De novo assembled databases enable species-specific protein-based stable isotope probing of microbiomes without prior knowledge of the community composition

Simon Klaes[1,2*], Christian White[1,3], Lisa Alvarez-Cohen[3], Lorenz Adrian[1,2], Chang Ding[1]

[1]Molecular Environmental Biotechnology, Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany

[2]Geobiotechnology, Technische Universität Berlin, Berlin, Germany

[3]Civil and Environmental Engineering, University of California, Berkeley, California, United States

Running title: denovo-SIP

*To whom correspondence should be addressed: Simon Klaes (simon.klaes(at)ufz.de)

Keywords: metaproteomics, *de novo* peptide sequencing, peptide databases, microbiome analysis, microbial communities, mass spectrometry, database assembly, isotopic labelling, microbial activity profiling, database quality optimization

Abstract

Background: Protein-based stable isotope probing (Protein-SIP) is a powerful approach that can directly link individual taxa to activity and substrate assimilation, elucidating metabolic pathways and trophic relationships within microbiomes. In Protein-SIP, peptides and corresponding taxa are identified by database matches. Thus, database quality is crucial for accurate Protein-SIP analyses. For samples with unknown community composition, Protein-SIP usually employs broad-spectrum or metagenome-derived databases. However, broad-spectrum databases require advanced post-processing strategies and metagenome-derived databases can be costly to acquire. Further, both can inflate database size, negatively impacting peptide identification.

Results: Here, an approach in which we use *de novo* assembled databases for Protein-SIP on microbiomes with unknown community compositions is introduced. This approach enables databases to be directly generated from the mass spectrometry raw data using *de novo* peptide sequencing. We then use the mass spectrometric data from labeled cultures to quantify isotope incorporation into specific peptides. We benchmark our approach against the canonical approach in which a sample-matching database is generated by DNA sequencing on three different datasets: 1) a proteome analysis from a defined microbial community containing ^{13}C-labeled *E. coli* cells, 2) time-course data of an anammox-dominated continuous reactor after feeding with ^{13}C-labeled bicarbonate, and 3) a model of the human distal gut simulating a high-protein and high-fiber diet cultivated in either $^{2}H_2O$ or $H_2{}^{18}O$. Our results show that *de novo* database assembly is applicable to different isotopes and detects similar amounts of labeled peptides compared to sample-matching databases. Furthermore, we show that peptide-centric Protein-SIP allows species-specific resolution, enabling the assessment of activity related to individual biological processes. Finally, we provide access to our modular Python pipeline to assist the *de novo* assembly of databases and subsequent peptide-centric Protein-SIP data analysis (https://git.ufz.de/meb/denovo-sip).

Conclusions: *De novo* assembled databases enable species-specific Protein-SIP of microbiomes without prior knowledge of the community composition.

Introduction

Microbiomes are microbial communities in a defined habitat with distinct physio-chemical properties [1]. As microbiomes are ubiquitous and crucial for the functioning and health of myriad ecosystems, microbiome studies are essential for understanding fundamental ecological processes and advancing knowledge of human health and the environment [2–4]. Microbiomes are frequently studied via omics approaches [5], which can provide insights into their state and dynamics [6]. However, the complexity of microbiomes often hinders attempts to link ecosystem functions to individual taxa. To establish direct links between substrates, metabolites, functions, and taxa, omics approaches can be combined with stable isotope probing (SIP) [7].

In SIP experiments, microbial communities are exposed to a substrate highly enriched in a stable isotope. Upon assimilation, microorganisms can incorporate the stable isotopes into biomolecules, such as DNA [8, 9], RNA [9, 10], phospholipid-derived fatty acids [11, 12], and proteins [13, 14]. Analyzing the relative amount of labeled biomolecules and their degree of isotope incorporation can establish a direct link of individual taxa to general (for ^{2}H or ^{18}O labeled water) or substrate-specific (for ^{13}C, ^{15}N, or ^{34}S labeled substrate) activity [8-19]. By establishing this link, SIP can identify key microorganisms [20, 21], elucidate metabolic pathways [22, 23], and unravel complex food webs in microbiomes [9, 24]. The choice of the investigated biomolecule affects the sensitivity and taxonomic resolution of SIP analyses [as reviewed in 7]. DNA and RNA-based SIP provide high taxonomic resolution but rely on the separation of heavy and light fractions via ultracentrifugation leading to low sensitivity and inaccuracy in isotope quantification. Fatty acid and protein-based SIP (Protein-SIP) employ mass spectrometry to accurately quantify isotope incorporation. While fatty acids only provide limited taxonomic information, proteins can characterize communities down to the strain level. Thus, Protein-SIP offers the advantage of accurately quantifying isotope incorporation while preserving high taxonomic resolution.

Protein-SIP is based on bottom-up metaproteomics. In brief, proteins are extracted from the sample and digested into peptides that are then analyzed by high-resolution tandem mass spectrometry (MS/MS). Obtained spectra are mapped back to peptides and relative isotope abundances (RIA) are calculated from shifts in precursor ion peaks [25–27]. In principle, spectra can be mapped to peptides by database or spectral library searches as well as *de novo* peptide sequencing. To date, only database searches are used for Protein-SIP. In database searches, acquired mass spectra are compared to theoretical spectra of peptides in a constructed target database to obtain peptide-spectrum matches (PSMs) that are then validated by a target-decoy approach. While database-dependent approaches are widely accepted for proteomics and Protein-SIP, they are heavily reliant on the quality of the database leading to several constraints. One key constraint is that database creation ideally requires prior knowledge of the community composition. For defined cultures, such as pure strains, sample-matching databases are readily available in public repositories. However, for samples of unknown composition, e.g., environmental samples, databases usually must be derived from metagenome sequencing, which can be a time-consuming and costly step. In addition, metagenome-derived databases often cover only a fraction of the protein-encoding genes in complex microbiomes, so that

uncovered proteins cannot be identified [28–30]. Alternatively, broad-spectrum databases such as NCBI non-redundant proteins (nr) or Uniprot can be used [13, 31–34]. However, broad-spectrum databases increase computational time substantially, require advanced filtering strategies, and often lead to lower taxonomic resolution [35–38]. Subsets of broad-spectrum databases focusing on taxonomic marker proteins were successfully applied to Protein-SIP, reducing computational time, but also limiting the taxonomic resolution [39]. Moreover, metagenome-derived and broad-spectrum databases contain sequences of proteins that are not expressed in the sample, inflating database size. Inflated databases can negatively impact peptide identification during target-decoy validation strategies [36, 40]. Furthermore, database-dependent search approaches are inherently biased towards the database content making it impossible to detect proteins missing in the database [37].

De novo peptide sequencing addresses the above-mentioned limitations of database-dependent searches by directly inferring peptides from MS/MS spectra without using references. Thereby, *de novo* peptide sequencing does not provide a link of peptides to proteins like database-dependent approaches, making it a peptide-centric approach, in which taxonomic and functional information of peptides can be inferred by text matching (Unipept [41]) or sequence similarity (BLASTP [42]). To date, *de novo* sequencing has been used to characterize microbial community compositions at the family level [43] and to assess database suitability [44]. Until recently *de novo* sequencing identified many fewer peptides than database-dependent approaches [45]. However, current advances in artificial intelligence and the increased amount of publicly available training data have massively improved *de novo* sequencing algorithms, challenging database-dependent approaches [46, 47].

Despite the advances in *de novo* sequencing, it remains unclear whether *de novo* algorithms can reliably identify peptides in partially labeled samples, where isotope peaks may cause signal interference. Moreover, *de novo* sequencing alone cannot identify labeled peptides or quantify isotope incorporation, necessitating their integration with SIP tools. To address this knowledge gap, we propose the creation of sample-matching databases derived from *de novo* sequencing of unlabeled peptides, enabling Protein-SIP analysis without prior knowledge of community compositions (Figure 1).

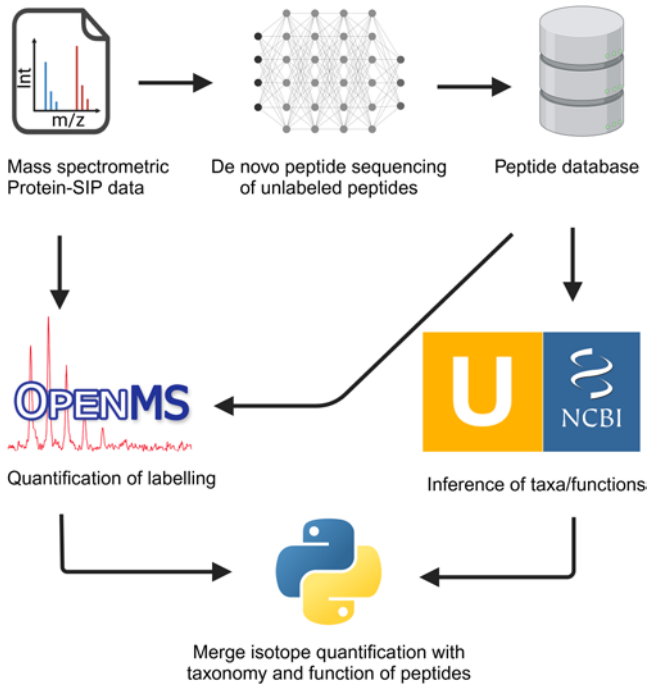


Figure 1: Proposed workflow for protein-based stable isotope probing (Protein-SIP) using *de novo* assembled peptide databases.

Here, we investigate whether *de novo* algorithms can accurately identify peptides in partially labeled samples and assess whether these identified peptides can be used to assemble sample-matching databases for Protein-SIP analysis using ^{2}H, ^{13}C, and ^{18}O labeled data. Additionally, we benchmark databases assembled with different state-of-the-art *de novo* sequencing tools with each other and with sample-matching reference databases.

Results

De novo sequencing of peptides in partially labeled samples

To investigate (1) if *de novo* algorithms can identify target peptides in partially labeled samples and (2) how the RIA affects *de novo* peptide sequencing, we analyzed previously published proteomic data (PRIDE accession number: PXD04141) of standard *E. coli* cultures cultivated with a defined ^{13}C RIA of 1.07–99% [27]. Specifically, we performed *de novo* peptide sequencing using the state-of-the-art engines PepNet and Casanovo and compared the number of *de novo* identified target and non-target peptides across the reported quality scores with a conventional database search using MS-GF+ and MetaProSIP (Figure 2).

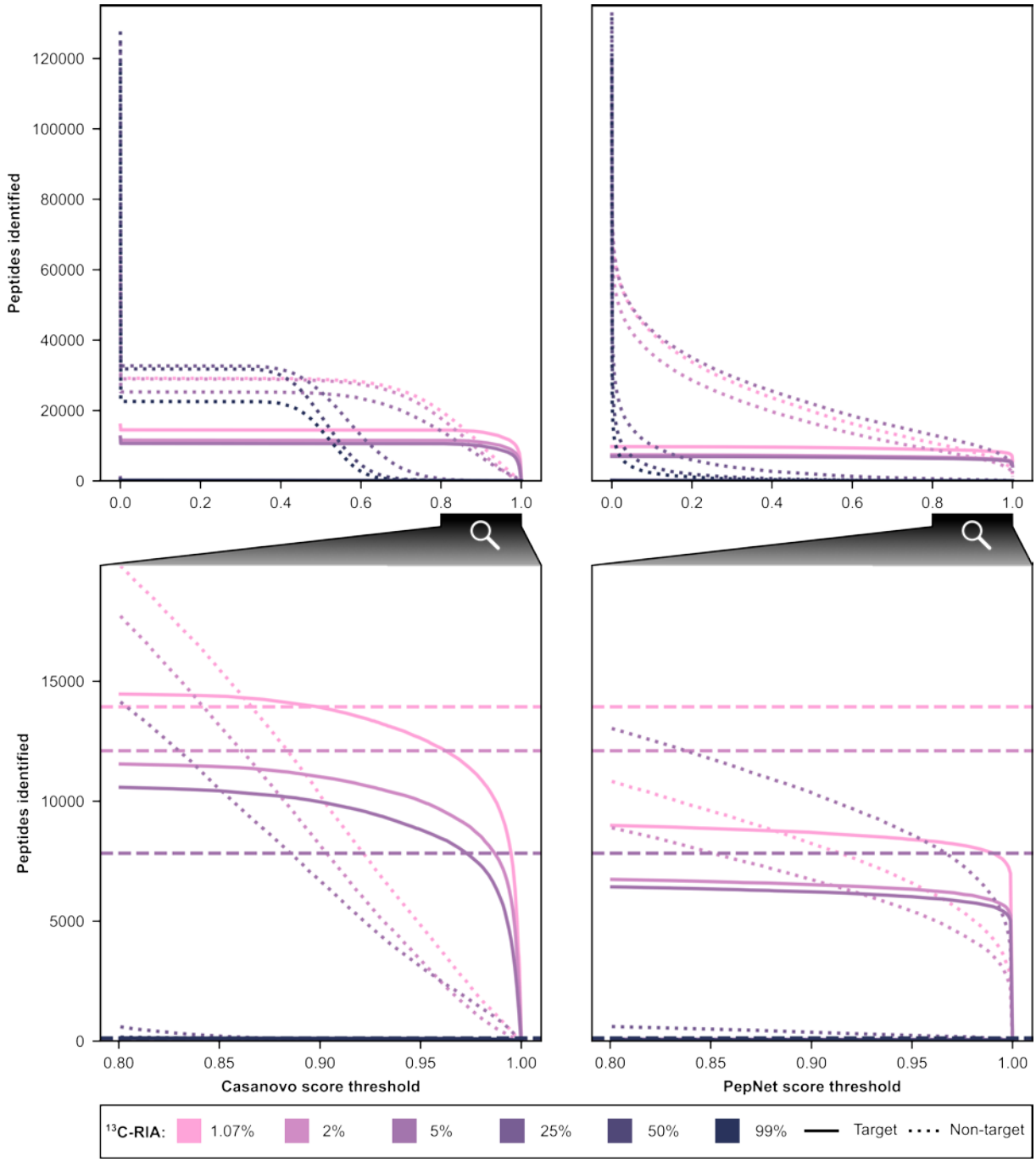


Figure 2: Comparison of the non-redundant number and quality score of peptides identified by *de novo* sequencing in standard *Escherichia coli* K12 cultures across biological triplicates cultivated with a defined ^{13}C relative isotope abundance (RIA) of 1.07–99%. *De novo* peptide sequencing was performed using Casanovo and PepNet. Negative quality scores are plotted as zero. Solid lines represent identified peptides matching the proteome of *E. coli* K12 + universal contaminants. Dotted lines indicate identified peptides not matching the proteome of *E. coli* K12 + universal contaminants. Dashed lines are only plotted in the bottom graphs and represent the number of identifications using the sample-matching reference database with MS-GF+ and MetaProSIP.

Casanovo and PepNet successfully identified 6,956–15,824 unique target peptides in samples containing a ^{13}C RIA of 1.07–5% ^{13}C. An increase in the ^{13}C RIA resulted in a reduced number of peptide

140 identifications. In samples with a ^{13}C RIA of 25–99%, both *de novo* algorithms as well as conventional
141 database search failed to detect a relevant number of target peptides. However, this limitation is not
142 critical, as SIP tools usually require an unlabeled reference for detecting labeled peptides. These
143 unlabeled references can then be used for de novo sequencing to create the peptide database.

144 Casanovo outperformed PepNet by identifying more target peptides and fewer non-target peptides.
145 Moreover, applying higher quality score thresholds led to a pronounced decrease in non-target
146 peptide identifications but only a marginal decrease in target peptide identifications. This indicates
147 effective separation of true and false-positive identifications by Casanovo at high quality scores.
148 Compared to conventional database searches, Casanovo identified 103.1–157.7% of the number of
149 unique target peptides.

150 **Benchmarking *de novo* assembled databases with a defined microbial community**

151 To test (1) if *de novo* algorithms can identify peptides from partially labeled samples, and (2) whether
152 identified peptides can be an adequate database for identifying labeled peptides, we analyzed
153 previously published proteomic data (PRIDE accession number: PXD024174) from a defined microbial
154 community containing unlabeled and ^{13}C-labeled *E. coli* cells in a 1:1 ratio [48].The community
155 consisted of 21 bacterial species belonging to 15 families, 1 eukaryote, 1 archaeon and 5 bacterial
156 viruses (Supplementary Table 1 in [49]). The labeled *E. coli* cells were cultivated with 10% fully labeled
157 (100% ^{13}C) and 90% unlabeled (1.1% ^{13}C) glucose. We performed *de novo* peptide sequencing with
158 PepNet and Casanovo. Subsequently, peptides were filtered applying different defined score
159 thresholds to assemble databases of 14,700 to 875,366 peptides (Supplementary Table S1) which were
160 then used to identify labeled peptides with a MetaProSIP workflow (Figure 3).

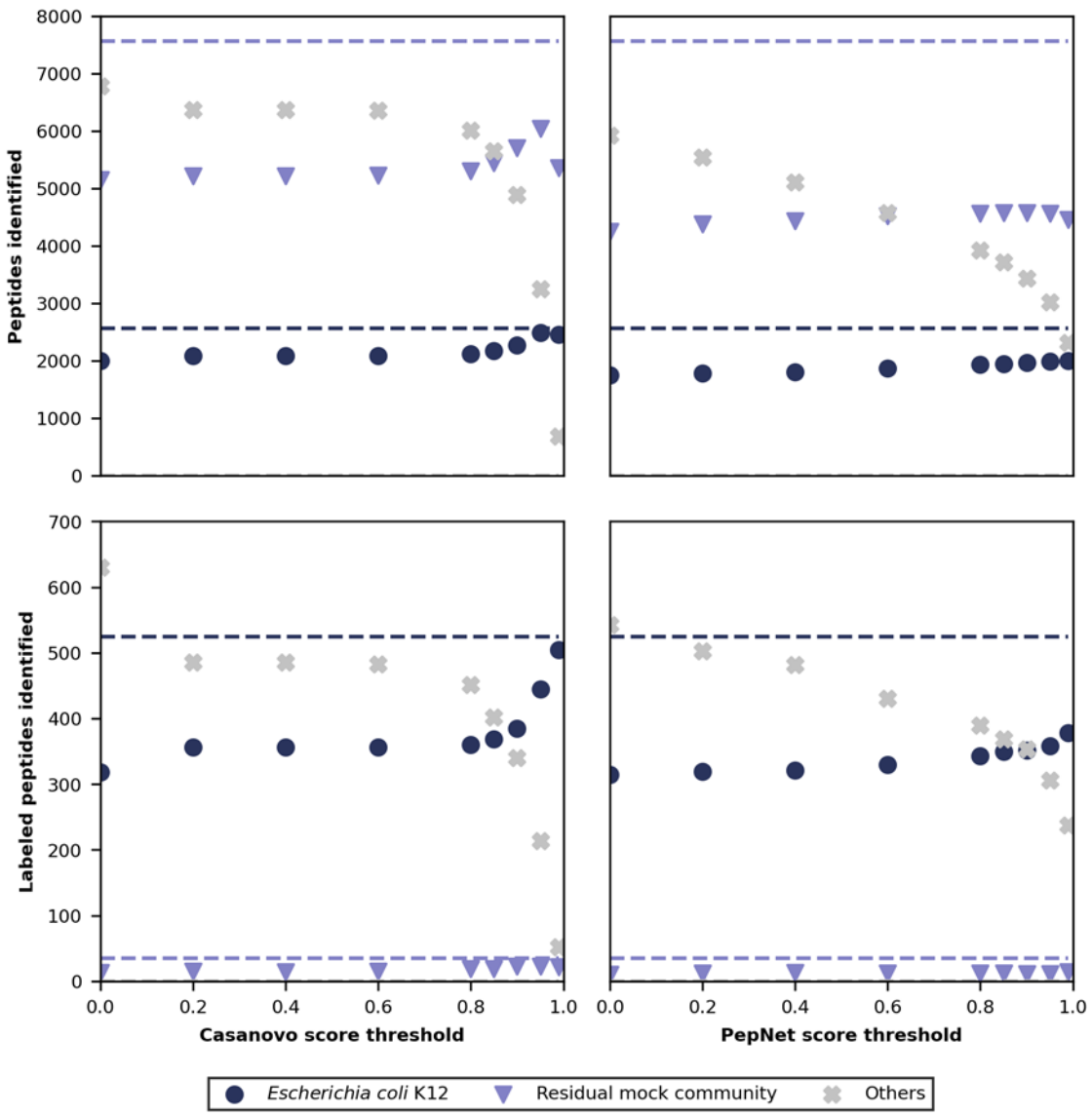
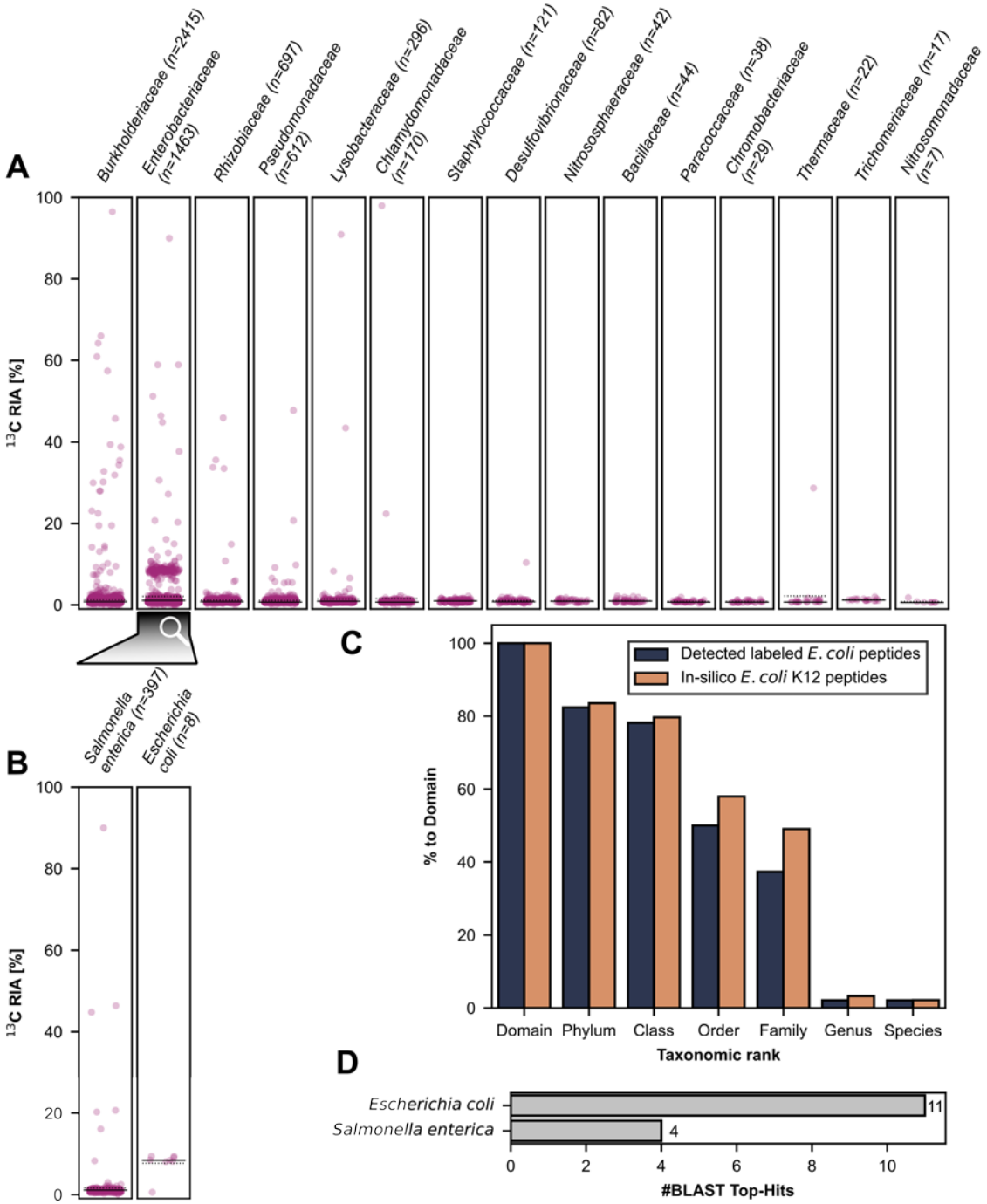


Figure 3: Comparison of *de novo* assembled databases for identifying unlabeled and ^{13}C-labeled peptides in a mock microbial community spiked with labeled *Escherichia coli* K12 cells obtained from [48] using a MetaProSIP workflow. *De novo* databases were assembled using Casanovo or PepNet and identifications were filtered with different thresholds for the quality score. The color represents whether the identified peptide is present in the proteome of *E. coli* K12 + universal contaminants (navy circle), of the residual mock community (purple rectangles), or in none of them (grey crosses). Dashed lines represent the number of identifications using the sample-matching reference database. Only peptides detected in at least two biological replicates are plotted.

We identified more peptides from the defined microbial community and fewer unrelated peptides using Casanovo-derived databases compared to those from PepNet. Casanovo identified 14,700 peptides with a score greater than 0.99, and a database built from these peptides (CS>0.99) yielded the highest number of labeled *E. coli* K12 peptides among the *de novo* assembled databases. Specifically, using the CS>0.99 database in our MetaProSIP workflow, we identified 8,496 peptides, including 505 labeled *E. coli* K12 peptides, just 20 fewer than with the reference database. Thus, the CS>0.99 database was used for further analyses. Using this database, we investigated the level of

taxonomic resolution that could be identified for the labeled organisms within the unlabeled residual community. For this purpose, we inferred the taxonomy of all identified peptides with a lowest-common-ancestor approach using Unipept and plotted the peptides with their assigned family against the calculated ^{13}C RIA (Figure 4A).



183 Figure 4: Analysis of ^{13}C relative isotope incorporation (RIA) into peptides of different taxa identified in a mock
184 microbial community with spike-in of ^{13}C-labeled *E. coli* cells. RIA values are shown for peptides detected in at
185 least two replicates and for taxa for which at least three peptides were detected. Dots indicate peptides assigned
186 to all detected families (**A**) or *Enterobacteriaceae* spp. (**B**). 'n' denotes the number of peptides identified for a
187 taxon across all replicates. Solid lines indicate medians. Dotted lines indicate means. Statistically significant
188 differences in RIA between species are marked with '**' based on Student's t-test for the means of two
189 independent samples with $p < 0.01$. **C**: Comparison of drop-off curves from *in silico* digestion of the *E.coli* K12
190 proteome and labeled peptides detected in at least two replicates of the mock microbial community with spike-
191 in of ^{13}C-labeled *E. coli* cells. **D**: Number of best hits to *Salmonella enterica* and *E. coli* when querying by BLASTP
192 the 31 labeled peptides not assigned to any lowest common ancestor by Unipept and detected in at least two
193 replicates of the mock microbial community with spike-in of ^{13}C-labeled *E. coli* cells. Sequence similarity search
194 was performed using BLASTP against the nr database. Organisms with only one hit and hits with e-value ≥ 0.001
195 were removed before plotting.

196 Our MetaProSIP workflow identified 11,497 non-redundant peptides out of which 8,449 peptides were
197 consistently detected in at least two replicates. From these 8,449 peptides, 7,897 were assigned to a
198 lowest common ancestor using Unipept. Additionally, 450 of the 8,449 peptides were labeled in at
199 least two sample out of which 419 were assigned to the lowest common ancestor via Unipept.

200 Overall, we identified 13 out of 15 bacterial families, as well as the eukaryotic and archaeal family in
201 the mock community. We did not detect any bacterial viruses. However, we consistently detected 6
202 peptides that were assigned to the fungal family *Trichomeriaceae* that was not expected in the mock
203 community. It is unclear whether these peptides are false-positives or originate from a contamination.
204 Most of the identified peptides cluster around a RIA of 1.1%, corresponding to the natural ^{13}C-
205 abundance indicative of unlabeled peptides. Only for *Enterobacteriaceae* we observed a second cluster
206 of peptides centered between a RIA of approximately 7% and 11%. *Salmonella enterica* and *E. coli* were
207 the only members of *Enterobacteriaceae* identified at higher taxonomic resolution, whereby peptides
208 assigned to *E. coli* exhibited a median RIA of 8.5%, which is significantly higher ($p < 0.001$) than that
209 found for peptides from *S. enterica,* indicating that only *E. coli* peptides were labeled (Figure 4B).

210 To test for hidden labeled populations, we applied three approaches. First, we compared the number
211 of labeled peptides assigned to *E coli* or superordinate taxa of its lineage with the number of labeled
212 peptides assigned to other taxa. Our results showed that most labeled peptides were assigned to E.
213 coli or superordinate taxa of its lineage, indicating a high population share of E. coli in the labeled
214 subpopulation of the sample (Supplementary Table S2). Secondly, we compared the theoretical drop-
215 off rate of *E. coli* K12 peptides with the observed drop-off rate of detected labeled peptides. Drop-off
216 rates describe the decline in taxon-specific peptide counts from higher to lower taxonomic levels [43].
217 By comparing observed and theoretical drop-off rates, it is possible to assess whether peptide counts
218 at a given taxonomic level can account for counts at higher levels [43]. Consequently, drop-off rates
219 can reveal hidden subpopulations not captured by the taxonomic database. Here, the observed drop-
220 off rate deviated only slightly from the theoretical one, indicating the absence of hidden side
221 populations (Figure 4C). Thirdly, we performed a sequence similarity search for the 31 labeled peptides

222 that Unipept could not assign to any lowest common ancestor (Figure 4D). For 11of those peptides,
223 BLASTP found highly similar sequences with an e-value < 0.001. After removing organisms with only
224 one hit, all highly similar sequences were related to *E. coli* or its close relative *S. enterica* [50].
225 Consequently, all performed tests indicate the absence of hidden labeled subpopulations in
226 accordance with the setup of the defined microbial community.

227 **Case study 1: Investigating trophic relations in a continuous flow reactor for the cultivation of**
228 **anammox bacteria by tracing ^{13}C incorporation into peptides**

229 To test the power and applicability of our approach on real samples, we analyzed proteomics data
230 from a continuous flow anammox reactor enriched with '*Candidatus* Kuenenia stuttgartiensis' strain
231 CSTR1 at three different time points after changing the carbon source from non-labeled to ^{13}C-labeled
232 bicarbonate while leaving nitrogen sources unchanged.

233 Again, it was possible to detect and subsequently quantify the RIA of peptides using *de novo* assembled
234 databases. The assembled databases had a size of 5,313 to 303,595 non-redundant peptides
235 (Supplementary Table S1). For samples acquired on day 21 (5.25 × hydraulic retention time), it was
236 necessary to mix labeled peptides with unlabeled peptides from day 0 to enable *de novo* peptide
237 sequencing and subsequent RIA calculation by MetaProSIP (Supplementary Table S3). Across all
238 samples, we detected more labeled '*Ca.* Kuenenia stuttgartiensis' strain CSTR1 peptides and less non-
239 CSTR1 labeled peptides with the Casanovo-derived databases compared to the PepNet-derived ones
240 (Supplementary Figure S1). Similarly, we compared different quality score thresholds for filtering *de*
241 *novo* identified peptides before creating the database for SIP analysis (Figure 5).

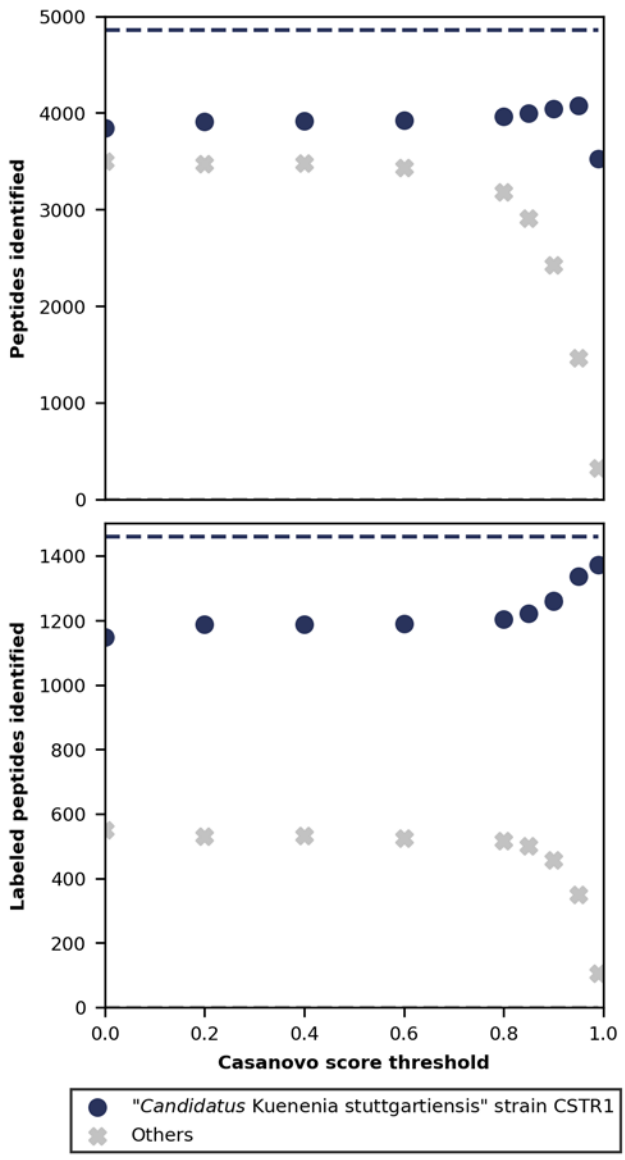
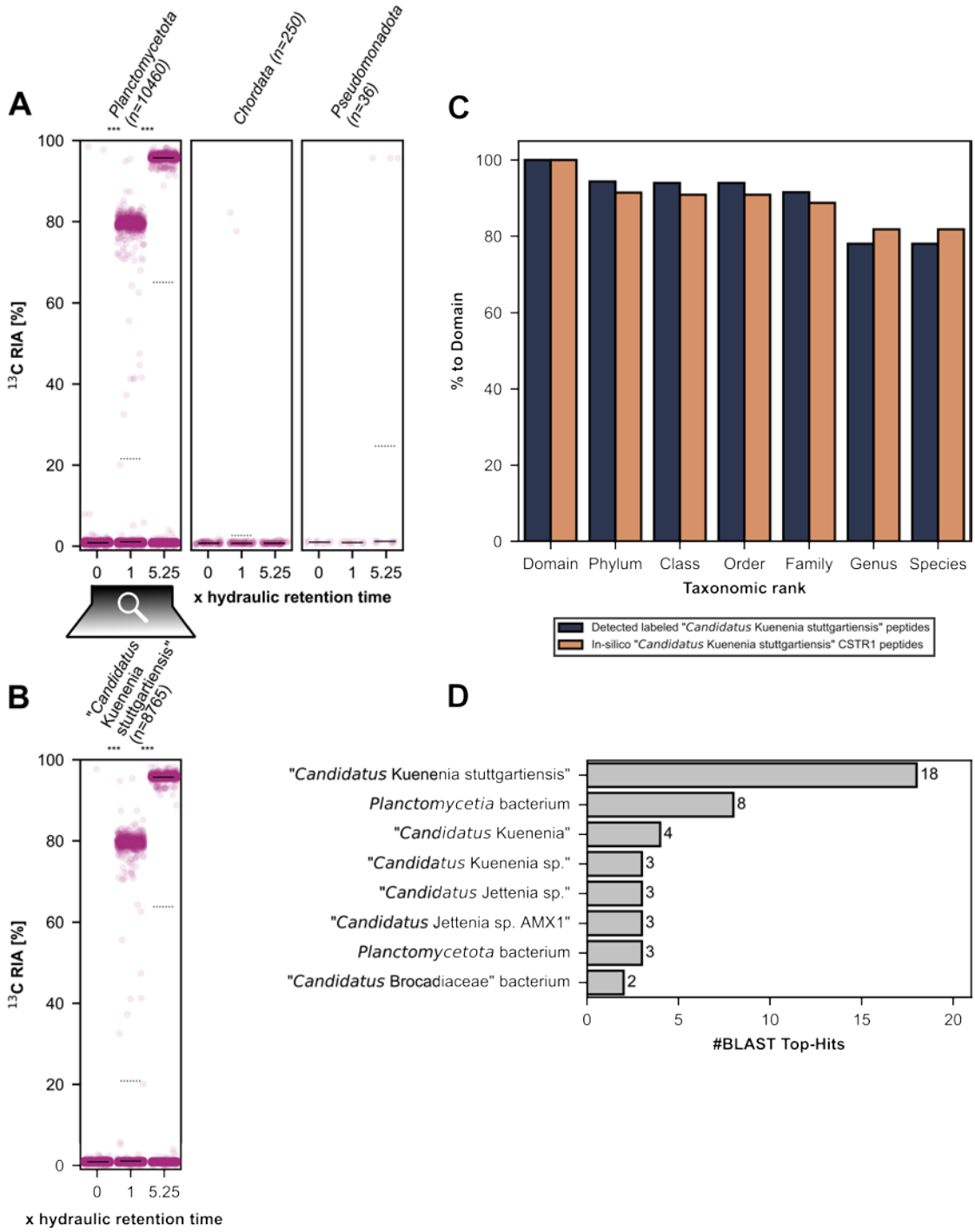


Figure 5: Comparison of *de novo* assembled databases for identifying unlabeled and ^{13}C-labeled peptides in an enrichment of '*Candidatus* Kuenenia stuttgartiensis' strain CSTR1 from a continuous flow reactor fed with ^{13}C-bicarbonate using a MetaProSIP workflow. *De novo* databases were assembled using Casanovo and filtering identifications with different thresholds for the quality score. Navy circles denote the number of identified peptides present in the proteome of '*Ca.* Kuenenia stuttgartiensis' + contaminants. Grey crosses denote the number of identified peptides absent in the proteome of '*Ca.* Kuenenia stuttgartiensis' + contaminants. Dashed lines represent the number of identifications using the sample-matching reference database derived from metagenome sequencing [51]. Only peptides detected in at least two biological replicates are plotted.

By applying a score threshold of 0.99 (CS>0.99), we identified 1372 labeled '*Ca.* Kuenenia stuttgartiensis' strain CSTR1 peptides, surpassing all other thresholds investigated. Additionally, the *de novo* assembled database demonstrated comparable performance to a genome-based one yielding 1459 labeled '*Ca.* Kuenenia stuttgartiensis' strain CSTR1 peptides. To investigate the ^{13}C incorporation

255 into taxa over time, we used the CS>0.99 database for the MetaProSIP workflow. Then, we analyzed
256 the RIA and taxonomy of detected peptides (Figure 6).



257

Figure 6: Distribution of the ^{13}C relative isotope abundance (RIA) in detected peptides across taxa identified in samples from a continuous flow reactor enriching '*Candidatus* Kuenenia stuttgartiensis' strain CSTR1 at different time points after switching to medium containing ^{13}C-bicarbonate. Dots indicate peptides assigned to detected phyla (**A**) or species (**B**). RIA values are shown for peptides detected in at least two replicates at each time point. Taxa are shown for which at least three peptides were detected. 'n' denotes the number of peptides assigned to a taxon across all replicates and time points. Solid lines indicate medians. Dotted lines indicate means. Statistically significant differences in RIA between time points are indicated with '***' based on Student's t-test for the means of two independent samples with $p < 0.001$. **C**: Comparison of drop-off curves from *in silico* digestion of the '*Candidatus* Kuenenia stuttgartiensis' strain CSTR1 proteome and labeled peptides detected in at least two samples of the continuous flow reactor enriching '*Ca*. Kuenenia stuttgartiensis' strain CSTR1. **D**: Top-hit species distribution after sequence similarity search of the 59 labeled peptides not assigned to any lowest common ancestor by Unipept but detected in at least two samples of the continuous flow reactor enriching '*Candidatus* Kuenenia stuttgartiensis' strain CSTR1. The sequence similarity search was performed with BLASTP against the nr database. Organisms with only one hit and hits with e-value ≥ 0.001 were removed before plotting.

The MetaProSIP workflow identified 4,440 non-redundant peptides out of which 1,947 peptides were consistently detected in at least two replicates of each time point. From these 1,947 peptides, 1,846 were successfully assigned to a lowest common ancestor using Unipept. Additionally, 1,232 of the 1,947 peptides were labeled in at least two sample out of which 1,173 were assigned to lowest common ancestor via Unipept.

At the phylum level, most detected peptides were assigned to *Plancomycetota*, consistent with their high abundance in the reactor (Figure 6A). Some peptides were assigned to the phylum *Chordata*, indicating human contamination, and a few were linked to *Pseudomonadota*, previously described as a subpopulation of the reactor [52]. At the species level, we only detected '*Ca.* Kuenenia stuttgartiensis' belonging to the phylum *Planctomycetota,* suggesting high enrichment of the culture (Figure 6B). Substantial incorporation of ^{13}C has been detected only for peptides assigned to '*Ca.* Kuenenia stuttgartiensis' or superordinate taxa of its lineage, indicating its activity and assimilation of ^{13}C-bicarbonate (Supplementary Table S4). Specifically, we observed a significant ($p < 0.001$) increase of ^{13}C RIA in '*Ca.* Kuenenia stuttgartiensis' peptides at each investigated time point, resulting in a median RIA of 95.7% after 5.25 × hydraulic retention time demonstrating almost complete substitution of ^{12}C with ^{13}C.

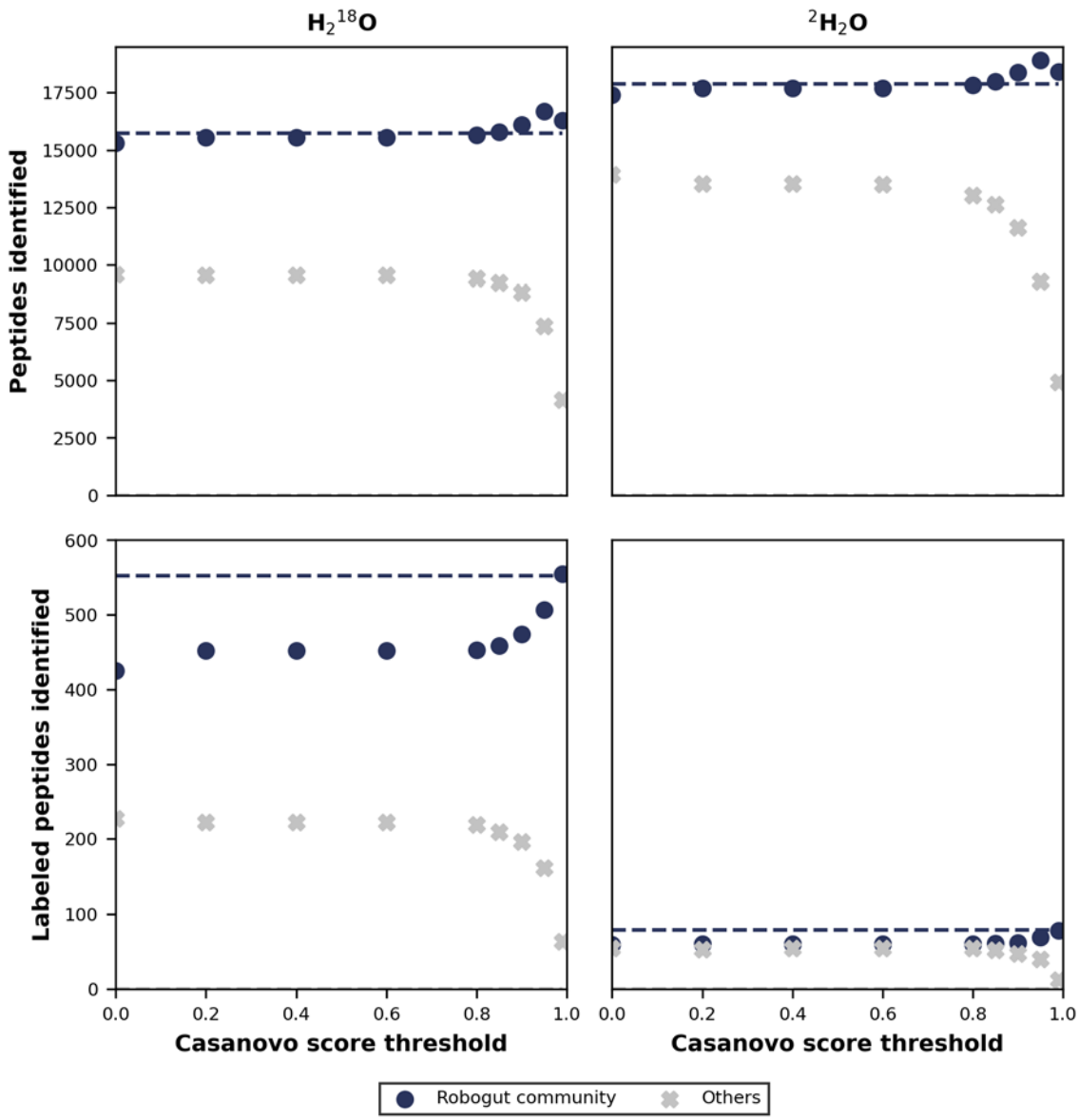
To investigate for labeled populations not covered by Unipept, we again compared observed and theoretical drop-off curves of peptides (Figure 6C). The drop-off curve of detected labeled peptides closely resembled the theoretical one, with a maximum deviation of only 4.9% (Supplementary Table S4). This suggests that all detected labeled peptides were derived from '*Ca.* Kuenenia stuttgartiensis'.

Furthermore, we searched for sequences similar to the 59 labeled peptides that Unipept could not assign to any taxon (Figure 6D). For 15 of those peptides, BLASTP identified highly similar sequences with e-value < 0.001. After removing organisms with only one hit, most of the highly similar sequences belonged to '*Ca.* Kuenenia stuttgartiensis' or superordinate taxa of its lineage. Three hits with high sequence similarity were affiliated with '*Candidatus* Jettenia sp.' and '*Candidatus* Jettenia sp. AMX1',

297 respectively, which are anammox bacteria. In brief, our results indicate that '*Ca.* Kuenenia
298 stuttgartiensis' also dominates the labeled population in the reactor.

299 **Case study 2: Investigating species-specific and function-specific activity between two different**
300 **simulated diets in a human distal gut microbiome model by tracing ^{2}H and ^{18}O incorporation**
301 **into peptides**

302 To test our strategy for comparative purposes and also for stable isotopes other than ^{13}C, we analyzed
303 previously published raw data (PRIDE accession number: PXD024291) from a model community of the
304 human gut microbiome ('Robogut community') cultivated for 12 h in a protein-rich or fiber-rich
305 medium simulating different diets [18, 48]. The medium contained unlabeled water (for controls) or
306 water in which 25% was substituted with either $^{2}H_2O$ or $H_2{}^{18}O$. First, we compared different *de novo*
307 assembled databases for identifying unlabeled and labeled peptides expected to be present in the
308 sample based on the published reference database (Figure 7).



309

310 Figure 7: Comparison of *de novo* assembled databases for identifying unlabeled and $^{18}O/^{2}H$-labeled peptides in
311 a model of the human gut microbiome (Robogut community) cultivated with $H_2{}^{18}O$ or $^{2}H_2O$, respectively, and
312 obtained from [18] using a MetaProSIP workflow. *De novo* databases were assembled using Casanovo after
313 filtering all identifications with defined thresholds for the quality score. Navy circles denote the number of
314 identified peptides present in the sample-matching reference proteome + contaminants. Grey crosses denote
315 the number of identified peptides absent in the sample-matching reference proteome + contaminants. Dashed
316 lines represent the number of identifications that we generated using the published sample-matching reference
317 database. Only peptides detected in at least two biological replicates are plotted.

318 The *de novo* assembled databases contained 48,715 to 715,031 non-redundant peptides
319 (Supplementary Table S1). We identified more Robogut peptides with databases created using
320 Casanovo compared to those created using PepNet, for both investigated stable isotopes ^{18}O and ^{2}H
321 (Supplementary Figure S2). The database for which peptides were *de novo* sequenced by Casanovo
322 and filtered with a score threshold of 0.99 (CS>0.99) led to the most identifications of labeled Robogut

peptides among all investigated databases. Furthermore, we identified more labeled peptides in samples cultivated with ^{18}O than with ^{2}H and more unlabeled peptides in the ^{2}H samples.

Next, we analyzed the RIA and taxonomy of peptides identified under different dietary conditions using the CS>0.99 database. Our MetaProSIP workflow identified 35,115 non-redundant peptides, out of which 8,099 and 8,999 peptides were detected in at least two replicates of both media compositions in presence of ^{18}O and ^{2}H, respectively. From these 8,099 and 8,999 peptides, 7,669 and 8,430 were assigned to a lowest common ancestor using Unipept, respectively.

Overall, the incorporation of ^{18}O was substantially higher than the incorporation of ^{2}H. Moreover, the identified peptides generally had significantly higher RIA ($p < 0.01$) in protein-rich compared to fiber-rich conditions (Supplementary Figure S3).

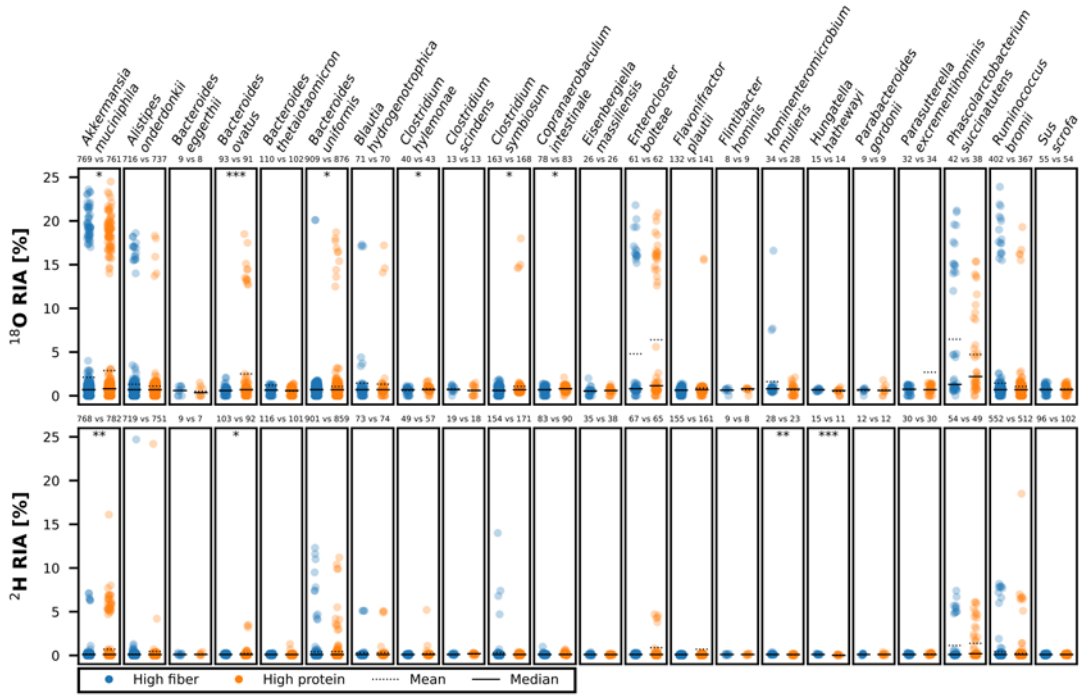


Figure 8: Distribution of the ^{18}O and ^{2}H relative isotope abundance (RIA) in detected peptides across species identified in a model of the human gut microbiome cultivated in either a high fiber or high protein media in the presence of either $H_2{}^{18}O$ or $^{2}H_2O$. RIA values are shown for peptides detected in at least two replicates of both media compositions. Species detected in the presence of $H_2{}^{18}O$ and $^{2}H_2O$ are plotted for which at least three peptides were detected. 'n' denotes the number of peptides assigned to a specific taxon across triplicates. Statistically significant differences in RIA between media are indicated with '*', '**', and '***' based on Student's t-test for the means of two independent samples with $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively. To improve readability, outliers were removed from the raw graph (Supplementary Figure S4).

After inferring the taxonomy of peptides via Unipept, we identified 22 species present in all tested setups (Figure 8). Notably, we detected *Copranaerobaculum intestinale* and *Hominenteromicrobium mulieris*, which were recently described [53, 54] but not reported by previous analyses of this dataset due to their absence in the originally used database [18, 48]. In ^{18}O incubated samples, more taxa were found to differ significantly in their RIA between the two simulated diets than ^{2}H incubated samples.

Specifically, we detected significantly higher RIA under high protein conditions compared to high fiber conditions for peptides assigned to *Akkermansia muciniphilia*, *Bacteroides ovatus*, *Bacteroides uniformis*, *Clostridium hylemonae*, *Clostridium symbiosum*, and *Copraanaerobaculum intestinale*, as well as significantly lower RIA for *Hominenteromicrobium mulieris* and *Hungatella hathewayi*.

To assess changes in activity related to individual biological processes, we assigned identified peptides and their RIA to biological processes via gene ontology (GO) numbers using Unipept. For an overview, GO numbers were first merged into GO slims of the prokaryote subset (Figure 9).

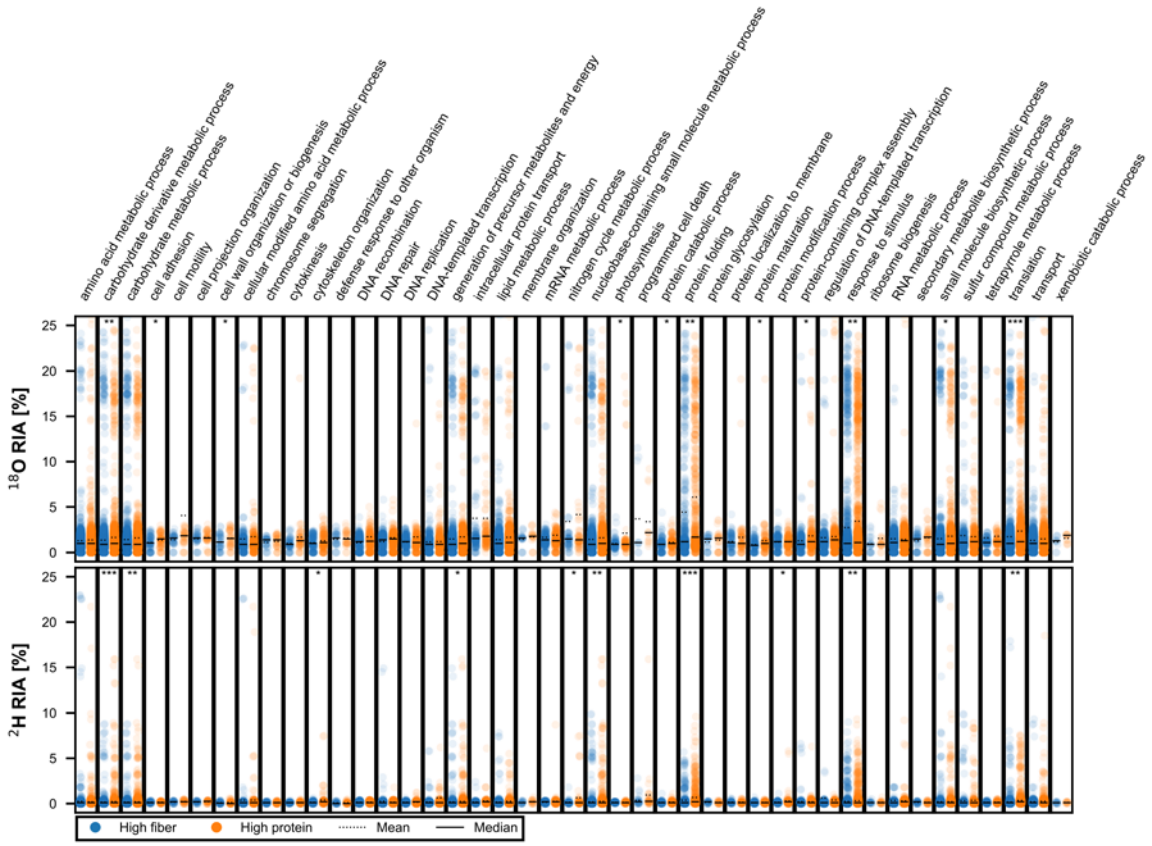


Figure 9: Distribution of the ^{18}O and ^{2}H relative isotope abundance (RIA) in detected peptides across general biological processes of prokaryotes (goslim_prokaryote) identified in a model of the human distal gut microbiome cultivated in either a high fiber or high protein medium in the presence of either $H_2{}^{18}O$ or $^{2}H_2O$. RIA values are shown for peptides detected in at least two replicates of both media and for biological processes detected in presence of $H_2{}^{18}O$ and $^{2}H_2O$ for which at least three peptides were detected (10 to 3457 peptides per strip). Statistically significant differences in RIA between media are indicated with '*', '**', and '***' based on Student's t-test for the means of two independent samples with $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively. To facilitate readability, outliers were removed from the raw graph (Supplementary Figure S5).

For most biological processes peptides exhibited similar RIA values between the diets, while they exhibited significantly different RIAs for other processes. Specifically, peptides assigned to carbohydrate derivative metabolic processes, protein folding, response to stimulus, and translation exhibited significantly higher ^{2}H and ^{18}O RIAs in the high protein medium. Peptides assigned to cell adhesion, cell wall organization or biogenesis, photosynthesis, protein catabolic processes, protein

368 maturation, protein-containing complex assembly, and small molecule biosynthetic processes only
369 exhibited significantly higher ^{18}O RIA in the protein-rich medium. Peptides assigned to carbohydrate
370 metabolic processes, cytoskeleton organization, generation of precursor metabolites and energy,
371 nitrogen cycle metabolic processes, nucleobase-containing small molecule metabolic processes, and
372 protein modification processes exhibited significantly higher ^{2}H RIA in the protein-rich medium.

373 To further investigate biological processes related to dietary differences, we specifically analyzed all
374 detected GO child terms of the parental GO slims of amino acid metabolic processes (GO:0006520),
375 carbohydrate derivative metabolic processes (GO:1901135), and carbohydrate metabolic processes
376 (GO:0005975). The identified peptides were assigned to a total of 160 biological processes related to
377 the above-mentioned GO slims, out of which 28 biological processes exhibited significant differences
378 ($p < 0.05$) in ^{2}H or ^{18}O RIA between both media compositions (for specific processes see Supplementary
379 Table S5). Differences in RIA related to amino acid or carbohydrate metabolism were isotope-specific.
380 In the protein-rich medium, significantly higher ^{18}O and ^{2}H RIA were observed in peptides associated
381 with L-histidine catabolism to glutamate and formate, as well as carbohydrate metabolism. The
382 protein-rich medium also showed significantly elevated ^{18}O RIA in peptides linked to
383 acetylglucosamine, arabinose, and fucose catabolisms, as well as fructose-6-phosphate metabolism,
384 while ^{2}H RIA was significantly higher in peptides involved in aminoacylation, biosynthesis of arginine,
385 UTP, and GMP, as well as aspartate, glucose, and fructose-1,6-bisphosphate metabolism. Conversely,
386 the fiber-rich medium exhibited significantly higher ^{18}O RIA in peptides related to amino acid
387 biosynthesis (particularly valine and isoleucine), and higher ^{2}H RIA in peptides associated with cellulose
388 catabolism, as well as arabinose and nucleoside metabolisms.

389

390 ## Discussion

391 Our results show that *de novo* peptide sequencing can generate sample-matching databases for ^{2}H,
392 ^{13}C, and ^{18}O Protein-SIP analyses without prior knowledge of community compositions. In our work,
393 databases derived from Casanovo allowed for the identification of consistently more target peptides
394 than PepNet-derived databases. Filtering *de novo* identified peptides to those with a quality score
395 > 0.99 before assembling the database (CS>0.99) enabled the identification of more labeled target
396 peptides than using a non-filtered database, supporting the conclusion that inflated database size can
397 challenge database search algorithms, as reported previously [36, 40]. By using the CS>0.99 databases,
398 we identified 72.6–103.6% of all target peptides and 94.0–100.5% of labeled target peptides compared
399 to the reference databases generated from (meta)genome information. These results indicate that the
400 CS>0.99 databases perform comparably to reference databases. Similarly, casanovo-derived databases
401 were recently reported to outperform the broad-spectrum database Swiss-Prot [55].

402 Based on the analyzed data and the obtained results, we propose a modular Denovo-SIP pipeline
403 (Figure 1). In the proposed Denovo-SIP pipeline, peptides are first *de novo* sequenced using Casanovo
404 [46] to create a peptide sequence database for subsequent database-dependent search. The database-
405 dependent search is carried out by MS-GF+ [56] and MetaProSIP [26] to validate peptides with a target-
406 decoy approach and infer isotope incorporation. Unipept [41] and BLASTP [42] can then retrieve taxa
407 and functions of peptides. Python scripts from the Denovo-SIP pipeline allow the integration of isotope
408 incorporation data with functional and taxonomic annotation for subsequent analysis. The modular
409 structure of the Denovo-SIP pipeline allows flexibility, enabling the integration of different engines and
410 the replacement or supplementation of modules with alternative tools to meet specific needs.. For
411 example, Sage could increase the speed of database search [57]; Sipros 4 could improve the accuracy
412 and sensitivity in detecting labeled peptides [27]; Calis-p could enhance the analysis of low amounts
413 of isotope incorporation [48]; and MetaX could further analyze peptide-associated taxa and functions
414 [58]. Furthermore, *de novo* peptide sequencing algorithms are developing fast: Engines such as π-
415 HelixNovo [59], PowerNovo [60], and NovoB [61], recently reported to outperform Casanovo on some
416 datasets, could be used in the Denovo-SIP pipeline. Moreover, rescoring algorithms could be employed
417 to increase sensitivity and precision of *de novo* peptide sequencing [62] as well as database searches
418 [63].

419 We first tested the Denovo-SIP pipeline by analyzing proteomic data from a defined microbial
420 community containing ^{13}C-labeled *E. coli* cells. Using the pipeline, we successfully identified the isotope
421 labeling in *E. coli* and showed that the residual community was unlabeled. We were unable to detect
422 the unlabeled families *Alteromonadaceae* and *Roseobacteriaceaee*, which may have been caused by
423 the raw data quality since even with the reference database, only 0 to 2 peptides of those families
424 could be isotopically quantified in most replicates (Supplementary Table S5 in [48]). Previous analyses
425 also suggest that raw data quality is a major bottleneck of *de novo* sequencing [45, 64]. We also did
426 not detect bacterial viruses in the mock community, despite their protein abundance of 0.08–0.14%
427 (Supplementary Table 3 in [49]). However, these bacterial viruses were also not reliably identified by
428 metaproteomics using a metagenome-derived database [49]. Similarly, another bottom-up proteomics

429 study reported difficulties in detecting viruses having a low relative abundance within a complex
430 microbiome [65].

431 By using the Denovo-SIP pipeline, we could also confirm that the anammox bacterium '*Candidatus*
432 Kuenenia stuttgartiensis' strain CSTR1 used bicarbonate as its sole carbon source. Furthermore, we did
433 not detect any other organisms in the reactor assimilating bicarbonate. This experiment demonstrated
434 that Denovo-SIP can detect highly enriched peptides with RIA of more than 95% as well as differentially
435 labeled peptides between time points. Therefore, Denovo-SIP can be applied to time-series studies.

436 We also employed the Denovo-SIP workflow to analyze previously published data from the human
437 distal gut model microbiome cultivated in a fiber-rich or protein-rich medium containing either ^{2}H- or
438 ^{18}O-labeled water. This approach is used to detect general activity as deuterium and oxygen from water
439 is always incorporated when a microorganism is active. Through this analysis, we showed that changes
440 in activity due to dietary compositions are species- and process-specific in agreement with previous
441 analyses [18, 48]. Previous work reported that a fiber-rich medium increases activity, thus RIA in
442 peptides, related to amino acid biosynthesis [18]. Our work specifically suggests that the biosynthesis
443 of valine and isoleucine, as well as the catabolism of histidine and the metabolism of carbohydrates
444 were affected by the medium composition.

445 In conclusion, our results show that *de novo* assembled databases perform comparably to sample-
446 matching reference databases in detecting labeled peptides. The main advantages of *de novo*
447 assembled databases are the saved time and costs associated with metagenome sequencing and the
448 ability to detect foreign peptides, which are typically absent in reference databases. Such foreign
449 peptides could originate from contamination [66], mutation [67], alternative splicing [68], or
450 nonspecific cleavage [69]. Moreover, *de novo* assembled databases allow the detection of peptides
451 from organisms not expected in the sample, as shown here for *C. intestinale* and *H. mulieris* in the
452 human distal gut model.

453 *De novo* assembled databases also have certain limitations. First, *de novo* sequencing was only
454 successful on peptides with RIA of up to 5%. Usually such low or unlabeled peptides are present in the
455 beginning or in controls of experiments allowing database creation for subsequent analyses of labeled
456 samples. This limitation is not specific to *de novo* sequencing. To date, all major SIP tools (SIPPER [25],
457 MetaProSIP [26], and Sipros [27]) require unlabeled reference samples to identify labeled peptides.
458 Secondly, *de novo* peptide sequencing is computationally intensive and often requires high-
459 performance computing graphics processing units (GPUs) to achieve inference speeds comparable to
460 database searches. Thirdly, *de novo* assembled databases contain amino acid sequences derived from
461 mass spectra. Amino acid sequences are less variable than nucleic acid sequences. Hence, *de novo*
462 assembled databases inherently result in lower taxonomic resolution than databases derived from
463 nucleic acid sequences, such as metagenome sequencing. Furthermore, taxonomy inference via a
464 lowest common ancestor approach requires the detection of taxon-specific peptides. The number of
465 taxon-specific peptides decreases with increasing taxonomic resolution [70]. Moreover, the
466 abundance of taxon-specific peptides varies between taxa [38]. Thus, the overall sensitivity and
467 taxonomic resolution of lowest common ancestor approaches is dependent on the specific taxa

468 present. Still, we demonstrated that *de novo* assembled databases can achieve species-specific
469 resolution. Another limitation is that although identifying peptides by *de novo* sequencing does not
470 require their presence in databases, interpreting the identified peptides by inferring proteins, taxa, or
471 functions still requires a reference database. We propose using Unipept for taxa/function inference.
472 However, the current Unipept version 5.1.5 can only infer information from perfect matches of tryptic
473 peptides. Recently, Unipept update 6.0 has been announced to allow the analysis of non-tryptic
474 peptides enhancing the synergy with *de novo* sequencing [71]. To get information about peptides that
475 do not perfectly match, taxonomic drop-off rates and BLAST can be integrated into our automated
476 workflow. In doing so, Denovo-SIP can assess the potential of unreferenced organisms. Estimating
477 false-discovery rates of *de novo*-identified peptides is challenging [72], and currently no tools for *de*
478 *novo* sequencing of labeled peptides exist. Therefore, our approach of assembling databases through
479 *de novo* peptide sequencing – followed by database searches to verify these peptides and calculate
480 isotope incorporation rates – shows promise. Denovo-SIP could advance the application of Protein-SIP
481 across various microbiological research fields, particularly when screening of many different samples
482 is necessary.

483 # Methods

484 **Cultivation and protein extraction**

485 Proteomics data of the anammox community were obtained with samples taken from a continuous
486 stirred-tank reactor cultivating a mixed culture containing predominantly the anammox bacterium
487 '*Candidatus* Kuenenia stuttgartiensis' strain CSTR1 [52]. The reactor had an effective volume of 500 mL
488 and was fed with two medium solutions containing 60 mM ammonium and 60 mM nitrite
489 (concentration after mixing with ratio 1:1), respectively. Sodium bicarbonate (10 mM, concentration
490 after mixing) was the only carbon source provided, provided with the nitrite medium. For a detailed
491 medium recipe, please refer to [52]. The reactor was running stably with non-labeled bicarbonate at
492 30 °C, constant stirring, and 4-day hydraulic retention time. The cell density was around 2.5×10^8 cells
493 mL^{-1}. At day 0, the nitrite medium containing non-labeled sodium bicarbonate was replaced with a
494 medium containing ^{13}C-labeled sodium bicarbonate, while all other medium components and reactor
495 operational conditions remained the same. Three samples (as triplicates) were taken from the reactor
496 on day 0 (non-labeled), day 4 (1 × hydraulic retention time), and day 21 (5.25 × hydraulic retention
497 time). Samples of 1.5 mL were centrifuged at 10,000 × *g*, 10 °C for 10 min, and the pellets were frozen
498 at -20 °C until proteomic analysis was done.

499 **Sample preparation and mass spectrometry**

500 Saved pellets from reactor liquid were resuspended in 50 µL of 100 mM ammonium bicarbonate
501 buffer, and went for a 3× freeze/thaw cycle to disrupt the cells (-80 °C followed by 1 min gentle shaking
502 at 40 °C). Two microliter of 40 ng μL^{-1} bovine serum albumin was added to each sample as internal
503 standard. Then, an equal volume of 10% (w/v) sodium deoxycholate (DOC) was added to each sample,
504 before 12 mM dithiothreitol (DTT) and 40 mM 2-iodoacetamide (IAA) (both were final concentrations)
505 were added sequentially to reduce and alkylate the disulfide bond (incubation conditions: 30 min,
506 37 °C on a rotary shaker at 400 rpm for DTT; and 45 min, room temperature, 400 rpm, in the dark for
507 IAA). Afterwards, 100 mM ammonium bicarbonate was added to dilute DOC to 1% (w/v). Sequencing-
508 grade trypsin (6.3 µL, 0.1 µg μL^{-1}, Promega, Madison, USA) was then added to digest the sample at
509 37 °C, 400 rpm overnight. After trypsin digestion, neat formic acid was added to a final concentration
510 of 2% (v/v), samples were vortexed, and two rounds of centrifugation at 16,000 × *g* for 10 min were
511 performed to completely remove precipitated DOC. Peptides in the supernatant were desalted using
512 100 µL C_{18} tips (Pierce, Thermo Fisher Scientific, Massachusetts, USA) following the manufacturer's
513 instruction. Eluted peptides were vacuum dried and resuspended in 50 µL of 0.1% (v/v) formic acid.
514 Peptide concentrations were determined based on the UV absorbance at 205 nm measured on a DS-
515 11 micro-volume spectrophotometer (DeNovix, Delaware, USA). Peptides from day 4 (1 × hydraulic
516 retention time) and day 21 (5.25 × hydraulic retention time) were mixed with peptides from day 0 in a
517 1:1 ratio before 3 µL were injected for measurement on nano-LC-MS/MS (Orbitrap) as described
518 previously [51] except that the MS2 resolution was lowered from 60,000 to 30,000. The proteomics

519 data were deposited at ProteomeXchange via the PRIDE partner repository under the accession
520 number PXD057215 (reviewer token: ejwbd9cUNtVw).

521 **Generation of *de novo* assembled peptide databases**

522 Mass spectrometric raw files (Supplementary Table S6) were converted into mgf format after peak-
523 picking, removing zero values in spectra, encoding in 64-bit precision, and zlib compression using
524 msConvert (ProteoWizard v3.0.23304) [73]. Mass spectra were translated to peptide sequence lists by
525 *de novo* sequencing using PepNet (v0.3.2) [47] and Casanovo (v3.5.0) [46] with the pre-trained models.
526 Peptide sequence lists were filtered by a threshold of the quality score (default: >0.99) reported by the
527 *de novo* engines. For each dataset, filtered peptide sequence lists were merged and converted to fasta
528 format using our DeNovo-SIP pipeline (https://git.ufz.de/meb/denovo-sip).

529 **Identification of isotopically labeled peptides**

530 Mass spectrometric raw files (Supplementary Table S6) were converted into mzML format after peak-
531 picking and encoding in 32-bit precision using msConvert (ProteoWizard v3.0.19046). [73] Next, mzML
532 files were analyzed using the *de novo* assembled peptide databases and for comparison also with the
533 sample-matching protein databases appended with universal contaminants [66] by a customized
534 MetaProSIP workflow [74] based on OpenMS v2.8 [75] on the Galaxy [76] instance of the Helmholtz
535 Centre for Environmental Research – UFZ. In brief, peptides were detected and validated by MS-GF+
536 [56] using a target-decoy approach, and peptide relative isotope abundances (RIA), as well as labeling
537 ratios (LR), were calculated by MetaProSIP [26]. The precursor mass tolerance was set to ± 10 ppm (for
538 the mock community with spike-ins of ^{13}C-labeled *Escherichia coli* [48] and the model of the human
539 gut microbiome [18]) or ± 5 ppm (for the '*Candidatus* Kuenenia stuttgartiensis' enrichment). We
540 allowed a peptide length of 6–100 amino acids, no (for peptide databases) or semi-tryptic (for protein
541 databases) cleavage and up to two of the following modifications: variable oxidation of methionine,
542 deamidation of asparagine or glutamine, N-terminal acetylation, carbamylation, or ammonia-loss and
543 fixed C-terminal carbamidomethylation. The weight merge window was set to ± 1.5 (for ^{2}H) and ± 5.0
544 (for ^{18}O and ^{13}C) in MetaProSIP. Identifications not assigned to a feature during pattern detection and
545 purely natural peptides were included in the report. The peptide-centric output of MetaProSIP was
546 used for further processing. Peptides for which MetaProSIP reported at least one non-natural weight
547 were considered as labeled. The taxonomy and gene ontology (GO) [77] terms of peptides were
548 inferred based on the lowest common ancestor approach by Unipept Desktop v2.0.2 [78] using UniProt
549 2024.03 [79]. Advanced missed cleavage handling and indistinguishability of leucine and isoleucine
550 were enabled. GO terms of biological processes were mapped to GO slims of the prokaryote subset
551 maintained by the GO Consortium [80] (goslim_prokaryote) using GOATOOLS [81]. Python scripts from
552 our Denovo-SIP pipeline (https://git.ufz.de/meb/denovo-sip) merged results from Unipept and
553 MetaProSIP for downstream analysis.

554 **Sequence similarity search of unassigned peptides**

555 Labeled peptides that could not be assigned to a lowest common ancestor by Unipept were searched
556 for similar sequences using BLASTP 2.16.0+ [42] with the nr database (accession date: 2024/11/01).
557 Hits with e-value ≥ 0.001 were removed before plotting top-hit species distributions using Blast2GO
558 6.0.3 [82].

559 **Taxonomy drop-off rates**

560 The number of taxon-specific peptides decreases as taxonomic resolution increases, resulting in a
561 'drop-off' from higher to lower taxonomic levels. Comparing observed and theoretical drop-off rates
562 within a lineage allows assessing whether the peptide count at a given taxonomic level (e.g., species
563 level) can explain the peptide count at higher taxonomic levels (e.g., phylum level) [43]. By this, drop-
564 off rates can reveal 'hidden' organisms not captured by a taxonomic database (here: Unipept). In this
565 manuscript, we use the term 'superordinate taxa' to refer to taxonomic ranks above a given taxon. For
566 example, the superordinate taxa of *Escherichia coli* (species) include *Escherichia* (genus),
567 *Enterobacteriaceae* (family), *Enterobacterales* (order), *Gammaproteobacteria* (class),
568 *Pseudomonadota* (phylum), and *Bacteria* (domain).

569 Labeled peptides were tested for hidden organisms comparing observed with theoretical taxonomy
570 drop-off rates of labeled taxa. To calculate theoretical drop-off rates, we extracted expressed proteins
571 from the proteome of *E. coli* K12 and '*Ca.* K. stuttgartiensis' strain CSTR1 based on previous studies
572 (Table S7 in [51] and Run4_Ecoli268_R3a_0_13C_400ng.pdResult in PXD023693 [83] from [48]).
573 Expressed proteins were *in-silico* digested into tryptic peptides with a length of 7–15 amino acids. The
574 detectability of peptides was assessed by DeepMSpeptide [84] and detectable peptides were
575 processed by Unipept as explained above. We accounted for false-positive Unipept annotations by
576 employing a sequence randomization strategy as described previously [43].

577 ## Availability of data and materials

578 The raw mass spectrometric proteomics data generated during this study have been deposited to the
579 ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner
580 repository [83] with the dataset identifier PXD057215 (reviewer token: ejwbd9cUNtVw). Raw mass
581 spectrometry proteomics data reanalyzed during this study is also available via the PRIDE partner
582 repository (Supplementary Table S6). Data obtained from our reanalyses has been deposited to
583 Zenodo (Supplementary Table S6). The Denovo-SIP pipeline's Python scripts are available via GitLab
584 (https://git.ufz.de/meb/denovo-sip). Any other relevant data will be made available upon request.

585 ## References

586 1. Berg G, Rybakova D, Fischer D, Cernava T, Vergès MCC, Charles T, et al. Microbiome definition re-
587 visited: old concepts and new challenges. Microbiome. 2020;8:103.

588 2. Layeghifard M, Hwang DM, Guttman DS. Disentangling interactions in the microbiome: a network
589 perspective. Trends Microbiol. 2017;25:217–28.

590 3. Lynch S V, Pedersen O. The human intestinal microbiome in health and disease. N Engl J Med.
591 2016;375:2369–79.

592 4. Falkowski PG, Fenchel T, Delong EF. The microbial engines that drive earth's biogeochemical cycles.
593 Science. 2008;320:1034–9.

594 5. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, et al. Best practices for analysing
595 microbiomes. Nat Rev Microbiol. 2018;16:410–22.

596 6. Zhang X, Li L, Butcher J, Stintzi A, Figeys D. Advancing functional and translational microbiome
597 research using meta-omics approaches. Microbiome. 2019;7:154.

598 7. Jehmlich N, Vogt C, Lünsmann V, Richnow HH, von Bergen M. Protein-SIP in environmental studies.
599 Curr Opin Biotechnol. 2016;41:26–33.

600 8. Radajewski S, Webster G, Reay DS, Morris SA, Ineson P, Nedwell DB, et al. Identification of active
601 methylotroph populations in an acidic forest soil by stable-isotope probing. Microbiology.
602 2002;148:2331–42.

603 9. Lueders T, Wagner B, Claus P, Friedrich MW. Stable isotope probing of rRNA and DNA reveals a
604 dynamic methylotroph community and trophic interactions with fungi and protozoa in oxic rice field
605 soil. Environ Microbiol. 2004;6:60–72.

606 10. Manefield M, Whiteley AS, Griffiths RI, Bailey MJ. RNA stable isotope probing, a novel means of
607 linking microbial community function to phylogeny. Appl Environ Microbiol. 2002;68:5367–73.

608 11. Treonis AM, Ostle NJ, Stott AW, Primrose R, Grayston SJ, Ineson P. Identification of groups of
609 metabolically-active rhizosphere microorganisms by stable isotope probing of PLFAs. Soil Biol Biochem.
610 2004;36:533–7.

611 12. Boschker HTS, De Graaf W, Köster M, Meyer-Reil LA, Cappenberg TE. Bacterial populations and
612 processes involved in acetate and propionate consumption in anoxic brackish sediment. FEMS
613 Microbiol Ecol. 2001;35:97–103.

614 13. Jehmlich N, Schmidt F, Von Bergen M, Richnow HH, Vogt C. Protein-based stable isotope probing
615 (Protein-SIP) reveals active species within anoxic mixed cultures. ISME J. 2008;2:1122–33.

616 14. Jehmlich N, Schmidt F, Hartwich M, von Bergen M, Richnow H-H, Vogt C. Incorporation of carbon
617 and nitrogen atoms into proteins measured by protein-based stable isotope probing (Protein-SIP).
618 Rapid Commun Mass Spectrom. 2008;22:2889–97.

619 15. Justice NB, Li Z, Wang Y, Spaudling SE, Mosier AC, Hettich RL, et al. ^{15}N- and ^{2}H proteomic stable
620 isotope probing links nitrogen flow to archaeal heterotrophic activity. Environ Microbiol.

621 2014;16:3224–37.

622 16. Berry D, Mader E, Lee TK, Woebken D, Wang Y, Zhu D, et al. Tracking heavy water (D_2O)
623 incorporation for identifying and sorting active microbial cells. Proc Natl Acad Sci U S A.
624 2015;112:E194–203.

625 17. Taubert M, Stöckel S, Geesink P, Girnus S, Jehmlich N, von Bergen M, et al. Tracking active
626 groundwater microbes with D_2O labelling to understand their ecosystem function. Environ Microbiol.
627 2018;20:369–84.

628 18. Starke R, Oliphant K, Jehmlich N, Schäpe SS, Sachsenberg T, Kohlbacher O, et al. Tracing
629 incorporation of heavy water into proteins for species-specific metabolic activity in complex
630 communities. J Proteomics. 2020;222:103791.

631 19. Pfeffer M, Maurer M, Köllensperger G, Hann S, Graf AB, Mattanovich D. Modeling and measuring
632 intracellular fluxes of secreted recombinant protein in *Pichia pastoris* with a novel ^{34}S labeling
633 procedure. Microb Cell Fact. 2011;10:47.

634 20. Kümmel S, Herbst FA, Bahr A, Duarte M, Pieper DH, Jehmlich N, et al. Anaerobic naphthalene
635 degradation by sulfate-reducing *Desulfobacteraceae* from various anoxic aquifers. FEMS Microbiol
636 Ecol. 2015;91:fiv006.

637 21. Ouyang WY, Su JQ, Richnow HH, Adrian L. Identification of dominant sulfamethoxazole-degraders
638 in pig farm-impacted soil by DNA and protein stable isotope probing. Environ Int. 2019;126
639 February:118–26.

640 22. Nowak KM, Miltner A, Gehre M, Schäffer A, Kästner M. Formation and fate of bound residues from
641 microbial biomass during 2, 4-D degradation in soil. Environ Sci Technol. 2011;45:999–1006.

642 23. Grob C, Taubert M, Howat AM, Burns OJ, Dixon JL, Richnow HH, et al. Combining metagenomics
643 with metaproteomics and stable isotope probing reveals metabolic pathways used by a naturally
644 occurring marine methylotroph. Environ Microbiol. 2015;17:4007–18.

645 24. Bastida F, Jechalke S, Bombach P, Franchini AG, Seifert J, von Bergen M, et al. Assimilation of
646 benzene carbon through multiple trophic levels traced by different stable isotope probing
647 methodologies. FEMS Microbiol Ecol. 2011;77:357–69.

648 25. Slysz GW, Steinke L, Ward DM, Klatt CG, Clauss TRW, Purvine SO, et al. Automated data extraction
649 from in situ protein-stable isotope probing studies. J Proteome Res. 2014;13:1200–10.

650 26. Sachsenberg T, Herbst FA, Taubert M, Kermer R, Jehmlich N, Von Bergen M, et al. MetaProSIP:
651 automated inference of stable isotope incorporation rates in proteins for functional metaproteomics.
652 J Proteome Res. 2015;14:619–27.

653 27. Xiong Y, Mueller RS, Feng S, Guo X, Pan C. Proteomic stable isotope probing with an upgraded
654 Sipros algorithm for improved identification and quantification of isotopically labeled proteins.
655 Microbiome. 2024;12:148.

656 28. Zhang X, Ning Z, Mayne J, Moore JI, Li J, Butcher J, et al. MetaPro-IQ: a universal metaproteomic
657 approach to studying human and mouse gut microbiota. Microbiome. 2016;4:1–12.

658 29. Mise K, Iwasaki W. Unexpected absence of ribosomal protein genes from metagenome-assembled
659 genomes. ISME Commun. 2022;2:118.

660 30. Rocha U, Kasmanas JC, Toscan R, Sanches DS, Magnusdottir S, Saraiva JP. Simulation of 69 microbial
661 communities indicates sequencing depth and false positives are major drivers of bias in prokaryotic
662 metagenome-assembled genome recovery. PLOS Comput Biol. 2024;20:e1012530.

663 31. Jechalke S, Franchini AG, Bastida F, Bombach P, Rosell M, Seifert J, et al. Analysis of structure,
664 function, and activity of a benzene-degrading microbial community. FEMS Microbiol Ecol. 2013;85:14–
665 26.

666 32. Starke R, Kermer R, Ullmann-Zeunert L, Baldwin IT, Seifert J, Bastida F, et al. Bacteria dominate the
667 short-term assimilation of plant-derived N in soil. Soil Biol Biochem. 2016;96:30–8.

33. Herbst FA, Bahr A, Duarte M, Pieper DH, Richnow HH, von Bergen M, et al. Elucidation of in situ polycyclic aromatic hydrocarbon degradation by functional metaproteomics (protein-SIP). Proteomics. 2013;13:2910–20.

34. Bozinovski D, Herrmann S, Richnow HH, Von Bergen M, Seifert J, Vogt C. Functional analysis of an anaerobic *m*-xylene-degrading enrichment culture using protein-based stable isotope probing. FEMS Microbiol Ecol. 2012;81:134–44.

35. Blakeley-Ruiz JA, Kleiner M. Considerations for constructing a protein sequence database for metaproteomics. Computational and Structural Biotechnology Journal. 2022;20:937–52.

36. Kumar D, Yadav AK, Dash D. Choosing an optimal database for protein identification from tandem mass spectrometry data. In: Keerthikumar S, Mathivanan S, editors. Proteome Bioinformatics. New York, NY: Springer New York; 2017. p. 17–29.

37. Knudsen GM, Chalkley RJ. The effect of using an inappropriate protein database for proteomic data analysis. PLoS One. 2011;6:e20873-.

38. Tanca A, Palomba A, Deligios M, Cubeddu T, Fraumene C, Biosa G, et al. Evaluating the impact of different sequence databases on metaproteome analysis: insights from a lab-assembled microbial mixture. PLoS One. 2013;8:e82981.

39. Klaes S, Madan S, Deobald D, Cooper M, Adrian L. Revealing taxonomy, activity, and substrate assimilation in mixed bacterial communities by GroEL-proteotyping-based stable isotope probing. iScience. 2024;27:111249.

40. Muth T, Kolmeder CA, Salojärvi J, Keskitalo S, Varjosalo M, Verdam FJ, et al. Navigating through metaproteomics data: a logbook of database searching. Proteomics. 2015;15:3439–53.

41. Mesuere B, Devreese B, Debyser G, Aerts M, Vandamme P, Dawyndt P. Unipept: Tryptic peptide-based biodiversity analysis of metaproteome samples. J Proteome Res. 2012;11:5773–80.

42. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.

43. Kleikamp HBC, Pronk M, Tugui C, Guedes da Silva L, Abbas B, Lin YM, et al. Database-independent de novo metaproteomics of complex microbial communities. Cell Syst. 2021;12:375-383.e5.

44. Johnson RS, Searle BC, Nunn BL, Gilmore JM, Phillips M, Amemiya CT, et al. Assessing protein sequence database suitability using de novo sequencing. Mol Cell Proteomics. 2020;19:198–208.

45. Muth T, Renard BY. Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? Brief Bioinform. 2018;19:954–70.

46. Yilmaz M, Fondrie WE, Bittremieux W, Melendez CF, Nelson R, Ananth V, et al. Sequence-to-sequence translation from mass spectra to peptides with a transformer model. Nat Commun. 2024;15:6427.

47. Liu K, Ye Y, Li S, Tang H. Accurate de novo peptide sequencing using fully convolutional neural networks. Nat Commun. 2023;14:7974.

48. Kleiner M, Kouris A, Violette M, D'Angelo G, Liu Y, Korenek A, et al. Ultra-sensitive isotope probing to quantify activity and substrate assimilation in microbiomes. Microbiome. 2023;11:24.

49. Kleiner M, Thorson E, Sharp CE, Dong X, Liu D, Li C, et al. Assessing species biomass contributions in microbial communities via metaproteomics. Nat Commun. 2017;8:1588.

50. Masao F, Kenichi K, Ryuji K. Phylogenetic analysis of *Salmonella*, *Shigella*, and *Escherichia coli* strains on the basis of the *gyrB* gene sequence. J Clin Microbiol. 2002;40:2779–85.

51. Ding C, Adrian L. Comparative genomics in '*Candidatus* Kuenenia stuttgartiensis' reveal high genomic plasticity in the overall genome structure, CRISPR loci and surface proteins. BMC Genomics. 2020;21:851.

52. Ding C, Enyi FO, Adrian L. Anaerobic ammonium oxidation (Anammox) with planktonic cells in a redox-stable semicontinuous stirred-tank reactor. Environ Sci Technol. 2018;52:5671–81.

715 53. Feng J-H, Chang Y-X, Li L-H, Li M-M, Huang Z-Y, Han N, et al. *Copranaerobaculum intestinale* gen.
716 nov., sp. nov., a novel anaerobic bacterium isolated from human faeces. Int J Syst Evol Microbiol.
717 2022;72.

718 54. Afrizal A, Hitch TCA, Viehof A, Treichel N, Riedel T, Abt B, et al. Anaerobic single-cell dispensing
719 facilitates the cultivation of human gut bacteria. Environ Microbiol. 2022;24:3861–81.

720 55. Chiang Y, Collins MJ. Orthrus: an AI-powered, cloud-ready, and open-source hybrid approach for
721 metaproteomics. bioRxiv. 2024;:2024.11.15.623814.

722 56. Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for
723 proteomics. Nat Commun. 2014;5:5277.

724 57. Lazear MR. Sage: an open-source tool for fast proteomics searching and quantification at scale. J
725 Proteome Res. 2023;22:3652–9.

726 58. Wu Q, Ning Z, Zhang A, Zhang X, Sun Z, Figeys D. MetaX: a peptide centric metaproteomic data
727 analysis platform using Operational Taxa-Functions (OTF). bioRxiv. 2024;:2024.04.19.590315.

728 59. Yang T, Ling T, Sun B, Liang Z, Xu F, Huang X, et al. Introducing π-HelixNovo for practical large-scale
729 de novo peptide sequencing. Brief Bioinform. 2024;25.

730 60. Petrovskiy D V, Nikolsky KS, Kulikova LI, Rudnev VR, Butkova T V, Malsagova KA, et al. PowerNovo:
731 de novo peptide sequencing via tandem mass spectrometry using an ensemble of transformer and
732 BERT models. Sci Rep. 2024;14:15000.

733 61. Lee S, Kim H. Bidirectional de novo peptide sequencing using a transformer model. PLoS Comput
734 Biol. 2024;20:e1011892.

735 62. Klaproth-Andrade D, Hingerl J, Bruns Y, Smith NH, Träuble J, Wilhelm M, et al. Deep learning-driven
736 fragment ion series classification enables highly precise and sensitive de novo peptide sequencing. Nat
737 Commun. 2024;15:151.

738 63. Ananth V, Sanders J, Yilmaz M, Wen B, Oh S, Noble WS. A learned score function improves the
739 power of mass spectrometry database search. Bioinformatics. 2024;40 Supplement_1:i410–7.

740 64. McDonnell K, Howley E, Abram F. The impact of noise and missing fragmentation cleavages on de
741 novo peptide identification algorithms. Comput Struct Biotechnol J. 2022;20:1402–12.

742 65. Lozano C, Pible O, Eschlimann M, Giraud M, Debroas S, Gaillard J-C, et al. Universal identification
743 of pathogenic viruses by liquid chromatography coupled to tandem mass spectrometry proteotyping.
744 Mol Cell Proteomics. 2024;:100822.

745 66. Frankenfield AM, Ni J, Ahmed M, Hao L. Protein contaminants matter: building universal protein
746 contaminant libraries for DDA and DIA proteomics. J Proteome Res. 2022;21:2104–13.

747 67. Sinitcyn P, Richards AL, Weatheritt RJ, Brademan DR, Marx H, Shishkova E, et al. Global detection
748 of human variants and isoforms by deep proteome sequencing. Nat Biotechnol. 2023;41:1776–86.

749 68. Faridi P, Li C, Ramarathinam SH, Vivian JP, Illing PT, Mifsud NA, et al. A subset of HLA-I peptides are
750 not genomically templated: evidence for cis- and trans-spliced peptide ligands. Sci Immunol.
751 2018;3:eaar3947.

752 69. Gattiker A, Bienvenut W V., Bairoch A, Gasteiger E. FindPept, a tool to identify unmatched masses
753 in peptide mass fingerprinting protein identification. Proteomics. 2002;2:1435–44.

754 70. Klaes S, Madan S, Deobald D, Cooper M, Adrian L. GroEL-proteotyping of bacterial communities
755 using tandem mass spectrometry. International Journal of Molecular Sciences. 2023;24:15692.

756 71. Vande Moortele T, Devlaminck B, Van de Vyver S, Van Den Bossche T, Martens L, Dawyndt P, et al.
757 Unipept 6.0: expanding metaproteomics analysis with support for missed cleavages, semi-tryptic and
758 non-tryptic peptides. bioRxiv. 2024;:2024.09.26.615136.

759 72. O'Bryon I, Jenson SC, Merkley ED. Flying blind, or just flying under the radar? The underappreciated
760 power of de novo methods of mass spectrometric peptide identification. Protein Sci. 2020;29:1864–
761 78.

762 73. Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid
763 proteomics tools development. Bioinformatics. 2008;24:2534–6.
764 74. Jehmlich N, von Bergen M. Protein-based stable isotope probing (Protein-SIP): applications for
765 studying aromatic hydrocarbon degradation in microbial communities. In: Anaerobic utilization of
766 hydrocarbons, oils, and lipids. Cham: Springer International Publishing; 2020. p. 277–84.
767 75. Röst HL, Sachsenberg T, Aiche S, Bielow C, Weisser H, Aicheler F, et al. OpenMS: a flexible open-
768 source software platform for mass spectrometry data analysis. Nat Methods. 2016;13:741–8.
769 76. The Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative
770 biomedical analyses: 2022 update. Nucleic Acids Res. 2022;50:W345–51.
771 77. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the
772 unification of biology. Nat Genet. 2000;25:25–9.
773 78. Verschaffelt P, Tanca A, Abbondio M, Van Den Bossche T, Moortele T Vande, Dawyndt P, et al.
774 Unipept Desktop 2.0: construction of targeted reference protein databases for metaproteogenomics
775 analyses. J Proteome Res. 2023;22:2620–8.
776 79. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res.
777 2023;51:D523–31.
778 80. Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, Ebert D, et al. The gene ontology
779 knowledgebase in 2023. Genetics. 2023;224:iyad031.
780 81. Klopfenstein D V., Zhang L, Pedersen BS, Ramírez F, Vesztrocy AW, Naldi A, et al. GOATOOLS: a
781 Python library for gene ontology analyses. Sci Rep. 2018;8:10872.
782 82. Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, et al. High-throughput
783 functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 2008;36:3420–35.
784 83. Perez-Riverol Y, Bai J, Bandla C, García-Seisdedos D, Hewapathirana S, Kamatchinathan S, et al. The
785 PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. Nucleic
786 Acids Res. 2022;50:D543–52.
787 84. Serrano G, Guruceaga E, Segura V. DeepMSPeptide: peptide detectability prediction using deep
788 learning. Bioinformatics. 2020;36:1279–80.
789
790

Acknowledgements

Proteomics raw data published in this study was acquired at UFZ at the Centre for Chemical Microscopy (ProVIS), which is supported by the European Regional Development Fund (EFRE) and the Helmholtz Association. *De novo* sequencing was performed at the High-Performance Computing (HPC) Cluster EVE, a joint effort of both the Helmholtz Centre for Environmental Research – UFZ (http://www.ufz.de/) and the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig (http://www.idiv-biodiversity.de/). We would like to thank Thomas Schnicke, Ben Langenberg, Guido Schramm, Toni Harzendorf, Tom Strempel, Lisa Schurack, and Christian Krause for maintaining EVE as well as Benjamin Scheer for technical assistance in protein mass spectrometry. We acknowledge the assistance of generative AI (ChatGPT and Claude) in code development and troubleshooting, with all code subsequently reviewed and verified by the authors. Figure 1 was created in BioRender (https://BioRender.com/o64a634). S.K. was funded by the German Research Foundation (DFG) grant number GRK 2032/2. C.W. received funding from the Erasmus+ ICM (KA171) program of the European Commission. Open Access funding was enabled by Project DEAL.

Author contributions

Conceptualization: S.K.; Data curation: S.K. and C.D.; Formal analysis: S.K. and C.D.; Funding acquisition: L.A.-C. and L.A.; Investigation: S.K. and C.W.; Methodology: S.K. and C.D.; Project administration: S.K. and C.D.; Resources: C.W., L.A.-C., and L.A..; Software: S.K. and C.D.; Supervision: C.D.; Validation: S.K. and C.D.; Visualization: S.K.; Writing – original draft: S.K.; Writing – review & editing: S.K., C.W., L.A.-C., L.A. and C.D.

Competing financial interests

The authors declare that they have no competing financial interest.

814 # Supplementary Material

815 Supplementary Table S1: Sizes of *de novo* assembled databases used in this study.

Non-redundant *de novo* identified peptides						
Dataset identifier	**PXD024174**		**AMX**		**PXD024291**	
Score threshold	Casanovo	PepNet	Casanovo	PepNet	Casanovo	PepNet
0	290,680	875,366	256,559	303,595	512,680	715,031
0.2	149,626	258,054	82,754	33,867	250,323	225,748
0.4	148,847	175,840	74,423	23,321	247,553	170,283
0.6	139,704	119,472	54,101	16,636	232,755	128,653
0.8	96,835	80,093	29,304	12,001	180,762	97,990
0.85	76,673	71,180	22,380	10,960	153,699	90,692
0.9	53,981	61,634	15,785	9,819	120,752	82,718
0.95	31,873	50,450	9,934	8,473	84,176	72,797
0.99	14,700	35,375	5,313	6,605	48,715	58,476

816

817 Supplementary Table S2: Drop-off data from *Escherichia coli*.

Labeled peptides	**Total submissions**	**Root**	**Superkingdom**	**Phylum**	**Class**	**Order**	**Family**	**Genus**	**Species**
in silico	3,499	3,499	1,344	1,123	1,071	779	659	44	29
main	450	419	150	120	112	71	53	3	3
others			0	1	7	9	9	9	7
random_main			8	3	1	0	0	0	0
random_others			5	7	8	10	9	9	7
main-random_main			142	117	111	71	53	3	3
others-random_others			-5	-6	-1	-1	0	0	0
purity [%]			103.6	105.4	100.9	101.4	100.0	100.0	100.0
main-random_main normalised			100.0	82.4	78.2	50.0	37.3	2.1	2.1
in silico normalised			100.0	83.6	79.7	58.0	49.0	3.3	2.2
drop off delta [%]				1.4	1.9	15.9	31.4	55.0	2.1

818

819

820 Supplementary Table S3: Comparison of the number of detected target peptides in unmixed and mixed
821 samples of case study 1.

x hydraulic retention time		**0**	**1**		**5.25**	
			unmixed	**mixed**	**unmixed**	**mixed**
Peptides identified by MetaProSIP with genome-based database		5,830	3,088	4,449	130	3,329
Target peptides identified by *de novo* sequencing	Casanovo	6,695	3,628	5,120	322	3,713
	PepNet	4,336	2,035	3,171	97	2,131

822

823 Supplementary Table S4: Drop-off data from '*Candidatus* Kuenenia stuttgartiensis'.

Labeled peptides	**Total submissions**	**Root**	**Domain**	**Phylum**	**Class**	**Order**	**Family**	**Genus**	**Species**
in silico	3,500	3,500	2,544	2,326	2,312	2,312	2,258	2,081	2,081
main	1,232	1,173	847	768	765	765	745	635	635
others		0	0	1	1	1	1	1	1
random_main		314	33	0	0	0	0	0	0
random_others		0	24	43	38	37	37	38	38
main-random_main		859	814	768	765	765	745	635	635
others-random_others		0	-24	-42	-37	-36	-36	-37	-37
purity [%]		100.0	103.0	105.8	105.1	104.9	105.1	106.2	106.2
main-random_main normalised			100.0	94.3	94.0	94.0	91.5	78.0	78.0
in silico normalised			100.0	91.4	90.9	90.9	88.8	81.8	81.8
drop off delta [%]			0.0	3.1	3.3	3.3	3.0	4.9	4.9

824

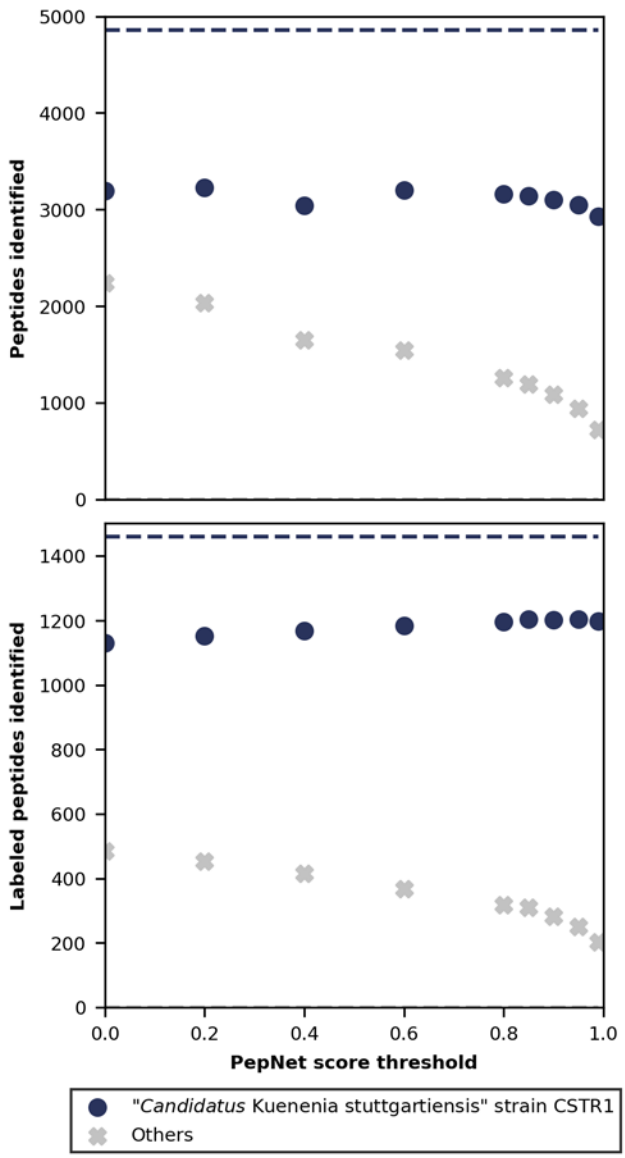
825 Supplementary Table S5. Relative isotope abundance of peptides assigned to gene ontology (GO) terms
826 of biological processes related to subterms of amino acid metabolic processes (GO:0006520),
827 carbohydrate derivative metabolic processes (GO:1901135), and carbohydrate metabolic processes
828 (GO:0005975). P-values were calculated by Student's t-test for the means of two independent samples.

829

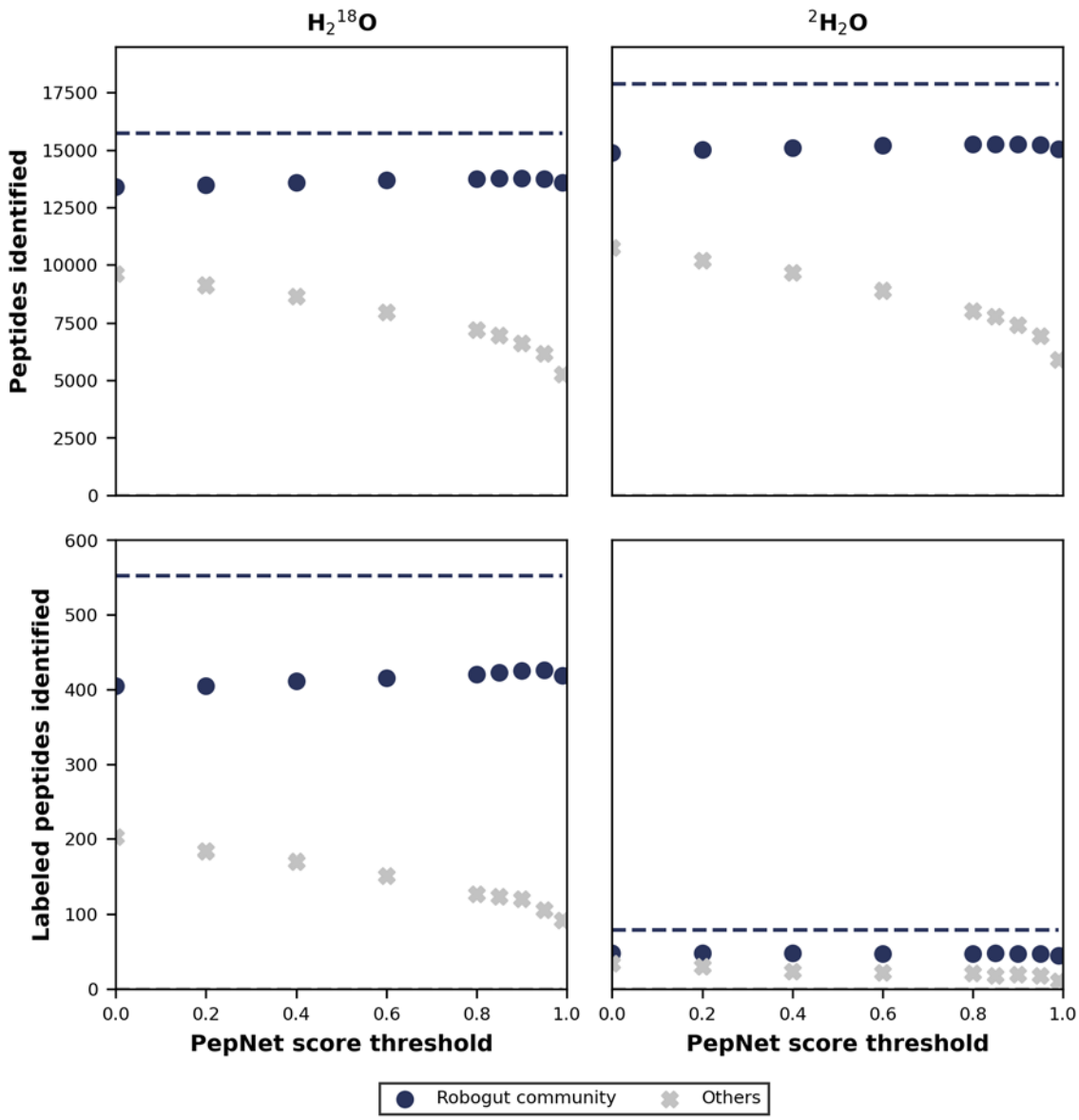
830 Supplementary Table S6: Overview of analyzed datasets

Dataset	Identifier	Files analyzed by MetaProSIP workflow	Analysis deposited to
Standard *Escherichia coli* K12 cultures cultivated with 1.07–99% ^{13}C [27]	PXD041414	none	https://zenodo.org/records/14030621?token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6ImE3YmQzOTE1LTkwMmUtNDUxOS04M2I1LTc3NjQyMTBkNGFjNSIsImRhdGEiOnt9LCJyYW5kb20iOiIyM2NkYjRmMmEyZDAxZWFmZjAxMzcwNzViNDExZGZiZCJ9.P7_Y1CciVFvJLeMB_8oFEVkOAbEMy5W5fE4Lz_9KoKuc8p8-mqYh3Epw2HkK2SwmwuFKwHNcUii7VXFxEZCG9A
Mock community with spike-in of ^{13}C-labeled *Escherichia coli* [48]	PXD024174	Run1_MockU2_EcoliR1_10_2000ng.raw Run1_MockU2_EcoliR2_10_2000ng.raw Run1_MockU2_EcoliR3_10_2000ng.raw	https://zenodo.org/records/14030389?token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6IjU3NDVlYWRkLTE3ZmYtNDVhOC05OTU1LWQ2ZTQxYWIyNTMzOSIsImRhdGEiOnt9LCJyYW5kb20iOiI3ZGQ3NjAxZDE4MjYwYzZkMjhjOTlhNzJjMTBkMGU3ZiJ9.dkyBgYoi3K4yZOJZCKCCy0FS1EFw5HsiedYpqDnKQXMebDMwRhQppR5cnVqelvFEknwAddDKNlaurQRBc2VwVA
'*Candidatus* Kuenenia stuttgartiensis' strain CSTR1 reactor	PXD057215	01_6_27_12C_01.raw 02_6_27_12C_02.raw 03_6_27_12C_03.raw 19_7_01_mix_01.raw 20_7_01_mix_02.raw 21_7_01_mix_03.raw 22_7_18_mix_01.raw 23_7_18_mix_02.raw 24_7_18_mix_03.raw	https://zenodo.org/records/14030687?token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6IjdmMGRiMGQ1LWJkMGMtNDBjOC1hOGFiLTZmODM2ZGU0ODE1YyIsImRhdGEiOnt9LCJyYW5kb20iOiI4Y2M3ZGViMWZjZGRiMjRhMWY2NDIzOGYzNzExNWYwNiJ9.NULxsUIzIhoxfapIhgyv-OaY5TPKPsXHpDbQeaEh98Dn84KhFK4nrInCFpLe0x5jZmgMJ9lxzHxYOmTc13YW3w
Model of the human distal gut microbiome [18, 48]	PXD024291	2017-02-07_RS_Robo_9.raw 2017-02-07_RS_Robo_10.raw 2017-02-07_RS_Robo_11.raw 2017-02-07_RS_Robo_12.raw 2017-02-07_RS_Robo_13.raw 2017-02-07_RS_Robo_14.raw 2017-02-07_RS_Robo_15.raw 2017-02-07_RS_Robo_16.raw 2017-02-07_RS_Robo_17.raw 2017-02-07_RS_Robo_18.raw 2017-02-07_RS_Robo_19.raw 2017-02-07_RS_Robo_20.raw	https://zenodo.org/records/14030664?token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6ImZlNWU0MDM4LTRkMjEtNDYxNC1hN2NkLWYzZDY1YjYwMDZiZSIsImRhdGEiOnt9LCJyYW5kb20iOiIyYmY4ZjgwODY5NDY5ZGRkMmI3OTNjZTRlNjdmOTlkYiJ9.CupCCNWTKUZ8ekCcyxHXkmuMNedLw75j0-LJm5ljf3cs24WbrQI38qbFXCy4m7CmadsWMuHIG17_lCvqWzRVxw

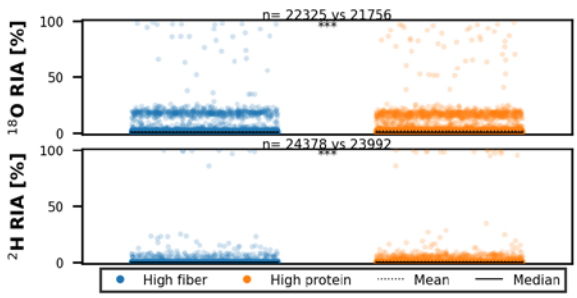
831



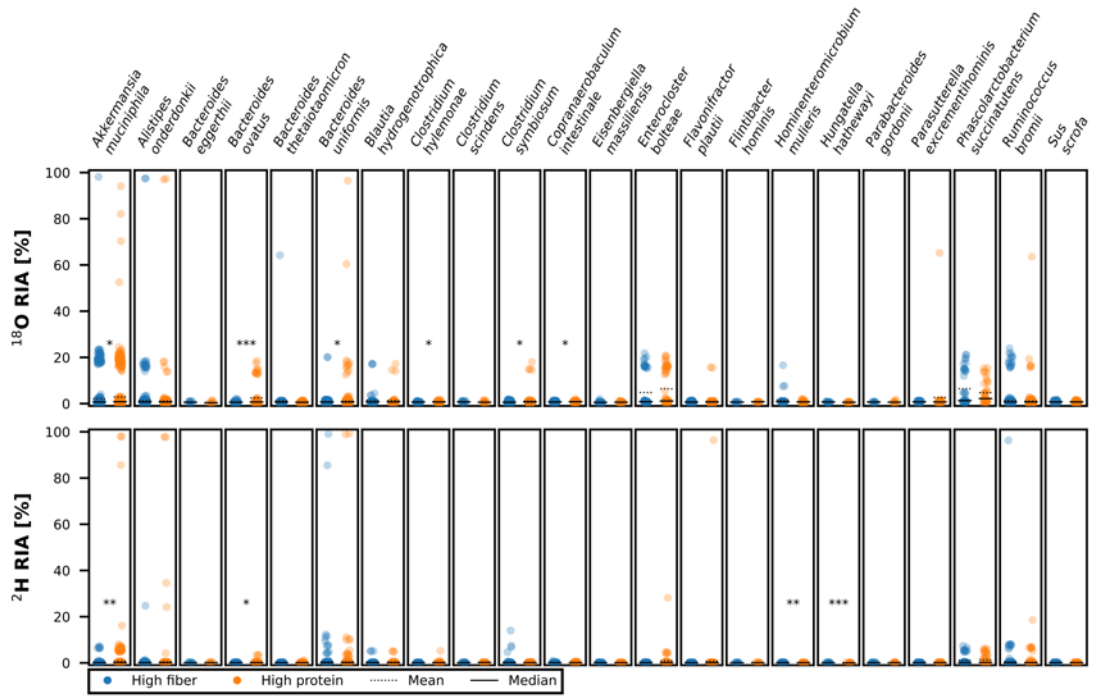
Supplementary Figure S1: Comparing *de novo* assembled databases for identifying unlabeled and ^{13}C-labeled peptides with a MetaProSIP workflow in a '*Candidatus* Kuenenia stuttgartiensis' strain CSTR1 enrichment from a continuous stir tank reactor fed with ^{13}C-bicarbonate. *De novo* databases were assembled using PepNet and filtering identifications with different thresholds for the quality score. Navy circles denote the number of identified peptides present in the proteome of '*Ca.* Kuenenia stuttgartiensis' + contaminants. Grey crosses denote the number of identified peptides absent in the proteome of '*Ca.* Kuenenia stuttgartiensis' + contaminants. Dashed lines represent the number of identifications using the sample-matching reference database derived from metagenome sequencing [51]. Only peptides detected in at least two biological replicates are plotted.



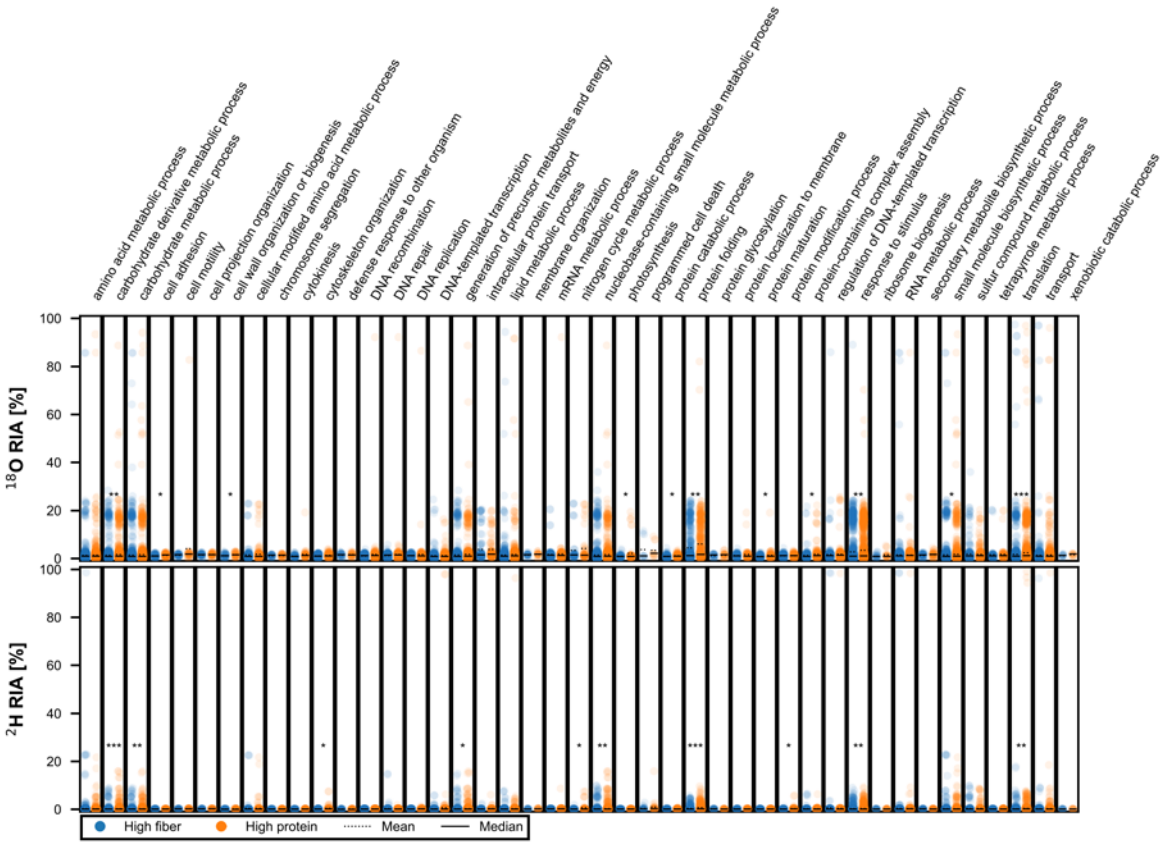
Supplementary Figure S2: Comparing *de novo* assembled databases for identifying unlabeled and $^2H/^{18}O$-labeled peptides with a MetaProSIP workflow in a model of the human gut microbiome (Robogut community) cultivated with ^2H_2O or $H_2{}^{18}O$, respectively, and obtained from [18]. *De novo* databases were assembled using PepNet and filtering identifications with different thresholds for the quality score. Navy circles denote the number of identified peptides present in the sample-matching reference proteome + contaminants. Grey crosses denote the number of identified peptides absent in the sample-matching reference proteome + contaminants. Dashed lines represent the number of identifications that we generated using the published sample-matching reference database. Only peptides detected in at least two biological replicates are plotted.



Supplementary Figure S3: Distribution of the ^{18}O and ^{2}H relative isotope abundance (RIA) in detected peptides identified in a model of the human gut microbiome cultivated in either a high fiber or high protein medium in the presence of either $H_2{}^{18}O$ or $^{2}H_2O$. RIA values are shown for peptides detected in at least two replicates of both media. 'n' denotes the number of peptides identified across triplicates. Statistically significant differences in RIA between media are indicated with '***' based on Student's t-test for the means of two independent samples with $p < 0.001$.



Supplementary Figure S4: Figure 8 without removal of outliers.



Supplementary Figure S5: Figure 9 without removal of outliers.