controls (Figure 3.5).

Consistent with these trends, the Sipros 5 WinnowNet filter gave the strongest overall spike-in recovery (Table 3.1). Under 1% label-FDR control with the negative-control constraint, WinnowNet recovered 9,048, 21,922, and 2,222 *E. coli* PSMs at 2%, 5%, and 50% ^{13}C , corresponding to 2,310, 5,105, and 747 peptides, respectively. The retained identifications showed high *E. coli* assignment percentages (90.00–95.69% of PSMs and 76.62–92.11% of peptides), and peptide-level ^{13}C medians matched the nominal spike-in levels (2.00%, 5.00%, and 50.00%). At the most difficult tested condition—2% ^{13}C in a 1:100 *E. coli*:fecal background—Sipros 5 retained label-FDR-controlled *E. coli* PSM, peptide, and protein identifications. Astral DIA showed the same specificity gain after SIP-aware filtering: *E. coli* PSM ratios increased from 14%, 29%, and 10% to 90%, 93%, and 94% at 2%, 5%, and 50% ^{13}C , while retaining 10,463, 73,033, and 24,561 *E. coli* PSMs and 2,359, 6,950, and 2,349 *E. coli* peptides.

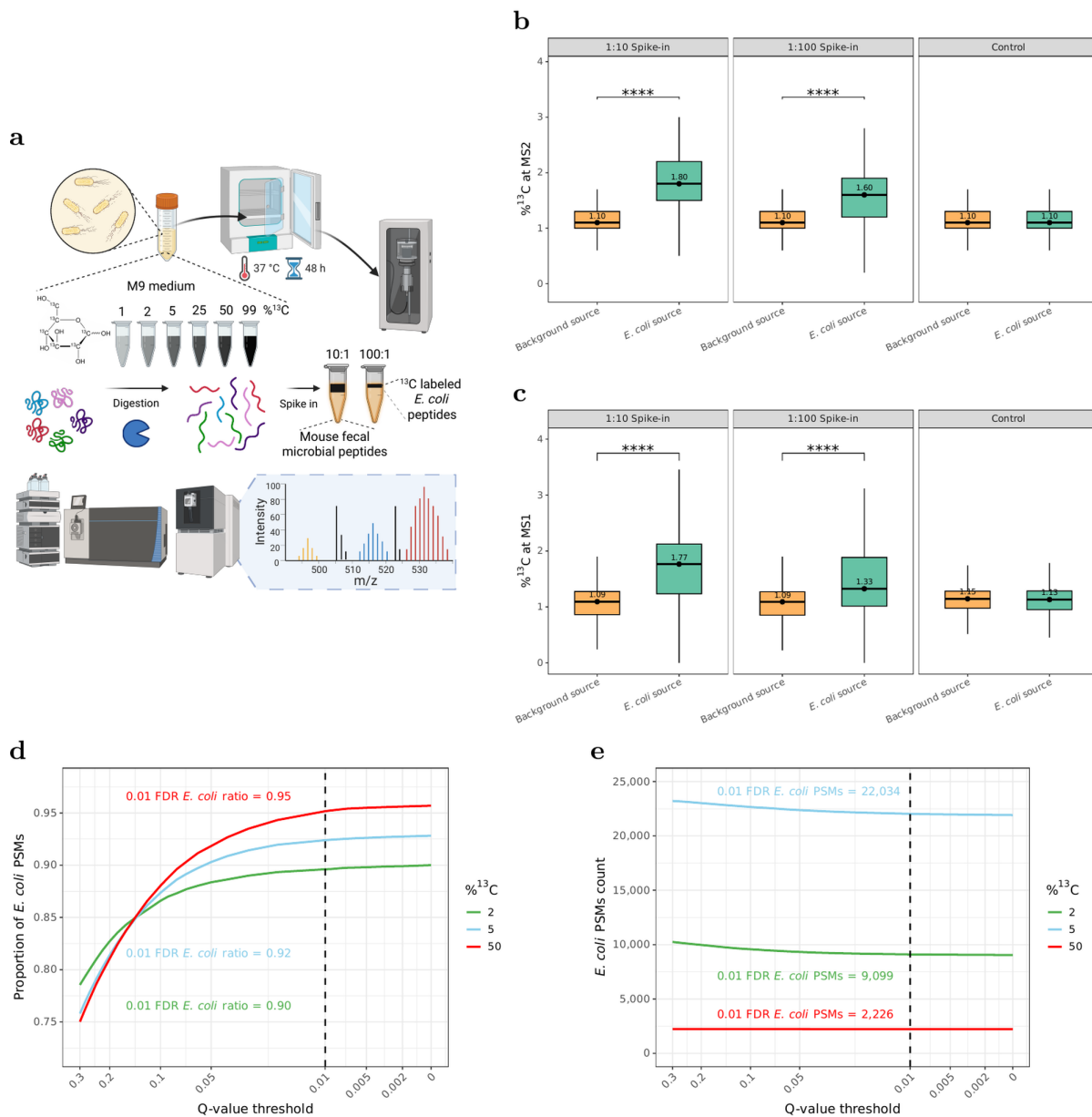


Figure 3.5: Preparation of standard ^{13}C -labeled *E. coli* samples and validation of Sipros 5 using ^{13}C -labeled *E. coli* peptide spike-ins in a mouse fecal peptidome background. (a) Schematic of the workflow. *E. coli* was cultured in M9 medium with defined ^{13}C -glucose enrichment, digested to peptides, and spiked into a mouse fecal microbial peptidome at 1:10 or 1:100 (*E. coli*:fecal peptide, w/w), with an unspiked fecal control included. Panels (b)–(i) show Orbitrap DDA results. (b, c) Boxplots for target PSMs from the 2% ^{13}C spike-in experiment before SIP-specific q-value filtering, comparing background-source and *E. coli*-source assignments across the 1:10 spike-in, 1:100 spike-in, and control samples. Inferred ^{13}C abundance is shown from MS2 evidence in (b) and MS1 evidence in (c); points and numbers denote group medians, and brackets show Wilcoxon-test significance. (d) WinnowNet q-value-threshold analysis for the 2%, 5%, and 50% ^{13}C spike-in datasets, showing the cumulative proportion of *E. coli* PSMs among retained PSMs. Dashed vertical lines indicate $q = 0.01$, with text annotations reporting the corresponding *E. coli* ratios. (e) Plot showing the cumulative number of *E. coli* PSMs across the q-value threshold for the 2%, 5%, and 50% ^{13}C spike-in datasets. The horizontal lines represent the number of *E. coli* PSMs at $q = 0.01$ for each dataset: 9,099 (2%), 22,034 (5%), and 2,226 (50%).

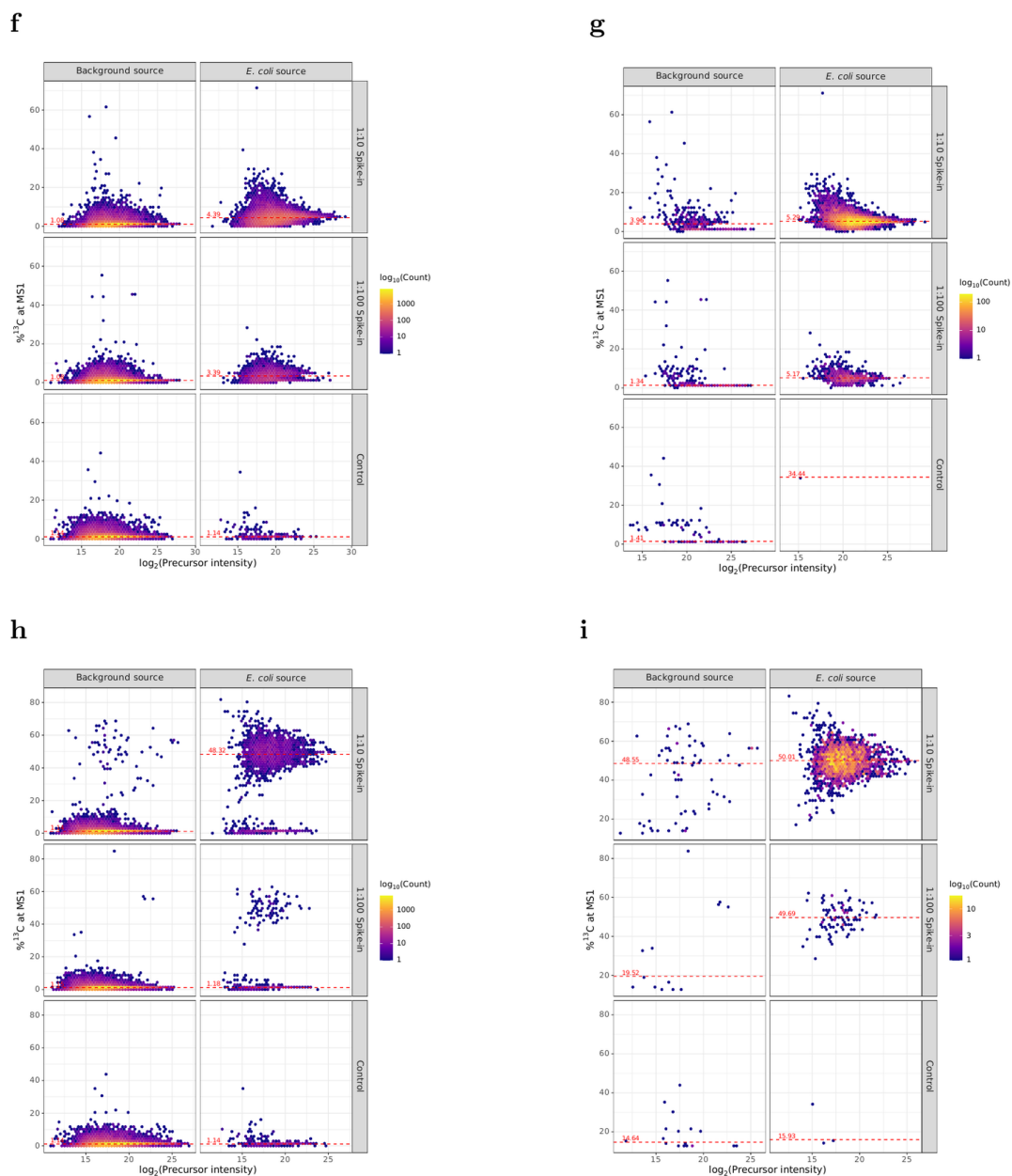


Figure 3.6: Validation of Sipros 5 using ^{13}C -labeled *E. coli* peptide spike-ins, continued. (e) WinnowNet q -value-threshold analysis for the 2%, 5%, and 50% ^{13}C spike-in datasets, showing the cumulative number of retained *E. coli* PSMs. Dashed vertical lines indicate $q = 0.01$, with text annotations reporting the corresponding *E. coli* PSM counts. (f–i) MS1 ^{13}C -abundance versus precursor-intensity density plots for 5% and 50% ^{13}C spike-ins before and after SIP-specific filtering: 5% before filtering (f), 5% after $q = 0.01$ filtering (g), 50% before filtering (h), and 50% after $q = 0.01$ filtering (i). Dashed median lines indicate the expected isotopic abundance restored after label-FDR control.

Table 3.1: Performance of filtering tools on Orbitrap DDA ^{13}C -labeled *E. coli* peptide spike-ins into mouse fecal peptides. PSM, peptide, and protein counts are *E. coli* assignments. ^{13}C med. and MAD are the peptide-level inferred ^{13}C median and median absolute deviation. “—” indicates missing values due to run failure or non-applicability.

Tool	$^{13}\text{C}\%$	PSMs	Peptides	Proteins	<i>E. coli</i> PSM%	<i>E. coli</i> Pep%	^{13}C med.	MAD
Sipros 5 WinnowNet	2	9,048	2,310	574	90.00	76.62	2.00	0.00
	5	21,922	5,105	998	92.82	83.98	5.00	2.00
	50	2,222	747	229	95.69	92.11	50.00	3.00
Sipros 5 Percolator	2	1,356	974	461	12.10	11.01	2.00	0.00
	5	8,984	2,660	707	93.30	84.18	6.00	1.48
	50	2,206	742	245	96.25	92.40	50.00	4.45
Sipros 5 Mokapot	2	—	—	—	—	—	—	—
	5	8,336	2,540	622	95.09	88.35	6.00	1.00
	50	2,192	741	228	89.54	77.11	50.00	4.00
Sipros 4	2	3,015	1,072	310	82.00	67.00	2.00	0.00
	5	6,387	2,265	562	88.00	77.00	5.00	1.48
	50	1,059	453	166	70.00	53.00	49.00	2.97
Sipros 3	2	177	142	81	52.00	47.00	2.00	0.00
	5	1,307	818	339	84.00	78.00	2.00	0.00
	50	86	37	22	31.00	17.00	49.00	5.93
Calisp	2	—	329	97	—	7.29	1.96	0.26
	5	—	217	99	—	10.07	4.72	0.52

The head-to-head software comparison was restricted to the defined *E. coli* spike-ins because these samples provide a known organismal source, nominal isotope enrichment, and controlled mixture ratio against which identification specificity and enrichment accuracy can be evaluated. The comparison proceeds in three stages: pure-culture standards isolate algorithm performance, *E. coli* peptides diluted into a mouse fecal peptidome test whether that performance persists in a complex matrix, and the cecal and SILAM datasets demonstrate biological application where no complete ground truth exists. Earlier Sipros versions could analyze DDA metaproteomic SIP samples, including the original analysis of the mouse SILAM experiment (Smyth et al., 2020); the mouse studies were therefore not impossible before Sipros 5. The advance evaluated here is deeper recovery from complex samples together with enrichment-aware MS1 and MS2 evidence, explicit

label-FDR control, and unified DDA/DIA support.

Metaproteomic SIP of mouse cecal contents with ^{13}C -glucose and ^{13}C -inulin

In ex vivo incubations of mouse cecal contents with U- ^{13}C -glucose in M9, Sipros 5 recovered extensive incorporation (median $>75\%$ ^{13}C) in dominant anaerobic genera by 24 h. Labeled peptide intensities and PSM recoveries closely tracked the corresponding peptide abundances in unlabeled controls, indicating minimal global expression distortion under these conditions and supporting interpretation of labeling primarily as substrate assimilation rather than broad proteome remodeling. Labeled proteins were enriched for translation and ribosome-associated processes.

In the ^{13}C -inulin time-course in diluted Gifu medium, the median MS1-inferred $^{13}\text{C}\%$ of PSMs increased only modestly (9.39% to 10.53%), whereas the number of ^{13}C -labeled PSMs increased sharply after 16 h, indicating a pronounced expansion of detectable labeled peptides in the later phase. The apparent difference between Figure 3.8c and e reflects the quantities plotted rather than a disagreement in the data. Panel c counts all filtered labeled PSMs: mean counts were similar at 4, 8, and 12 h (780, 725, and 773, respectively) before increasing to 1,579, 3,010, and 3,620 at 16, 20, and 24 h. Panel e instead displays precursor-intensity contributions within a curated enzyme set and scales each enzyme row across time. Consequently, the identities and relative intensities of the labeled proteins can shift during the first 12 h even while the total labeled-PSM count remains stable.

Genus-level dynamics showed a temporal shift: *Lactobacillus* exhibited early and strong ^{13}C incorporation (4–12 h), while *Blautia* increased later in the incubation (around 16 h). Bayesian estimates of genus-level ^{13}C kinetics recapitulated this succession: *Lactobacillus* exhibited a significantly shorter ^{13}C turnover half-life, whereas *Blautia* reached the highest asymptotic ^{13}C relative isotope abundance (RIA). Inulin is a fructan composed predominantly of fructose residues; cleavage by GH32 β -fructofuranosidases releases fructose, which explains why the labeled response includes fructose transport, phosphorylation, and glycolytic-entry functions rather than a glucose-specific pathway (Zhu et al., 2021; Vega-Sagardía et al., 2024). Early in the incubation, in-

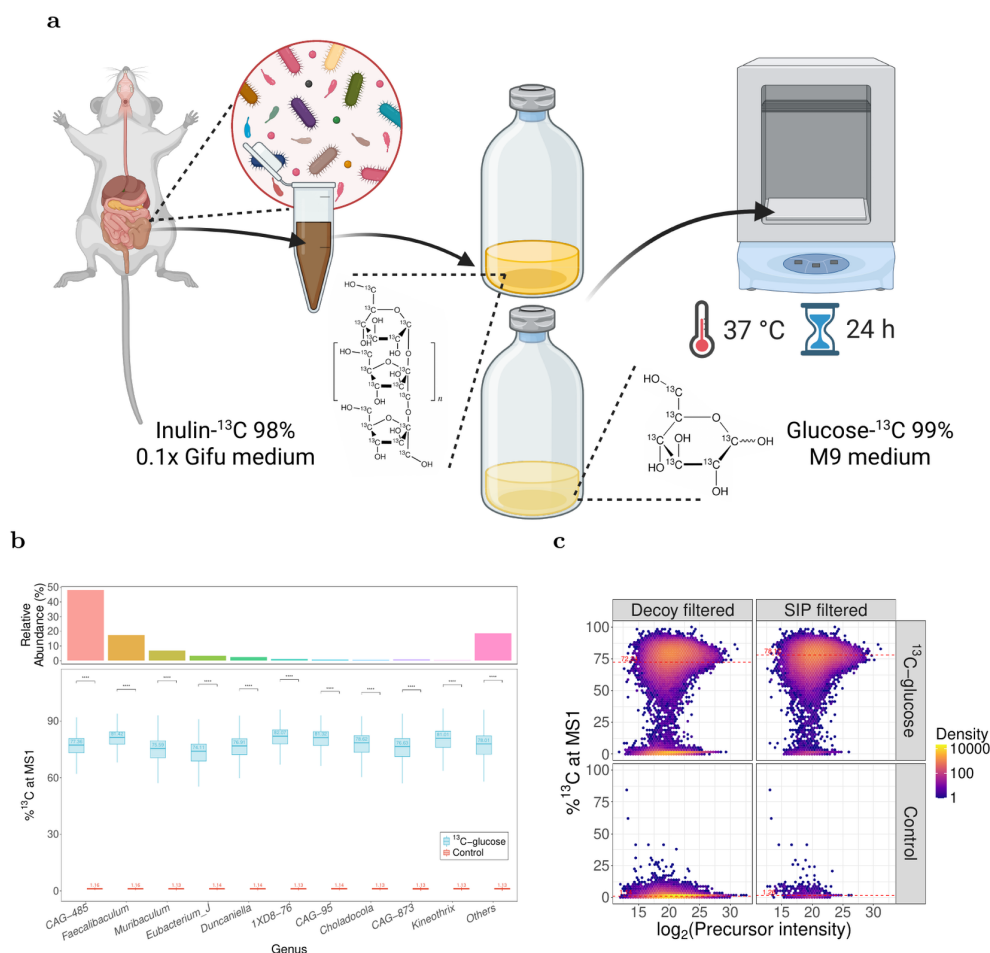


Figure 3.7: Performance verification of Sipros 5 in real-world metaproteomic SIP experiments using ex vivo incubation of mouse cecum contents with ^{13}C -glucose in M9 medium or ^{13}C -inulin in Gifu medium. (a) Schematic of the incubation and LC-MS/MS workflow. Fresh cecum contents collected from euthanized mice were incubated with ^{13}C -glucose-M9 or ^{13}C -inulin-Gifu before proteomic analysis. (b) Genus-level summary of the ^{13}C -glucose-M9 incubation. Top, relative precursor-intensity abundance of SIP-filtered ^{13}C -labeled PSMs assigned to the ten most frequently represented genera and to all remaining genera (Others); bottom, MS1-inferred $^{13}\text{C}\%$ for PSMs assigned to the same genera in the ^{13}C -glucose treatment and unlabeled control. Box labels show median values, and asterisks denote Benjamini-Hochberg-adjusted Wilcoxon rank-sum significance. (c) Hexbin density plots of PSMs showing $\log_2$ precursor intensity versus inferred $^{13}\text{C}\%$ after decoy filtering and after SIP negative-control filtering, for the ^{13}C -glucose sample and unlabeled control. Color denotes $\log_{10}(\text{PSM count})$ density, and red dashed lines indicate median $^{13}\text{C}\%$ values. (d) Relationship between $\log_2$ peptide abundance in the unlabeled control and the number of ^{13}C -labeled PSMs recovered for the corresponding microbial peptide after ^{13}C -glucose-M9 incubation; hexbin color denotes $\log_{10}(\text{peptide count})$ density, and the black curve shows a LOESS fit. (e) Relationship between $\log_2$ peptide abundance in the unlabeled control and $\log_2$ abundance of the corresponding SIP-labeled microbial peptide after ^{13}C -glucose-M9 incubation. Points are colored by median MS1-inferred $^{13}\text{C}\%$, the black line shows a linear fit, and marginal histograms show abundance distributions; insets report correlation statistics and P values.

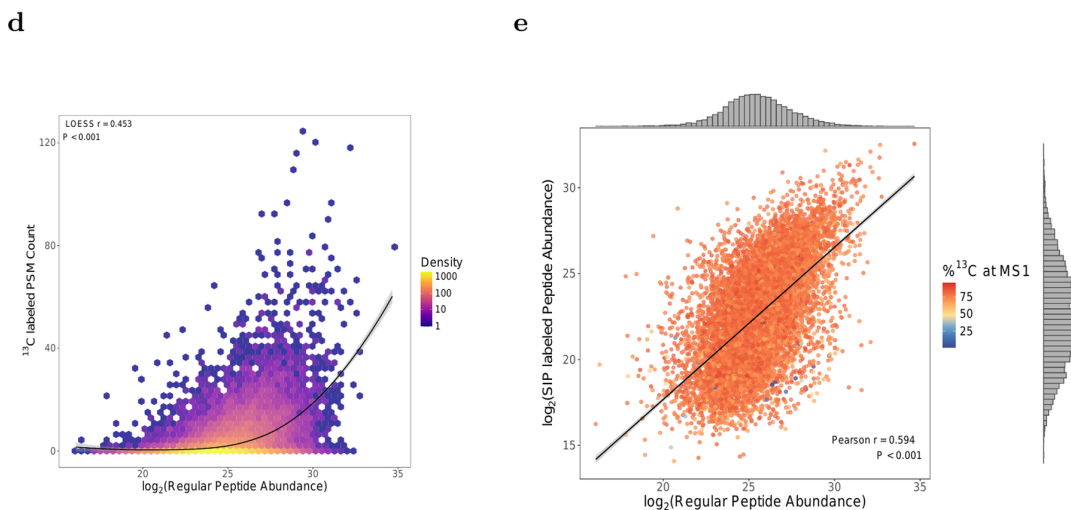


Figure 3.7: Performance verification of Sipros 5 in real-world metaproteomic SIP experiments, continued: panels (d) and (e).

ulin hydrolysis and fructose uptake/phosphorylation functions (Glyco-hydro 32C, EIHA-man, PTS EIHA 1, EII-Sor, PTS EIHA 2, EIID-AGA) were predominantly assigned to *Lactobacillus*; after 16 h, *Blautia* and phosphofructokinase-associated entry of fructose-6-phosphate into glycolysis increased. This functional redistribution precedes the later rise in total labeled PSMs and supports a temporal handoff in fructan use.

Divergent ^{15}N turnover dynamics between gut microbiome and host proteins

We reanalyzed raw files from a mouse stool metaproteomic SIP time course collected during 43 days of ^{15}N stable isotope labeling of mammals (SILAM) feeding. The original DDA study established that isotope incorporation could be traced in this host-associated metaproteome (Smyth et al., 2020); the Sipros 5 reanalysis extends that capability through increased identification depth, combined MS1/MS2 isotope evidence, and explicit label-FDR control sufficient for dense protein-level kinetic modeling. Across the dataset, Sipros 5 identified 7,036 ^{15}N -labeled proteins, 14,210 peptides, and 173,137 filtered PSMs. The relative abundance of ^{15}N -labeled proteins from the dominant gut microbial families was largely stable over the time course, with a transient increase

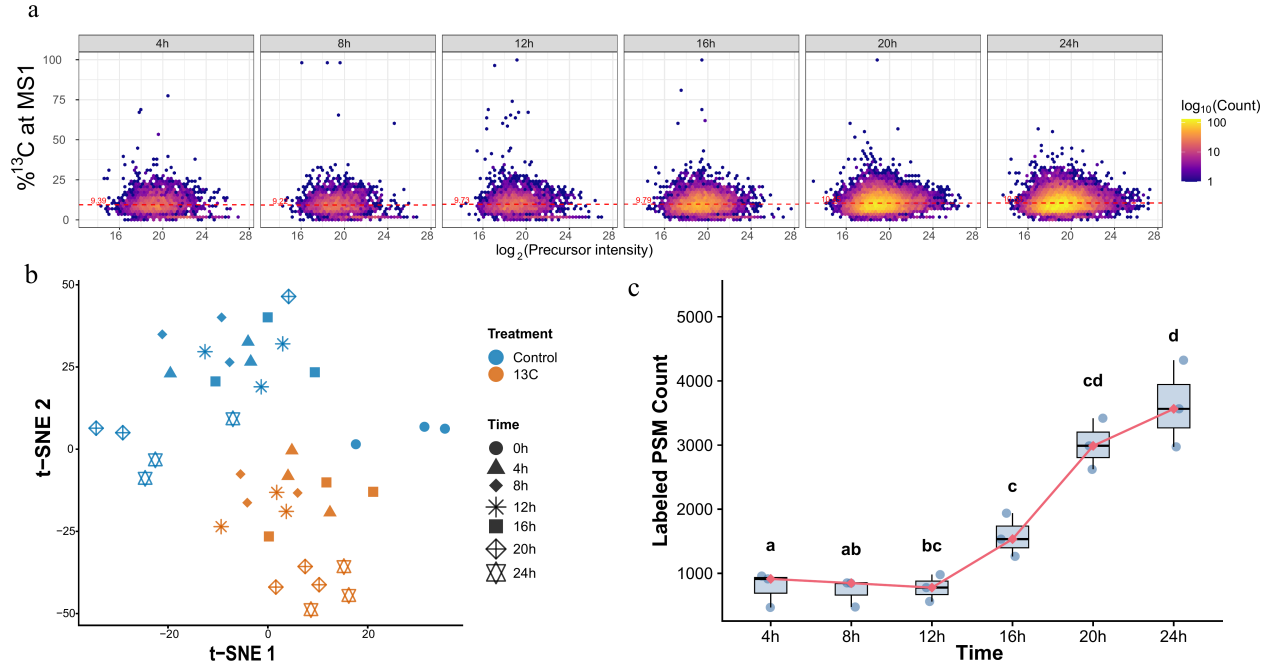


Figure 3.8: Time-resolved ^{13}C -inulin assimilation by the mouse cecal microbiome identified with Sipros 5. (a) Two-dimensional density plots of PSMs showing $\log_2$ MS1 precursor intensity versus inferred $\%^{13}\text{C}$ at 4, 8, 12, 16, 20, and 24 h. Color denotes $\log_{10}(\text{PSM count})$ density, and the red dashed line denotes the median $\%^{13}\text{C}$ of PSMs at each time point. (b) t -SNE projection of peptide abundance stratified by $\%^{13}\text{C}$ enrichment, with ^{13}C -labeled incubations and unlabeled negative controls shown by treatment color and time points indicated by distinct symbols. (c) Numbers of ^{13}C -labeled PSMs ($\%^{13}\text{C} \geq 5\%$) at each time point. Bars annotated with different letters indicate significant differences among time points ($P < 0.05$, Tukey HSD test). (d) Genus-level alluvial plot summarizing the relative abundance of the top ten genera contributing ^{13}C -labeled proteins at each time point, with all remaining genera grouped as Others. (e) Heatmap of ^{13}C -labeled enzyme abundances involved in stepwise inulin degradation across genera, normalized across time points, with enzyme names shown on the y-axis. (f) Genus co-occurrence network derived from ^{13}C -inulin-labeled proteins, where node size is proportional to genus-level ^{13}C -labeled protein abundance, node color denotes phylum, and edges represent inferred associations between genera. (g) Estimated ^{13}C turnover half-life for the dominant genera, ordered from shortest to longest. (h) Asymptotic ^{13}C relative isotope abundance (RIA) for the same genera, ordered from lowest to highest. Bars with different letters indicate significant differences ($P < 0.05$, Benjamini–Hochberg-adjusted Wilcoxon rank-sum test).

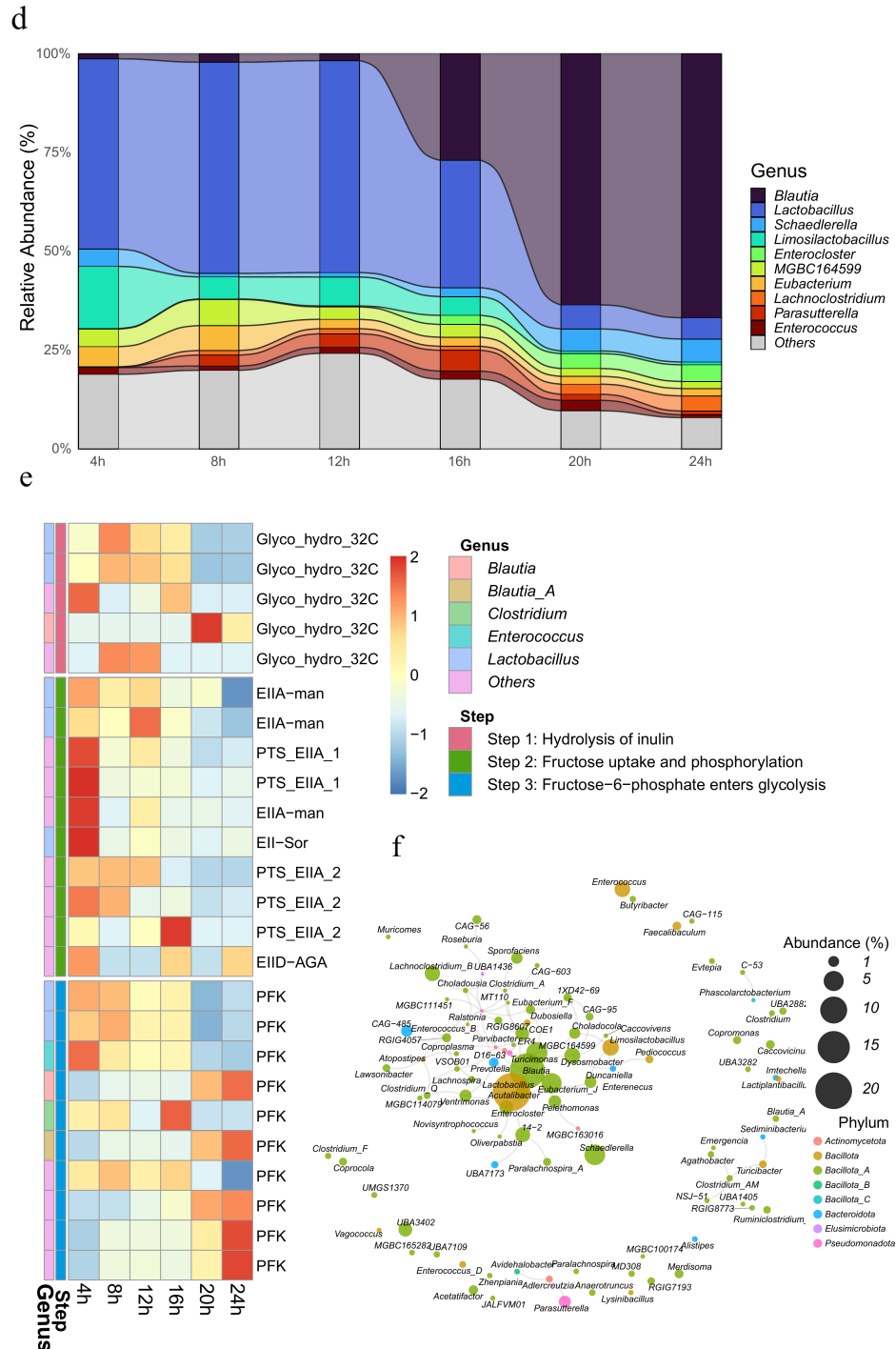


Figure 3.8: Time-resolved ^{13}C -inulin assimilation by the mouse cecal microbiome, continued: panels (d)–(f).

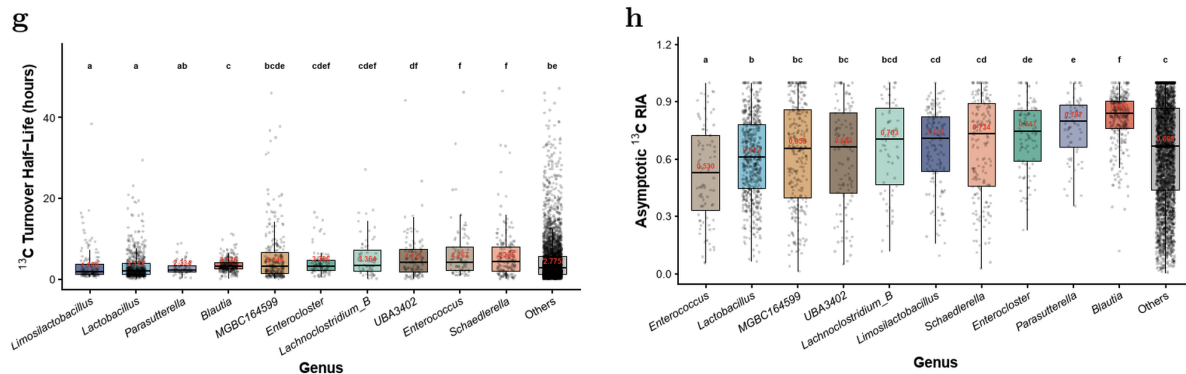


Figure 3.8: Time-resolved ^{13}C -inulin assimilation by the mouse cecal microbiome, continued: panels (g) and (h).

in Oscillospiraceae and Lachnospiraceae on days 4–8 followed by recovery after day 12. Isotopic incorporation differed markedly: Lactobacillaceae and Enterococcaceae exceeded 90% ^{15}N incorporation after the first day of feeding, whereas the remaining families approached comparable incorporation only after ~ 19 days. At the proteome level, ^{15}N incorporation into microbial proteins reached a median ^{15}N incorporation of 46% at day 1 and $\sim 90\%$ by day 29. Mouse host proteins exhibited a sustained lag throughout the 43 days. Bayesian modeling quantified this difference: gut microbial proteins reached 50% and 90% conversion from ^{15}N -natural-abundance to ^{15}N in ~ 3 and ~ 10 days, whereas host proteins required ~ 5 and ~ 17 days, respectively. Functional enrichment of ^{15}N -labeled microbial proteins highlighted pathways for anaerobic carbon flux (PflD, Por/NifJ, Fhs, PckA), nitrogen assimilation (GdhA), protein synthesis (Tuf, FusA, GroEL, L7/L12), nutrient acquisition (TupA, NgcE, ExbB), and cell-surface remodeling (Atl, VanY, ENGASE, FliC). In the host compartment, ^{15}N -labeled proteins were enriched for mucosal defense/barrier programs and cornified-envelope formation.

Discussion

Algorithmic advances enable robust SIP identification

Sipros 5 mitigates the broadening effect by modeling precursor isotope envelopes across arbitrary enrichment levels using a binomial estimation framework, incorporating convolution-based

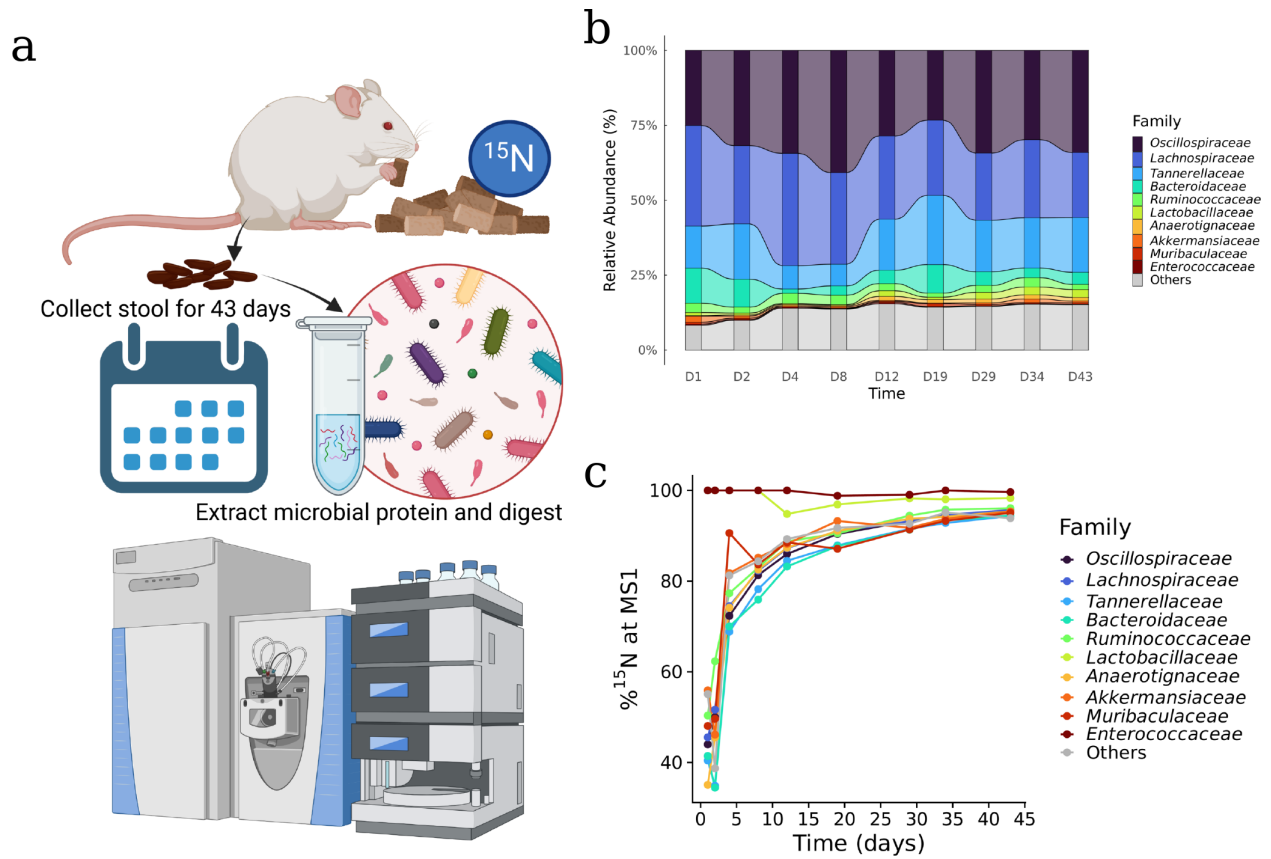


Figure 3.9: Label-FDR-controlled, taxon-, protein-, and host-compartment-resolved analysis of mouse stool during a ^{15}N -SILAM diet using Sipros 5. (a) Experimental workflow of ^{15}N -SILAM feeding, serial feces collection over 43 days, protein extraction, digestion, and LC-MS/MS analysis. (b) Alluvial summary of the relative abundance of the ten most abundant gut microbial families contributing ^{15}N -labeled proteins across the 43-day time course, with remaining families grouped as Others. (c) Time-resolved median MS1-derived $^{15}\text{N}\%$ trajectories for the same microbial families. (d) Distribution of PSM-level $^{15}\text{N}\%$ for the ten most abundant microbial families across all samples. Horizontal bars denote medians; points represent individual ^{15}N -labeled PSMs. Different letters indicate significant differences among families based on pairwise Wilcoxon tests with Benjamini–Hochberg correction. (e) Comparison of PSM-level $^{15}\text{N}\%$ between mouse host proteins and gut microbial proteins across the time course. Horizontal bars denote medians; points represent individual ^{15}N -labeled PSMs. ****, $P < 0.001$ (two-sided Wilcoxon rank-sum test with Benjamini–Hochberg correction). (f, g) Box plots summarizing protein turnover half-life estimated from ^{15}N incorporation (f) and time to 90% ^{15}N incorporation (g) for gut microbiome versus mouse host proteins; red numbers denote medians. ****, $P < 0.001$ (two-sided Wilcoxon rank-sum test). (h) Significantly enriched KEGG KO terms among ^{15}N -labeled microbial proteins; bar length indicates $-\log_{10}(\text{adjusted } P \text{ value})$ and labels indicate the number of proteins mapped to each term. (i) Significantly enriched Reactome pathways among ^{15}N -labeled host proteins; bar length indicates the number of proteins mapped to each pathway and color encodes enrichment significance.

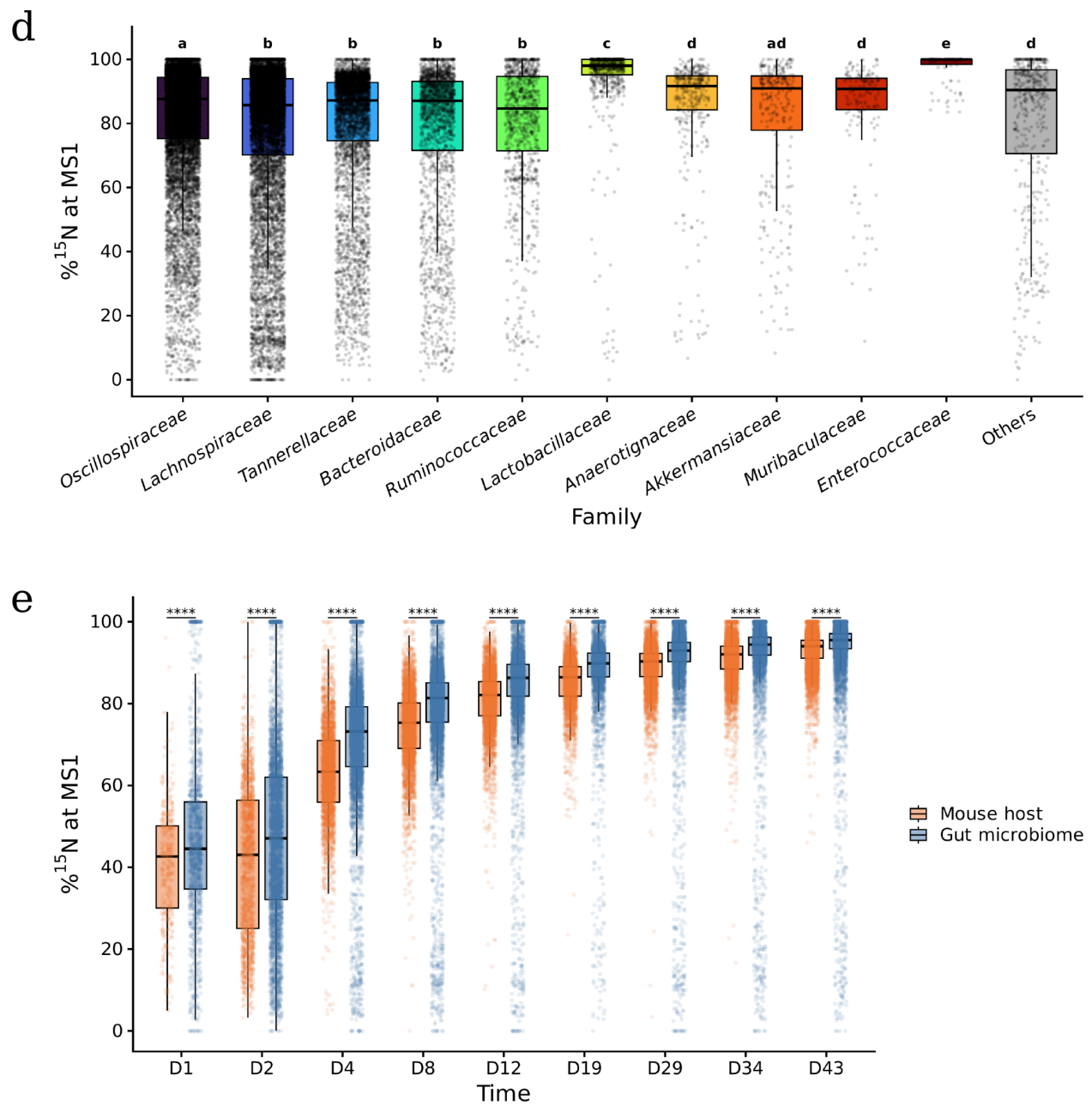


Figure 3.9: Label-FDR-controlled, taxon-, protein-, and host-compartment-resolved analysis of mouse stool during a ¹⁵N-SILAM diet using Sipros 5, continued: panels (d) and (e).

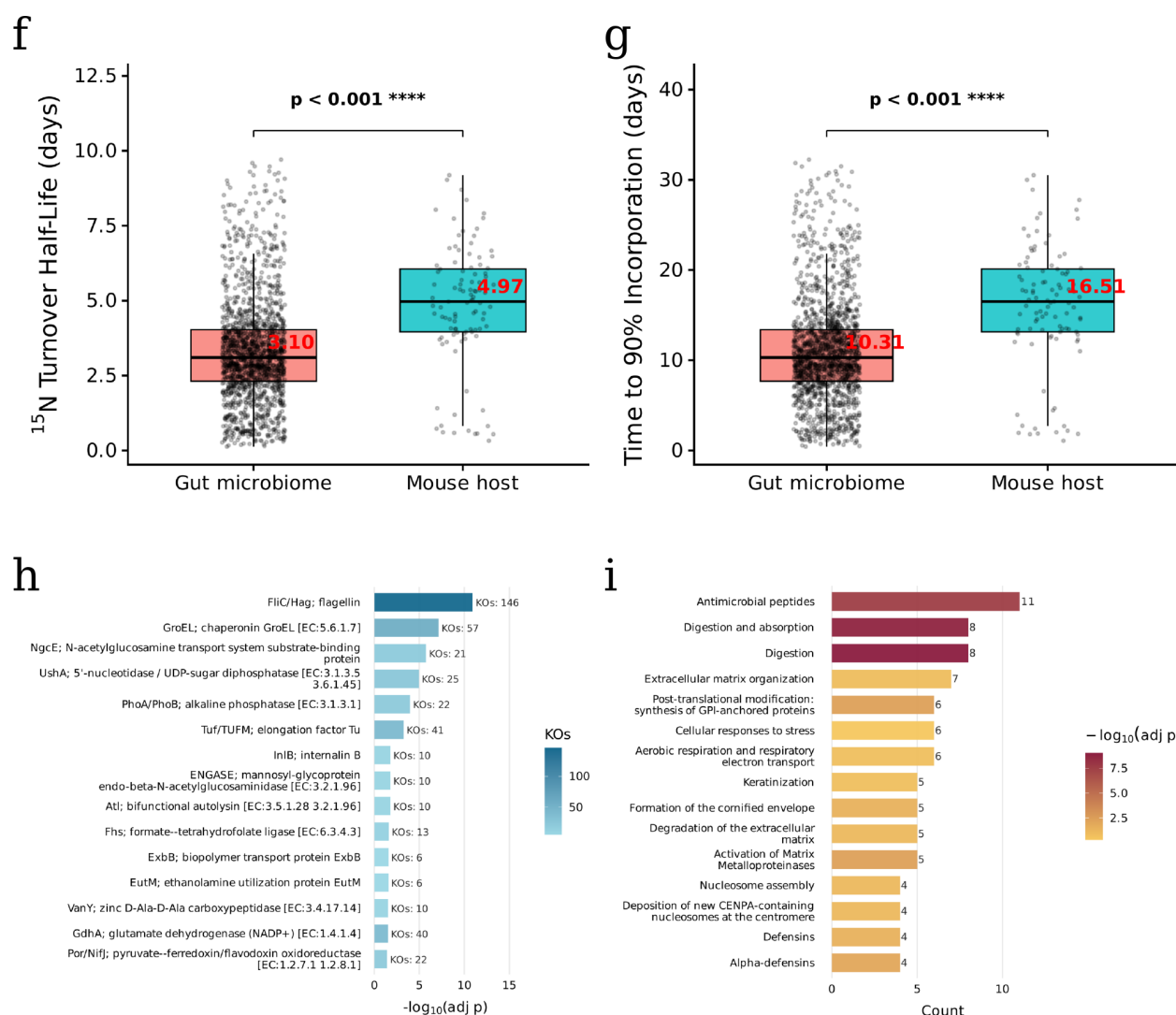


Figure 3.9: Label-FDR-controlled, taxon-, protein-, and host-compartment-resolved analysis of mouse stool during a ^{15}N -SILAM diet using Sipros 5, continued: panels (f)–(i).

fragment-ion isotope patterns, and extracting enrichment-aware MS1 and MS2 features. A major practical advance is extending SIP analysis from DDA to DIA. However, DIA introduces a SIP-specific bias: because isolation windows are fixed, they can intersect different regions of the precursor isotope envelope and transmit a non-representative subset of isotopologues into MS2, providing a direct explanation for systematic discrepancies between MS1- and MS2-derived enrichment estimates under pre-defined DIA window schemes.

Dual validation provides principled label-FDR control

Sipros 5 combines conventional target–decoy filtering for identification confidence with negative-control label validation, in which labeled PSMs detected in unlabeled controls serve as empirical “label decoys” to constrain label-FDR. Spike-in experiments illustrate both the utility and the limits of this strategy: at low enrichment (2% ^{13}C), labeled and background populations overlap substantially, limiting separability at the individual-PSM level. Applying label-FDR filtering enriched the retained labeled set for the true spike-in population and restored the expected median enrichment, but at a sensitivity cost most pronounced at low enrichment.

Substrate-specific incubations reveal contrasting labeling dynamics

Glucose and inulin labeling exposed contrasting modes of carbon flow. ^{13}C -glucose produced rapid, broad labeling enriched for translation. ^{13}C -inulin produced slower assimilation: early labeling was dominated by *Lactobacillus*, whereas *Blautia* became more prominent later, with a shift from fructan-processing and fructose-assimilation functions to glycolytic-entry functions. The co-occurrence network placed *Lactobacillus* and *Blautia* at the center of SIP-labeling-linked associations, and the combination of faster protein turnover but lower ^{13}C assimilation in *Lactobacillus* relative to *Blautia* supports complementary niches and cross-feeding during inulin processing.

^{15}N turnover distinguishes microbiome vs. host renewal kinetics

The ^{15}N -SILAM reanalysis demonstrates that Sipros 5 resolves protein turnover across the coupled host–microbiome system. Microbial proteins incorporated ^{15}N substantially faster than host proteins, consistent with the shorter generation times and more rapid proteome renewal expected for gut bacteria relative to mammalian tissues. Functional separation of labeled microbial proteins into central carbon metabolism and fermentative pathways versus host proteins into digestion, epithelial barrier, and mucosal defense highlights the value of SIP metaproteomics for dissecting complementary physiology in interacting compartments. Overall, Sipros 5 improves SIP identification across enrichment regimes and acquisition modes and introduces explicit label-FDR con-

trol using unlabeled negative controls. Remaining challenges include reduced sensitivity at very low incorporation and intensity dilution at intermediate enrichment, particularly in DIA, where isolation-window placement can bias MS2-derived enrichment estimates. Future developments that jointly model MS1 and MS2 isotope evidence and integrate multi-scan reconstruction along chromatographic elution profiles should further improve robustness and accuracy.

Methods

Sample preparation and labeling

The defined ^{13}C -labeled bacterial standards and spike-in peptides were prepared from *E. coli* DH5 α obtained from New England Biolabs, following the protocol detailed in Chapter 2. Briefly, 5 μL of washed inoculum was transferred into 10 mL M9 medium containing the specified mixture of unlabeled and U- ^{13}C -glucose as the sole carbon source. Three replicate cultures per enrichment level were grown at 37 °C for 2 days to $\text{OD}_{600} > 0.5$, and a pellet from 5 mL culture was used for each protein extraction.

Fresh mouse fecal pellets were collected from cages and pooled to obtain 10 g of mixed material for preparation of the fecal microbial peptide background used in the spike-in experiments. Separately, for each cecal incubation experiment, three 1-year-old mice were euthanized and their cecal contents were collected immediately into 2-mL microcentrifuge tubes within an anaerobic container (BD GasPak system). Mixed cecal contents (0.2 g) were suspended in 1.0 mL pre-reduced culture solution and homogenized by shaking for 5 min. Large debris was removed by centrifugation at $300 \times g$ for 2 min; the supernatant was transferred and the low-speed centrifugation was repeated for a total of three rounds. Microbial cells were then collected at $14,000 \times g$ for 1 min, washed once with pre-reduced culture solution, and re-pelleted at $14,000 \times g$ for 1 min. Each washed microbial-cell pellet was resuspended in 1 mL pre-reduced solution and used to inoculate a 100-mL serum bottle containing 10 mL sterilized, deoxygenated medium. Bottles were sealed and equilibrated with a 75% N_2 /25% CO_2 headspace. Incubations were performed with three biological replicates per condition at 37 °C for 24 h, with sampling at 0, 4, 8, 12, 16, 20, and 24 h. For

SIP labeling, M9 medium was supplemented with 0.3 g/L sodium thioglycolate, and glucose was replaced with 100% U-¹³C-glucose (Cambridge Isotope Laboratories). Gifu medium was diluted 10-fold and supplemented with 4 g/L inulin (Thermo Scientific), 0.3 g/L sodium thioglycolate, and cysteine; for SIP, inulin was replaced with 100% U-¹³C-inulin (IsoLife).

Protein extraction, digestion, and spike-in

Microbial-cell pellets for protein extraction were obtained from either 0.2 g of direct fecal/cecal material or 1.0 mL of incubation culture by centrifugation at 4 °C and 10,000 ×g for 1 min. Cells were not enumerated; the extraction input was standardized by starting mass or culture volume rather than by a measured cell count. Each pellet was washed twice with 1 mL 10 mM Tris–HCl (pH 7.0) and resuspended in 0.5 mL lysis buffer containing 100 mM Tris–HCl, 4% sodium dodecyl sulfate (SDS), and 0.1 M freshly prepared dithiothreitol (DTT). Cells were sonicated on ice for five 30-s pulsed cycles. The lysate was clarified at 4 °C and 10,000 ×g for 1 min; the resulting supernatant, rather than the unclarified whole lysate, was mixed with 0.5 mL chilled (–20 °C) 50% (w/v) trichloroacetic acid to reach a final concentration of 25% and incubated at –20 °C overnight.

The precipitated proteins were collected at 20,800 ×g for 20 min, washed twice with 1 mL chilled 80% acetone and once with 1 mL chilled acetone, and re-pelleted at 20,000 ×g for 20 min after each wash. Pellets were vacuum-dried and resolubilized in 600 μL 8 M urea in Tris–HCl containing 20 mM iodoacetamide (pH 8.0). Protein concentration was measured using the bicinchoninic acid (BCA) assay. Fifty micrograms of protein per sample was digested by filter-aided sample preparation (FASP; Vivacon 500, Sartorius) with 2 μg sequencing-grade trypsin (V5113, Promega) overnight at 37 °C, and peptides were desalted using Pierce spin columns (Thermo Scientific). For spike-in evaluation, 2 μg of mouse fecal microbiome peptides were mixed with either 0.2 μg (10:1) or 0.02 μg (100:1) of *E. coli* peptides at 2%, 5%, or 50% ¹³C in triplicate.

LC–MS/MS acquisition

For DDA, peptides (1 μ g) were separated on an XSelect CSH C18 column (2.5 μ m, 150 $\times$ 0.075 mm; Waters) on an UltiMate 3000 RSLCnano (Thermo) with a 90-min gradient (98% A/2% B to 65% A/35% B; A: 0.1% formic acid, 0.5% acetonitrile; B: 0.1% formic acid in 99.9% acetonitrile). Peptides were ionized by electrospray (2.4 kV) and analyzed on an Orbitrap Fusion Tribrid in DDA mode: MS1 in the Orbitrap (profile, 120,000 resolution, m/z 375–1500); MS2 in the Orbitrap (centroid, 30,000) after higher-energy collisional dissociation (HCD) with a 5 m/z isolation window. For DIA, 500 ng peptides were separated on an EvoSep One using the 30-samples-per-day method. Orbitrap Astral was operated in DIA: MS1 (Orbitrap, profile, 240,000 resolution, m/z 380–980); MS2 (Astral, profile, 6 ms per event) using 149 custom isolation windows of 4 m/z spanning m/z 380–980. Total method run time was 35 minutes. Raw files were deposited at ProteomeXchange (PXD075964 Astral DIA; PXD075589 *E. coli* spike-in; PXD075720 13 C-glucose SIP; PXD076077 13 C-inulin SIP). Previously published data were re-analyzed from PXD041414 (*E. coli* standards) and PXD017451 (15 N mouse diet).

SIP proteomics search and validation

SIP proteomic data (DDA and DIA) were analyzed using Sipros 5 (<https://github.com/theipanlab/sipros5>), developed in our laboratory. Thermo raw files were converted to MS1- and MS2-level FT1 and FT2 formats using Raxport 5 (<https://github.com/theipanlab/Raxport.net>). Sipros 5 identifies candidate precursor ions within isolation windows based on intensity and evaluates isotopic patterns using a binomial-distribution model to match observed and theoretical peptide isotope envelopes (Xiong et al., 2025). Fragment-ion isotope distributions are modeled using convolution-based theoretical isotope patterns to improve matching accuracy (Xiong et al., 2024, 2025). Peptide-spectrum matches (PSMs) are scored using weighted dot product (WDP) (Xiong et al., 2024), cross-correlation (Xcorr) (Eng et al., 2013), and multivariate hypergeometric (MVH) (Tabb et al., 2007); other feature-level evidence, including precursor and fragment isotope-abundance ratios, is computed for the top candidate PSMs per scan.

Initial PSM confidence was controlled by conventional target–decoy validation with Percolator (Käll et al., 2007). SIP-labeling confidence was then controlled using a second, label-specific target–decoy strategy: labeled PSMs detected in SIP-enriched samples served as label targets, whereas labeled PSMs detected in unlabeled controls served as empirical label decoys. The resulting label-target/label-decoy sets were filtered with Percolator (Käll et al., 2007) or Mokapot (Fondrie and Noble, 2021) using semi-supervised learning, or with the CNN-based WinnowNet classifier (Feng et al., 2025) (<https://github.com/theapanlab/WinnowNet>) to constrain the SIP-labeling false discovery rate to $\leq 1\%$. Filtered peptide identifications, together with SIP-labeling calls, isotope-abundance estimates, and intensity annotations, were supplied to Philosopher (da Veiga Leprevost et al., 2020) for protein assembly. Philosopher performed parsimony-based protein grouping from the filtered peptide evidence and summarized SIP-labeled peptide support at the protein level. Sipros 5 outputs isotopic-abundance and intensity estimates at the PSM, peptide, and protein levels. SIP-labeled and unlabeled peptide PSMs were visualized using Aerith (Xiong et al., 2025).

Metaproteomic DDA preparation (FragPipe MSFragger-DDA+)

Orbitrap DDA raw files from mouse gut metaproteomes and cecum incubations were also processed using FragPipe v23 (Yu et al., 2023). A metaproteomics database-preparation workflow (DDA+ mode) was used to reduce FASTA size for subsequent SIP searches on ^{13}C - or ^{15}N -labeled samples and unlabeled controls. Protein sequences were derived from the EBI MGnify mouse gut genomes species catalogue (v1.0) (Richardson et al., 2023). Database searches were performed against a target–decoy database with trypsin specificity (maximum 1 missed cleavage). Cysteine carboxymethylation was set as a fixed modification; protein N-terminal acetylation and methionine oxidation were set as variable modifications. Precursor and fragment mass tolerances were set to 20 ppm, and peptide- and protein-level FDRs were controlled at 1 % using MSBooster, Percolator, ProteinProphet, and Philosopher. Label-free quantification was performed with IonQuant, with match-between-runs enabled where appropriate. Species present in samples were inferred

from unique peptides identified during the FragPipe metaproteomics database-preparation step. Proteins from inferred species genomes were used for a subsequent MSFragger-DDA+ search, which performs full isolation-window searching, extracts targeted precursor extracted-ion chromatograms (XICs), and rescores them to resolve co-fragmented peptides as described previously (Yu et al., 2025). Proteins identified by MSFragger-DDA+ were then used as the input database for SIP searches with Sipros 5.

For the CNN-based SIP-labeling filter, each target or decoy/control PSM was encoded as a seven-channel tensor. Matched peaks were ranked by raw intensity. The top 256 peaks were retained, ordered by experimental m/z and zero-padded when fewer peaks were available. For each peak, the seven channels encoded experimental m/z , mass error relative to the theoretical peak, monoisotopic m/z of the associated ion envelope, normalized experimental intensity, normalized theoretical intensity, the number of SIP atoms in the ion, and an isotope-envelope enrichment ratio. The network used three one-dimensional convolutional blocks with batch normalization and rectified linear unit (ReLU) activation, followed by global max and mean pooling. Pooled features were passed through a shared dropout-regularized fully connected layer and two output heads for label target/decoy classification and percent-labeling regression. The model was trained on approximately 1.31 million Sipros 5–derived PSMs: 612,000 target PSMs from the 2%, 5%, 25%, 50%, and 99% isotope-labeling datasets and 698,000 control PSMs from unlabeled controls. To minimize peptide-level information leakage, PSMs were partitioned by peptide group into peptide-disjoint training, validation, and test sets using an approximately 8:1:1 split. Training optimized cross-entropy loss for classification plus a weighted SmoothL1 loss on $\log(1 + \text{percent labeling})$, with percent-balanced sampling to reduce enrichment-level imbalance. During prediction, raw spectra were converted to tensors and retained when their softmax label-target probabilities passed the threshold required for decoy-based false-discovery control.

Bayesian protein turnover model

Protein turnover rates were estimated using a Bayesian exponential-incorporation model. For each protein, isotopic abundance (%) or relative isotope abundance (RIA) at time t was modeled as

$$y(t) = f_0 + (f_\infty - f_0) \left[1 - e^{-kt} \right] + \varepsilon, \quad (3.1)$$

where f_0 is the baseline natural isotope abundance, f_∞ is the asymptotic isotope fraction at steady state, and k is the first-order turnover rate constant per hour or per day. Baseline values were fixed according to the labeling scheme: 0.368 % for ^{15}N , 1.07 % for ^{13}C , and 0 for RIA, based on unlabeled controls (Hellerstein and Neese, 1999). The residual error term was modeled as $\varepsilon \sim \mathcal{N}(0, \sigma^2/w)$, where w is the normalized PSM precursor-abundance weight. RIA was defined as the ratio of precursor abundance from PSMs with >1 % ^{13}C or ^{15}N incorporation to that from PSMs with ≤ 1 % incorporation.

Only proteins with at least four distinct observed time points were analyzed. PSM-level MS1 isotopic abundances were used as independent observations without averaging across proteins or time points. Each PSM was weighted by precursor intensity, and weights were normalized within each protein, such that higher-abundance PSMs contributed more strongly to parameter estimation. Posterior distributions were sampled in RStan using Hamiltonian Monte Carlo with four chains and 2,000 iterations per chain, including 1,000 warm-up iterations. Protein half-life was calculated as $\ln(2)/k$ and time to 90 % incorporation as $\ln(10)/k$. Proteins with convergence failures or implausible half-lives (>24 h for ^{13}C cecum-content labeling, >10 d for ^{15}N mouse fecal labeling) were excluded (Gupta et al., 2024; Li et al., 2025; Reddien, 2024).

Acknowledgments and funding

The authors acknowledge the IDeA National Resource for Quantitative Proteomics at the University of Arkansas for Medical Sciences and, separately, the Proteomics and Metabolomics Core Facility at Duke University School of Medicine for conducting LC–MS-based proteomic analyses;

the University of Oklahoma Supercomputer Center for Education and Research (OSCER) for providing high-performance computing resources; and the National Institutes of Health, specifically the National Center for Complementary & Integrative Health and the National Institute of General Medical Sciences (Grant R01AT011618), for funding support. Graphical figures were generated using `ggplot2` (v3.4.0) (Wickham, 2016) and `patchwork` (v1.1.2) (Pedersen, 2019) within the R statistical computing environment (v4.3.0) (R Core Team, 2023).

Supplementary Material

The supplementary material accompanying this chapter is reproduced or summarized in Appendix B. Appendix Figure B.1 shows the Astral DIA validation of Sipros 5 on the *E. coli* peptide spike-ins, including MS2 and MS1 ^{13}C abundance distributions, WinnowNet q-value-threshold curves, and labeled-PSM density before and after filtering. Appendix Figures B.2 and B.4 provide genus-level annotations of unlabeled and ^{13}C -labeled proteins in the ^{13}C -glucose and ^{13}C -inulin cecal incubations, respectively; Appendix Figures B.3 and B.5 give functional enrichment for the two substrates. Appendix Figure B.6 summarizes precursor intensity versus MS1 ^{15}N abundance across the SILAM time course, Appendix Figure B.7 summarizes functional enrichment of labeled microbiome and host proteins, and Appendix Figure B.8 shows protein-turnover half-lives across enriched KEGG orthology terms. Appendix Tables B.1–B.16 reproduce or summarize the benchmark, spike-in, medium-composition, substrate-assimilation, and SILAM workbooks; several source workbooks are divided into multiple appendix tables so that each displayed table has its own B-series number.

Aerith: Visualization and Annotation of Isotopic Enrichment Patterns of Peptides and Metabolites with Stable Isotope Labeling from Proteomics and Metabolomics

Published as Xiong et al. (2025) in Analytical Chemistry 97, 24916–24921 (2025).

Reproduced with permission under the journal's open-access policy.

Abstract

Stable isotope probing (SIP) traces the metabolism of biological cells using isotopically heavy substrates (e.g., ^{13}C , ^{15}N , or ^2H). Confident identification of the metabolic products of isotopic labeling remains a challenge due to the difficulties in simulating, visualizing, and annotating the isotopic patterns of partially labeled peptides and metabolites found in mass spectrometry (MS) data. Here, we present **Aerith**, an R package designed to visualize simulated and observed isotopic envelopes of peptides and metabolites with user-defined formulas and atom% enrichment levels. Aerith models the isotopic distributions of the fragment ion series of a peptide by sequentially convolving isotopic envelopes of monomeric units using a convolution algorithm. Aerith simulates fine isotopic structures of a compound using Monte Carlo simulation via the multinomial distribution and calculates the isotopic envelopes of metabolites with known chemical formulas using a fast Fourier transform (FFT)-based algorithm. These algorithms provide accurate simulation of the isotopic envelopes of SIP-labeled peptides and metabolites with high computational efficiency. Aerith evaluates peptide-spectrum matches through multiple robust and commonly used scoring functions. These algorithms were implemented in C++ and accessed in R via Rcpp (Eddelbuettel and François, 2011) to ensure real-time interactivity. We present case studies to demonstrate Aerith's utility in resolving isotopic fine structures and envelopes for glucose, penicillin, and microbial peptides containing natural and enriched isotopes.

The metabolite case studies in Figure 4.1 show the two complementary simulation modes described in the Aerith paper. Glucose and penicillin demonstrate Monte Carlo simulation of isotopic fine structures, where partially labeled molecules separate into closely spaced isotopologue peaks that can be annotated by composition. In contrast, the gamma-aminobutyric acid examples show the faster FFT-based envelope calculation, which is appropriate when the integer-spaced isotope envelope is sufficient for comparing natural-abundance and SIP-labeled spectra.

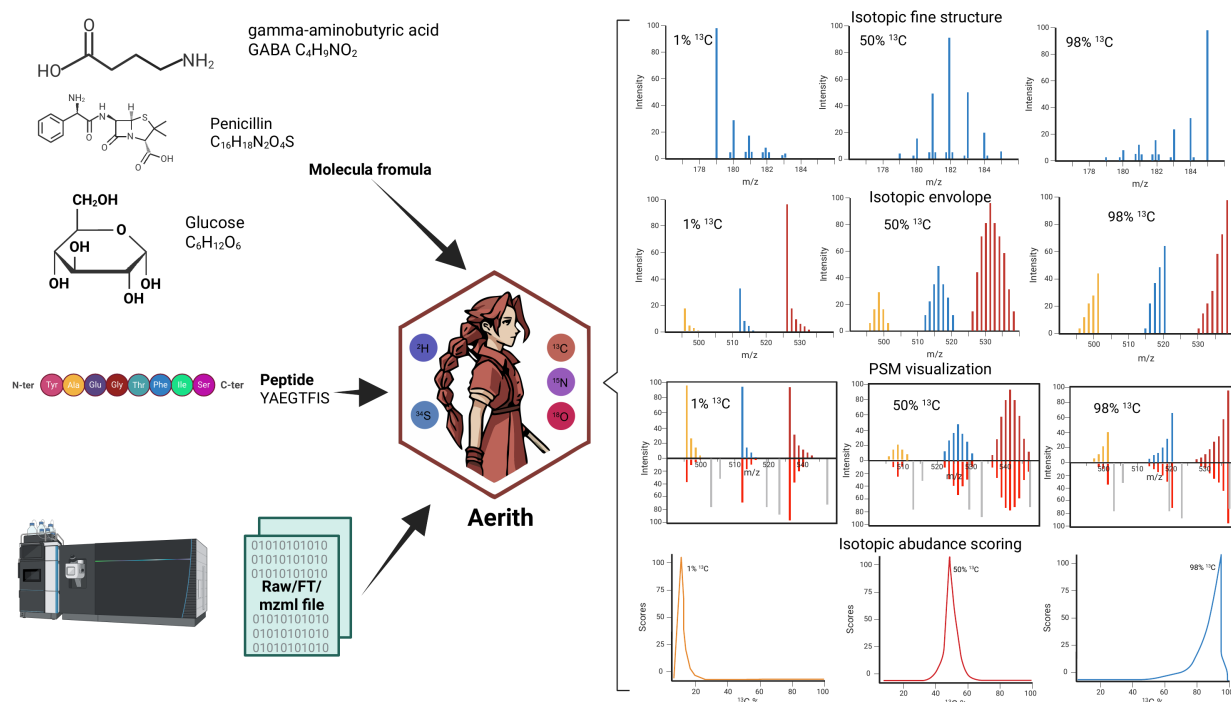


Figure 4.1: Simulated isotopic fine structures and envelopes of SIP-labeled metabolites. The isotopic fine structures of negative-ion-mode glucose ($[M-H]^-$) are simulated at 1.07% natural abundance (A) and 50% ^{13}C enrichment (B), which can be compared with measured spectra from MZcloud (Thermo Fisher; Figure C.1). Adjacent isotopic peaks with minor m/z differences (< 0.03 Da) were computationally offset while preserving true positional relationships for visual clarity. High-abundance isotopologues are annotated in red with their isotopic compositions. The isotopic fine structures of positive-ion-mode penicillin ($[M+H]^+$) at 1.07% natural abundance (C) and 50% ^{13}C enrichment (D) were validated against reference spectra from MZcloud (Figure C.2). The isotopic envelopes, ignoring fine structure, of positive-ion-mode gamma-aminobutyric acid ($[M+H]^+$) at 1.07% natural abundance (E) and 50% ^{13}C enrichment (F) were generated using the FFT algorithm in comparison with experimental metabolomics data from natural-abundance and SIP-labeled samples.

Introduction

Stable isotopes can be used as tracers for metabolic fluxes in biological and biogeochemical systems, enabling their quantification in complex biological contexts. Modern mass spectrometry, including Fourier-transform mass spectrometry (FTMS) and time-of-flight mass spectrometry (TOF-MS), allows precise measurement of isotope distributions and fine structures at high resolution. Several R packages have been developed for stable isotope fingerprinting (SIF) and isotope ratio mass spectrometry (IRMS); for example, `simmr` employs a Bayesian framework to model stable isotope mixing for ecological tracer analysis, the `isoverse` suite integrates `isoreader` for data parsing and `isoorbi` for isotope calculations from Orbitrap data, `assignR` facilitates geographic provenance mapping using isotopic signatures, and `IsoSpec2` provides rapid simulation of isotopic fine structures under natural-abundance conditions but does not support analysis of compounds with partial isotope incorporation from SIP experiments.

SIP has provided novel insights into the diversity and functions of microbial communities within various biogeochemical cycles (Li et al., 2019; Kleiner et al., 2017). The `Miso` R package supports metabolomic analyses through multi-isotope labeling workflows but requires the isotopic envelope of the known 100% isotopically enriched molecular tracer. In contrast, proteomic SIP tracks stable-label incorporation into proteins via LC–MS/MS and enrichment-resolved database searches. Several proteomic SIP tools—`Sipros` (Pan et al., 2011; Xiong et al., 2024), `MetaProSIP` (Sachsenberg et al., 2015), and `Calisp` (Kleiner et al., 2018)—have been developed to analyze partially labeled peptide mixtures; however, they only report estimated labeling percentages without further interpretation. `PDV` enables visualization of proteomic search results in `pepXML` and `mzid` formats but does not allow user-defined amino acid sequences or molecular formulas for isotopically labeled peptides and compounds.

To address these limitations, we developed `Aerith`, a computational tool for simulation of theoretical isotopic peaks, fine structures, and spectral envelopes of user-defined chemical formulas or peptide sequences at both the MS and MS/MS levels. The software allows user-defined isotopic

abundances of SIP-labeled elements across all enrichment levels. Aerith enables visualization and manual validation of unlabeled and SIP-labeled peptide-spectrum matches (PSMs) from proteomic SIP data.

Feature and Implementation

Input data format

Aerith accepts spectral data files in multiple formats, including Raxport-processed FT2, mzML, and MGF formats, as well as pepXML and PIN files (Percolator outputs), TSV files from the Sipros search engine (Pan et al., 2011; Xiong et al., 2024) for peptide amino acid sequences or PSM input, and FASTA files. The mzR package enables direct parsing of mzML and MGF files, and all imported spectra are available for downstream analysis. Figure 4.2 summarizes key metadata of input spectra processed by Aerith. Users may also input peptide sequences or chemical formulas to simulate isotopic envelopes with adjustable isotopic abundances (e.g., ^{13}C , ^2H , ^{18}O , ^{15}N , ^{34}S).

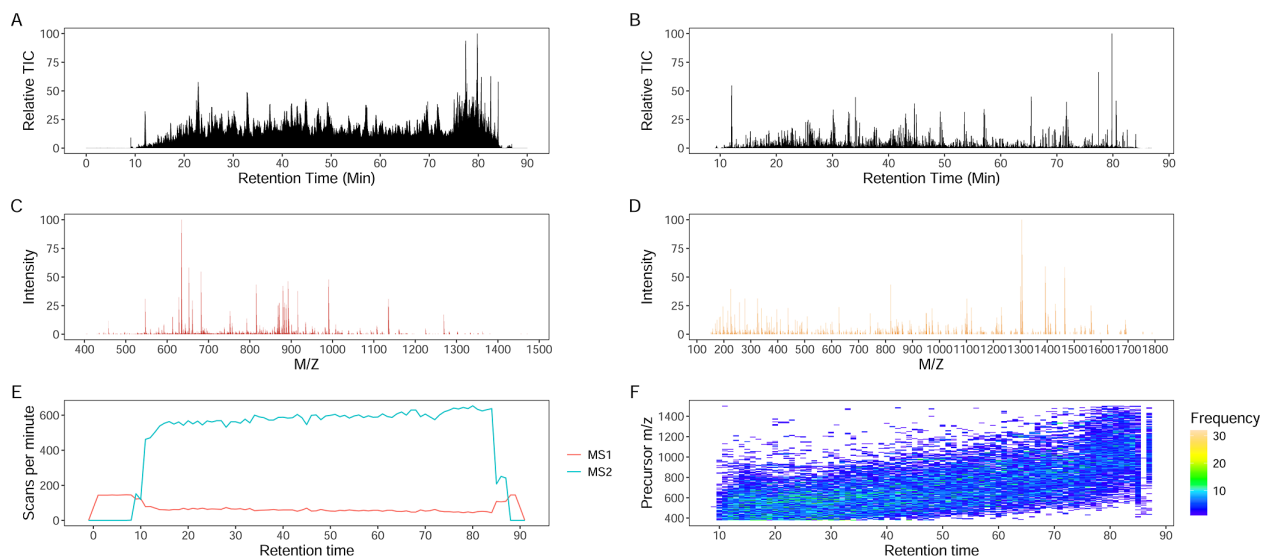


Figure 4.2: Visualization of an LC–MS/MS run in Aerith. An unlabeled *E. coli* sample was analyzed in DDA mode. (A) Total ion chromatogram (TIC) of MS1 scans. (B) TIC of MS2 scans. (C) MS1 scan at the retention time of 74.04 min. (D) Data-dependent MS2 scan on m/z 950.9609 with a 5 Da isolation window of the MS1 scan shown in (C). (E) Scan acquisition rates for MS1 (red) and MS2 (blue) levels. (F) Precursor selection frequencies across m/z and retention time.

Theoretical spectra generation for SIP-labeled metabolites

Theoretical spectra with isotopic fine structures are generated using a Monte Carlo algorithm that repeatedly simulates multinomial isotope distributions and captures low-abundance fine-structure variants. To obtain better computational efficiency, Aerith also incorporates an FFT-based algorithm to produce isotopic envelopes. This algorithm retains only the main peaks separated by integer neutron differences, transforms each elemental distribution via FFT, multiplies them pointwise according to stoichiometry, and applies an inverse FFT to obtain the aggregate envelope. Users may select either method. Both algorithms accept user-defined chemical formulas and specified stable-isotope abundances from 0–100 % as input. The resulting theoretical spectra can be exported as tabulated data, including m/z , charge state, and intensity values.

The metabolite case studies in Figure 4.3 show the two complementary simulation modes described in the Aerith paper (Xiong et al., 2025). Glucose and penicillin demonstrate Monte Carlo simulation of isotopic fine structures, where partially labeled molecules separate into closely spaced isotopologue peaks that can be annotated by composition. In contrast, the γ -aminobutyric acid examples show the faster FFT-based envelope calculation, which is appropriate when the integer-spaced isotope envelope is sufficient for comparing natural-abundance and SIP-labeled spectra.

Theoretical spectra of SIP-labeled peptides in bottom-up proteomics

Aerith simulates the isotopic envelopes of peptide precursors and B/Y fragment ions using a convolution algorithm that combines the mass distributions of isotopologues from individual amino acid residues. Aerith also implements three PSM scoring functions: multivariate hypergeometric (MVH), cross-correlation (Xcorr), and weighted dot product (WDP, originally developed for the Sipros algorithm (Pan et al., 2011; Xiong et al., 2024)). All algorithms were implemented in C++ for computational efficiency and integrated into R via Rcpp (Eddelbuettel and François, 2011). Given the amino acid sequence of a peptide and its isotopic abundances of SIP-labeled elements, theoretical spectra without isotopic fine structures are generated using a convolution-based

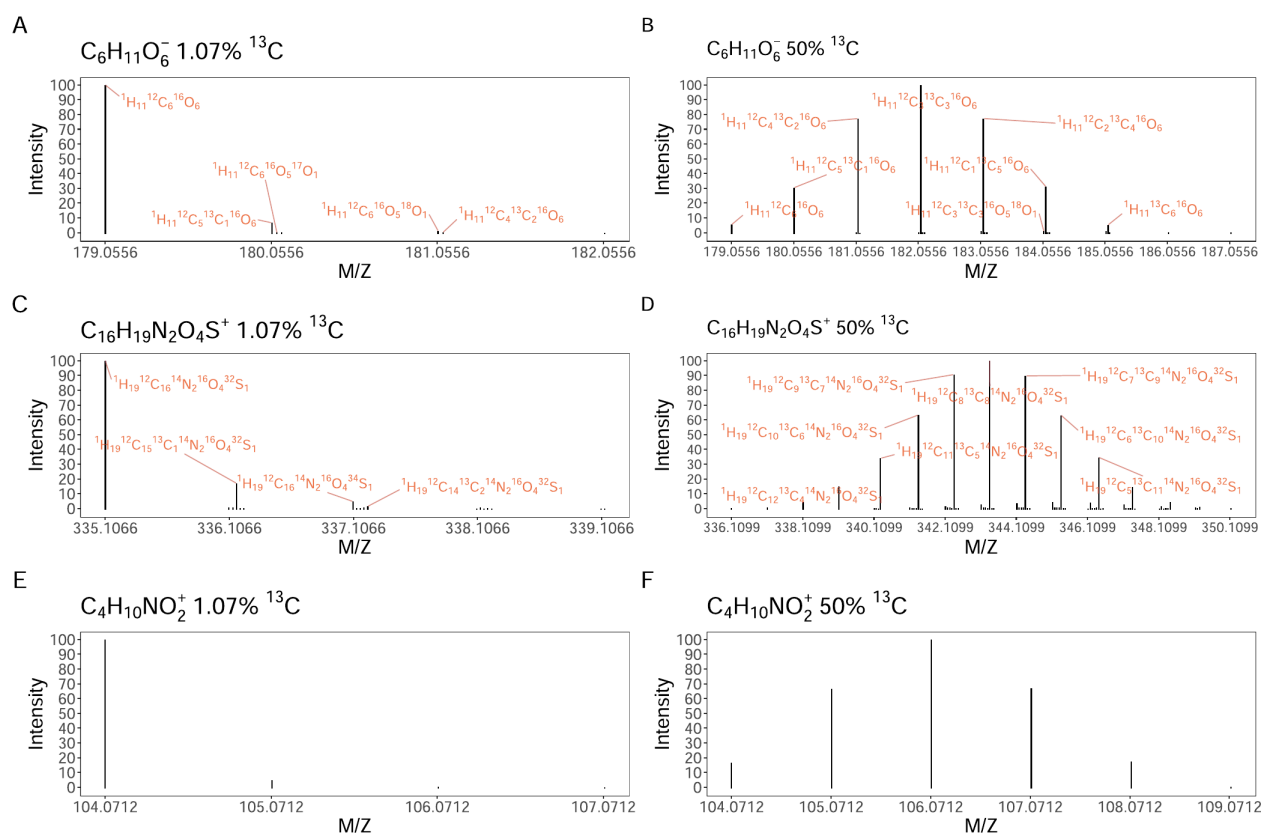


Figure 4.3: Simulated isotopic fine structures and envelopes of SIP-labeled metabolites. The isotopic fine structures of negative-ion-mode glucose ($[M-H]^-$) are simulated at 1.07% natural abundance (A) and 50% ^{13}C enrichment (B), which can be compared with measured spectra from MZcloud (Thermo Fisher; Figure C.1). Adjacent isotopic peaks with minor m/z differences (< 0.03 Da) were computationally offset while preserving true positional relationships for visual clarity. High-abundance isotopologues are annotated in red with their isotopic compositions. The isotopic fine structures of positive-ion-mode penicillin ($[M+H]^+$) at 1.07% natural abundance (C) and 50% ^{13}C enrichment (D) were validated against reference spectra from MZcloud (Figure C.2). The isotopic envelopes, ignoring fine structure, of positive-ion-mode γ -aminobutyric acid ($[M+H]^+$) at 1.07% natural abundance (E) and 50% ^{13}C enrichment (F) were generated using the FFT algorithm in comparison with experimental metabolomics data from natural-abundance and SIP-labeled samples.

algorithm combined with binomial distribution estimation.

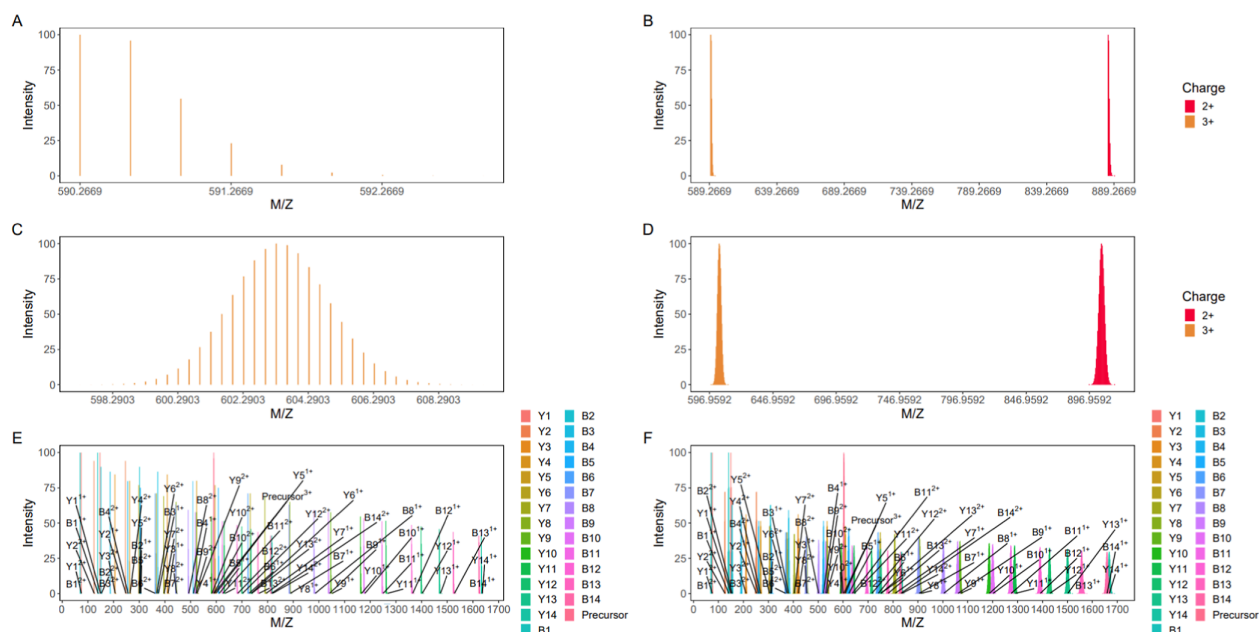


Figure 4.4: Simulated isotopic envelopes of SIP-labeled peptide precursors and fragment ions for HYAHVDCPGHADYVK. (A) Precursor ion (charge state +3) at 1.07% ^{13}C . (B) Precursor ions (charge states +2 [red] and +3 [yellow]) at 1.07% ^{13}C . (C) Precursor ion (+3) at 50% ^{13}C . (D) Precursor ions (+2 red, +3 yellow) at 50% ^{13}C . (E) B/Y fragment ions (charge states +1/+2) at 1.07% ^{13}C . (F) B/Y fragment ions (+1/+2) at 50% ^{13}C .

Visualization of SIP-labeled and unlabeled PSMs

Aerith employs ggplot2 (Wickham, 2016) to provide publication-quality visualization of PSMs. The visualization requires two spectral inputs: observed spectra (optionally preprocessed via a denoising function) and theoretical spectra. The output is a composite S4 object of PSM containing matched-ion metadata, spectral data, and customizable ggplot2 layers in the R environment.

Performance evaluation of PSM scoring functions

The PSM can be quantitatively assessed using MVH, Xcorr, WDP, and spectral entropy. Among these, the WDP scoring function originally developed for the Sipros algorithm demonstrates the narrowest peak width in score distributions across 0–100% ^{13}C abundance levels, indicating superior resolution in distinguishing true matches. Users may also implement custom scoring algo-

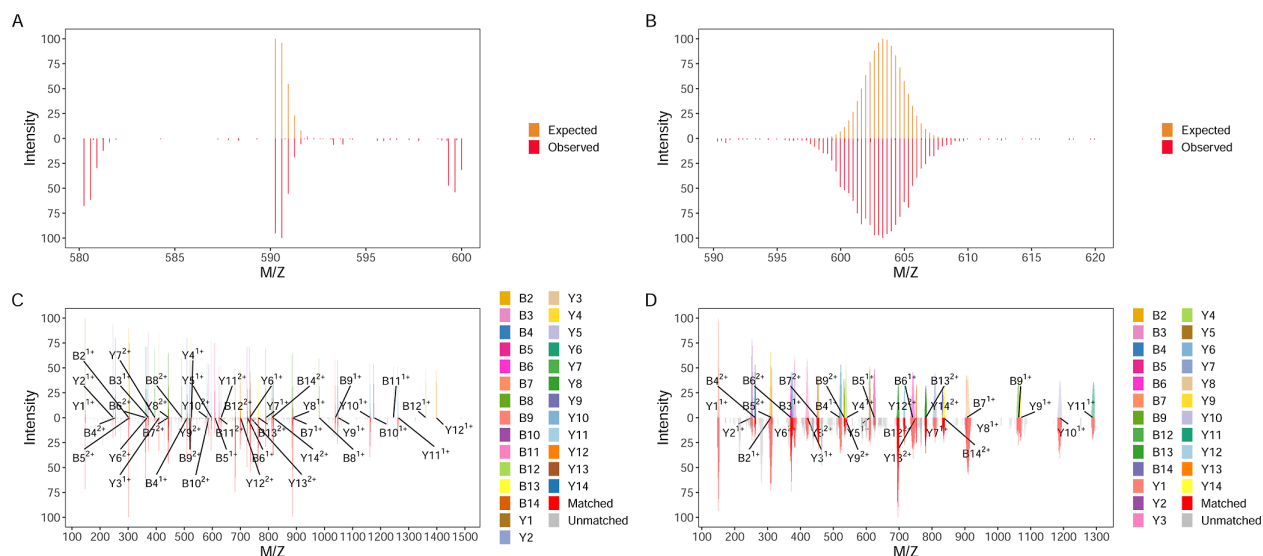


Figure 4.5: Visualization of PSMs with observed spectra in the lower panel and theoretical isotopic peaks in the upper panel. The highest-matching isotopic peaks within each envelope are highlighted. B/Y fragment ions are annotated based on user preferences. (A) Precursor PSM at 1.07% ^{13}C natural abundance, with theoretical envelope (yellow) and observed peaks (red). (B) Precursor PSM at 50% ^{13}C showing the mass shift from isotopic enrichment. (C) B/Y ion PSM at 1.07% ^{13}C ; unmatched peaks are annotated in gray. (D) B/Y ion PSM at 50% ^{13}C .

rithms via the package interface.

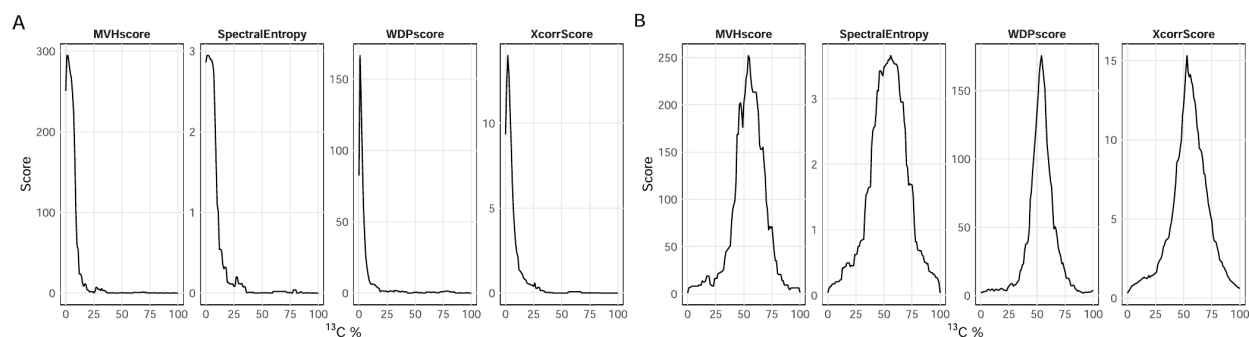


Figure 4.6: Comparison of scoring functions for estimation of ^{13}C enrichment levels from PSMs. (A) Scores (MVH, Xcorr, WDP, spectral entropy) for the peptide HYAHVDCPGHADYVK at 1.07% ^{13}C (natural abundance). (B) Scores for the peptide at 50% ^{13}C enrichment. WDP shows the sharpest separation between true and false matches across the full enrichment range.

Visualization of SIP proteomic results

The SIP enrichment levels of PSMs, peptides, and proteins generated by Sipros can be analyzed and visualized using the enrichment-visualization function. This visualization can be presented in

three dimensions: SIP-labeled isotopic abundance, number of spectra, and intensity.

Results and Discussion

Aerith establishes a robust computational framework for simulating isotopic patterns with a modular C++/R architecture. The package provides three distinct methods for theoretical isotopic envelope generation, each suited to specific applications. The FFT method achieves the highest computational speed with the lowest algorithmic complexity but does not account for isotopic fine structure, which can be resolved experimentally via high-resolution mass spectrometry (HRMS) or simulated using the Monte Carlo method. For small molecules with known chemical formulas, the FFT method is optimal. The Monte Carlo method is preferable when isotopic fine structure must be considered, such as in clumped isotopes. For peptides, the convolution-based method efficiently computes fragment envelopes.

Aerith enhances proteomic SIP workflows by enabling systematic validation of database search outputs. For scoring peptide-spectrum matches, the weighted dot product (WDP) function outperforms other methods for SIP-labeled PSMs (Figure 4.6) due to its comprehensive evaluation of all isotopic peaks in fragment ions (Appendix Section C.4.1). In contrast, when analyzing low-resolution mass spectra (e.g., ion-trap or time-of-flight MS2 measurements of peptide fragments), incomplete isotopic envelopes favor the Xcorr and MVH scoring functions. MS2 scans obtained in data-independent acquisition (DIA) mode can also result in incomplete isotopic envelopes. These two scoring functions are particularly suited to this situation. This preference arises because both methods prioritize the most abundant isotopic peak and account for noise-induced random matches, which are also common in TOF or Astral (Stewart et al., 2023) scans (Appendix Sections C.4.2 and C.4.3). MS2 scans obtained in DIA can be processed alongside DDA spectra using the same visualization pipeline, providing a uniform manual-validation tool for the Sipros 4 and Sipros 5 outputs presented in Chapters 2 and 3.

Three case studies illustrate Aerith’s utility: (1) negative-ion glucose $[M-H]^-$ fine structures at natural abundance and 50% ^{13}C validated against MZcloud; (2) positive-ion penicillin

envelopes against reference spectra and FFT-based γ -aminobutyric acid envelopes against experimental metabolomics data; and (3) the peptide HYAHVDCPGHADYVK precursor and B/Y fragments at natural abundance and 50% ^{13}C .

Conclusions

Few software tools focus specifically on isotopic mass spectra. The SIP-labeled spectra generation methods presented here provide an extensible isotopic-spectra generation framework that can be integrated into other software through the underlying C++ classes or imported directly into other R packages. Aerith facilitates the analysis of SIP-labeled proteomic and metabolomic data, provides interactive visualization for manual validation of labeled identifications, and complements the upstream Sipros 4 and Sipros 5 search engines. Beyond proteomic SIP, the same C++ kernel can simulate SIP-labeled metabolite spectra at arbitrary isotopic enrichment levels for any user-supplied chemical formula, suggesting natural extension to SIP-metabolomics studies. Prior work has demonstrated spectral entropy as a critical feature for scoring small-molecule spectral identifications (Li et al., 2021), and Aerith’s spectral-entropy implementation (Appendix Section C.6) could advance metabolite identification in SIP-metabolomics. Aerith also streamlines user-defined and automatic mass-spectral annotation, supporting validation of critical findings in proteomic SIP studies, especially under the broader isotopic envelopes encountered in narrow-window Astral DIA workflows (Stewart et al., 2023).

Software and data availability

The resources associated with this work are publicly available. The Aerith R package for spectral analysis is accessible at <https://github.com/thePANlab/Aerith>. The raw-file conversion tool Raxport can be downloaded from <https://github.com/xyz1396/Raxport.net>. For SIP proteomics searches, the Sipros program is hosted on GitHub at <https://github.com/thePANlab/Sipros4> and on Bioconda at <https://anaconda.org/bioconda/sipros>. Case-study RAW files are deposited in the PRIDE Archive under accession PXD041414 (<https://www.ebi.ac>.

uk/pride/), and the corresponding analysis scripts are embedded within the Aerith package vignettes. Graphical representations were generated using the ggplot2 (Wickham, 2016) package integrated with the patchwork (Pedersen, 2019) library within the R statistical computing environment (R Core Team, 2023).

Acknowledgments and funding

The authors acknowledge the IDeA National Resource for Quantitative Proteomics for conducting proteomic analyses utilizing LC–MS instrumentation and the OU Supercomputer Center for Education and Research (OSCER) for providing computational resources. This work was funded through the National Center for Complementary & Integrative Health and the National Institute of General Medical Sciences at the National Institutes of Health (Grant R01AT011618). The authors declare no competing financial interest.

Supplementary Material

The supplementary information accompanying this chapter (Xiong et al., 2025) is reproduced in Appendix C and reorganized under dissertation appendix numbering. Appendix Section C.1 presents isotopic fine-structure calculation for chemical formulas by Monte Carlo simulation; Section C.2 presents fast Fourier transform envelope calculation; Section C.3 provides reference spectra and convolution-based peptide envelopes; Section C.4 derives the WDP, MVH, and Xcorr scores; Section C.5 presents binomial precursor-mass estimation; and Section C.6 presents the spectral-entropy calculation for PSMs.

Summary and Output

Synthesis

The three data chapters of this dissertation address complementary bottlenecks in proteomic and metaproteomic SIP. Chapter 2 delivered Sipros 4, an enrichment-resolved search engine that combines a vectorized convolution algorithm with an updated scoring function to identify labeled PSMs across the full 1–99 atom% range with more than twenty-fold higher throughput than Sipros 3 and substantially better sensitivity in real-world soil metaproteomes. Chapter 3 extended this framework into a unified DDA/DIA platform (Sipros 5) that explicitly controls label-specific FDR using unlabeled negative controls and captures the fivefold sensitivity gain that narrow-window Astral DIA provides, demonstrated on ^{13}C -glucose, time-resolved ^{13}C -inulin, and 43-day ^{15}N -SILAM mouse gut microbiome experiments. Chapter 4 introduced Aerith, an R package that generates accurate theoretical spectra for arbitrary peptides and chemical formulas at any enrichment level and renders interactive publication-quality plots for manual validation of SIP-labeled PSMs.

Together, these tools form an end-to-end analytical pipeline for proteomic SIP studies that span single organisms, microbiome metaproteomes (soil, rhizosphere, gut), and tracer-based metabolomics, allowing investigators to move from raw MS files to label-FDR-controlled, manually validated lists of labeled peptides and proteins with substantially less manual effort than previously possible. The kinetic phenotyping demonstrated in Chapter 3 extends proteomic SIP from endpoint substrate tracing toward quantitative measurement of protein renewal in single organisms and host-associated microbial ecosystems.

Future directions

Three directions extend the work most naturally. First, the Sipros 5 deep-learning rescorer creates an opening for integrating prior labeled datasets across labs and substrates into a shared model

that learns the distribution of true SIP-labeled PSMs in a fully transferable way; future scoring strategies could explicitly model isotopic convolution and use neural-network–based fragment predictors. Second, linking Sipros 5 with single-cell proteomics workflows on the Orbitrap Astral platform could enable proteomic SIP at the cell-population level where current bulk SIP loses resolution; jointly modeling MS1 and MS2 isotope evidence and integrating multi-scan reconstruction along chromatographic elution profiles should further improve robustness and accuracy. Third, Aerith’s visualization framework can be generalized to multi-isotope and clumped-isotope experiments, providing a unified manual-validation tool across proteomic and metabolomic SIP.

Output of this dissertation

Peer-reviewed journal papers published or accepted

1. Xiong, Yi; et al. “Proteomic stable isotope probing with an upgraded Sipros algorithm for improved identification and quantification of isotopically labeled proteins.” *Microbiome* 12, 148 (2024). (Xiong et al., 2024)
2. Xiong, Yi; et al. “Aerith: Visualization and annotation of isotopic enrichment patterns of peptides and metabolites with stable isotope labeling from proteomics and metabolomics.” *Analytical Chemistry* 97, 24916–24921 (2025). (Xiong et al., 2025)
3. Feng, Shichao; Zhang, Bailu; Wang, Huan; Xiong, Yi; et al. “Enhancing peptide identification in metaproteomics through curriculum learning in deep learning.” *Nature Communications* 16, 8934 (2025). (Feng et al., 2025)
4. Zhang, Bailu; Feng, Shichao; Parajuli, Manushi; Xiong, Yi; et al. “SEMQuant: Extending Sipros-Ensemble with match-between-runs for comprehensive quantitative metaproteomics.” In *ISBRA*, Springer (2024). (Zhang et al., 2024)
5. Li, Jiancheng; Xiong, Yi; et al. “CloudProteoAnalyzer: Scalable processing of big data from proteomics using cloud computing.” *Bioinformatics Advances* (2024). (Li et al., 2024)
6. Dean, Danya A.; Roach, Jarrod; Ulrich vonBargen, Rebecca; Xiong, Yi; et al. *ACS Infectious*

Diseases (2023). (Dean et al., 2023)

7. Collins, Quincy P.; Grunsted, Michael J.; Arcila, Dahiana; Xiong, Yi; et al. *G3* (2023). (Collins et al., 2023)
8. Zhang, Li; Jonscher, Karen R.; Zhang, Zuyuan; Xiong, Yi; et al. “Islet autoantibody sero-conversion in type-1 diabetes is associated with metagenome-assembled genomes in infant gut microbiomes.” *Nature Communications* (2022). (Zhang et al., 2022)
9. Saha, Prasenjit; Fan, Lin; Thibivilliers, Sandra; Xiong, Yi; et al. *Frontiers in Plant Science* (2021). (Saha et al., 2021)

Manuscripts in preparation

1. Xiong, Yi; Feng, Shichao; Pan, Chongle. “Sipros 5 significantly enhances proteomic isotope probing in DDA Orbitrap and narrow-window DIA Astral mass spectrometry analyses.” Under review (2026). (Xiong et al., 2026)

References

- Berry, D., Stecher, B., Schintlmeister, A., Reichert, J., Brugiroux, S., Wild, B., Wanek, W., Richter, A., Rauch, I., Decker, T., Loy, A., and Wagner, M. (2013). Host-compound foraging by intestinal microbiota revealed by single-cell stable isotope probing. *PNAS*, 110:4720–4725.
- Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIA-MOND. *Nat. Methods*, 12:59–60.
- Chen, F. Y. H. et al. (2020). Converting *Escherichia coli* to a synthetic methylotroph growing solely on methanol. *Cell*, 182:933–946.
- Chen, S. et al. (2019). Root-associated microbiomes of wheat under the combined effect of plant development and nitrogen fertilization. *Microbiome*, 7:136.
- Collins, Q. P. et al. (2023). Transcriptomic analysis provides insight into the mechanism of IKK β -mediated suppression of HPV18E6-induced cellular abnormalities. *G3: Genes, Genomes, Genetics*.
- Coyotzi, S., Pratscher, J., Murrell, J. C., and Neufeld, J. D. (2016). Targeted metagenomics of active microbial populations with stable-isotope probing. *Curr. Opin. Biotechnol.*, 41:1–8.
- da Veiga Leprevost, F. et al. (2020). Philosopher: a versatile toolkit for shotgun proteomics data analysis. *Nat. Methods*, 17:869–870.
- Dean, D. A. et al. (2023). Persistent biofluid small-molecule alterations induced by *Trypanosoma cruzi* infection are not restored by parasite elimination. *ACS Infectious Diseases*.
- Dumont, M. G. and Murrell, J. C. (2005). Stable isotope probing — linking microbial identity to function. *Nature Reviews Microbiology*, 3:499–504.
- Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *J. Stat. Softw.*, 40:1–18.
- Eng, J. K., Jahan, T. A., and Hoopmann, M. R. (2013). Comet: an open-source MS/MS sequence database search tool. *Proteomics*, 13:22–24.
- Eng, J. K., McCormack, A. L., and Yates, J. R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.*, 5:976–989.
- Feng, S. et al. (2025). Enhancing peptide identification in metaproteomics through curriculum learning in deep learning. *Nature Communications*, 16:8934.
- Fondrie, W. E. and Noble, W. S. (2021). mokapot: Fast and flexible semisupervised learning for peptide detection. *J. Proteome Res.*, 20:1966–1971.

- Grayston, S. J. et al. (1998). Selective influence of plant species on microbial diversity in the rhizosphere. *Soil Biol. Biochem.*, 30:369–378.
- Guo, X., Li, Z., Yao, Q., Mueller, R. S., Eng, J. K., Tabb, D. L., Hervey IV, W. J., and Pan, C. (2018). Sipros Ensemble improves database searching and filtering for complex metaproteomics. *Bioinformatics*, 34:795–802.
- Gupta, M. et al. (2024). Global protein turnover quantification in *Escherichia coli* reveals cytoplasmic recycling under nitrogen limitation. *Nature Communications*, 15:5890.
- Guzman, U. H. et al. (2024). Ultra-fast label-free quantification and comprehensive proteome coverage with narrow-window data-independent acquisition. *Nature Biotechnology*, 42:1855–1866.
- Haichar, F. E. Z. et al. (2008). Plant host habitat and root exudates shape soil bacterial community structure. *ISME J.*, 2:1221–1230.
- Haichar, F. E. Z. et al. (2014). Root exudates mediated interactions belowground. *Soil Biol. Biochem.*, 77:69–80.
- Heil, L. R. et al. (2023). Evaluating the astral mass analyzer for quantitative proteomics using data-independent acquisition. *J. Proteome Res.*, 22:3290–3300.
- Hellerstein, M. K. and Neese, R. A. (1999). Mass isotopomer distribution analysis at eight years: theoretical, analytic, and experimental considerations. *Am. J. Physiol.*, 276:E1146–E1170.
- Huerta-Cepas, J. et al. (2017). Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.*, 34:2115–2122.
- Jehmlich, N., Schmidt, F., Taubert, M., Seifert, J., Bastida, F., von Bergen, M., Richnow, H.-H., and Vogt, C. (2010). Protein-based stable isotope probing. *Nature Protocols*, 5:1957–1966.
- Jehmlich, N., Schmidt, F., von Bergen, M., Richnow, H.-H., and Vogt, C. (2008). Protein-based stable isotope probing reveals active species within anoxic mixed cultures. *ISME J.*, 2:1122–1133.
- Jeon, C. O., Park, W., Padmanabhan, P., DeRito, C., Snape, J. R., and Madsen, E. L. (2003). Discovery of a bacterium, with distinctive dioxygenase, that is responsible for in situ biodegradation in contaminated sediment. *PNAS*, 100:13591–13596.
- Justice, N. B. et al. (2014). ¹⁵N- and ²H-proteomic stable isotope probing links nitrogen flow to archaeal heterotrophic activity. *Environ. Microbiol.*, 16(10):3224–3237.
- Käll, L. et al. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods*, 4:923–925.
- Keltjens, J. T. et al. (2014). PQQ-dependent methanol dehydrogenases: rare-earth elements make a difference. *Appl. Microbiol. Biotechnol.*, 98:6163–6183.

- Kieft, B., Li, Z., Bryson, S., Hettich, R. L., Pan, C., Mayali, X., and Mueller, R. S. (2021). Phytoplankton exudates and lysates support distinct microbial consortia with specialized metabolic and ecophysiological traits. *PNAS*, 118(41):e2101178118.
- Kleiner, M., Dong, X., Hinzke, T., Wippler, J., Thorson, E., Mayer, B., and Strous, M. (2018). Metaproteomics method to determine carbon sources and assimilation pathways of species in microbial communities. *PNAS*, 115:E5576–E5584.
- Kleiner, M. et al. (2017). Assessing species biomass contributions in microbial communities via metaproteomics. *Nature Communications*, 8:1558.
- Lane, N. (2020). How energy flow shapes cell evolution. *Curr. Biol.*, 30:R471–R476.
- Lane, N. and Martin, W. (2010). The energetics of genome complexity. *Nature*, 467:929–934.
- Li, J., Xiong, Y., et al. (2024). CloudProteoAnalyzer: Scalable processing of big data from proteomics using cloud computing. *Bioinformatics Advances*.
- Li, W. et al. (2025). Turnover atlas of proteome and phosphoproteome across mouse tissues and brain regions. *Cell*, 188:2267–2287.
- Li, Y., Kind, T., Folz, J., Vaniya, A., Mehta, S. S., and Fiehn, O. (2021). Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. *Nat. Methods*, 18:1524–1531.
- Li, Z., Yao, Q., Guo, X., Crits-Christoph, A., Mayes, M. A., Hervey IV, W. J., Lebeis, S. L., Banfield, J. F., Hurst, G. B., Hettich, R. L., and Pan, C. (2019). Genome-resolved proteomic stable isotope probing of soil microbial communities using $^{13}\text{CO}_2$ and ^{13}C -methanol. *Frontiers in Microbiology*, 10:2706.
- Macey, M. C. et al. (2020). Impact of plants on the diversity and activity of methylophs in soil. *Microbiome*, 8:31.
- Manefield, M., Whiteley, A. S., Griffiths, R. I., and Bailey, M. J. (2002). RNA stable isotope probing, a novel means of linking microbial community function to phylogeny. *Appl. Environ. Microbiol.*, 68:5367–5373.
- Mordant, A., Blakeley-Ruiz, J. A., and Kleiner, M. (2025). Metaproteomics-based stable isotope fingerprinting links intestinal bacteria to their carbon source and captures diet-induced substrate switching. *ISME J.*, 19:wraf127.
- Neufeld, J. D., Dumont, M. G., Vohra, J., and Murrell, J. C. (2007). Methodological considerations for the use of stable isotope probing in microbial ecology. *Microbial Ecology*, 53:435–442.
- Oldham, K. B., Myland, J. C., and Spanier, J. (2009). The Error Function $\text{erf}(x)$ and its complement $\text{erfc}(x)$. In *An Atlas of Functions: with Equator, the Atlas Function Calculator*, pages 405–415. Springer, New York.

- Pan, C., Fischer, C. R., Hyatt, D., Bowen, B. P., Hettich, R. L., and Banfield, J. F. (2011). Quantitative tracking of isotope flows in proteomes of microbial communities. *Mol. Cell. Proteomics*, 10(4):M110.006049.
- Pedersen, T. L. (2019). Patchwork: The composer of plots. CRAN.
- Picone, N. and Op den Camp, H. J. M. (2019). Role of rare earth elements in methanol oxidation. *Curr. Opin. Chem. Biol.*, 49:39–44.
- R Core Team (2023). R: A language and environment for statistical computing.
- Radajewski, S., Ineson, P., Parekh, N. R., and Murrell, J. C. (2000). Stable-isotope probing as a tool in microbial ecology. *Nature*, 403:646–649.
- Reddien, P. W. (2024). The purpose and ubiquity of turnover. *Cell*, 187:2657–2681.
- Richardson, L. et al. (2023). MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.*, 51:D753–D759.
- Rockwood, A. L. and Haimi, P. (2006). Efficient calculation of accurate masses of isotopic peaks. *J. Am. Soc. Mass Spectrom.*, 17:415–419.
- Sachsenberg, T., Herbst, F.-A., Taubert, M., Kermer, R., Jehmlich, N., von Bergen, M., Seifert, J., and Kohlbacher, O. (2015). MetaProSIP: Automated inference of stable isotope incorporation rates in proteins for functional metaproteomics. *J. Proteome Res.*, 14:619–627.
- Saha, P., Fan, L., Thibivilliers, S., Xiong, Y., et al. (2021). Phenylpropanoid biosynthesis gene expression precedes lignin accumulation during shoot development in lowland and upland switchgrass genotypes. *Frontiers in Plant Science*.
- Seifert, J., Taubert, M., Jehmlich, N., Schmidt, F., Voelker, U., Vogt, C., Richnow, H.-H., and von Bergen, M. (2012). Protein-based stable isotope probing (protein-SIP) in functional metaproteomics. *Mass Spectrometry Reviews*, 31:683–697.
- Senko, M. W., Beu, S. C., and McLafferty, F. W. (1995). Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.*, 6:229–233.
- Smyth, P., Zhang, X., Ning, Z., Mayne, J., Moore, J. I., Walker, K., Lavallée-Adam, M., and Figeys, D. (2020). Studying the temporal dynamics of the gut microbiota using metabolic stable isotope labeling and metaproteomics. *Analytical Chemistry*, 92(24):15711–15718.
- Stewart, D. P. C. and Methereell, A. K. (1999). Carbon (^{13}C) uptake and allocation in pasture plants following field pulse-labelling. *Plant Soil*, 210:61–73.
- Stewart, H. I., Grinfeld, D., Giannakopoulos, A., Petzoldt, J., Shanley, T., Garland, M., Denisov, E., Peterson, A. C., Damoc, E., Zeller, M., et al. (2023). Parallelized acquisition of Orbitrap and Astral analyzers enables high-throughput quantitative analysis. *Analytical Chemistry*, 95:15656–15664.

- Tabb, D. L., Fernando, C. G., and Chambers, M. C. (2007). MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. Proteome Res.*, 6:654–661.
- Vega-Sagardía, M. et al. (2024). Screening microbial interactions during inulin utilization reveals strong competition and proteomic changes in *Lactocaseibacillus paracasei* M38. *Probiotics Antimicrob. Proteins*, 16:993–1011.
- von Bergen, M. et al. (2013). Insights from quantitative metaproteomics and protein-stable isotope probing into microbial ecology. *ISME J.*, 7(10):1877–1885.
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York.
- Wiśniewski, J. R., Zougman, A., Nagaraj, N., and Mann, M. (2009). Universal sample preparation method for proteome analysis. *Nat. Methods*, 6:359–362.
- Xiong, Y., Feng, S., Pan, C., et al. (2026). Proteomic stable isotope probing to quantify gut microbial activity and protein turnover (Sipros 5). Under review.
- Xiong, Y., Mueller, R. S., Feng, S., Guo, X., and Pan, C. (2024). Proteomic stable isotope probing with an upgraded Sipros algorithm for improved identification and quantification of isotopically labeled proteins. *Microbiome*, 12(1):148.
- Xiong, Y., Mueller, R. S., Feng, S., Guo, X., and Pan, C. (2025). Aerith: Visualization and annotation of isotopic enrichment patterns of peptides and metabolites with stable isotope labeling from proteomics and metabolomics. *Analytical Chemistry*, 97(45):24916–24921.
- Yu, F., Deng, Y., and Nesvizhskii, A. I. (2025). MSFragger-DDA+ enhances peptide identification sensitivity with full isolation window search. *Nature Communications*, 16:3329.
- Yu, F. et al. (2023). Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform. *Nature Communications*, 14:4154.
- Zhang, B., Feng, S., Parajuli, M., Xiong, Y., et al. (2024). SEMQuant: Extending Sipros-ensemble with match-between-runs for comprehensive quantitative metaproteomics. In *ISBRA*. Springer.
- Zhang, L., Jonscher, K. R., Zhang, Z., Xiong, Y., et al. (2022). Islet autoantibody seroconversion in type-1 diabetes is associated with metagenome-assembled genomes in infant gut microbiomes. *Nature Communications*.
- Zhu, Y.-t. et al. (2021). Prebiotics inulin metabolism by lactic acid bacteria from young rabbits. *Front. Vet. Sci.*, 8.

Appendix A: Supplementary Material for Chapter 2

This appendix reproduces the supplementary content accompanying Xiong et al. (2024) and summarizes the legacy Sipros 4 labeled-search workflow as implemented in the Sipros4 repository.

A.1 Supplementary tables

Table A.1: Comparison of different isolation-window sizes for 50% ^{13}C -labeled *E. coli*. MAD denotes median absolute deviation. FDRs at the peptide and protein levels were controlled at 1 %.

Level	Metric	0.8 Da	1.5 Da	3 Da	5 Da	7 Da
PSM	median $^{13}\text{C}\%$	50	50	50	50	50
	MAD $^{13}\text{C}\%$	4.45	4.45	4.45	4.45	2.97
	count	6,123	9,740	9,962	16,137	9,878
Peptide	median $^{13}\text{C}\%$	50	50	50	50	49.5
	MAD $^{13}\text{C}\%$	2.97	2.97	2.97	2.97	2.22
	count	2,589	4,045	4,173	6,200	4,073
Protein/group	median $^{13}\text{C}\%$	50	50	50	50	50
	MAD $^{13}\text{C}\%$	1.48	1.48	1.48	1.48	1.48
	count	403	616	659	894	705

Table A.2: Comparison of Sipros 4 and Sipros 3 using standard ^{13}C -labeled *E. coli* samples. Running time is wall-clock time in minutes on the 24-core compute server. FDRs at the peptide and protein levels were controlled at 1 %.

Expected $^{13}\text{C}\%$	PSMs		Peptides		Proteins/groups		Time (min)	
	Sipros 4	Sipros 3	Sipros 4	Sipros 3	Sipros 4	Sipros 3	Sipros 4	Sipros 3
1%	64,113	63,305	18,147	18,003	1,815	1,834	30.2	983.8
2%	65,131	66,050	14,523	14,824	1,458	1,546	25.7	932.9
5%	46,665	47,855	11,297	12,420	1,210	1,428	32.7	1026.2
25%	43,526	20,758	8,188	5,478	973	871	32.2	952.1
50%	37,659	6,138	7,275	2,218	893	476	32.7	955.6
99%	26,583	27,356	9,901	10,042	1,490	1,491	31.6	980.6

Table A.3: Unlabeled regular search results of Sipros Ensemble, Proteome Discoverer 3.0, and MaxQuant 2.0 on standard ^{13}C -labeled *E. coli* samples. NA indicates that database searching failed to produce any identification. FDRs at peptide and protein levels were controlled at 1 %.

$^{13}\text{C}\%$	PSMs			Peptides			Proteins/groups		
	SE	PD3.0	MQ2.0	SE	PD3.0	MQ2.0	SE	PD3.0	MQ2.0
1%	87,623	69,966	63,581	28,423	16,774	14,113	1,977	1,988	1,869
2%	79,967	62,523	57,333	22,787	14,043	11,457	1,646	1,716	1,588
5%	36,702	26,234	17,222	16,208	9,642	5,698	1,533	1,566	1,291
25%	NA	NA	NA	NA	NA	NA	NA	NA	NA
50%	NA	NA	NA	NA	NA	NA	NA	NA	NA
99%	NA	NA	NA	NA	NA	NA	NA	NA	NA

SE = Sipros Ensemble; PD3.0 = Proteome Discoverer 3.0; MQ2.0 = MaxQuant 2.0.

Table A.4: Comparison of Sipros 4 and Sipros 3 using the standard ^{15}N -labeled acid mine drainage (AMD) samples. FDRs at peptide and protein level were controlled at 1 %.

Expected $^{15}\text{N}\%$	PSMs		Peptides		Proteins/groups		median $^{15}\text{N}\%$	
	Sipros 4	Sipros 3	Sipros 4	Sipros 3	Sipros 4	Sipros 3	Sipros 4	Sipros 3
0.37% (nat. ab.)	43,289	39,546	20,156	19,324	2,730	2,961	0.4%	0%
50%	29,626	18,172	12,085	8,799	1,643	1,516	47%	49%
98%	41,976	42,286	19,811	20,366	2,586	2,608	96%	99%

Table A.5: Unlabeled PSMs, peptides, and proteins identified in the initial soil, ^{13}C -methanol SIP soil, and $^{13}\text{CO}_2$ SIP soil. FDRs at peptide and protein levels were controlled at 1 %.

SIP treatment	Organisms	# PSMs	# Peptides	# Proteins/groups
Initial soil	Microbes	111,787	27,377	11,392
^{13}C -methanol SIP	Microbes	178,481	61,221	14,437
$^{13}\text{CO}_2$ SIP	Microbes	663,752	61,906	20,166
	Plants	22,973	4,311	686

Table A.6: Recipes for ^{13}C -labeled M9 media used for standard *E. coli* cultures (1 mL of 20 % D-glucose solution, M9 minimal salts, MgSO_4 and CaCl_2 stocks). Final 40-mL working solution: 20 mL of M9 $2\times$ salts, 0.8 mL of 20 % glucose, 0.08 mL of 1 M MgSO_4 , 0.004 mL of 1 M CaCl_2 , 0.04 mL of $1000\times$ vitamin, and 19.076 mL of H_2O .

Target $^{13}\text{C}\%$	Reagent	Formula	MW	1 mL usage (g)
$\sim 1\%$	D-Glucose	$\text{C}_6\text{H}_{12}\text{O}_6$	180.16	0.2000
	D-Glucose- $^{13}\text{C}_6$	$^{13}\text{C}_6\text{H}_{12}\text{O}_6$	186.11	0.0000
$\sim 2\%$	D-Glucose	$\text{C}_6\text{H}_{12}\text{O}_6$	180.16	0.1980
	D-Glucose- $^{13}\text{C}_6$	$^{13}\text{C}_6\text{H}_{12}\text{O}_6$	186.11	0.0020
$\sim 5\%$	D-Glucose	$\text{C}_6\text{H}_{12}\text{O}_6$	180.16	0.1917
	D-Glucose- $^{13}\text{C}_6$	$^{13}\text{C}_6\text{H}_{12}\text{O}_6$	186.11	0.0083
$\sim 25\%$	D-Glucose	$\text{C}_6\text{H}_{12}\text{O}_6$	180.16	0.1500
	D-Glucose- $^{13}\text{C}_6$	$^{13}\text{C}_6\text{H}_{12}\text{O}_6$	186.11	0.0500
$\sim 50\%$	D-Glucose	$\text{C}_6\text{H}_{12}\text{O}_6$	180.16	0.0989
	D-Glucose- $^{13}\text{C}_6$	$^{13}\text{C}_6\text{H}_{12}\text{O}_6$	186.11	0.1011
$\sim 99\%$	D-Glucose	$\text{C}_6\text{H}_{12}\text{O}_6$	180.16	0.0000
	D-Glucose- $^{13}\text{C}_6$	$^{13}\text{C}_6\text{H}_{12}\text{O}_6$	186.11	0.2000

M9 minimal salts $2\times$ (per 250 mL): 6.4 g $\text{Na}_2\text{HPO}_4\cdot 7\text{H}_2\text{O}$, 1.5 g KH_2PO_4 , 0.25 g NaCl , 0.5 g NH_4Cl ; adjust pH 6.6–7.0. 1 M MgSO_4 : 1.2037 g per 10 mL. 1 M CaCl_2 : 0.11098 g per 1 mL.

Table A.7: Genome-resolved proteomic SIP results from the soil community samples. Each row corresponds to one labeled protein assigned to a metagenome-assembled genome (MAG). The full gene ontology (GO) term list per row has been omitted for space; the full spreadsheet (24 rows $\times$ 35 columns including all GO terms) is available in the publisher deposit. This excerpt shows representative rows from both ^{13}C -methanol and $^{13}\text{CO}_2$ SIP, with key columns retained for the dissertation: SIP treatment, time, plant host, MAG, enrichment level, PSM count, short description, gene, enzyme commission (EC) number, KEGG orthology (KO), phylum, and order. “_” indicates not applicable; “#N/A” indicates no functional or taxonomic annotation was returned.

SIP	Time	Plant	MAG	$^{13}\text{C}\%$	PSMs	Description (short)	Gene	EC	KO	Phylum	Order
^{13}C -methanol	Day 8	–	MAG21	5	2	Isocitrate dehydrogenase (NADP)	icd	1.1.1.42	K00031	#N/A	#N/A
^{13}C -methanol	Day 8	–	MAG13	87	1	Enolase (glycolysis)	eno	4.2.1.11	K01689	Proteobacteria	Rhizobiales
^{13}C -methanol	Day 8	–	MAG31	7	1	–	–	–	–	Acidobacteriota	UBA5066
^{13}C -methanol	Day 3	–	MAG14	5	1	Acyl carrier protein	acpP	–	K02078	Proteobacteria	Sphingomonadales
^{13}C -methanol	Day 8	–	MAG04	5	1	Sulfatase	–	3.1.6.1	K01130	#N/A	#N/A
^{13}C -methanol	Day 8	–	MAG13	5	3	Endostatin-domain protein	–	–	–	Proteobacteria	Rhizobiales
^{13}C -methanol	Day 8	–	MAG05	6	1	Ni-containing superoxide dismutase	sodN	1.15.1.1	K00518	Actinobacteriota	Propionibacteriales
$^{13}\text{CO}_2$	–	Arabidopsis	MAG48	5	1	XoxF (methanol dehydrogenase, PQQ)	xoxF	1.1.2.8	K00114	Proteobacteria	Rhizobiales
$^{13}\text{CO}_2$	–	Wheat	MAG01	50	1	Cold-shock protein	cspA_1	–	K03704	Actinobacteriota	Mycobacteriales
$^{13}\text{CO}_2$	–	Maize	MAG46	17	1	Extracellular solute-binding protein	sugB	–	K02027	Actinobacteriota	Actinomycetales
$^{13}\text{CO}_2$	–	Maize	MAG44	5	1	Aconitase B	acnB	4.2.1.3	K01682	Proteobacteria	Pseudomonadales
$^{13}\text{CO}_2$	–	Arabidopsis	MAG05	7	1	30S ribosomal protein S1	rpsA	–	K02945	Actinobacteriota	#N/A
$^{13}\text{CO}_2$	–	Maize	MAG01	5	1	S-adenosylmethionine synthetase	metK	2.5.1.6	K00789	#N/A	#N/A
$^{13}\text{CO}_2$	–	Arabidopsis	MAG45	7	1	Glyoxylate carboligase	gcl	4.1.1.47	K01608	Proteobacteria	Burkholderiales

Continued on next page

Table A.7: Genome-resolved proteomic SIP results from the soil community samples, continued.

SIP	Time	Plant	MAG	¹³ C%	PSMs	Description (short)	Gene	EC	KO	Phylum	Order
¹³ CO ₂	–	Wheat	MAG46	27.5	2	Succinyl-CoA synthetase β	sucC	6.2.1.5	K01903	Actinobacteriota	Actinomycetales
¹³ CO ₂	–	Wheat	MAG46	24.5	4	Dihydrolipoyl dehydrogenase	lpdA	1.8.1.4	K00382	Actinobacteriota	Actinomycetales
¹³ CO ₂	–	Arabidopsis	MAG54	21.75	4	Aldehyde-oxidase/xanthine dehydrogenase	iorB	1.3.99.16	K07303	Proteobacteria	Rhizobiales
¹³ CO ₂	–	Maize	MAG20	21	1	GAPDH	gapA	1.2.1.12	K00134	#N/A	#N/A
¹³ CO ₂	–	Maize	MAG42	11	1	Enoyl-(ACP) reductase	–	1.1.1.100/ 47	K00034	Proteobacteria	Xanthomonadales
¹³ CO ₂	–	Maize	MAG44	16	1	Phosphate-binding protein	pstS	–	K02040	Proteobacteria	Pseudomonadales

Table A.8: Abstract summary of the Appendix A supplementary workbooks that are too large to reproduce fully in the dissertation. The complete spreadsheets remain available from the open-access publisher deposit.

Workbook	Source sheet or workbook	Representative row	Representative columns	Abstracted content
Publisher source S5 workbook	Mouse stool peptide workbook	One nonredundant ¹⁵ N-labeled peptide entry from mouse stool	Peptide sequence; PSM count; protein identifier; label percentage	Nonredundant reference peptides and labeled peptide identifications from mouse stool, plus the peptide- and protein-level summaries used in the chapter.
Publisher source S8 workbook	Soil SIP labeled-protein workbook	One labeled protein assigned to a MAG	SIP treatment; time; plant host; MAG; gene; EC; KO; phylum; order	Labeled protein identifications from soil SIP experiments; the full workbook also contains the omitted GO-term columns and the complete MAG annotations.

A.2 Supplementary figures

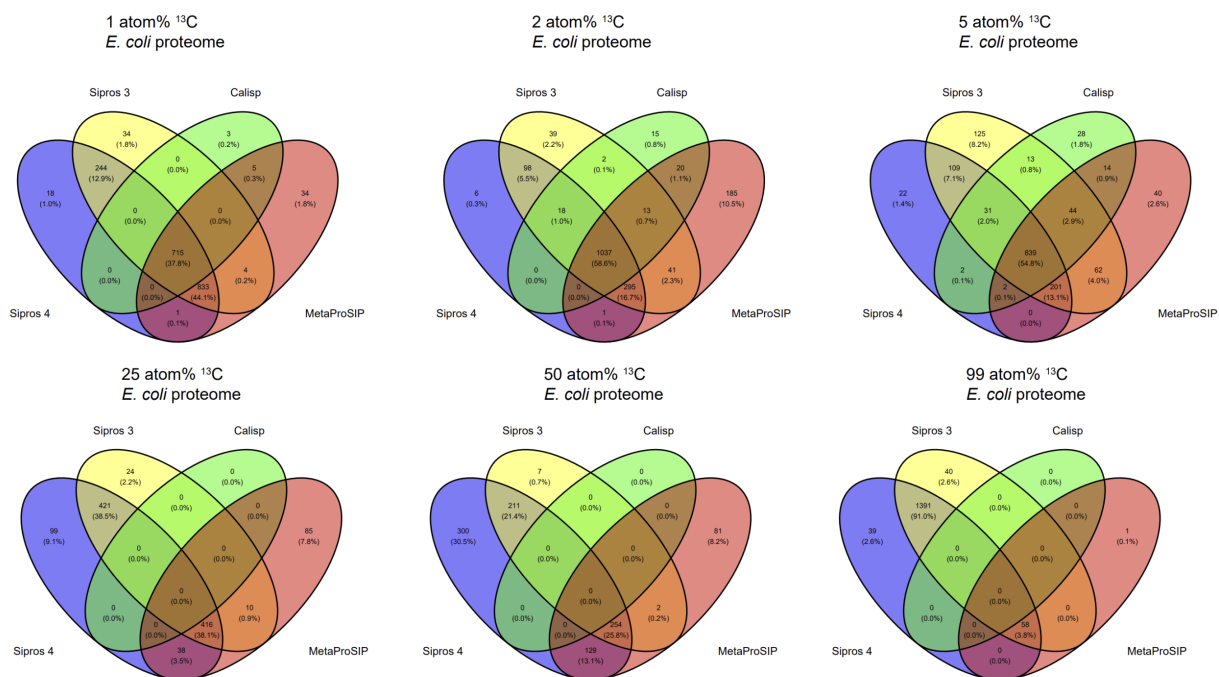


Figure A.1: Venn diagrams of the proteins identified by Sipro 3, Sipro 4, Calisp, and MetaProSIP on the standard *E. coli* samples. Across all six enrichment levels, Sipro 4 captures the largest unique-identification set. At 25 and 50 atom% ^{13}C , the Sipro 4 set entirely encompasses the (much smaller) MetaProSIP set, while Calisp returns no identifications.

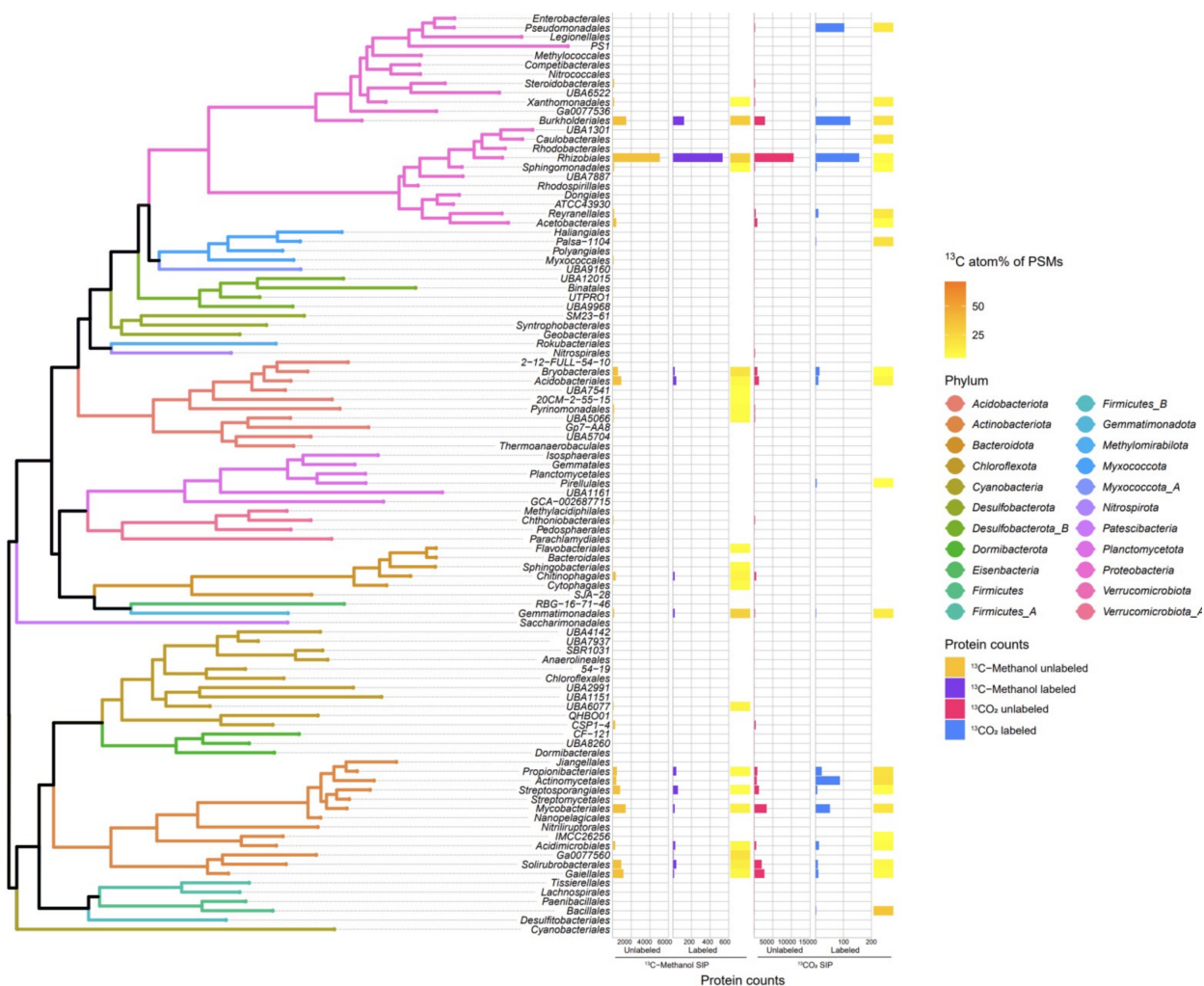


Figure A.2: Taxonomic tree of the microbial proteins identified in the initial soils, the ^{13}C -methanol SIP soils, and the $^{13}\text{CO}_2$ SIP soils. Tree tips represent inferred Orders; branches are colored by Phylum. The bar charts show numbers of unlabeled and labeled proteins per Order in the two SIP experiments. Heatmap columns encode the average enrichment levels of labeled proteins per Order in the two experiments.

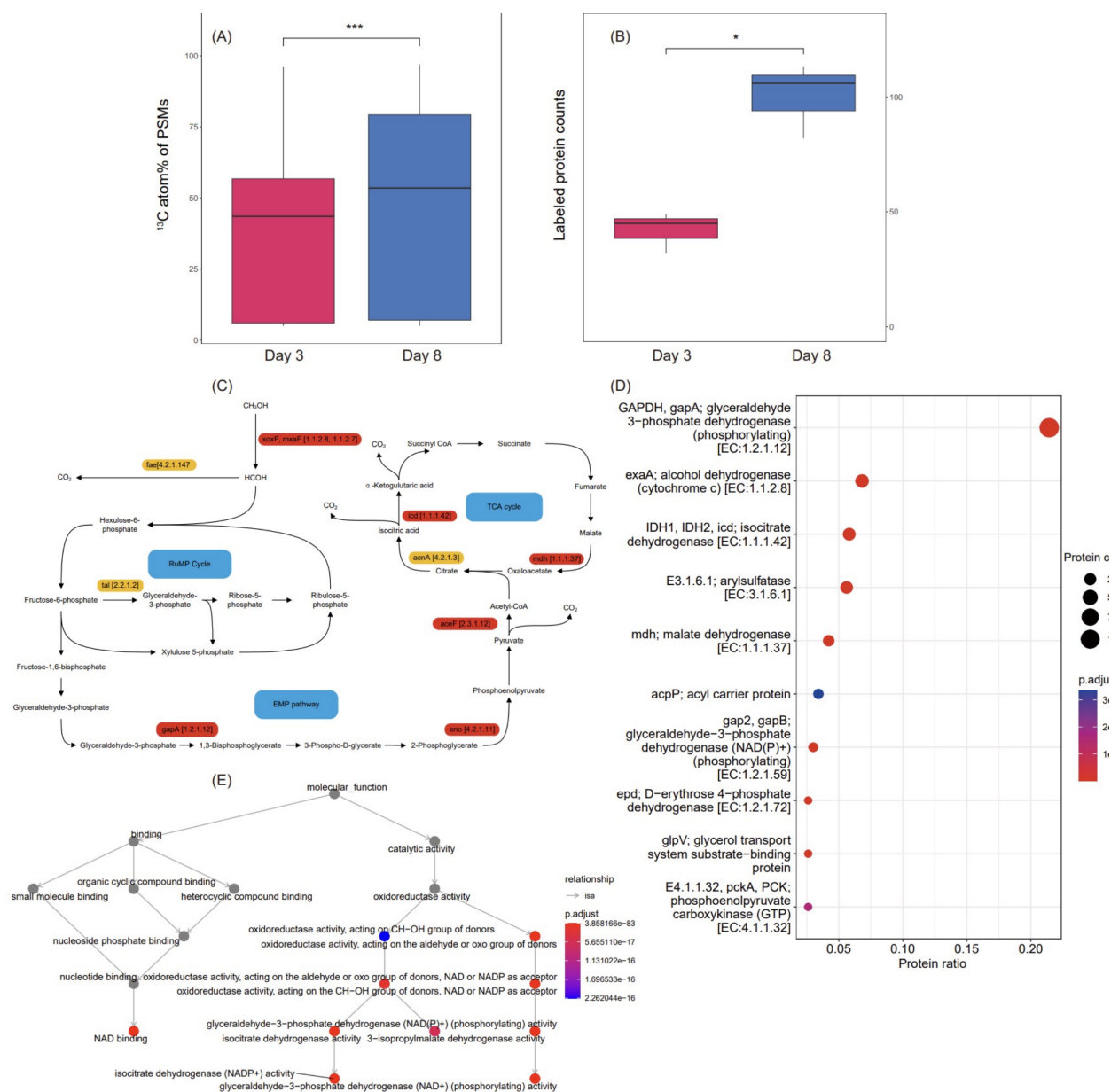


Figure A.3: Functional analysis of the ^{13}C -methanol SIP results. (A) Boxplot of ^{13}C enrichment levels of PSMs identified at day 3 versus day 8 ($p < 0.001$). (B) Boxplot of labeled-protein counts at day 3 vs. day 8 ($p < 0.05$). (C) ^{13}C -labeled enzymes involved in methanol degradation; EC numbers in yellow for identified enzymes, red for significantly enriched. (D) Top-10 enriched KEGG Orthology (KO) terms ($p_{\text{adj}} < 0.01$). (E) Enriched molecular-function GO terms ($p_{\text{adj}} < 0.01$).

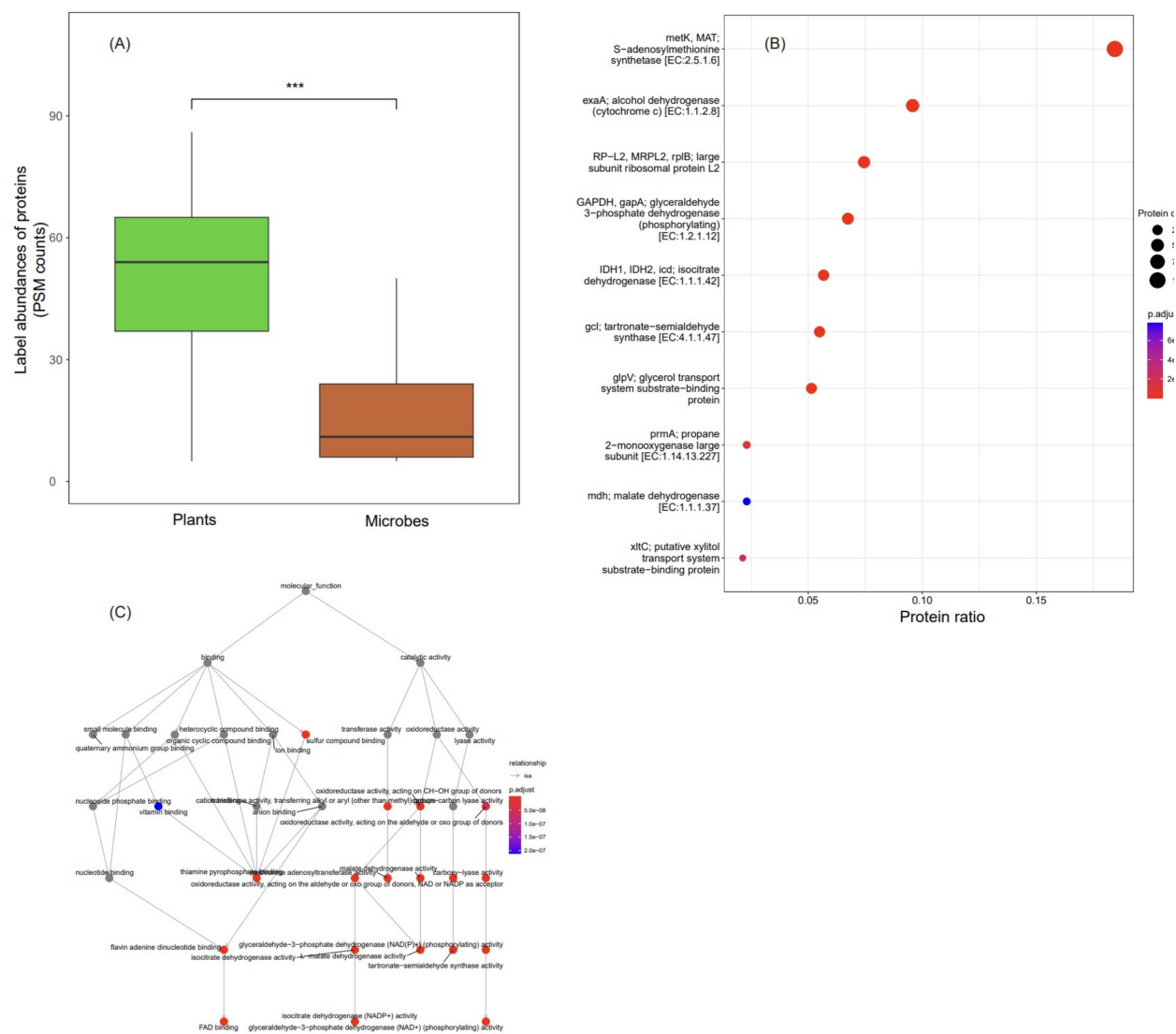


Figure A.4: Functional analysis of the $^{13}\text{CO}_2$ SIP results. (A) Total label abundances of plant proteins vs. microbial proteins ($p < 0.001$). (B) Top-10 enriched KO terms ($p_{\text{adj}} < 0.01$). (C) Enriched molecular-function GO terms ($p_{\text{adj}} < 0.01$).

A.3 Sipros 4 labeled-search workflow

The control flow enumerates a set of generated `.cfg` files, one per configured enrichment level or search condition, and each search invocation produces a matching `.sip` file. The wrapper then hands those `.sip` files to the filtering and assembly scripts, which merge the evidence back into the standard PSM, peptide, protein, and clustering outputs. In the dissertation text below, the enrichment ceiling is described on the human-readable 0–100% scale; the historical configuration file stores the same ceiling as a thousand-fold scaled integer, so the user-facing 100% corresponds to the legacy value 100000 after dividing by 1000.

The main search template keeps the search space focused on isotope labeling rather than open modification discovery. It sets `Search_Type = SIP`, uses HCD fragmentation, opens parent-ion windows at `-2, -1, 0, 1, 2`, and applies 0.01 Da tolerances to both parent and fragment ions. Digestion is tryptic, with cleavage after K/R, up to two missed cleavages, try-first-methionine cleavage enabled, and peptide lengths restricted to 6–60 residues. The template also disables an explicit PTM search with `Max_PTM_Count = 0`; that keeps the search focused on the isotopic broadening introduced by carbon labeling rather than on unrelated mass shifts. On the labeling side, the stable-isotope block fixes the labeled element to carbon, searches 0–100% enrichment in 1% increments, and keeps the clustering threshold at 20% enrichment with at least one PSM per isotopic cluster. The legacy configuration additionally records `minValue = 0.005` and `fold = 4`, which tune the isotope-search deduction coefficient used by the labeled-search code.

Taken together, those choices make Sipros 4 a practical enrichment-resolved search engine. The search engine explores a deliberately small modification space, evaluates candidate peptides at each enrichment level, and hands a compact, charge-aware set of candidate matches to the post-processing scripts that control FDR and protein inference.

A.3.1 Search-space preparation and FT conversion

The first stage of the workflow prepares the search environment. The helper script creates dedicated `configs`, `raw`, `ft`, `sip`, and `fasta` directories, then uses `configGenerator` to materialize labeled-search configuration files from the template. For workflows that start from a previously analyzed unlabeled run, `makeDBforLabelSearch.R` converts the earlier protein calls into a reduced FASTA database so that the labeled search is not forced to rediscover proteins that are already irrelevant to the sample. This reduction step matters in SIP because every candidate peptide must be considered over many enrichment levels; trimming the search space early pays off in both throughput and specificity.

The same preparation stage also converts the RAW files to FT1/FT2 format through Raxport. Sipros 4 consumes the FT files as the substrate for the enrichment-resolved search, and the FT conversion determines which precursor and fragment scans are available to the search engine. In the legacy script, the conversion and search stages are deliberately separated so that the same FT files can be reused for multiple downstream filtering and clustering passes.

Pseudocode

Input:

vendor RAW files, reference FASTA, prior unlabeled `.pro.txt` files, SIP template

Output:

FT files and a labeled-search FASTA

1. Prepare working directories

create `configs`, `raw`, `ft`, `sip`, and `fasta`

2. Build search configuration

generate SIP config files with `configGenerator`

for each generated `.cfg` file, set the search mode to SIP and keep the enrichment ceiling on the human-readable 0–100% scale

3. Reduce the protein database when prior results exist

if unlabeled protein calls are available:

build a reduced labeled-search FASTA with `makeDBforLabelSearch.R`

else:

use the full reference FASTA

4. Convert vendor files

convert RAW to FT1/FT2 with Raxport
index the FT files by run so they can be searched and post-processed together

A.3.2 Enrichment-resolved database search

Once the FT files are available, SiprosV40MP performs the actual labeled database search. The binary is invoked once per generated configuration file, and each invocation writes one corresponding .sip result file. For each observed spectrum, the engine enumerates candidate peptides from the reference database, evaluates them across the configured enrichment grid, and builds theoretical isotope envelopes for the precursor and fragment ions before scoring the match. In the legacy workflow, the enrichment grid spans 0–100% in 1% increments, which is the range that matters for the ^{13}C -labeled *E. coli* standards and the soil SIP experiments reported in the chapter. The search is intentionally exhaustive within that range: a candidate peptide that matches well at one enrichment level can match poorly at another, so the reported PSM is the best-scoring candidate after the enrichment scan has completed.

The wrapper script in `V4Scripts/SIPcmd.sh` reflects that one-config/one-output structure directly. It creates the configuration directory, generates the per-enrichment config files, runs the search once for each config, and then dispatches the resulting .sip files to filtering, assembly, clustering, and protein-FDR refinement. That design makes the search stage easy to parallelize: each configuration is independent, so the wrapper can hand different configs to different workers while preserving a clean filename-to-result mapping.

Pseudocode

Input:

generated .cfg files, FT1/FT2 spectra, labeled FASTA, search tolerances

Output:

one .sip file per configuration file

1. Enumerate configuration files
 - collect all generated .cfg files in the configs directory
 - sort them so the searches are reproducible
2. Search each configuration independently
 - for each config file c:

```
load c into the Sipros search engine
search the available FT1/FT2 scans against the labeled FASTA
evaluate candidate peptides across the configured enrichment grid
write one matching .sip output for c
```

3. Return the result set
hand the full collection of .sip files to the filtering stage

A.3.3 PSM and peptide filtering

The raw search output is still too permissive to use directly. `sipros_peptides_filtering.py` groups candidate matches by spectrum, separates the evidence by charge state, and computes score cutoffs from target and decoy counts at the configured FDR threshold. The script writes temporary working files first, then emits the final `.psm.txt` and `.pep.txt` tables once the cutoffs are known. This is the point where the search transitions from a ranked list of candidates to a conservative set of accepted identifications.

A subtle but important feature of the filtering stage is that it preserves the best evidence across repeated observations of the same peptide. The peptide table is not simply a deduplicated copy of the PSM table; it consolidates repeated hits while retaining the score and charge information needed for later protein assembly. That makes the peptide report the natural bridge between the spectrum-level search and the protein-level inference.

Pseudocode

Input:

`.sip` search files and the Sipros config

Output:

`.psm.txt` and `.pep.txt` filtered at the requested FDR

1. Load search outputs
read all SIP result files in the working directory
collect per-spectrum candidates, peptide strings, scores, and target/decoy labels
2. Estimate score cutoffs
for each charge state bucket:
sort the candidate scores
scan for the smallest score that satisfies the configured FDR threshold
store the charge-specific cutoff

3. Emit filtered tables

- write PSMs that meet the cutoff to `.psm.txt`
- collapse PSMs by peptide and write the accepted peptides to `.pep.txt`

A.3.4 Protein assembly and clustering

After peptide filtering, `sipros_peptides_assembling.py` maps each accepted peptide back to the FASTA database and builds the peptide-to-protein and protein-to-peptide relationships needed for protein inference. The script then merges indistinguishable proteins that explain exactly the same peptide set and applies a greedy selection routine that prefers proteins with broader peptide support and unique evidence. The output is a standard Sipros protein report set: `.pro.txt` at the protein level and the supporting `.pro2pep.txt` and `.pro2psm.txt` linkage tables.

The later `ClusterSip.py` step is specific to labeled workflows. Once proteins have been inferred, Sipros 4 clusters the SIP abundance estimates within each protein group so that enrichment-related variation can be summarized at the cluster level rather than at the raw PSM level. The configured clustering threshold of 20% and the minimum one-PSM requirement keep the clusters from becoming so permissive that weak, noisy identifications dominate the label estimate.

Pseudocode

Input:

- filtered peptide and PSM tables, reference FASTA, protein-level thresholds

Output:

- protein groups and enrichment clusters

1. Build protein evidence maps

- for each accepted peptide:
 - map the peptide to all compatible proteins in the FASTA database
 - update the protein-to-peptide and peptide-to-protein graphs

2. Merge indistinguishable proteins

- find proteins that are supported by the exact same peptide set
- collapse them into one protein group

3. Greedy protein inference

- repeatedly select the protein that covers the largest remaining peptide set
- remove peptides already explained by the selected protein from the unresolved pool
- stop when the minimum peptide and unique-peptide thresholds are no longer satisfied

4. Cluster SIP evidence
 - group protein-level label estimates by enrichment proximity
 - retain clusters that satisfy the minimum PSM-per-cluster requirement

A.3.5 Protein-FDR refinement and label quantification

The final cleanup step uses `refineProteinFDR.R` to apply the requested protein-level FDR threshold to the assembled protein groups, producing the refined protein list used for downstream interpretation. The companion script, `getLabelPCTinEachFT.R`, reports the label percentage in each FT file from the refined protein groups and the configured abundance threshold, avoiding spurious low-support estimates. In the dissertation chapter, these summaries turn the search-engine output into biologically interpretable counts of labeled proteins across enrichment conditions.

At this point the workflow is complete: the search has been filtered at the spectrum, peptide, and protein levels; ambiguous protein assignments have been collapsed; enrichment clusters have been stabilized; and the final per-run label percentages are ready for comparison across experiments.

Pseudocode

Input:

- protein groups, accepted PSMs, and per-run enrichment summaries

Output:

- refined protein FDR calls and FT-specific label percentages

1. Refine protein groups
 - run `refineProteinFDR.R` on the assembled protein table
 - apply the requested protein FDR threshold and export the refined protein set
2. Summarize label percentage per FT file
 - run `getLabelPCTinEachFT.R` on the refined protein table and the PSM table
 - filter to protein groups above the abundance threshold
 - report the per-file label percentage and the associated support counts
3. Return the final outputs
 - emit the refined protein groups and the run-level SIP summaries

Appendix B: Supplementary Material for Chapter 3

This appendix reproduces the supplementary content accompanying the Sipros 5 manuscript (Xiong et al., 2026) and summarizes the Sipros 5 workflow as implemented in the Sipros5 repository.

B.1 Supplementary tables

The original Sipros 5 supplementary tables are distributed as Excel workbooks. Compact tables below reproduce the small summaries and selected excerpt tables used in the dissertation; where a workbook contains hundreds to thousands of rows, the caption identifies the omitted full-sheet scope rather than implying that the excerpt is complete.

Table B.1: Benchmark summary of Sipros 5 vs. Sipros 4 on the ^{13}C -glucose *E. coli* isotope-enrichment standards. PSM, peptide and protein counts are reported for three workflows: Sipros 4 with Orbitrap DDA, Sipros 5 with Orbitrap DDA, and Sipros 5 with Astral DIA. Sipros 5–Astral DIA consistently produces $2\text{--}3\times$ more PSMs in Orbitrap DDA and approximately $5\times$ more in Astral DIA than Sipros 4. Data extracted from the bar charts of Figure 3.2.

$^{13}\text{C}\%$	PSMs			Peptides			Proteins		
	S4-DDA	S5-DDA	S5-DIA	S4-DDA	S5-DDA	S5-DIA	S4-DDA	S5-DDA	S5-DIA
1%	64,113	162,635	995,445	18,147	19,301	30,360	1,815	2,052	2,383
2%	65,131	160,827	674,057	14,523	17,215	26,389	1,458	1,796	2,168
5%	46,665	142,078	660,338	11,297	16,477	22,499	1,210	1,803	1,958
25%	43,526	105,516	430,942	8,188	11,338	16,046	973	1,389	1,542
50%	37,659	94,937	399,666	7,275	10,263	13,036	893	1,241	1,381
99%	26,583	111,804	768,793	9,901	16,486	30,043	1,490	1,985	2,323

S4 = Sipros 4; S5 = Sipros 5. DDA = data-dependent acquisition on Orbitrap Fusion (90-min run); DIA = narrow-window data-independent acquisition on Orbitrap Astral (35-min run).

Table B.2: Performance of filtering tools on Astral DIA ^{13}C -labeled *E. coli* peptide spike-ins into mouse fecal peptides. The *E. coli* PSM and peptide ratios indicate the fraction of identifications assigned to spike-in *E. coli* after each filter. Peptide ^{13}C median and median absolute deviation (MAD) are peptide-level inferred ^{13}C abundance summaries. A dash indicates a zero value, missing value due to run failure, or non-applicability.

Tool	$^{13}\text{C}\%$	PSMs	Peptides	Proteins	<i>E. coli</i> PSM ratio	<i>E. coli</i> peptide ratio	^{13}C med.	MAD
Sipros 5 WinnowNet	2	10,463	2,359	596	0.90	0.77	2.00	1.00
	5	73,033	6,950	1,274	0.93	0.75	5.00	2.00
	50	24,561	2,349	586	0.94	0.84	49.00	5.00
Sipros 5 Percolator	2	—	—	—	—	—	—	—
	5	24,377	3,560	757	0.94	0.83	7.00	2.00
	50	24,761	2,367	599	0.93	0.77	49.00	5.00
Sipros 5 Mokapot	2	—	—	—	—	—	—	—
	5	11,588	2,222	516	0.93	0.86	6.00	2.00
	50	24,762	2,367	599	0.93	0.77	49.00	5.00
No SIP filter	2	41,720	5,128	1,143	0.14	0.10	2.00	—
	5	91,456	7,880	1,504	0.29	0.16	4.00	1.48
	50	26,623	2,635	792	0.10	0.05	48.50	3.71

MAD = median absolute deviation, $\text{MAD} = \text{median}(|x_i - \text{median}(x)|)$.

Table B.3: Gifu anaerobic broth composition and working-medium preparation used for the ^{13}C -inulin time-course cecal incubation. Gifu medium was diluted 10-fold ($0.1\times$) and supplemented with 4 g/L inulin (Thermo Scientific), replacing the standard fructan source. For SIP experiments, inulin was substituted with 100% uniformly labeled U- ^{13}C -inulin (IsoLife). Sodium thioglycolate (0.3 g/L) and cysteine were added to maintain anaerobiosis.

Component	Formulation	Concentration
Pancreatic digest of casein	Gifu base	5.0 g/L
Soy peptone	Gifu base	3.0 g/L
Yeast extract	Gifu base	10.0 g/L
Meat extract	Gifu base	2.2 g/L
Liver extract	Gifu base	1.2 g/L
Glucose	Carbon source	3.0 g/L
Soluble starch	Carbon source	5.0 g/L
L-Cysteine HCl	Reducing agent	0.3 g/L
Sodium chloride	Salts	3.0 g/L
Dipotassium phosphate	Buffer	2.5 g/L
Working dilution	10-fold ($0.1\times$ Gifu)	—
SIP fructan substitution	U- ^{13}C -inulin (IsoLife)	4.0 g/L
Anaerobic supplement	Sodium thioglycolate	0.3 g/L
Anaerobic supplement	L-Cysteine (additional)	to anaerobiosis
Headspace	75% N_2 / 25% CO_2	—
Final pH	adjusted	7.3

Approximate composition; exact lot-specific values are listed in the source spreadsheet. Cecal incubations were performed in 100-mL serum bottles containing 10 mL of working medium at 37 °C for 24 h with sampling at 0, 4, 8, 12, 16, 20, and 24 h.

Table B.4: Top-12 gene ontology (GO) biological-process terms enriched among ^{13}C -incorporating proteins from the M9 ^{13}C -glucose cecal incubation, ranked by Benjamini–Hochberg-adjusted P value. “Count” is the number of labeled proteins mapped to each term; “Ratio” is labeled-protein count divided by the labeled-protein background. The full enrichment workbook accompanies the Sipros 5 manuscript.

Term ID	Description	Count	Ratio	p_{adj}
GO:0009058	biosynthetic process	321	321/479	1.8e-07
GO:0043604	amide biosynthetic process	180	180/479	3.7e-05
GO:0009059	macromolecule biosynthetic process	187	187/479	4.6e-05
GO:0043043	peptide biosynthetic process	171	171/479	4.6e-05
GO:0006412	translation	171	171/479	4.6e-05
GO:0043603	amide metabolic process	186	186/479	1.1e-04
GO:0010467	gene expression	192	192/479	1.2e-04
GO:0006518	peptide metabolic process	175	175/479	1.2e-04
GO:0000028	ribosomal small subunit assembly	25	25/479	0.011
GO:0006520	amino acid metabolic process	106	106/479	0.012
GO:0016070	RNA metabolic process	81	81/479	0.029
GO:0043170	macromolecule metabolic process	230	230/479	0.033

Table B.5: Top-6 GO molecular-function terms enriched among ^{13}C -incorporating proteins from the M9 ^{13}C -glucose incubation. “Count” is the labeled-protein count and “Ratio” is the labeled-protein ratio against the enrichment background.

Term ID	Description	Count	Ratio	p_{adj}
GO:0003735	structural constituent of ribosome	154	154/520	4.3e-05
GO:0005198	structural molecule activity	155	155/520	0.022
GO:0016874	ligase activity	59	59/520	0.224
GO:0140098	catalytic activity, acting on RNA	56	56/520	0.224
GO:0004812	aminoacyl-tRNA ligase activity	40	40/520	0.224
GO:0016875	ligase activity, forming carbon-oxygen bonds	40	40/520	0.224

Table B.6: Top-10 KEGG orthology (KO) terms enriched among ^{13}C -incorporating proteins from the M9 ^{13}C -glucose incubation, ranked by adjusted P value.

KO ID	Description	Count	p_{adj}
K04077	groEL, HSPD1, HSP60, CPN60; chaperonin GroEL [EC:5.6.1.7]	73	0.033
K01738	cysK; cysteine synthase [EC:2.5.1.47]	30	0.033
K01610	pckA; phosphoenolpyruvate carboxykinase (ATP) [EC:4.1.1.49]	67	0.187
K02358	tuf, TUFM; elongation factor Tu	117	0.281
K00134	GAPDH, gapA; glyceraldehyde 3-phosphate dehydrogenase (phosphorylat...	86	0.281
K02935	RP-L7, MRPL12, rplL; large subunit ribosomal protein L7/L12	35	0.281
K02886	RP-L2, MRPL2, RML2, rplB; large subunit ribosomal protein L2	28	0.281
K02967	RP-S2, MRPS2, rpsB; small subunit ribosomal protein S2	25	0.281
K02897	RP-L25, rplY; large subunit ribosomal protein L25	12	0.281
K02948	RP-S11, MRPS11, rpsK; small subunit ribosomal protein S11	14	0.309

Table B.7: Top-12 GO biological-process terms enriched among ^{13}C -inulin-incorporating proteins, ranked by adjusted P value. The inulin response is dominated by glycolytic, pyruvate, nucleoside-diphosphate, and pyridine/NAD-cofactor metabolism, consistent with fermentative central carbon flow from fructose entry. The full enrichment workbook and curated inulin-enzyme set accompany the Sipros 5 manuscript.

Term ID	Description	Count	P_{adj}
GO:0009135	purine nucleoside diphosphate metabolic process	30	2.5e-04
GO:0009179	purine ribonucleoside diphosphate metabolic pro...	30	2.5e-04
GO:0009185	ribonucleoside diphosphate metabolic process	30	2.5e-04
GO:0046031	ADP metabolic process	29	2.5e-04
GO:0006096	glycolytic process	29	2.5e-04
GO:0009166	nucleotide catabolic process	29	2.5e-04
GO:0042866	pyruvate biosynthetic process	29	2.5e-04
GO:0006090	pyruvate metabolic process	29	2.5e-04
GO:0009132	nucleoside diphosphate metabolic process	31	2.6e-04
GO:0019359	nicotinamide nucleotide biosynthetic process	30	2.6e-04
GO:0019363	pyridine nucleotide biosynthetic process	30	2.6e-04
GO:0072525	pyridine-containing compound biosynthetic process	30	2.6e-04

Table B.8: Top-6 GO molecular-function terms enriched among ^{13}C -inulin-incorporating proteins.

Term ID	Description	Count	P_{adj}
GO:0050840	extracellular matrix binding	13	2.2e-06
GO:0001968	fibronectin binding	13	2.2e-06
GO:0043236	laminin binding	13	2.2e-06
GO:2001065	mannan binding	30	0.022
GO:0004634	phosphopyruvate hydratase activity	7	0.022
GO:0035375	zymogen binding	7	0.022

Table B.9: Top-10 KEGG orthology terms enriched among ^{13}C -inulin-incorporating proteins. Elongation factor Tu (gene *tuf*) leads the list with $p_{\text{adj}} = 4.8 \times 10^{-17}$; downstream highly enriched terms include pyruvate–ferredoxin oxidoreductase (genes *por/nifJ*), GAPDH (*gapA*), chaperonin GroEL, pyruvate–orthophosphate dikinase (*ppdK*), enolase, elongation factor G (*fusA*), bifunctional GlcN-1-P transferase (*glmU*), formate dehydrogenase (*fdoG/fdhF/fdwA*), and anaerobic CO dehydrogenase (*cooS/acsA*).

KO ID	Description	Count	P_{adj}
K02358	tuf, TUFM; elongation factor Tu	71	4.8e-17
K03737	por, nifJ; pyruvate-ferredoxin/flavodoxin oxidoreductase [EC:1.2.7....	47	1.0e-06
K00134	GAPDH, gapA; glyceraldehyde 3-phosphate dehydrogenase (phosphorylat...	47	6.2e-06
K04077	groEL, HSPD1, HSP60, CPN60; chaperonin GroEL [EC:5.6.1.7]	41	1.4e-04
K01006	ppdK; pyruvate, orthophosphate dikinase [EC:2.7.9.1]	40	1.4e-04
K01689	ENO1_2_3, eno; enolase 1/2/3 [EC:4.2.1.11]	21	6.6e-04
K02355	fusA, GFM, EFG; elongation factor G	50	0.005
K04042	glmU; bifunctional UDP-N-acetylglucosamine pyrophosphorylase / gluc...	7	0.013
K00123	fdoG, fdhF, fdwA; formate dehydrogenase major subunit [EC:1.17.1.9]	10	0.013
K00198	cooS, acsA; anaerobic carbon-monoxide dehydrogenase catalytic subun...	15	0.013

Table B.10: Time-resolved abundance of curated inulin-utilization enzyme steps (% of curated enzyme pool) from the 24-h ^{13}C -inulin cecal incubation. Step 1 (GH32 β -fructofuranosidase, EC 3.2.1.26) carries the early inulin hydrolysis; Step 2 (PTS fructose/mannose EIIA-D family) drives early/intermediate fructose uptake and phosphorylation; Step 3 (ATP- and PPi-dependent 6-phosphofructokinase, EC 2.7.1.11/2.7.1.90) takes over at later glycolytic entry. The shift from Step 2 to Step 3 between 16 and 24 h captures the *Lactobacillus*→*Blautia* succession.

Step	4 h	8 h	12 h	16 h	20 h	24 h
Step 1: inulin hydrolysis	6.07 %	11.11 %	9.35 %	8.24 %	1.58 %	1.45 %
Step 2: fructose uptake & phosphor.	35.43 %	23.64 %	33.22 %	20.33 %	16.49 %	8.92 %
Step 3: F6P enters glycolysis (PFK)	58.50 %	65.25 %	57.43 %	71.43 %	81.93 %	89.63 %

Table B.11: Counts of total terms tested versus significant terms ($FDR \leq 0.05$) across microbiome and mouse-host ^{15}N -labeled proteins, by annotation source and class.

Compartment	Annotation source	Annotation class	Total tested	Significant ($FDR \leq 0.05$)
Microbiome	Gene Ontology	Biological Process	216	31
Microbiome	Gene Ontology	Cellular Component	27	0
Microbiome	Gene Ontology	Molecular Function	50	6
Microbiome	KEGG	KEGG Module	113	3
Microbiome	KEGG	KEGG Orthology	159	15
Microbiome	KEGG	KEGG Pathway	223	0
Microbiome	KEGG	KEGG Reaction	236	17
Mouse host	Gene Ontology	Biological Process	1,838	128
Mouse host	Gene Ontology	Cellular Component	255	68
Mouse host	Gene Ontology	Molecular Function	334	73
Mouse host	KEGG	KEGG Pathway	143	24
Mouse host	Reactome	Reactome Pathway	255	9

Table B.12: Top-10 GO biological-process terms enriched among ^{15}N -labeled microbiome proteins in the 43-day ^{15}N -SILAM mouse-stool time course, ranked by adjusted P value.

Term ID	Description	Count	p_{adj}
GO:0019752	carboxylic acid metabolic process	28	3.5e-05
GO:0006082	organic acid metabolic process	28	3.5e-05
GO:0043436	oxoacid metabolic process	28	3.5e-05
GO:0006520	amino acid metabolic process	19	8.2e-05
GO:0044281	small molecule metabolic process	36	0.002
GO:0044283	small molecule biosynthetic process	14	0.033
GO:0016052	carbohydrate catabolic process	9	0.033
GO:0005996	monosaccharide metabolic process	9	0.033
GO:0046496	nicotinamide nucleotide metabolic process	8	0.033
GO:0034655	nucleobase-containing compound catabolic process	8	0.033

Table B.13: Top-10 KEGG-orthology terms enriched among ^{15}N -labeled microbiome proteins. Flagellin (genes *fliC/hag*) is the most strongly enriched ($p_{\text{adj}} = 1.3 \times 10^{-11}$), followed by chaperonin GroEL, N-acetylglucosamine transport-system substrate-binding protein (*ngcE*), 5'-nucleotidase (*ushA*), alkaline phosphatase (*phoA/phoB*), elongation factor Tu (*tuf*), mannosyl-glycoprotein endo- β -N-acetylglucosaminidase (ENGASE), bifunctional autolysin (*atl*), internalin B (*inlB*), and formate-tetrahydrofolate ligase (*fhs*). Parenthetical lowercase italic symbols are gene aliases reproduced from the KEGG descriptions.

KO ID	Description	Count	p_{adj}
K02406	fliC, hag; flagellin	146	1.3e-11
K04077	groEL, HSPD1, HSP60, CPN60; chaperonin GroEL [EC:5.6.1.7]	57	7.1e-08
K10200	ngcE; N-acetylglucosamine transport system substrate-binding p...	21	1.8e-06
K11751	ushA; 5'-nucleotidase / UDP-sugar diphosphatase [EC:3.1.3.5 3....	25	1.1e-05
K01077	E3.1.3.1, phoA, phoB; alkaline phosphatase [EC:3.1.3.1]	22	1.1e-04
K02358	tuf, TUFM; elongation factor Tu	41	5.4e-04
K01227	ENGASE; mannosyl-glycoprotein endo-beta-N-acetylglucosaminidas...	10	0.016
K13714	atl; bifunctional autolysin [EC:3.5.1.28 3.2.1.96]	10	0.016
K13731	inlB; internalin B	10	0.016
K01938	fhs; formate-tetrahydrofolate ligase [EC:6.3.4.3]	13	0.025

Table B.14: Top-10 GO biological-process terms enriched among ^{15}N -labeled mouse host proteins. Digestion (GO:0007586), regulation of body-fluid levels, response to metal ion (GO:0010038), regulation of wound healing, acute-phase response, microvillus organization, and antimicrobial humoral response are the strongest signals.

Term ID	Description	Count	p_{adj}
GO:0007586	digestion	11	4.3e-08
GO:0050878	regulation of body fluid levels	12	3.3e-05
GO:0010038	response to metal ion	12	3.3e-05
GO:0061041	regulation of wound healing	8	3.3e-05
GO:0006953	acute-phase response	6	3.3e-05
GO:0032528	microvillus organization	5	3.3e-05
GO:0010288	response to lead ion	5	3.3e-05
GO:0019730	antimicrobial humoral response	9	7.0e-05
GO:0032536	regulation of cell projection size	4	8.1e-05
GO:0032530	regulation of microvillus organization	4	8.1e-05

Table B.15: Top-10 Reactome pathways enriched among ^{15}N -labeled mouse host proteins. Digestion (R-MMU-8935690), Digestion and absorption (R-MMU-8963743), and Antimicrobial peptides (R-MMU-6803157) dominate the list ($p_{\text{adj}} < 10^{-7}$), consistent with the gut-epithelial barrier and luminal immune defense.

Reactome ID	Description	Count	p_{adj}
R-MMU-8935690	Digestion	8	9.2e-10
R-MMU-8963743	Digestion and absorption	8	1.8e-09
R-MMU-6803157	Antimicrobial peptides	11	2.9e-08
R-MMU-163125	Post-translational modification: synthesis of GPI-anchore...	6	0.004
R-MMU-1592389	Activation of Matrix Metalloproteinases	5	0.004
R-MMU-1462054	Alpha-defensins	4	0.009
R-MMU-2022377	Metabolism of Angiotensinogen to Angiotensins	3	0.009
R-MMU-6809371	Formation of the cornified envelope	5	0.032
R-MMU-1461973	Defensins	4	0.032
R-MMU-1474228	Degradation of the extracellular matrix	5	0.078

Table B.16: Abstract summary of the large Sipros 5 supplementary workbook that is too large to reproduce fully in the dissertation. The complete spreadsheet remains available from the manuscript supplement.

Workbook	Source sheet or workbook	Representative row	Representative columns	Abstracted content
Publisher source S5 workbook	Mouse fecal ^{15}N time-course workbook	One labeled PSM row from a mouse fecal sample at a specific time point	PSM count; peptide sequence; protein-level summary; time point	Raw ^{15}N -labeled mouse fecal PSM-level data with peptide- and protein-level summaries across nine time points; the dissertation preserves the aggregate statistics and the family-level trajectories.

B.2 Supplementary figures

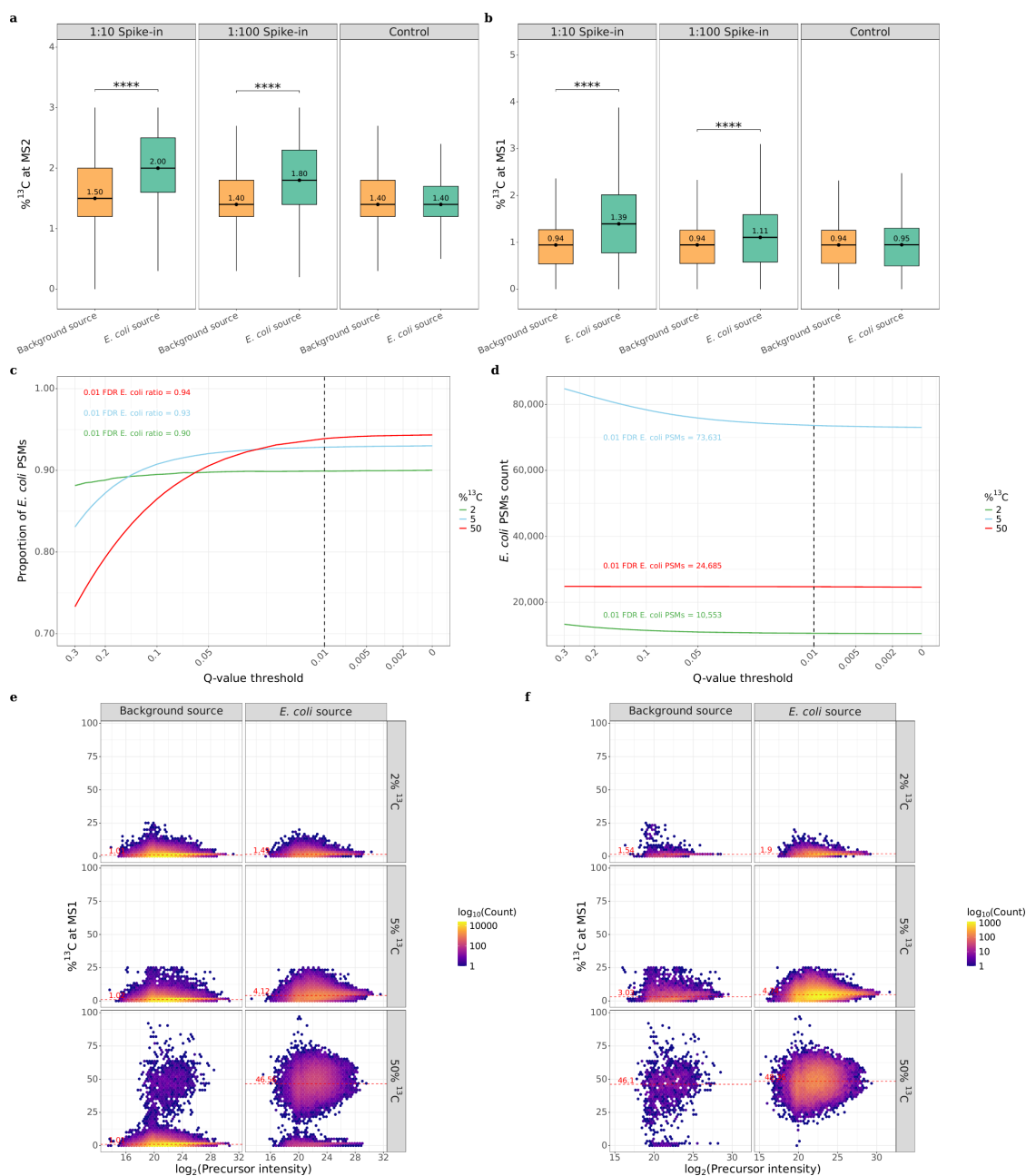


Figure B.1: Astral DIA validation of Sipros 5 using ^{13}C -labeled *E. coli* peptide spike-ins in a mouse fecal peptidome background. Panels (a) and (b) compare MS2- and MS1-derived ^{13}C abundance for 2% spike-ins before SIP-specific q-value filtering. Panels (c) and (d) show WinnowNet q-value-threshold curves for the 2%, 5%, and 50% spike-in datasets. Panels (e) and (f) show labeled Astral DIA PSM density before SIP-specific filtering and after WinnowNet filtering.

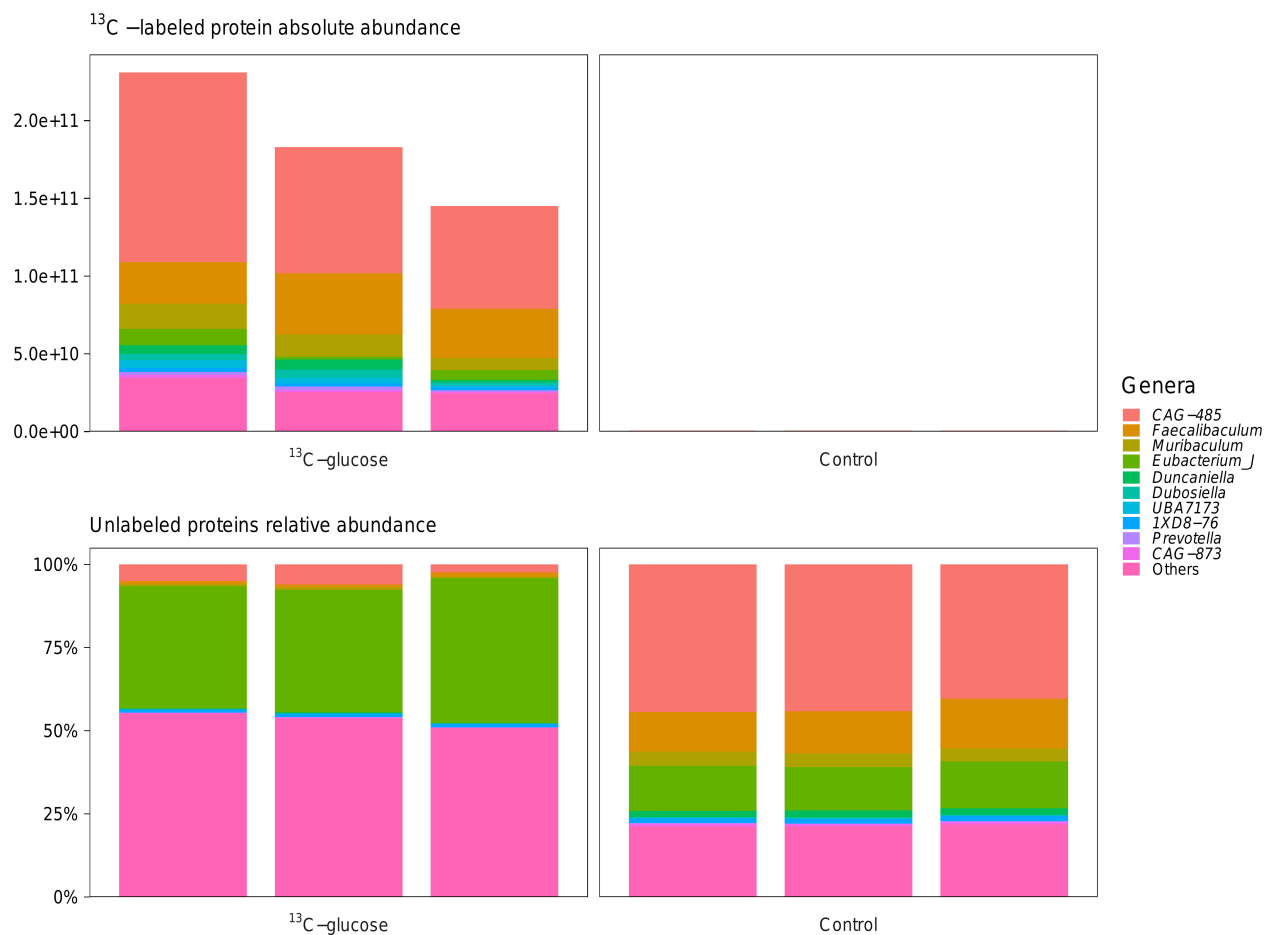


Figure B.2: Genus-level annotation of unlabeled and ^{13}C -labeled proteins in cecum samples from control and ^{13}C -glucose-incubation treatment. The upper panel shows the relative abundance of unlabeled proteins, and the lower panel shows the absolute abundance of ^{13}C -labeled proteins with a scientific-notation y axis. Genera are ranked by combined abundance across both panels, with the top genera shown individually and all remaining genera grouped as Others.

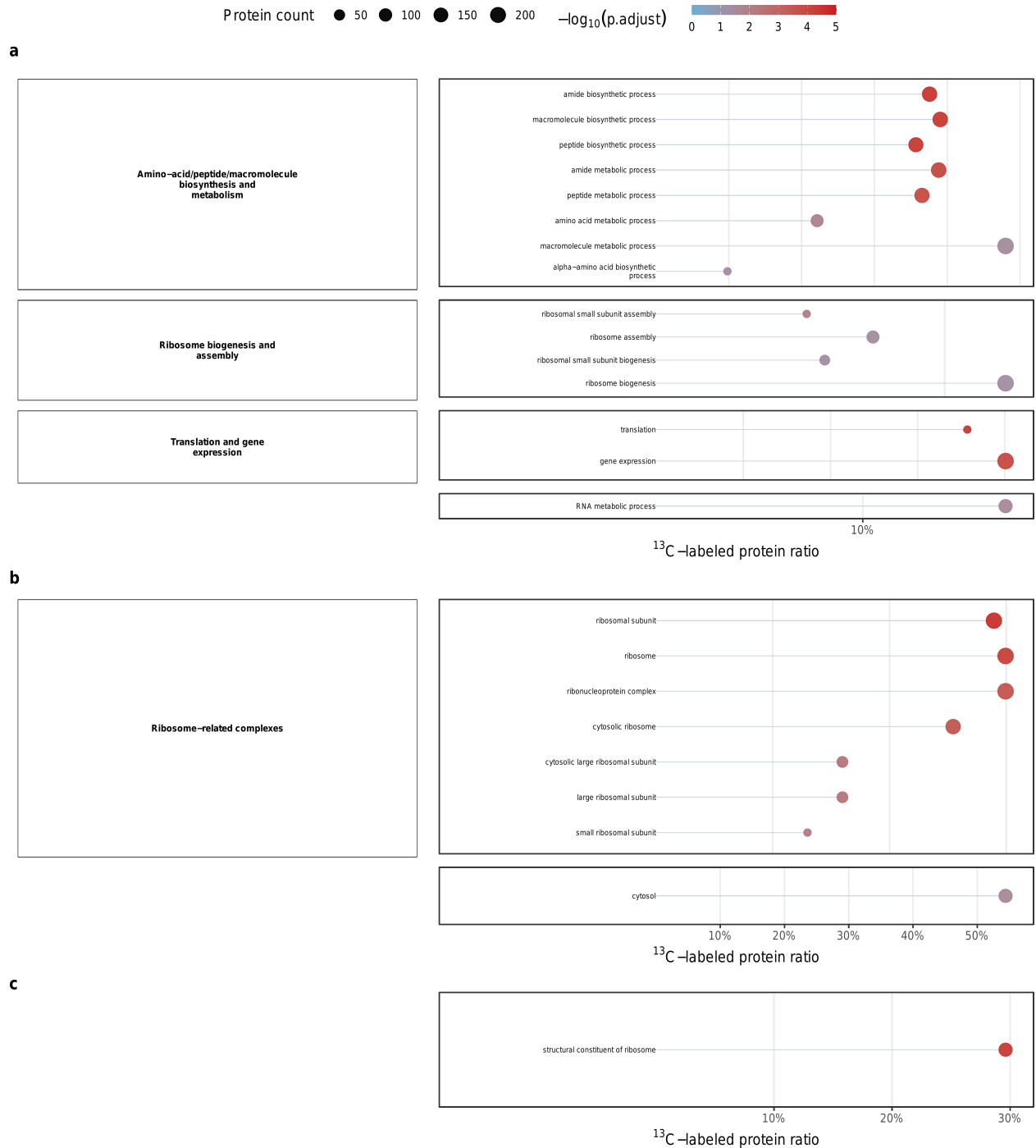


Figure B.3: GO enrichment among ^{13}C -incorporating proteins from the M9 glucose incubation. Panel (a) shows enrichment of biological processes associated with translation, gene expression, ribosome biogenesis or assembly, RNA metabolism, and amino-acid, peptide, amide, and broader macromolecule biosynthetic or metabolic processes. Panel (b) shows enrichment of cellular components centered on the cytosol and ribosome-related complexes, including ribosomes, cytosolic ribosomes, ribosomal subunits, and ribonucleoprotein complexes. Panel (c) shows GO molecular-function enrichment for structural constituent of ribosome. Point color indicates $-\log_{10}(\text{adjusted } P \text{ value})$, point size indicates protein count, and the x-axis shows the ^{13}C -labeled protein ratio.

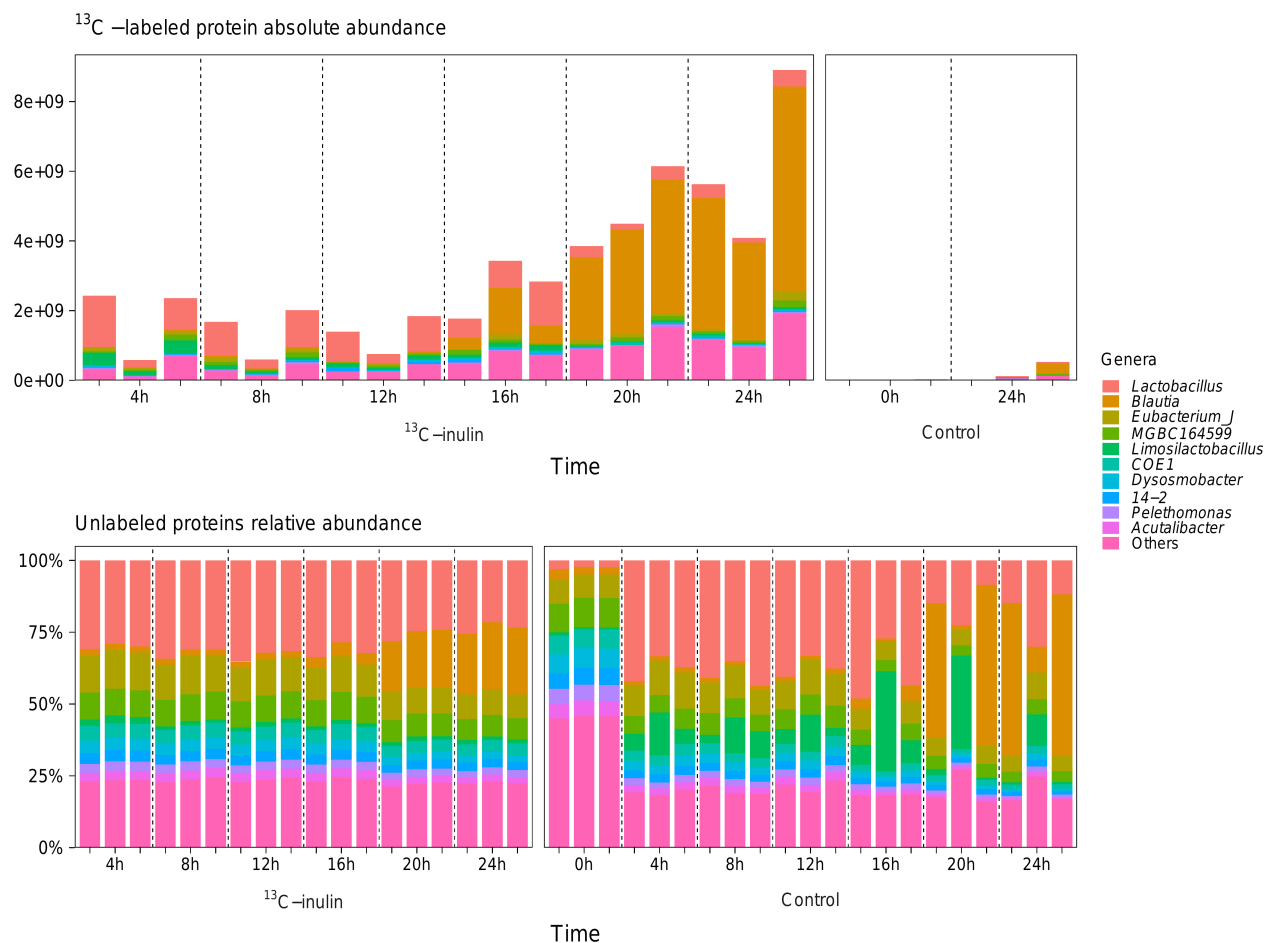


Figure B.4: Genus-level annotation of unlabeled and ^{13}C -labeled proteins in cecum samples from control and ^{13}C -inulin treatment across time. The upper panel shows the relative abundance of unlabeled proteins, and the lower panel shows the absolute abundance of ^{13}C -labeled proteins for PSMs with MS2 isotopic abundance greater than 4. Replicates are shown as separate bars, with time labels centered beneath each triplicate, and the top genera displayed individually while all remaining genera are grouped as Others.

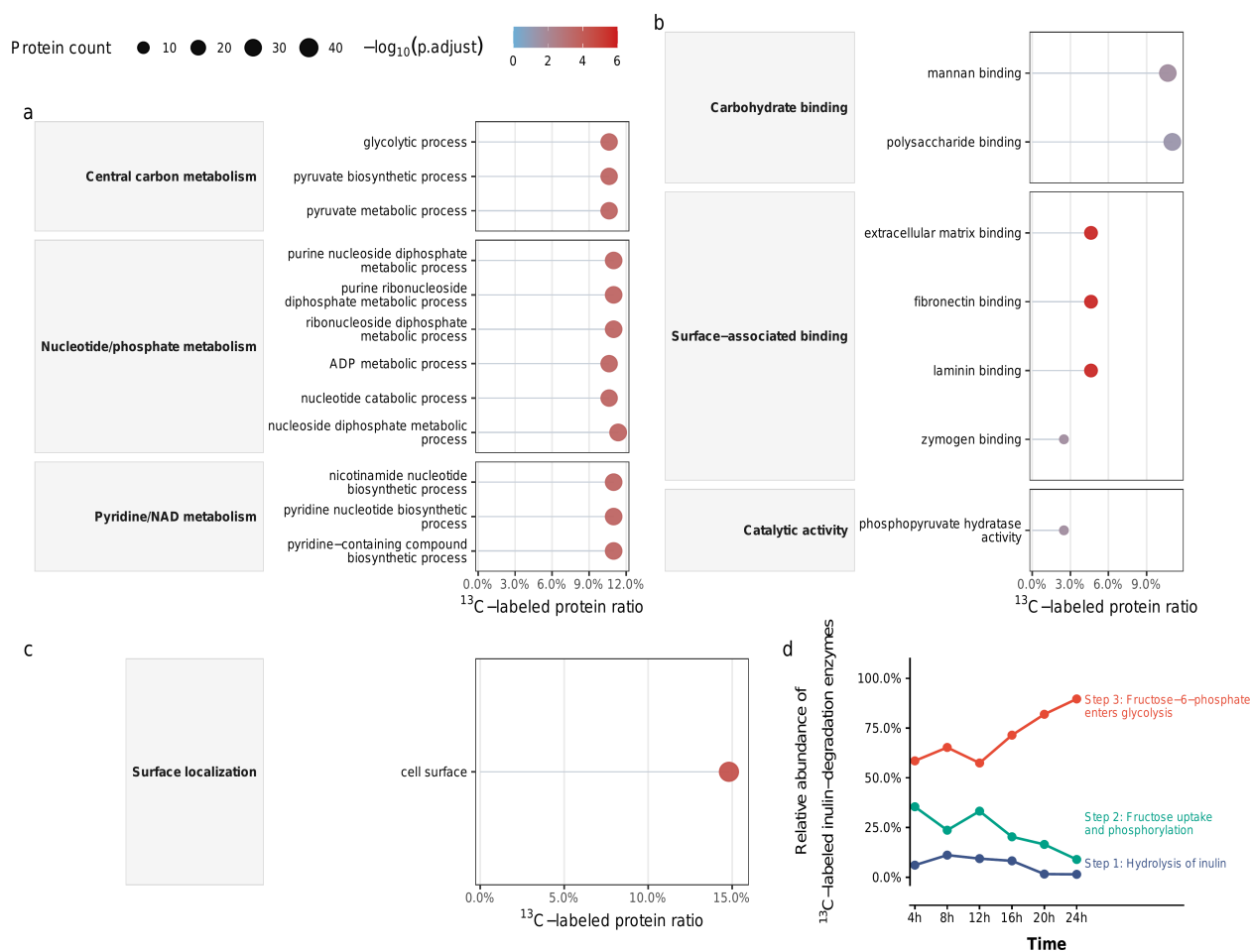


Figure B.5: GO enrichment and enzyme-level functional shift among ^{13}C -inulin-incorporating proteins. Panels (a)–(c) summarize significant GO biological-process, molecular-function, and cellular-component terms. GO biological-process enrichment highlights central carbon metabolism, pyruvate and glycolytic functions, nucleotide/phosphate metabolism, and pyridine/NAD-linked metabolism. GO molecular-function enrichment highlights carbohydrate or polysaccharide binding together with surface-associated binding terms. The cellular-component signal is limited and is dominated by cell-surface enrichment. Panel (d) summarizes the curated inulin-utilization enzyme set across labeled time points and supports a shift from early fructan-processing and fructose-assimilation functions toward later glycolytic-entry functions, including phosphofructokinase-associated steps.

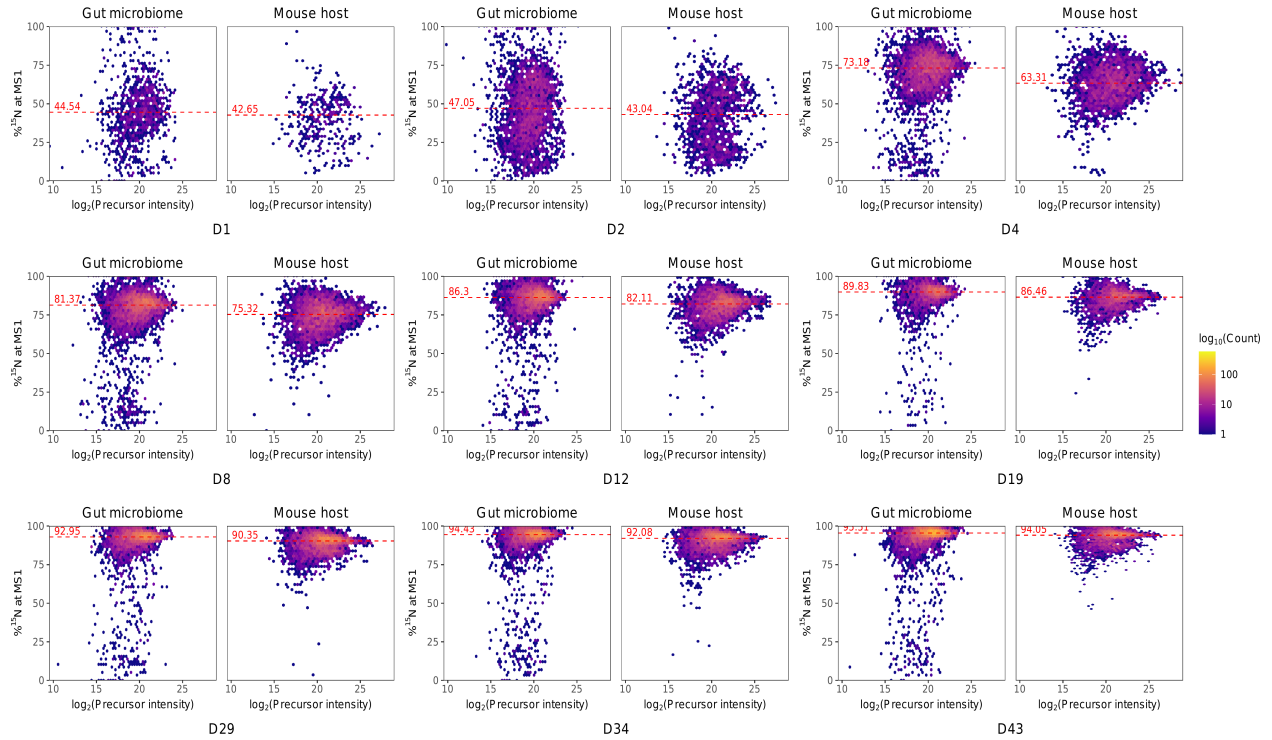


Figure B.6: Hexbin density plots showing the relationship between precursor intensity and MS1-level ¹⁵N abundance for labeled PSMs detected in mouse fecal samples across nine time points (D1, D2, D4, D8, D12, D19, D29, D34, and D43). For each time point, gut-microbiome-derived labeled PSMs are shown on the left and mouse-host-derived labeled PSMs are shown on the right. Time-point labels are centered beneath each paired subplot. PSMs were filtered to retain labeled identifications with MS2 isotopic abundance greater than 12 and to exclude D19 Ref4 and D19 Ref5. Red dashed lines indicate the median MS1 ¹⁵N abundance for each panel, and hexagon color represents the log₁₀-transformed PSM count.

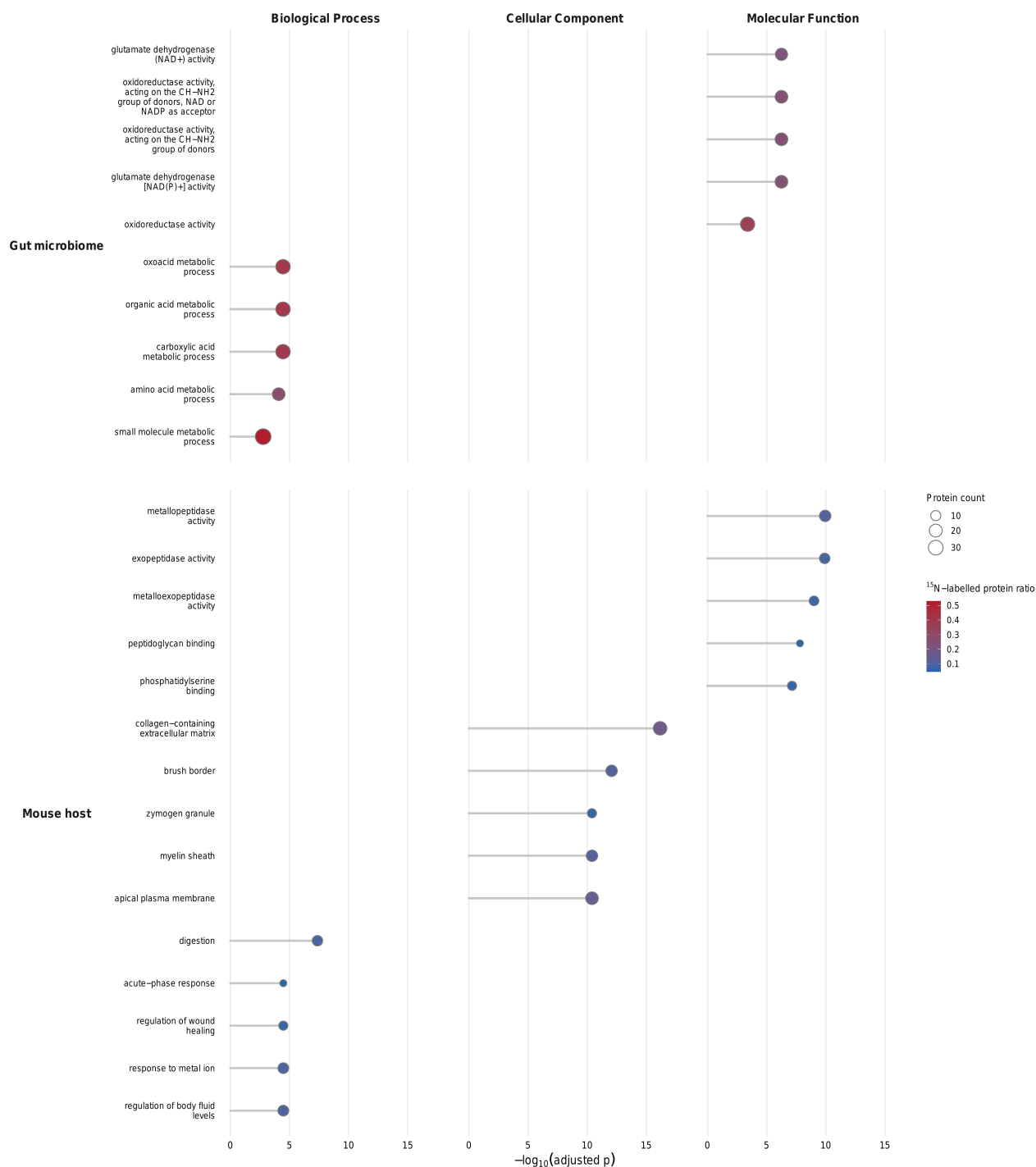


Figure B.7: GO enrichment landscape of ¹⁵N-labeled gut microbiome and mouse host proteins. The faceted lollipop-bubble plot summarizes significant GO enrichment results. Points encode enrichment strength, term size, and ¹⁵N-labeled protein ratio across gut microbiome and mouse host compartments.

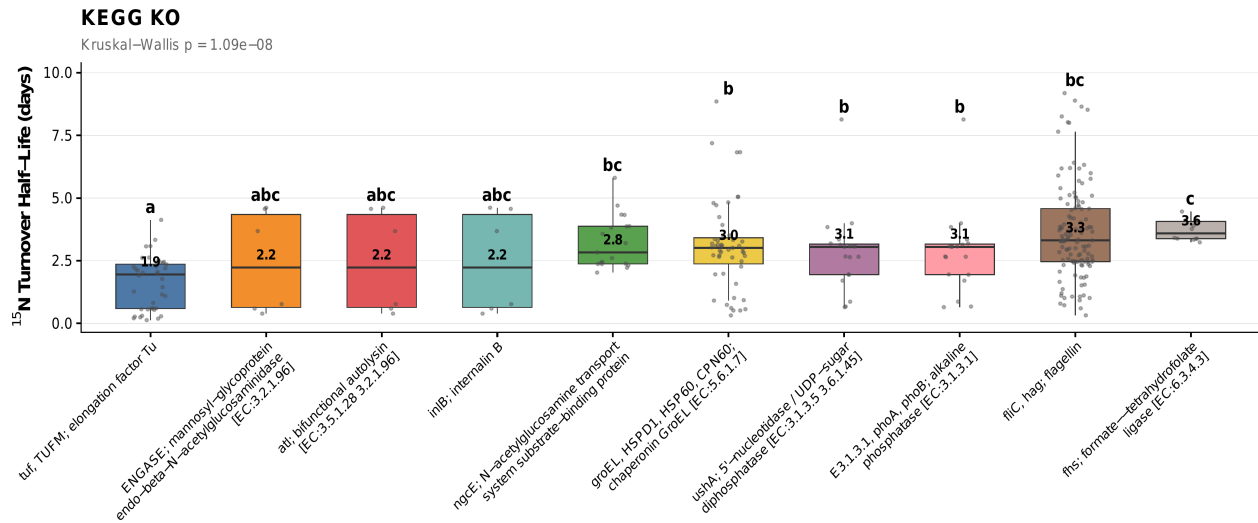


Figure B.8: ^{15}N protein-turnover half-life across enriched KEGG orthology terms. Box plots summarize protein-level posterior mean half-lives for microbiome proteins assigned to each enriched KEGG orthology term; jittered points represent individual proteins.

B.3 Sipros 5 workflow

Sipros 5 is a coordinated workflow that binds together raw-file conversion, config-file generation, enrichment-aware database searching, feature extraction, Percolator filtering, SIP-specific negative-control filtering, and protein inference. The user-facing entry point is `siproswf`, which accepts either a single RAW/mzML file or a directory of files, builds the search environment, and then runs the target and decoy branches in parallel. The same coordinator is used for regular and labeled searches; the difference is whether the workflow is invoked with the regular-search mode `-e R` or with a label such as `-e C13`, `-e N15`, `-e O18`, `-e H2`, `-e S33`, or `-e S34`.

Sipros 5 separates regular and labeled searches at the command-line level. A regular search can be run directly on a RAW file, and the resulting protein table can optionally be used to extract a reduced FASTA before the labeled search begins. In practice, that optional reduction step matters in metaproteomics because the same labeled search may need to be repeated over many enrichment levels and a smaller database shortens both search time and downstream filtering. The label-search mode then runs across the requested enrichment sweep, keeping the peptide database fixed while changing the isotopic envelope model and the parent-mass windows to match the expected label percentage.

At the config level, Sipros 5 keeps the SIP search intentionally narrow. The shipped `SIP.cfg` template uses `Search_Type = SIP`, `Search_Name = SIP`, `Fragmentation_Method = HCD`, parent-mass windows of -2, -1, 0, 1, 2, and 0.01 Da tolerances for both parent and fragment ions. The digestion rules are tryptic, with cleavage after K and R, up to two missed cleavages, a minimum peptide length of 7, and a maximum peptide length of 60. The SIP block fixes the label element to carbon, uses `SIP_Element_Isotope = 13`, and searches enrichment levels from 0 to 100 % in 1 % increments.

Sipros 5 is config-file-driven at two levels. First, the workflow uses `configGenerator` to materialize one or more search configs from a template. Second, the C++ search driver enumerates every `.cfg` file it finds in the config directory and runs the search once for each scan file and each

config file. For labeled searches, this means the workflow does not search one generic enrichment model and then post-fit the result. It explicitly evaluates a sweep of enrichment-specific searches, each with its own Search_Name and its own parent-mass window pattern. That design is the key to understanding Sipros 5: the search stage is a discrete grid over enrichment levels.

Input:

- input files – one RAW/mzML file or a directory of such files
- fasta – target protein FASTA
- element – SIP label element or R for regular search
- range – enrichment range, e.g. 0-100
- step – enrichment step, e.g. 1
- output – output directory
- negative_control – optional unlabeled file names

Output:

- per-sample .Spe2Pep.txt, .pin, filtered PSM TSVs, pepXML, protein tables

1. Reverse the target FASTA to create a decoy database
 - decoy.faa ← reverse each protein sequence and prefix headers with Decoy_
2. Generate search configs from the template
 - if element = R:
 - write one Regular.cfg with the requested MS1/MS2 tolerances
 - else:
 - for pct from lower to upper in increments of step:
 - write one SIP_element_pctPct.cfg
 - set Search_Type = SIP
 - set Search_Name to the percentage-specific tag
 - set SIP_Element, SIP_Element_Isotope, and Element_Percent
 - update Parent_Mass_Windows to match the enrichment width
3. Convert RAW files to FT2 unless the FT2 files already exist
 - for each RAW file:
 - run Raxport into a per-sample ft/ folder
 - reuse pre-existing FT2 files if the conversion folder already contains them
4. For each sample and each config, search target and decoy databases
 - for each config file and each scan file:
 - run Sipros once against the target FASTA
 - run Sipros once against the decoy FASTA
 - write sample.Search_Name.Spe2Pep.txt
5. Extract PSM features from the smallest positive enrichment config
 - build a .pin file per sample from target and decoy search outputs plus FT features

6. Run Percolator on each sample and keep 1% q-value PSMs
 write `sample_target_psms.tsv`, `sample_decoy_psms.tsv`, and `sample_filtered_psms.tsv`
7. If the run is a SIP search, apply the label filter
 keep only `Label = 1`, `q-value <= 0.01`, and `MS2IsotopicAbundances >= label_threshold`
 treat negative-control samples as decoys and rerun Percolator on `SIP.pin`
 clip MS1 isotopic abundances to the 0–100 range and write `SIP_filtered_psms.tsv`
8. Combine pepXML files, infer proteins, and write per-sample protein reports
 merge target and decoy FASTA into `targetDecoy.faa`
 run Philosopher protein inference at 1% protein FDR
 write `protein.tsv`, `protein_with_PSM.tsv`, and, for SIP runs, `protein_with_SIP_filtered_PSM.tsv`

B.3.1 Search-space preparation and config generation

The workflow begins by normalizing the input file set and the search space. The Python wrapper accepts either a directory or a comma-separated list of raw files, resolves each path, and separates RAW and mzML inputs into their own lists. It also verifies that any declared negative-control file names are actually present in the input list before the search proceeds. That guard is small, but it matters: the negative-control filter in Sipros 5 is only meaningful if the unlabeled files are truly part of the same batch and can legitimately provide a decoy-like estimate of the label-specific false discovery rate.

Once the file list is known, the wrapper creates a per-sample directory tree with `ft`, `target`, and `decoy` subdirectories. The raw-to-FT conversion is handled by Raxport, and the workflow skips the conversion if the relevant FT2 files are already present. That behavior keeps repeated searches inexpensive, which is useful when the user wants to try a different label threshold, rerun Percolator, or adjust the negative-control set without repeating the heavy raw-file conversion step.

The config generator then builds the actual search grid. For a regular search, it writes a single `Regular.cfg`. For a labeled search, it iterates from the lower bound to the upper bound in the requested step size and writes one config per enrichment level. The generated file names encode both the isotope and the percentage, so the search name itself becomes part of the output filename. A request such as `-e C13 -r 0-100 -p 1` therefore produces a complete sweep of percentage-

specific configs rather than one generic search file. The generator also adapts the parent-mass windows as the enrichment changes. In the current implementation, the window width is largest around intermediate enrichment, where the isotope envelope is broadest, and smaller near the extremes, where the envelope is more compact. This is a sensible compromise: the search needs enough mass-window coverage to catch broadened labeled precursors, but it should not waste effort on unnecessarily wide windows where the label is near natural abundance or near full labeling.

The remaining template settings are deliberately conservative. `Max_PTM_Count = 0` in `SIP.cfg` keeps the labeled search focused on isotope broadening rather than on open-ended modification discovery. By contrast, the regular-search template allows a small number of common variable modifications. This difference is one of the reasons Sipros 5 can support both discovery-style regular searches and label-specific searches within the same codebase: the workflow changes the template rather than hard-coding a second search engine.

Pseudocode

Input:

workflow arguments: input path, FASTA, element, range, precision, tolerances, output path
optional arguments: negative-control basenames, scan-splitting size, thread count

Output:

decoy FASTA, generated configs, per-sample `ft`, `target`, and `decoy` folders

1. Initialize workflow paths
 - resolve tool locations from `workflow.cfg`, or use Sipros 5 defaults
 - create the requested output directory and workflow log
2. Prepare target and decoy search space
 - read the target FASTA
 - write decoy `.faa` by reversing each protein sequence and prefixing headers with `Decoy_`
3. Generate config files
 - if no element is supplied, set element to R and use `Regular.cfg`
 - else use `SIP.cfg` with the supplied isotope element, range, and step
 - run `configGenerator` with MS1/MS2 tolerances and write configs under `configs/`
4. Normalize input files
 - if input is a directory, enumerate its files; otherwise split the comma-separated file list
 - resolve each file to an absolute path and keep only `.raw` and `.mzml` inputs
 - record each sample basename for later feature extraction, filtering, and assembly

5. Validate negative controls and create sample folders
if negative-control basenames are supplied, require each one to appear in the input basenames
for each sample basename, create `ft/`, `target/`, and `decoy/` subdirectories
6. Convert vendor RAW files
for each RAW file, run Raxport into the sample `ft/` directory
if FT2 files already exist in that directory, skip the conversion

B.3.2 Search execution and output naming

After configuration, the C++ search driver enumerates every `.cfg` file in the generated config directory and every scan file in the input directory. It supports both FT2 and mzML inputs, and it processes one scan file at a time across the config sweep. This is the point where the workflow becomes easy to misdescribe if one only reads the command-line help. The real engine does not run one search and then loop over percent labels afterward; it runs a distinct search for each config file, and each config file carries its own `Search_Name`. In other words, the file name tells you which enrichment hypothesis was used when the search ran.

The output naming follows that rule exactly. The C++ core inserts `Search_Name` into the result filename, producing a `baseName.Search_Name.Spe2Pep.txt` file for each scan file and config pair. For regular search, Sipros 5 uses SE as the search name. For labeled search, the config generator rewrites the name with the percentage tag, so a carbon label search might produce names such as `SIP_C13_001.000Pct` or `SIP_C13_050.000Pct`. The Python workflow then organizes those files into `target` and `decoy` directories so that downstream filtering can compare the same enrichment hypothesis against the reversed decoy database.

The search stage also respects Sipros 5's basic performance safeguards. If the FT2 files already exist, the converter is skipped. If a search output already exists and is large enough to be clearly complete, the wrapper can skip rerunning the same search. The binaries themselves use OpenMP, while the Python wrapper uses process pools to distribute file-level jobs. That division of labor is a good fit for proteomic SIP, because the inner search loop is CPU intensive and the outer workflow is mostly file orchestration.

The search kernel itself is label aware. In SIP mode, the C++ code dispatches to the WDP-based labeled search path rather than the regular MVH path. The dissertation does not need to reproduce the scorer line by line, but the important conceptual point is that the engine is searching isotope-resolved peptide hypotheses, not merely adding a fixed mass shift to a conventional peptide matcher. The labeled search model is what lets the workflow accommodate broadened precursor envelopes and isotopically shifted fragments across the full enrichment sweep.

Pseudocode

Input:

generated config files, sample FT2 or mzML files, target FASTA, decoy FASTA, output folders

Output:

`baseName.Search_Name.Spe2Pep.txt` files in each sample's `target/` and `decoy/` folders

1. Build the search command list in the Python workflow
 - collect every `.cfg` file from `configs/`
 - for each config and each RAW-derived FT2 file, add a target FASTA search command
 - for each config and each mzML file, add a target FASTA search command
 - repeat the same command construction with `decoy.faa` and the sample `decoy/` folder
2. Skip completed searches when possible
 - for each command, infer the expected `.Spe2Pep.txt` filename
 - if the expected output exists and is larger than the completion threshold, do not rerun it
3. Run searches in parallel at the workflow level
 - set `OMP_NUM_THREADS` for each Sipros process
 - dispatch search commands through a process pool sized from the requested thread count
4. Inside the C++ Sipros executable
 - parse either one scan file with `-f` or a scan directory with `-w`
 - parse either one config file with `-c` or a config directory with `-g`
 - enumerate scan files with extensions `.ft2`, `.FT2`, `.mzml`, and `.mzML`
 - enumerate config files with extensions `.cfg` and `.CFG`
5. Search each scan/config pair
 - for each scan file:
 - for each config file:
 - load the config with `ProNovoConfig::setFilename`
 - override the FASTA path if `-fasta` was supplied
 - load the scan data into `MS2ScanVector`
 - if `Search_Type = SIP`, call `startProcessingWdpSip`
 - else call `startProcessingMvh`

B.3.3 Feature extraction and decoy-controlled PSM filtering

Once the search stage has written the `.Spe2Pep.txt` outputs, the workflow moves to feature extraction and PSM filtering. The Sipros 5 feature extractor reads the search results together with the FT files and converts them into `.pin` features for Percolator. The code deliberately chooses the smallest positive-enrichment config when a labeled search has more than one config, which keeps the feature-extraction step tied to a real SIP search without forcing it to handle the entire enrichment sweep at once. The resulting `.pin` file retains the isotopic-abundance features, including `MS1IsotopicAbundances` and `MS2IsotopicAbundances`, which later stages use to quantify enrichment.

Percolator then performs the first decoy-controlled filter. For each sample, the workflow runs Percolator in `-only-psms` mode on the sample-specific `.pin` file and writes target and decoy PSM tables. The dissertation can describe that step simply as a 1% q-value filter, but the code is a little more specific: it integrates the Percolator output back into the original feature table, keeps the rows with `q-value <= 0.01`, and writes a sample-level `*_filtered_psms.tsv`. The workflow also exports a pepXML representation so that Philosopher can read the evidence during protein inference.

There is also an optional `ignorePCT` branch in the filter step. If that flag is set, the workflow drops the isotopic-abundance columns before Percolator rescoring. The default SIP workflow keeps the percentage columns because they are central to the downstream label-specific analysis.

The practical meaning of this stage is straightforward. The search engine proposes candidate labeled and unlabeled PSMs, the feature extractor packages the candidate evidence into a learned representation, and Percolator removes the weaker matches while preserving the score and posterior error probability needed for later inference. In the final filtered PSM tables, the dissertation retains the label percentages because the enrichment estimate is not merely a side annotation; it is the quantitative readout that makes proteomic SIP distinct from conventional proteomics.

Pseudocode

Input:

per-sample target and decoy `.Spe2Pep.txt` files, FT files, generated configs, Percolator path

Output:

per-sample `.pin`, target PSM TSV, decoy PSM TSV, and filtered PSM evidence

1. Choose the feature-extraction config
 - list all generated `.cfg` files
 - if there is only one config, use it
 - else parse percentage tags such as `001.000Pct` and choose the smallest value greater than zero
2. Extract Percolator features
 - for each sample basename:
 - run `aerithFeatureExtractor` on the sample target/ and decoy/ folders
 - provide the sample `ft/` folder, selected config, and output `sample.pin`
 - set `OMP_NUM_THREADS` and `OMP_STACKSIZE` for the extractor
3. Optionally remove isotope-percentage features
 - if `ignorePCT` is set:
 - read `sample.pin`
 - drop `MS1IsotopicAbundances` and `MS2IsotopicAbundances`
 - write `sample_NoPCT.pin`
 - else keep the original `sample.pin`
4. Run Percolator for ordinary PSM-level FDR control
 - for each sample basename:
 - run Percolator with `-only-psms` and `-no-terminate`
 - write `sample_target_psms.tsv`
 - write `sample_decoy_psms.tsv`
 - use the original or `_NoPCT` PIN file according to the selected option
5. Hand filtered evidence to assembly
 - preserve q-values, posterior error probabilities, and isotope-abundance columns for downstream SIP filtering and protein inference

B.3.4 SIP-specific filtering, negative controls, and protein inference

The SIP-specific branch adds a second, stricter layer of filtering on top of the ordinary PSM FDR control. After the sample-level q-value filter, the workflow collects all filtered PSMs, keeps only those with `Label = 1`, and requires that the MS2 isotopic abundance exceed the user-set label threshold. In the Sipros 5 default workflow, that threshold is 2.0 percentage points. The resulting PSMs are then split into target and negative-control sets. Negative-control samples are relabeled as

decoys, merged with the target SIP PSMs, and reanalyzed by Percolator as a combined `SIP.pin` file. That second pass is the Sipros 5 way of estimating label-specific error rates in a way that is anchored to unlabeled material rather than to the usual target-decoy structure alone.

The output of that stage is `SIP_filtered_psms.tsv`. Sipros 5 clips the MS1 isotopic abundance values to the 0–100 range before writing the table, which protects downstream summary code from the occasional non-carbon-label edge case in which the raw abundance estimate can drift slightly outside the nominal percentage interval. Sipros 5 keeps both isotope abundance columns because they carry complementary information: MS1 isotopic abundances are more sensitive, whereas MS2 isotopic abundances are more accurate. That distinction explains why the workflow keeps both measurements rather than collapsing them into a single percentage field.

Protein inference is handled by Philosopher. The workflow combines target and decoy FASTA sequences into `targetDecoy.faa`, copies the per-sample pepXML files into a shared `pepxmls` directory, and then runs Philosopher’s `proteinprophet`, `database`, `filter`, and `report` steps. Sipros 5 uses a 1% protein-level FDR threshold, keeps the picked-decoy strategy, and applies razor peptide assignment during filtering. After the global inference pass, the workflow repeats a sample-specific Philosopher filter and report step so that each raw file gets its own protein-level output. The result is a conventional protein-inference table, but one that can be traced back to the label-aware PSM evidence produced earlier in the workflow.

This is also the point where the negative-control logic propagates into the protein output. If SIP negative controls were supplied, the workflow attaches the SIP-filtered peptide evidence back to the protein table and writes `protein_with_SIP_filtered_PSM.tsv`. That table is especially useful in time-course or multi-sample studies because it lets the user see which proteins persist across samples and which peptide evidence actually survives the label-specific filter. The companion `protein_with_PSM.tsv` captures the ordinary, decoy-filtered PSM evidence for each sample regardless of SIP status.

Pseudocode

Input:

per-sample Percolator PSM tables, per-sample .pin files, target FASTA, decoy FASTA
negative-control sample names, isotope label mode, label threshold

Output:

SIP_filtered_psms.tsv, protein.tsv, protein_with_PSM.tsv, and SIP-filtered protein evidence

1. Merge target and decoy FASTA files

write targetDecoy.faa by concatenating the target FASTA and decoy.faa

2. Integrate sample-level PSM filters with feature tables

for each sample basename:

read sample_target_psms.tsv, sample_decoy_psms.tsv, and sample.pin

keep target and decoy PSMs with q-value ≤ 0.01 for the filtered PSM table

merge Percolator scores, q-values, and posterior error probabilities back to PIN features

write sample_filtered_psms.tsv

create sample.pep.xml from PSMs with posterior_error_prob < 0.5

3. Apply SIP-specific negative-control filtering

if negative-control samples are supplied:

for each sample, keep PSMs with Label = 1, q-value ≤ 0.01 , and MS2IsotopicAbundances $\geq$ label_threshold

mark PSMs from negative-control samples with Label = -1

merge the target and negative-control SIP PSMs into SIP.pin

run Percolator on SIP.pin to write SIP target and decoy PSM tables

keep SIP PSMs at q-value ≤ 0.01 and clip MS1 abundances to 0–100

write SIP_filtered_psms.tsv

4. Infer proteins with Philosopher

copy all per-sample pepXML files into pepxmls/

initialize the Philosopher workspace

run proteinprophet on all pepXML files to produce combined.prot.xml

annotate targetDecoy.faa with the Decoy_ prefix

run Philosopher filter and report with 1% protein FDR, picked decoys, and razor peptides

5. Attach PSM evidence to protein outputs

map ordinary filtered PSMs back to proteins and write protein_with_PSM.tsv

if SIP-filtered PSMs exist, map them by sample and protein

write protein_with_SIP_filtered_PSM.tsv

repeat Philosopher filtering and reporting inside each sample directory for per-sample protein.tsv

B.3.5 Quantification outputs and interpretation

At the end of the workflow, Sipros 5 leaves a compact but information-rich set of outputs. For each sample there is a .pin file, target and decoy PSM TSVs, a merged filtered-PSM TSV, a

pepXML file, a per-sample `protein.tsv`, and, if SIP filtering was activated, sample-specific peptide and protein summaries that retain the isotope-abundance annotations. Across the entire batch, the workflow additionally writes `SIP_filtered_psms.tsv` and `protein_with_SIP_filtered_PSM.tsv`. These files are the bridge between the search engine and the biological interpretation that appears in the main chapter: isotope incorporation at the PSM level, protein-level assembly, and downstream taxonomic or functional summaries.

Sipros 5 treats label percentage as a first-class quantitative variable throughout the workflow. It enters at config generation, shapes the search windows, influences the output filename, survives feature extraction, and is preserved through filtering and protein inference. That design is what makes Sipros 5 suitable for both DDA Orbitrap data and narrow-window DIA Astral data, because the same workflow can carry the enrichment information from the raw scan all the way to the protein table.

In short, the Sipros 5 workflow is a layered system. The config generator writes one search hypothesis per enrichment level, the search driver enumerates those hypotheses and writes one output per config file, the feature extractor converts the resulting evidence into Percolator features, the filter stage controls ordinary PSM FDR, the SIP branch controls label-specific FDR with negative controls, and Philosopher produces the final protein tables.

Appendix C: Supplementary Material for Chapter 4

This appendix reproduces the supporting information of Xiong et al. (2025), which provides additional methods, algorithms, reference spectra, and scoring details that complement the chapter.

C.1 Isotopic fine-structure calculation for chemical formulas

This algorithm simulates the natural or SIP-modified isotopic distribution of a molecule given its chemical formula. The formula is parsed into element counts (e.g., $\text{C}_6\text{H}_{12}\text{O}_6 \rightarrow \{\text{C}:6, \text{H}:12, \text{O}:6\}$). For each element E with n atoms and isotopes $i_1, i_2, \dots, i_k$ with abundances $p_1, p_2, \dots, p_k$, the count of each isotope in a random sample follows a multinomial distribution,

$$[n_1, n_2, \dots, n_k] \sim \text{Multinomial}(n, [p_1, p_2, \dots, p_k]), \quad n_1 + \dots + n_k = n.$$

The mass of a specific isotopic composition is the sum $\text{Mass} = \sum n_i \cdot m(i_i)$, where $m(i_i)$ is the isotope mass; the abundance of each unique composition is its count divided by the total number of Monte Carlo trials. User-defined SIP overrides (e.g., $p_{^{13}\text{C}} = 0.5$) replace the natural abundance of one isotope and proportionally redistribute the remaining fraction among siblings. The default Monte Carlo trial count is 10,000.

Pseudocode

Input:

formula_str – string, e.g. “C6H12O6”
N – integer, number of Monte Carlo trials (default 10000)
params – named overrides, e.g. {C13 = 0.5, N15 = 0.2}

Output:

isotope_numbers – data frame of simulated isotope counts, masses, and abundances

1. Save original abundances

original_abundances $\leftarrow$ copy(shared_env\$isotopic_abundances)

2. Apply SIP overrides to abundances

```

for each (iso_name, new_abund) in params:
    if iso_name ∈ names(shared_env$isotopic_abundances):
        # set target isotope to user value
        shared_env$isotopic_abundances[iso_name] ← new_abund

        # redistribute the remaining fraction among its sibling isotopes
        element_symbol ← strip_digits(iso_name)
        siblings ← supported_isotopes[element_symbol] \ {iso_name}
        remaining ← 1 – new_abund
        # scale siblings to sum to remaining
        total_sib_abund ← sum(original_abundances[siblings])
        for each sib in siblings:
            shared_env$isotopic_abundances[sib] ←
                original_abundances[sib] / total_sib_abund * remaining
    else:
        warning(“Unsupported isotope override:”, iso_name)

3. Run Monte Carlo simulation
    # this calls cal_isotope_numbers() which parses the formula,
    # draws random isotope counts per element based on shared_env,
    # then summarizes into counts, masses, abundances
    isotope_numbers ← cal_isotope_numbers(formula_str, N)

4. Restore original abundances
    shared_env$isotopic_abundances ← original_abundances

```

C.2 Isotopic envelope calculation using FFT

For a molecule with multiple elements, the isotopic distribution is the convolution of element-level distributions raised to their respective atom counts. Let $P_X(i)$ denote the probability that element X has i additional neutrons and n_X the count of atoms of element X . The aggregate molecular distribution is

$$P_{\text{mol}}(k) = \underbrace{P_C * P_C * \dots * P_C}_{n_C \text{ times}} * \underbrace{P_H * P_H * \dots * P_H}_{n_H \text{ times}} * \dots$$

By the convolution theorem,

$$\mathcal{F}[P_{\text{mol}}] = \mathcal{F}[P_C]^{n_C} \cdot \mathcal{F}[P_H]^{n_H} \dots, \quad P_{\text{mol}} = \mathcal{F}^{-1}[\mathcal{F}[P_C]^{n_C} \cdot \mathcal{F}[P_H]^{n_H} \dots].$$

This algorithm is substantially more efficient than direct convolution for large molecules.

Pseudocode

Input:

formula_str – string, e.g. “C6H6O”
N_width – integer, FFT vector length (must cover full isotope envelope)
min_abundance – float, filter threshold for low-abundance peaks
... – named overrides for individual isotope abundances

Output:

mass_bins – array of peak m/z or mass values
probabilities – array of normalized abundances

1. Parse the formula into element counts

element_array $\leftarrow$ parse_chemical_formula(formula_str)
e.g. {“C”:6, “H”:6, “O”:1}

2. (Optional) override single-isotope abundances via ... parameters

for each param in ...:
 if param matches an isotope name:
 adjust shared_env\$isotopic_abundances[param] and rescale its element’s other isotopes

3. Compute two key mass constants

monoisotopic_mass $\leftarrow \sum[\text{element, count}] (\text{count} \times \text{mass_of_most_abundant_isotope}(\text{element}))$
average_neutron_mass $\leftarrow$ calculate_average_neutron_mass(element_array)

4. Build the FFT-based abundance envelope

abundance_fft $\leftarrow$ 1
for each (element, count) in element_array:
 if count > 0 and element has >1 isotope:
 iso_vec $\leftarrow$ shared_env\$isotopic_abundances[all isotopes of element]
 padded_vec $\leftarrow$ pad_with_zeros(iso_vec, N_width)
 abundance_fft $\leftarrow$ abundance_fft $\times$ [FFT(padded_vec)]^count
raw $\leftarrow$ IFFT(abundance_fft)
probs $\leftarrow$ Re(raw) / sum(Re(raw))

5. Translate neutron-count bins into masses

for k from 0 to length(probs)-1:
 mass_bins[k] $\leftarrow$ monoisotopic_mass + k $\times$ average_neutron_mass

6. Filter out tiny peaks and return

keep $\leftarrow$ probs > min_abundance
return mass_bins[keep], probs[keep]

C.3 Reference spectra and convolution-based peptide envelopes

Appendix Figures C.1 and C.2 provide reference spectra from MZcloud (Thermo Fisher) for negative-ion-mode glucose ($[M-H]^-$) and positive-ion-mode penicillin ($[M+H]^+$) at 1.07 % natural-abundance ^{13}C for the case studies in Figure 4.3. The remainder of Section C.3 derives the convolution-based isotope-envelope algorithm for peptide sequences: starting from per-residue isotopic envelopes computed by the FFT method, the full peptide envelope is obtained by sequential convolution along the peptide; B and Y fragments are then derived by sliding-window subsequence convolution from the parent envelope.

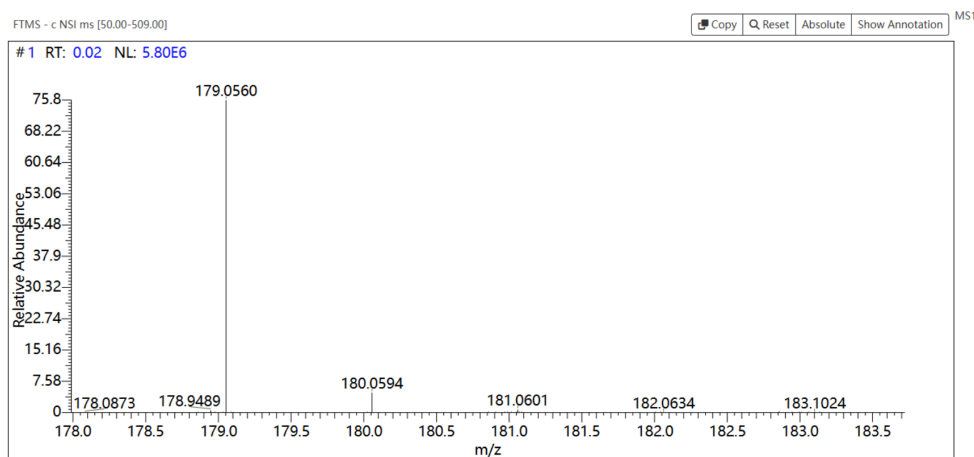


Figure C.1: Reference spectrum from MZcloud (Thermo Fisher) of negative-ion-mode glucose ($[M-H]^-$) at 1.07% natural-abundance ^{13}C . The main peak at m/z 179.0560 is accompanied by natural-abundance ^{13}C isotopologues at 180.0594, 181.0601, 182.0634, and 183.1024. Used as a validation target for the Aerith Monte Carlo and FFT-based isotopic-envelope simulations of glucose in Figure 4.3A,B.

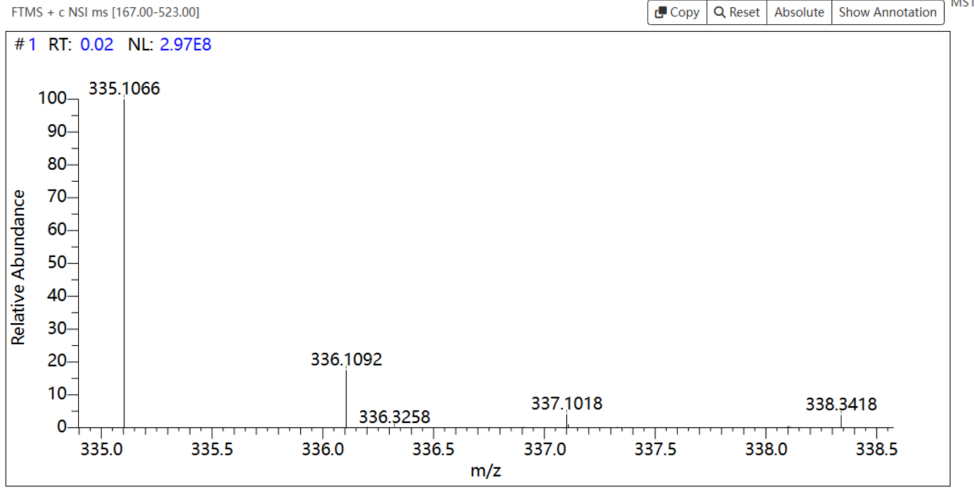


Figure C.2: Reference spectrum from MZcloud (Thermo Fisher) of positive-ion-mode penicillin ($[M+H]^+$) at 1.07% natural-abundance ^{13}C . The main peak at m/z 335.1066 is accompanied by natural-abundance ^{13}C and ^{34}S isotopologues at 336.1092, 336.3258, 337.1018, and 338.3418. Used as a validation target for the Aerith Monte Carlo and FFT-based isotopic-envelope simulations of penicillin in Figure 4.3C,D.

For two isotope distributions D_0 and D_1 , the combined distribution is

$$D_{\text{sum}} = D_0 \otimes D_1.$$

The probability and mass at peak k are computed by summing over all overlapping peak pairs,

$$p_{\text{sum}}(k) = \sum_{i=\max(0, k-n_1+1)}^{\min(n_0-1, k)} p_0(i) p_1(k-i),$$

$$m_{\text{sum}}(k) = \frac{\sum_{i=\max(0, k-n_1+1)}^{\min(n_0-1, k)} p_0(i) p_1(k-i) [m_0(i) + m_1(k-i)]}{\sum_{i=\max(0, k-n_1+1)}^{\min(n_0-1, k)} p_0(i) p_1(k-i)}.$$

Repeated self-convolution is written as $D_{\text{product}} = D_0^{\otimes c}$, where c is the number of identical atoms or residues. For a peptide sequence $S = [r_1, r_2, \dots, r_L]$, the complete envelope is

$$D_{\text{peptide}} = D_{N-\text{term}} \otimes D_{C-\text{term}} \otimes D_{r_1} \otimes D_{r_2} \otimes \dots \otimes D_{r_L},$$

and the B- and Y-ion envelopes are obtained from the prefix and suffix convolutions of that se-

quence.

Pseudocode

```
function computeIsotopicDistribution(atomicComposition, resultDistribution):
    # Input: atomicComposition, vector of counts for each atom type (C, H, O, ...)
    # Output: resultDistribution, final convolved isotopic distribution
    sumDistribution ← multiply(vAtomIsotopicDistribution[0], atomicComposition[0])
    for i from 1 to atomicComposition.length-1:
        currentDistribution ← multiply(vAtomIsotopicDistribution[i], atomicComposition[i])
        sumDistribution ← sum(sumDistribution, currentDistribution)
    resultDistribution ← sumDistribution

function sum(distribution0, distribution1):
    newSize ← distribution0.size + distribution1.size - 1
    sumDistribution ← new IsotopeDistribution(newSize)
    for k from 0 to newSize-1:
        sumWeight ← 0; weightedMassSum ← 0
        start ← max(0, k - distribution1.size + 1)
        end ← min(distribution0.size - 1, k)
        for i from start to end:
            j ← k - i
            w ← distribution0.prob[i] * distribution1.prob[j]
            sumWeight ← sumWeight + w
            weightedMassSum ← weightedMassSum + w * (distribution0.mass[i] + distribution1.mass[j])
        sumDistribution.prob[k] ← sumWeight
        sumDistribution.mass[k] ← weightedMassSum / sumWeight
    return sumDistribution

function multiply(distribution, count):
    result ← distribution
    for copy from 2 to count:
        result ← sum(result, distribution)
    return result
```

C.4 PSM scoring functions

C.4.1 Weighted dot product (WDP)

The WDP score evaluates the quality of a PSM by analyzing the isotopic envelopes of matched fragment ions, combining mass accuracy, isotopic pattern similarity, charge-state correctness, and

presence of complementary ion pairs. For each matched ion isotopic envelope k :

$$\text{WDP}_k = s_k \cdot c_k \cdot h_k \cdot g_k \quad (\text{C.1})$$

and the total WDP score is $\text{WDP} = \sum_k \text{WDP}_k$. The mass-accuracy score is

$$h_k = 2 \cdot \left[1 - 0.5 \cdot \left(1 + \text{erf} \left(\frac{\overline{\Delta m}}{(t_{\text{acc}}/2)\sqrt{2}} \right) \right) \right], \quad (\text{C.2})$$

where $\overline{\Delta m} = \frac{1}{n} \sum_{i=1}^n |m/z_{\text{obs},i} - m/z_{\text{exp},i}|$ and $t_{\text{acc}} = \text{tolerancePPM} \cdot 10^{-6} \cdot m/z_{\text{exp}}$. The isotopic envelope score is

$$s_k = 1 + \sum_{i=1}^{n-1} e_i, \quad e_i = 0.5 - 0.5 \cdot \text{erf} \left(\frac{|I_{\text{exp},i}^{\text{rel}} - I_{\text{obs},i}^{\text{rel}}|}{\sqrt{(I_{\text{exp},i}^{\text{rel}})^2 + (I_{\text{obs},i}^{\text{rel}})^2}} \right), \quad (\text{C.3})$$

with relative intensities $I_{\bullet,i}^{\text{rel}} = I_{\bullet,i}/I_{\bullet,0}$ (peak intensity normalized to the monoisotopic peak). The charge-state penalty is $c_k = 1$ if the expected and observed charge states match and 0.5 otherwise; the complementary-fragment bonus is $g_k = 2$ when the complementary fragment exists and 1 otherwise. WDP was originally developed for Sipros (Pan et al., 2011; Xiong et al., 2024) and has the narrowest score-peak width across the full enrichment range, indicating superior resolution in distinguishing true matches.

Pseudocode

Input:

matchedIonIsotopicEnvelopes – map from (ionType, residuePos) to matched peaks
 tolerancePPM – float
 peptideLength – integer

Output:

WDPscore – float

Initialize total score

WDPscore $\leftarrow$ 0.0

For each (ionType, residuePos), matches in matchedIonIsotopicEnvelopes:

if matches is empty: continue

sumMassError $\leftarrow$ 0.0

```

for each match in matches:
    sumMassError ← sumMassError + |match.observedMZ – match.expectedMZ|
avgMassError ← sumMassError / matches.size()
massAccuracyThreshold ← tolerancePPM * 1e-6 * matches[0].expectedMZ
h_k ← mass-error weight from the error function
s_k ← 1.0
for i from 1 to matches.size()-1:
    expRel ← matches[i].expectedIntensity / matches[0].expectedIntensity
    obsRel ← matches[i].observedIntensity / matches[0].observedIntensity
    e_i ← intensity-shape agreement score
    s_k ← s_k + e_i
if expected and observed charge states match: c_k ← 1.0; else c_k ← 0.5
compType ← (ionType == "B") ? "Y" : "B"
compPos ← peptideLength – residuePos
if (compType, compPos) in matchedIonIsotopicEnvelopes: g_k ← 2.0; else g_k ← 1.0
WDPscore ← WDPscore + s_k * c_k * h_k * g_k

return WDPscore

```

C.4.2 Multivariate hypergeometric (MVH)

The MVH score (Tabb et al., 2007) computes the probability of the observed peak matches under a hypergeometric null model and reports the negative logarithm of the tail probability.

Pseudocode

Input: realScan→mz, expectedMZs, residuePositions, ionKinds, matchedIndices, tolerancePPM

Output: MVHscore

```

maxMz ← max(realScan→mz); minMz ← min(realScan→mz)
toleranceWidth ← ((maxMz + minMz)/2) * tolerancePPM/1e6 * 2
N ← (maxMz – minMz) / toleranceWidth
n_theo ← 0; n_match ← 0; n_obs ← len(realScan→mz)
topN ← 3; ix ← 0; nISO ← 0
while ix < len(expectedMZs):
    reset nISO when the residue position changes
    if minMz ≤ expectedMZs[ix] ≤ maxMz and ionKinds[ix] ∈ {B, Y}: n_theo ← n_theo + topN
    if matchedIndices[ix] ≠ -1 and nISO < topN: n_match ← n_match + 1; nISO ← nISO + 1
    ix ← ix + 1
pValue ←  $\sum_{j=n\_match}^{n\_theo}$  hypergeomProbability(n_theo, n_obs, j, N)
if pValue < 1e-300: pValue ← 1e-300
MVHscore ← -log10(pValue)

```

C.4.3 Cross-correlation (*Xcorr*)

The *Xcorr* score (Eng et al., 1994, 2013) bins observed and theoretical spectra, computes the zero-shift correlation, subtracts the average shifted correlation, and floors the final score at zero.

Pseudocode

Input: observedSpectrum, peptideSequence, tolerancePPM

Output: xcorr

```
function calculateXCorr(observedSpectrum, peptideSequence, tolerancePPM):
    convert observed peaks to single-charge m/z and collect intensities
    normalize observed intensities to [0, 1]
    generate theoretical fragment spectrum for peptideSequence and normalize it
    binWidth ← tolerancePPM/1e6 * (minMz + maxMz)/2
    shiftWindow ← 75
    theoBins ← binSpectrum(theoMz, theoIntensity, binWidth, minMz, maxMz)
    obsBins ← binSpectrum(obsMz, obsIntensity, binWidth, minMz, maxMz)
    cc0 ← correlationAtShift(theoBins, obsBins, shift=0)
    avgShift ← average correlationAtShift over shifts -75...75, excluding 0
    xcorr ← max(0, cc0 - avgShift)
    return xcorr
```

C.5 Binomial precursor-mass estimation

For a SIP-labeled peptide with n_E atoms of the labeled element E at enrichment p , the labeled-atom count is modeled as $n_{\text{labeled}} \sim \text{Binomial}(n_E, p/100)$. The modal count of a binomial distribution is $\lfloor (n+1)p \rfloor$, and Aerith approximates the precursor mass as

$$M_{\text{precursor}}(P) = M_{\text{base}}(P) + \text{round}(n_k(P) p_k^{\text{SIP}}) \Delta m_k,$$

where M_{base} is the monoisotopic base mass, $n_k(P)$ is the atom count for the labeled element, p_k^{SIP} is the enrichment level, and Δm_k is the neutron-mass shift. Hydrogen, oxygen, nitrogen, and sulfur labeling add the corresponding natural carbon-13 background term before the mass shift is applied. This approximation provides a fast way to place the precursor within the correct isotopic envelope.

Pseudocode

Input:

baseMass – monoisotopic base mass of the peptide
atomCount – number of atoms of the SIP element in the peptide
neutronMass – mass shift per neutron
SIPelement – one of “C”, “H”, “O”, “N”, “S”

Output:

estimated precursor mass

```
function estimatePrecursorMassbyNP(baseMass, atomCount, neutronMass):
  if SIPelement == "C":
    expected ← round(atomCount * C13Abundance)
  else if SIPelement == "H":
    expected ← round(atomCount * H2Abundance + pepAtomCounts["C"] * C13Abundance)
  else if SIPelement == "O":
    expected ← round(2 * atomCount * O18Abundance + pepAtomCounts["C"] * C13Abundance)
  else if SIPelement == "N":
    expected ← round(atomCount * N15Abundance + pepAtomCounts["C"] * C13Abundance)
  else if SIPelement == "S":
    expected ← round(2 * atomCount * S34Abundance + pepAtomCounts["C"] * C13Abundance)
  return baseMass + expected * neutronMass
```

C.6 Spectral entropy

The spectral entropy of a peak list $\{I_i\}$ with relative intensities $\hat{I}_i = I_i / \sum_j I_j$ is

$$H = - \sum_i \hat{I}_i \log \hat{I}_i.$$

The spectral-entropy similarity between an observed and a theoretical spectrum is

$$S = 1 - \frac{H_{\text{merged}} - (H_{\text{obs}} + H_{\text{theo}})}{\ln 4},$$

where H_{merged} is the entropy of the merged peak set. Aerith reports spectral entropy alongside MVH, Xcorr, and WDP so that the user can compare the score-peak resolution of each function across enrichment levels (Figure 4.6).

Pseudocode

Input:

real_intensities – array of observed peak intensities
matched_indices – array of matched observed-peak indices

Output:

```

entropy_score ← matchedSpectraEntropyScore

Initialize accumulators
totalIntensity ← 0.0
entropy ← 0.0

for i from 0 to length(real_intensities) – 1 do
    totalIntensity ← totalIntensity + real_intensities[i]
end for

for i from 0 to length(matched_indices) – 1 do
    if matched_indices[i] ≠ –1 then
        intensity ← real_intensities[matched_indices[i]]
        prob ← intensity / totalIntensity
        entropy ← entropy + (–prob * log(prob))
    end if
end for

entropy_score ← entropy

```