Uncovering hidden members and functions of the soil microbiome using *de novo* metaproteomics

Joon-Yong Lee[1], Hugh D. Mitchell[1], Meagan C. Burnet[1], Ruonan Wu[1], Sarah C. Jenson[2], Eric D. Merkley[2], Ernesto S. Nakayasu[1], Carrie D. Nicora[1], Janet K. Jansson[1*], Kristin E. Burnum-Johnson[3*], Samuel H. Payne[4, 5*]

1. Biological Sciences Division, Pacific Northwest National Laboratory, Richland WA
2. Signature Sciences and Technology Division, Pacific Northwest National Laboratory, Richland WA
3. Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland WA
4. Biology Department, Brigham Young University, Provo UT
5. Lead contact

* Correspondence: sam_payne@byu.edu, kristin.burnum-johnson@pnnl.gov, janet.jansson@pnnl.gov

Abstract

Metaproteomics has been increasingly utilized for high-throughput molecular characterization in complex environments and has been demonstrated to provide insights into microbial composition and functional roles in soil systems. Despite its potential for the study of microbiomes, significant challenges remain in data analysis, including the creation of a sample-specific protein sequence database as the taxonomic composition of soil is often unknown. Almost all metaproteome analysis tools require this database and their accuracy and sensitivity suffer when the database is incomplete or contains extraneous sequences from organisms which are not present. Here, we leverage a *de novo* peptide sequencing approach to identify sample composition directly from metaproteomic data. First, we created a deep learning model, Kaiko, to predict the peptide sequences from mass spectrometry data, and trained it on 5 million peptide-spectrum matches from 55 phylogenetically diverse bacteria. After training, Kaiko successfully identified unsequenced soil isolates directly from proteomics data. Finally, we created a pipeline for metaproteome database generation using Kaiko. We tested the pipeline on native soils collected in Kansas, showing that the *de novo* sequencing model can be employed to construct the sample-specific protein database instead of relying on (un)matched metagenomes. Our pipeline identified all highly abundant taxa from 16S ribosomal RNA sequencing of the soil samples and also uncovered several additional species which were strongly represented only in proteomic data. Our pipeline offers an alternative and complementary method for metaproteomic data analysis by creating a protein database directly from proteomic data, thus removing the need for metagenomic sequencing.

Significance Statement

Proteomic characterization of environmental samples, or metaproteomics, reveals microbial activity critical to our understanding of climate, nutrient cycling and human health. Metaproteomic samples originate from diverse environs, such as soil and oceans. One option for data analysis is a *de novo* interpretation of the mass spectra. Unfortunately, the current generation of *de novo* algorithms were primarily trained on data originating from human proteins. Therefore, these algorithms struggle with data from environmental samples, limiting our ability to analyze metaproteomics data. To address this challenge, we trained a new algorithm with data from dozens of diverse environmental bacteria and achieved significant improvements in accuracy across a broad range of organisms. This generality opens proteomics to the world of natural isolates and microbiomes.

Introduction

The soil microbiome is responsible for carrying out many functions that are important on a global scale, including cycling of carbon and other nutrients and support of plant growth. Over the last few decades high throughput sequencing technologies have made great strides in revealing the soil microbial community composition in a variety of soil habitats and how those communities are impacted by environmental change. Amplicon sequencing has revealed that soil and sediment microorganisms have a very high diversity; much more so than other ecosystems[1]. In addition, metagenome sequencing has proven to be an extremely useful tool for not only determining the composition of soil microbiomes, but also their putative functions. However, not all genes detected in a metagenome survey are actively expressed and significant challenges remain in understanding the biological functions that are carried out by active members of the soil microbiome. Other meta-omics technologies, such as metatranscriptomics and metaproteomics, have helped to close this current knowledge gap. Metatranscriptomics provides information on community transcription and is often used as a proxy for assigning metabolically active members of a soil microbiome. However, metatranscriptomics can only provide a snapshot of gene expression at the moment of sampling. A significant amount of post-transcriptional regulation affects protein abundance and activity[2]. Therefore, metaproteomics provides an essential layer of information about microbiome activity by revealing which proteins are actually produced and have passed transcriptional and translational regulation points.

Despite the promise of metaproteomics for elucidating functions of elusive soil microorganisms, significant challenges remain. An important assumption in most mass spectrometry proteomics identification algorithms is that the set of potential proteins is known, and thus a database of these protein sequences is a typical requirement[3,4]. In environmental samples, however, obtaining an accurate catalog of organisms and their proteins is a challenge, as it is not possible to know the organisms present in the sample beforehand. Amplicon and metagenome sequencing of a matched sample is often used to identify community membership; however, many species might not be observable by sequencing[5–10].

Here, we present a new method to generate a protein database directly from metaproteomic data as an alternative and orthogonal method of identifying soil microbe composition. The method starts with analyzing mass spectra *de novo* (without a database)[11], identifying species from the observed peptides and then gathering full proteomic databases for these species. As currently available software tools for *de novo* identification were not sufficiently accurate for environmental samples, we first trained a new deep learning model on spectra from 55 bacteria in nine phyla. After confirming that the new model could successfully identify natural soil isolates, we applied the model to metaproteomics samples. Using a metaproteomics dataset from Kansas soil, our pipeline identified all abundant taxa identified in traditional 16S data as well as identifying new abundant organisms in the soil. Using the identified organisms, we

re-analyzed the metaproteomics data and identified differential metabolic functions between species in the microbiome.

Results

A new model for *de novo* MS/MS identification

Using a large and environmentally diverse set of mass spectrometry proteomics data, we sought to improve on peptide/spectrum identification where no protein sequence database is available. We adapted a deep neural network structure[12] and trained a new model called Kaiko, after the Japanese deep ocean submersible used to explore the Marianas Trench. For training and validation, we used 4,604,540 spectra and 927,316 peptides from 51 distinct bacteria (Fig. 1A, Supplementary Table 1). Deep neural networks, like Kaiko, require very large training datasets for parameter optimization. For our neural network architecture, training events with less than 3 million spectra resulted in severely overfit models (Supplementary Figs. 1 and 2). After training and optimization, we evaluated the accuracy of Kaiko against spectra in the test dataset consisting of spectra from four additional organisms not used in model training (511,765 spectra and 90,048 peptides). Kaiko achieved an average accuracy of 33% over all testing files and organisms, a significant improvement over other *de novo* algorithms (Fig. 1B). When considering the top five spectrum annotations, average accuracy exceeded 41%.

We next looked at model performance as a function of peptide length (Fig. 1C). Most algorithms performed well with short peptides, length < 8. Unfortunately, these peptides are infrequent in bottom-up proteomics data samples (Supplementary Fig 3). Kaiko exhibited significantly improved accuracy at all lengths, but especially for the most common peptide lengths (10-15 residues), where it achieved an accuracy of ~30-60%. We note that Kaiko had high accuracy at very long peptide lengths of 15 and above. Although these peptides are extremely difficult to annotate *de novo*, they are valuable for predicting phylogeny as the long sequences are more likely to be uniquely mapped to a small taxonomy range.

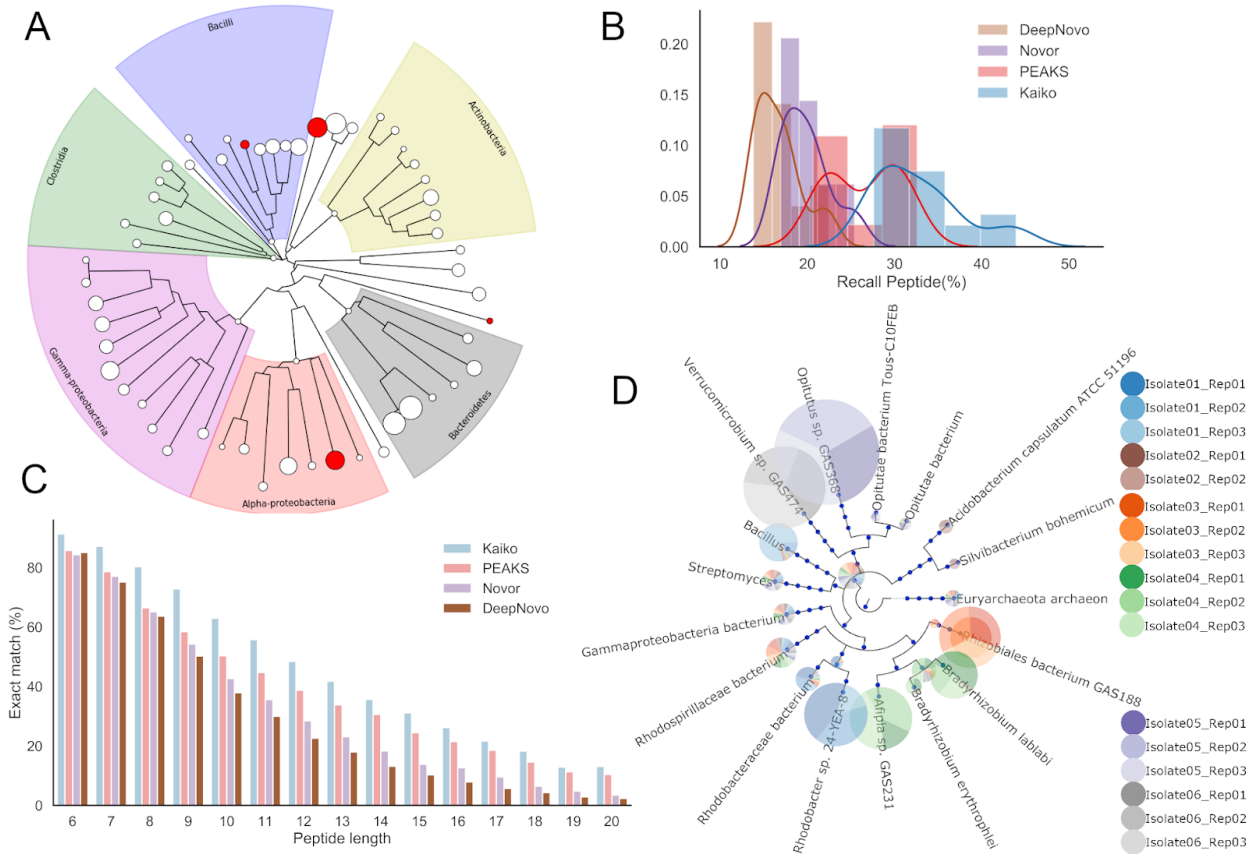


Figure 1. Training, validation and testing of a new *de novo* peptide identification algorithm. (A) Bacteria represented in training and testing data and shown in a phylogenetic tree built from the multiple sequence alignment of rplB is shown for all organisms in the training (white nodes) and testing datasets (red nodes). The size of the node is scaled to represent the number of spectra used. (B) Accuracy of spectrum annotation for four de novo spectrum annotation tools. (C) For each peptide sequence length, the accuracy of spectrum annotation is shown for each of the four algorithms. (D) For each of the six natural isolates, replicate proteomics data was annotated with Kaiko and identified peptides are visualized on a phylogenetic tree. The size of the pie wedge is scaled to represent the number of spectra matching that taxon. For each sample, the top 5 taxa according to the number of peptide hits was included in the visualization.

Identification of soil isolates via proteomics

Proteomics analysis of natural bacterial isolates from soil often requires *de novo* spectrum annotation. To demonstrate the ability of our deep learning-based algorithm to annotate spectra from an unknown organism and also to accurately identify the unknown organism, we obtained bottom-up proteomics data from six microbes isolated from soil and attempted to identify the sample. For each sample, we annotated the proteomics data with Kaiko and used DIAMOND[13]

to identify the closest sequences in the UniProt database[14] (see Methods). We then plotted the organisms which had the most matching spectra and inferred the organism for the sample.

For four samples, a matched proteome database became public during our investigation; however, this was still blinded from our analysis. In each of these cases, we identified the exact species as the source of the sample (Fig. 1D). This included two Verrucomicrobia for which Kaiko's training data had nothing in the same phylum: *Opitutus* sp. GAS368 and *Verrucomicrobium* sp. GAS474. The other two isolates with a matched genome were from the order Rhizobiales: *Afipia* sp. GAS231 and *Rhizobiales* bacterium GAS188. The *Afipia* sample also contained spectra which mapped to neighboring *Bradyrhizobium* species, which could be from shared gene content, contamination or previously unidentified co-culturing.

For two samples, there were no matched proteomes in UniProt and we attempted to derive the true sample identity by 16S sequencing. Isolate 02 cannot be definitively assigned to a genus within NCBI's taxonomy based on 16S sequencing, but is close to multiple genera within the family Acidobacteriaceae. Using Kaiko's peptide annotations, we identified two potential candidates for the sample: *Acidobacterium capsulatum* and *Silvibacterium bohemicum* (both Acidobacteriaceae). However, both species had significantly fewer peptide hits matching their proteome and therefore, were weaker matches than expected. This weak alignment to a single organism and splitting between organisms within the same family is consistent with the isolate's ambiguous taxonomic assignment. The final sample, Isolate 01, was suggested to be a *Gemmobacter* species by 16S sequencing. Peptide hits from Kaiko identified this sample as *Rhodobacter* sp. 24-YEA-8, which is within the same family as *Gemmobacter* (Rhodobacteraceae). With the difficulties surrounding bacterial taxonomic classification and the uncertainty of species designation, this is still a close match.

Building a protein database without metagenomics

In metaproteomics data analysis, constructing a protein sequence database is a critical component for protein identification[15,16], as identification sensitivity suffers as database size increases[17]. It is therefore essential to identify organisms present in a sample with taxonomic precision, so that databases include as few species as possible. We present a new solution that derives the organisms present in a sample directly from Kaiko's analysis of metaproteomics data, thus enabling metagenome-free peptide identification (Fig. 2).

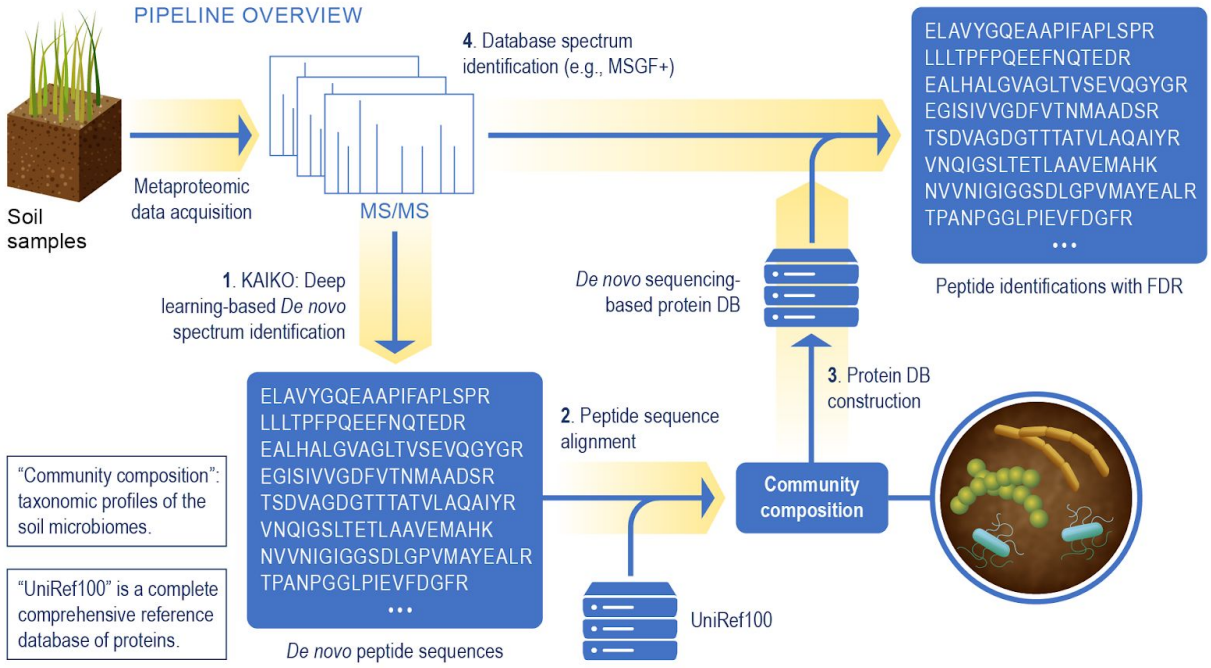


Figure 2. **Overview of the metaproteomics data analysis leveraging *de novo* spectrum identification based on the Kaiko model**. Peptides are identified using Kaiko, and used to infer community composition (steps 1-3). In step 4, the spectra are re-analyzed using a database search algorithm, e.g. MSGF+, and the protein sequence database created in step 3. This yields a final list of peptide identifications which can be used for functional analysis.

To demonstrate this *de novo*-based metaproteomics pipeline, we analyzed metaproteomic data acquired from pooled samples of native soils collected in three sites located in Kansas[18,19]. To identify species, the Kaiko model and DIAMOND were employed to determine the most dominant organisms, and whole proteomes were retrieved from UniProt (See Methods). 6,410 unique taxa IDs were identified in total and 224 taxa had more than 5 matched peptides. These taxa included well-known bacterial phylotypes consistently detected as a core component of soil ecosystem such as Proteobacteria, Actinobacteria, Acidobacteria, Planctomycetes, Chloroflexi, Verrucomicrobia, Bacteroidetes, Gemmatimonadetes, Firmicutes and Armatimonadetes[20,21]. In addition, our pipeline revealed globally abundant fungal classes such as Agaricomycetes, Sordariomycetes, Eurotiomycetes, Leotiomycetes and Mortierellomycetes[21,22].

Table 1. **Relative abundance of the top 20 bacterial phyla detected from 16S and Kaiko**. A dash in the table represents the corresponding phylum was 'not detected'. The asterisk(*) Indicates that some taxa in the corresponding phyla are used to construct the protein DB.

Phylum	Read counts By 16S	Peptide counts By Kaiko	Relative read counts % total reads at the phylum level	Relative Peptide counts % By Kaiko at the phylum level
Proteobacteria*	40778	4903	34.6	38.1
Actinobacteria*	16501	3949	14.0	30.7
Acidobacteria*	18562	1010	15.7	7.8
Firmicutes*	6761	634	5.7	4.9
Chloroflexi*	767	479	0.7	3.7
Bacteroidetes*	9712	467	8.2	3.6
Planctomycetes*	11427	321	9.7	2.5
Candidatus Rokubacteria*	-	266	-	2.1
Verrucomicrobia*	11841	237	10.0	1.8
Cyanobacteria	489	162	0.4	1.3
Gemmatimonadetes*	869	61	0.7	0.5
Nitrospirae*	18	44	-	0.3
Candidatus Tectomicrobia*	-	43	-	0.3
Deinococcus-Thermus	-	32	-	0.2
Spirochaetes	-	32	-	0.2
Elusimicrobia	-	15	-	0.1
Tenericutes	99	15	0.1	0.1
Armatimonadetes	75	13	0.1	0.1
Ignavibacteriae	16	6	0.01	0.05
Chlamydiae	2	4	0.00	0.03

To evaluate the taxa annotation from the Kaiko model, we also identified taxa using 16S rRNA data from the same samples (See Methods). 243 unique taxa IDs were determined for 3,693 OTUs. All of the highly abundant phyla detected by 16S were also detected by Kaiko (Table 1). Several phyla uniquely found by Kaiko are known to be present in environmental soils[23–28]. For example, Candidatus Rokubacteria is distributed globally in diverse terrestrial ecosystems, including soils and the rhizosphere[23] and Candidatus Tectomicrobia has also been detected in soils[24].

To construct the protein database from the identified organisms, we selected the 100 most abundant bacterial taxa, resulting in a protein database containing 17,448,135 protein clusters (UniRef sequences) from 12 bacterial phyla. We note that the 100 taxa identified by proteome data consist of 91 species, 1 genus, 7 strains, and the remaining 1 had no phylogenetic rank. Unfortunately, the 16S taxa annotations were often resolved only to a phylum or class level; relatively few taxa from 16S data were able to be narrowly identified at the level of genus or species. Creating a protein sequence database for all species within a broad taxonomic category, such as phylum, would dramatically increase the size of the protein sequence database and reduce the sensitivity of the proteomics data analysis.

Soil metaproteomic data analysis

Using the protein database generated by Kaiko, we re-analyzed the mass spectra from the soil samples using the database search tool MSGF+ and identified 30,762 unique peptides from 31,848 PSMs with 5% peptide FDR (see Methods). We performed functional annotations with these identified peptides using Unipept[29], and found 1,760 Enzyme Commission (EC) numbers matched to 11,646 peptides (42%). Functions in the top 20 EC numbers (Supplementary Table 2) included various enzymatic functions for transcription and translation, energy production and signaling. 787 EC numbers were mapped to KEGG metabolic pathways, extensively covering carbohydrate and amino acid metabolism, as well as the metabolism of cofactors, vitamins and xenobiotics.

Among identified peptides, 14,028 peptides were highly conserved sequences and therefore were assigned to bacterial phyla. 3,228 of these phyla-affiliated peptides were linked to 708 EC numbers (Supplementary Figure 4). These highly conserved peptides were assigned to ubiquitous bacterial functions commonly detected across most phyla, such as DNA-directed RNA polymerase (EC:2.7.7.6) and H(+)-transporting two-sector ATPase (EC:7.1.2.2), With NAD(+) or NADP(+) as acceptor (EC:1.2.1.-), Acting on ATP (EC:3.6.4.-), Protein-synthesizing GTPase (EC:3.6.4.-). In particular, EC numbers of highly-ranked peptide counts were mainly detected in abundant phyla (Proteobacteria, Actinobacteria, and Acidobacteria) and functional information was biased by the common and abundant proteins.

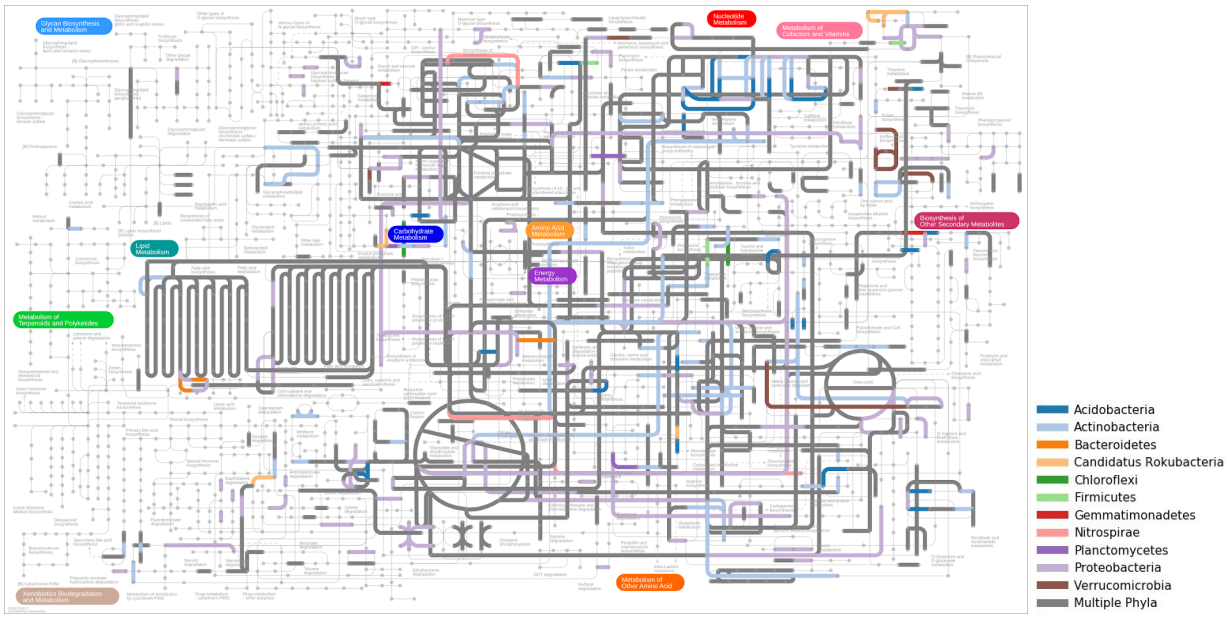


Figure 3. Distribution of bacterial functions in the metabolic pathway map. Several metabolic steps are shared among multiple phyla (dark grey). Other colors indicate unique EC numbers and their associated metabolic function found only in a specific phylum.

We next examined the mapped EC numbers to identify metabolic functions for specific taxa (Fig. 3). By mapping the taxonomic affiliation of the enzymatic reactions within metabolic pathways it was possible to determine which metabolic pathways were shared or unique among the represented phyla. EC numbers involved in carbon metabolism were often found in organisms from multiple phyla, and represent basic functions from glycolysis, carbon fixation, the TCA cycle, etc. Enzymes and metabolic functions for 427 EC numbers were represented by only a single phylum, and are shown with different colors in Figure 3. The two most abundant phyla detected in the metaproteomics data were Proteobacteria and Actinobacteria. It is clear from the functional mapping of peptides that these two phyla utilize distinct metabolic routes. For example, purine metabolism contains numerous enzymes which are exclusively found in either Actinobacteria or Proteobacteria (Supplementary Fig. 5). Significant divergence between these two dominant taxa was also seen in enzymes related to amino acid metabolism. Unique metabolic capacity is also observed for organisms with less proteomic coverage. Despite having relatively few identified peptides, Verrucomicrobia were the only species with enzymes for folate metabolism.

Finally, we examined the peptides and biological functions associated with species unique to the Kaiko database, i.e. species not found in the 16S rRNA sequences. 266 peptides were identified in Candidatus Rokubacteria and mapped to EC numbers. Biological functions associated with six EC numbers were exclusive to Candidatus Rokubacteria; 4-hydroxy-tetrahydrodipicolinate reductase (EC:1.17.1.8, lysine biosynthesis and monobactam biosynthesis), pyrroloquinoline-quinone synthase (EC:1.3.3.11), thioredoxin-disulfide reductase (EC:1.8.1.9, selenocompound metabolism), 3-oxoadipate enol-lactonase (EC:3.1.1.24,

benzoate degradation), inositol-phosphate phosphatase (EC:3.1.3.25, inositol phosphate metabolism and streptomycin biosynthesis), and aminopyrimidine aminohydrolase (EC:3.5.99.2, thiamine metabolism).

Discussion

Although genome and metagenome sequencing have greatly expanded the number of species that contain a sequenced genome and therefore an annotated proteome, there are still significant practical and financial barriers that prevent labs from always having an assembled and well-annotated genome for samples taken from nature. Yet metaproteome spectrum identification tools rely on a protein sequence database. Therefore tools which can create a proteome database for environmental samples without requiring sequencing data are a significant benefit to the microbiome community. One option for creating a proteome database without using sequencing data utilizes a *de novo* interpretation of metaproteomics data to identify organisms present in the sample. A significant drawback of current *de novo* tools is their poor performance on spectra from diverse organisms (see Figure 1). Algorithms which are only exposed to a limited number of organisms[11,12], or those that focus only on human data[30], will be inadequate when faced with the vast sequence diversity of microbial proteins found in soil and environmental samples.

To assist in the analysis of metaproteomic data, we have created a pipeline for generating the proteome sequence database directly from the metaproteomic data. A key element in our pipeline is a new *de novo* spectrum annotation tool, Kaiko, which has significantly improved accuracy compared to other *de novo* algorithms. This improvement comes from a deliberate focus on training the algorithm with mass spectrometry data from dozens of diverse environmental bacteria. Moreover, our training dataset size is dramatically larger than comparable *de novo* tools in terms of the number of peptides and spectra, which was essential for overcoming an overfit model. We evaluated Kaiko by using it to identify the taxonomy of bacterial soil isolates, including samples from phyla where no training data existed. Thus, it is better equipped for evaluating metaproteomics data where identifying spectra from diverse organisms is essential.

When using Kaiko as part of our database generation pipeline to identify soil community composition, we were able to identify all abundant species from 16S data, and also new species with significant proteomic evidence which were not seen in the sequencing data. Indeed, five of the top sixteen taxa (>30%) identified in the metaproteomics data were not identified in sequencing data. These ‘hidden microbes’ represent bacteria that are known to play an important role in community metabolism and function[23], including secondary metabolite biosynthesis[31,32] as was seen in our Candidatus Rokubacteria data. We also note that the metaproteomics pipeline was able to identify fungi in the soil, which are entirely absent in 16S data.

A second significant advantage of inferring community composition directly from metaproteome data is the level of taxon specificity. Using metaproteome data, we could narrow taxon identification to species or strain (98%). However taxa identified using 16S data for these same samples frequently were only able to distinguish broad taxonomic levels. Unfortunately, spectrum identification algorithms generally suffer a significant sensitivity loss when working with large protein databases[17]. Therefore, methods which specify community composition in broad taxonomic terms will yield poor results, compared to a method which is able to narrowly define organisms present in the community.

As metaproteomics data analysis continues to mature, progress will happen in multiple areas, e.g. more sensitive peptide ID algorithms, improved protein inference for multi-organism mapped peptides and functional analysis of pathways with multiple participating organisms. But a central feature in all of this work is the original identification of spectra, and currently the best algorithms require a protein database. Thus the creation of a protein sequence database is a pivotal step in metaproteomics data analysis. The most important future improvement in creating a protein sequence database will come from greater coverage and greater specificity in the identification of community membership. *De novo* proteomics offers one avenue for this, which is independent of advances made in sequencing technologies. Improving the accuracy of *de novo* tools, especially with regard to diverse environmental sequences, will be a significant benefit to metaproteomics.

Acknowledgements

The authors thank Court Corley and Nathan Hodas (PNNL) for insightful discussions. We thank Kristen DeAngelis and Grace Pold (University of Massachusetts Amherst) for natural isolate samples. Funding for this project was provided by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Early Career Research Program (to SHP and KEBJ) and PNNL's Deep Learning for Scientific Discovery initiative. Proteomics data used in this manuscript were generated in the Environmental Molecular Science Laboratory, a DOE national scientific user facility at Pacific Northwest National Laboratory (PNNL) in Richland, WA. Battelle operates PNNL for the DOE under contract DE-AC05-76RLO01830.

Methods

Data generation for Kaiko

Cell culture and sample preparation. The growth, sample preparation and data collection was reported previously[33]. Cells were harvested by centrifuging at 3,500 x *g* for 5 min at room temperature and washed twice with 5 mL PBS by centrifuging at the same conditions. Cells were lysed in a Bullet Blender (Next Advance) for 4 minutes at speed 8 in 200 µL of 100 mM ammonium bicarbonate (NH_4HCO_3) and approximately 100 µL 0.1 mm zirconia/silica beads at

4° C. Lysates were transferred into clean tubes and the remaining beads were washed with 200 µL of 100 mM NH_4HCO_3. The supernatants from the washing step were collected and combined with the cell lysate. Resulting protein extract was assayed by bicinchoninic acid (BCA) assay (Thermo Fisher Scientific, San Jose, CA) following manufacturer instructions. Aliquots of 300 µg of proteins were denatured and reduced using 8M urea and 5 mM DTT, and incubated at 60° C for 30 min with 850 rpm shaking. Samples were then diluted 10 fold in 100 mM NH_4HCO_3 and $CaCl_2$ was added to a final concentration of 1 mM using a 1M stock. Trypsin was added at 1/50 of the protein concentration and the digestion was carried out for 3 h at 37° C. Digestion products were desalted in 1-mL C18 cartridges (50 mg beads, Strata, Phenomenex). Cartridges were activated with 3 mL of methanol and equilibrated with 2 mL of 0.1% TFA before loading the samples. After sample loading, the cartridges were washed with 4 mL of 5% acetonitrile (ACN)/0.1% TFA and peptides were eluted with 1 mL of 80% ACN/0.1% TFA. Peptides were dried in a vacuum centrifuge, resuspended in water and assayed using a BCA assay. Peptide concentrations were normalized to 0.1 µg/µL before randomization and analysis by liquid chromatography-tandem mass spectrometry (LC-MS/MS).

LC-MS/MS data acquisition. The data acquisition was performed as previously described in detail[33] using a Waters nanoEquity™ UPLC system (Millford, MA) coupled with a Q Exactive Plus mass spectrometer from Thermo Fisher Scientific (San Jose, CA). The LC was configured to load the sample first on a solid phase extraction (SPE) column followed by separation on an analytical column. 500 ng of peptides were loaded into the SPE column (5 cm x 360 µm OD x 150 µm ID fused silica capillary tubing (Polymicro, Phoenix, AZ); packed with 3.6-µm Aeries C18 particles (Phenomenex, Torrance, CA) and the separation was carried out in a capillary column (70 cm x 360 µm OD x 75 µm ID packed with 3-µm Jupiter C18 stationary phase particles (Phenomenex). The elution was performed at 300 nl/min flow rate and the following gradient of acetonitrile (ACN) in water, both containing 0.1% formic acid: 1-8% ACN solvent in 2 min, 8-12% ACN in 18 min, 12-30% ACN in 55 min, 30-45% ACN in 22 min, 45-95% ACN in 3 min, hold for 5 min in 95% ACN and 99-1% ACN in 10 min. Eluting peptides were directly analyzed in the mass spectrometer by electrospray using etched silica fused tips[34]. Full MS spectra were acquired at a scan range of 400-2000 m/z and a resolution of 35,000 at m/z 400. Tandem mass spectra were collected for the top 12 most intense ions with ≥ 2 charges using high-collision energy (HCD) fragmentation from collision with N_2 at a normalized collision energy of 30% and a resolution of 17,500 at m/z 400. Each parent ion was targeted once for fragmentation and then dynamically excluded for 30 s.

Peptide identification for training/testing the Kaiko model. In the training and test set, the true source\taxonomy of each sample is known. To create the ground truth of spectrum identifications, we used the correct organism's protein sequence database and annotated spectra with the MSGF+ algorithm, as previously described[33]. PSM results from MSGF+ were filtered using a q-value threshold of 0.001. The PSMs passing this filter were considered the ground truth for the deep neural network training and testing. Because our use of this data is for *de novo* spectrum annotation, we limited peptides/spectrum matches further to exclude peptides

longer than 30 residues as these were unlikely to have complete peptide fragment peaks, which are important for a *de novo* solution. We also filtered peptides with a precursor mass >3000 Da. After filtering, the total number of distinct peptides was 1,013,498 from 5,116,305 spectra. Peptide sequences are highly specific to each organism, and the overlap between organisms was very low. Except for the pairs of organisms within the same genus or species (i.e. the two different strains of *B. subtilis* or the two different species within *Bifidobacterium*), the average amount of shared peptides between any two organisms was ~0.17%. These arise from highly conserved proteins like EF-Tu or RpoC for which peptides can be found conserved across phyla.

Training Kaiko

Codebase. Kaiko is based on DeepNovo, a deep neural network algorithm for peptide/spectrum matching[12]. We downloaded the source code for DeepNovo (https://github.com/nh2tran/DeepNovo) and its pre-trained model, which is publicly available at https://drive.google.com/open?id=0By9IxqHK5MdWalJLSGliWW1RY2c. As described below, we modified the original DeepNovo codebase, keeping with Python 2.7 and TensorFlow 1.2 as used in the original. First, we modified the codebase to accept multiple input files for training and testing. Our training and testing data came from over 250 mass spectrometry files, but the original DeepNovo was designed for only a single input file. Therefore, we added extra command-line options (e.g., --multi_decode and --multi_train) and the associated wrapper methods to allow for multi-file execution. A second change was done to avoid rebuilding the Cython codes on every parameter adjustment. For this, we replaced the Cython with the python *numba* package without any loss of performance and speed. Finally, we changed the code for spectral modeling based on domain knowledge. Specifically, we corrected the mass calculation of doubly charged ions and changed the bins used for isotopic profiles within the ion-CNN model.

We trained multiple models for Kaiko, which differed primarily in the number of peptides/spectra used during training: ~300K spectra, 1M spectra, 2M spectra, 3M spectra and the final models trained with all spectra. When training the final model on the full dataset, we adjusted the learning rate to 10^{-4} rather than using the default value (10^{-3}) of AdamOptimizer in DeepNovo. Training our final model requires very significant computational resources and time. With the hardware used in this project, training took ~12 hours per epoch; our final model was achieved after 60 epochs. All training and testing was performed on PNNL's Marianas cluster, a machine learning platform that is part of PNNL's Institutional Computing. System specifications on the nodes used in this training were: Dual Intel Broadwell E5-2620 v4 @ 2.10GHz CPUs (16 cores per node), 64 GB 2133Mhz DDR4 memory, and Dual NVIDIA P100 12GB PCI-e based GPUs.

Experimental Design and Statistical Rationale

Given that Deep Neural Networks are very sensitive to overfit during the training procedure, we anticipated that a very large amount of data would be required to make useful models. The original DeepNovo was trained on 50,000 spectra and we believed that a significantly larger

amount of data would be necessary. As described below in “Assessing Progress” we were able to quickly determine that a model with only 300,000 spectra was overfit. We therefore determined that we would aim for 5,000,000 spectra representing about 1,000,000 peptides in order to have sufficient data for training the very large neural networks that comprise Kaiko. During training we were able to determine that this number was more than sufficient to produce a generalized model that did not overfit to training data. Spectra included in the training, validation and testing set are assessed as described above in the “Data Generation” section.

Assessing Progress. The training regimen for deep learning is pragmatically broken up into several rounds of iteration over the training data, called epochs. During each epoch, a mini-batch stochastic optimization was employed, in which each batch of 128 spectra is randomly chosen and training proceeds on each batch one at a time. The model is trained by updating the parameters within the neural network (weights and biases) after each batch is compared to the true labels. While training, the error associated with the model can be calculated as a cross-entropy loss for the probabilities of correctly predicting the amino acid letters on the training data. After each batch, we also randomly sample 15,000 spectra from the validation dataset (~1% of total testing data) and compute the loss error, which we call the validation error. Importantly, model performance on this validation set is **not used** to update the model parameters; we simply use it to independently evaluate model performance and make a checkpoint to track the best models. The training and validation error after each batch for 20 epochs of training is shown in Supplementary Fig.2.

By comparing the training and validation error, we clearly see when the model has started to overfit. This happens when the training error crosses over (becomes smaller than) the validation error and continues to decrease as the validation error levels off. This is a result of the model learning specific features of the training data that are not generalizable. In models built with more than 3 million spectra, no overfitting is seen yet; models built with less than 3 million spectra quickly overfit to the training data.

Comparing Kaiko to other *de novo* tools

To compare the performance of Kaiko to state-of-the-art *de novo* tools, we analyzed all files in the testing data sets using DeepNovo[12], PEAKS[35] and Novor[30]. As mentioned above, we used a pre-trained model for the DeepNovo to predict peptide sequences for the test files using a ‘decode’ option. PEAKS Studio version 8.5 was run using default data refinement options on mzML formatted data[36]. *De novo* settings were as follows: precursor error tolerance - 20ppm, fragment ion error tolerance - 0.02 Da. Oxidation of methionine was set as a variable modification. For Novor the spectral files were converted from mzML to Mascot generic format (MGF) using MSConvert[37]. Novor version 1.05 was run using the following settings: fragmentation - HCD, massAnalyzer - FT, precursor error tolerance - 20ppm, fragment ion error tolerance - 0.02 Da. Oxidation of methionine was set as a variable modification. All other settings were left at their defaults. Only the best peptide spectrum match was used in the evaluation. Please refer to

https://github.com/PNNL-Comp-Mass-Spec/Kaiko_Publication/analysis/for_novor and /for_peaks for specific implementation details.

Assigning taxonomy to unknown samples

Proteomics data from six bacterial soil isolates was acquired using the same sample preparation and LC-MS/MS method as described above. The isolates are from the natural isolate collection at the Kristen DeAngelis laboratory at the University of Massachusetts Amherst, and researchers at PNNL were blinded to the identity of these isolates until after both data generation and analysis were finished. Kaiko's top-scoring peptide sequence for each spectrum was used for species identification. We filtered these peptide/spectrum matches to include only the top 25% according to Kaiko's quality prediction score. We then exclude sequences shorter than 10 and longer than 17 residues. The resulting sequences were used to search the Uniref100 protein database [https://www.uniprot.org/uniref/] using DIAMOND[13] to identify an organism(s) containing that peptide sequence. Only database matches of 100% were retained for species prediction. Taxon scoring then proceeded using a two-pass procedure. In the first pass, for each peptide sequence, all taxa possessing a 100% match were assigned 1 hit, such that multiple taxa were often assigned a hit from a single peptide sequence. Taxa were then ranked by the total number of hits assigned. In the second pass, hits were only assigned to the highest-ranking taxon with a 100% match to each predicted sequence. In this way, scoring is assigned to the candidate most likely to be correct.

Metaproteomics data analysis

Sample preparation from soils

Kansas prairie soil was quickly thawed and weighed into 10 g aliquots in 50 mL methanol/chloroform compatible tubes (Genesee Scientific, San Diego, CA) along with 10 mL of 0.9–2.0mm stainless steel beads, 0.1mm zirconia beads and 0.1mm garnet beads. All beads had previously been washed with chloroform and methanol and dried in a fume hood. Protein extraction occurred using a modified method of the Folch extraction[38] specifically for soil called Soil MPLex (Metabolite, protein, lipid extraction)[39]. Here, 4 mL of ice-cold ultrapure "Type 1" water (Millipore, Billerica, MA) was added to each sample and transferred to an ice bucket in a fume hood. Using a 25 mL glass serological pipette, −20 °C 2:1 chloroform: methanol (v/v) (Sigma-Aldrich, St. Louis, MO), was added to the sample in a 5:1 ratio over sample volume (20 mL) and vigorously mixed (by vortexing). The tubes were attached to a 50 mL tube vortex-attachment and horizontally mixed for 10 min at 4 °C and placed inside a −80 °C freezer for 5 min. Using a probe sonicator (model FB505, Thermo Fisher Scientific, Waltham, MA) inside a fume hood, each sample was sonicated with a 6mm probe (20 kHz fixed ultrasonic frequency) at 60% of the maximum amplitude for 30 s on ice, allowed to cool on ice, then sonicated once more. Samples were allowed to cool for 5 min. at −80 °C, then mixed for 60 s and centrifuged at 4,500 *xg* for 10 min at 4 °C. The upper aqueous phase was removed and the interphase containing proteins that partitioned between the methanol and chloroform phases was collected into a separate tube and precipitated through addition of 5 mL of −20 °C 100%

methanol. Following methanol addition, the tube was mixed then centrifuged at 4,500 *xg* for 5 min at 4 °C in order to pellet the proteins. The supernatant was decanted and the protein pellet dried upside down. Meanwhile, the bottom organic phase was removed, and 5 mL of −20 °C 100% methanol was added to the bottom debris pellet, mixed and centrifuged at 4,500 *xg* for 5 min at 4 °C. The supernatant was removed, and the protein pellet was dried upside down. Protein pellets from both the debris and interphase were frozen and lyophilized for 2 h.

Proteins from the interphase were solubilized by addition of 10 mL of SDS-Tris buffer containing 4% sodium dodecyl sulfate (SDS), 100mM DL-dithiothreitol (DTT) in 100mM Tris-HCl, pH 8.0, (Sigma-Aldrich, St Louis, MO), briefly probe sonicated at 20% amplitude, then incubated on a lab tube rotator for 30 min at 300 rpm, 50 °C. Proteins from the debris pellet were solubilized in 20 mL of SDS buffer, horizontally vortexed for 10 min. to lyse any remaining intact cells, then combined with the interphase proteins and mixed on the rotator assembly for the time remaining (approximately 20 min). Following mixing, the tubes were centrifuged at 4,500 *xg* for 10 min., and the supernatant from each tube were combined into a single 50 mL tube. The proteins were precipitated by adding up to 25% trichloroacetic acid (TCA; Sigma-Aldrich, St. Louis, MO), mixed and placed at −20 °C overnight. The proteins were thawed and centrifuged at 4,500 *xg* at 4 °C for 10 min to collect the precipitated proteins. The supernatant was gently decanted, and the protein pellet washed through addition of 2 mL of −20 °C acetone, mixed, then placed at −80 °C for 5 min. Proteins were pelleted by centrifugation for 10 min at 4,500 *xg* at 4 °C. The acetone was removed by gently decanting, and the wash step was repeated 2 more times. The washed pellet was then air dried by inverting the tube. After drying, 100 µl–200 µl of SDS-Tris buffer was added and the solution was transferred into 1.5 mL tubes and incubated at 95 °C for 5 min, then cooled at 4 °C for 10 min. The samples were centrifuged at 15,000 *xg* for 10 min to pellet any remaining debris and transferred into fresh 1.5 mL tubes in preparation for digestion using the Filter-Aided-Sample-Preparation (FASP) digestion method[40]. For protein digestion, up to 30 µl of proteins in SDS-Tris buffer were transferred to a 30,000 Da molecular weight cut off (MWCO) 500 µl spin filter provided in the Expedeon FASP kit (Expedeon LTD, Cambridgeshire, UK) along with 400 µl of 8 M urea solution. The spin filter was centrifuged at 14,000 *xg* for 30 min. The waste was removed from the collection tubes and 400 µl of 8M urea solution was added to each sample and centrifuged as described above, then repeated for a total of 3 urea additions. 400 µl of 25mM NH_4HCO_3, pH 8, was added and centrifuged as described above, then repeated for a total of 2 ammonium bicarbonate washes. The spin column was transferred into a fresh-labeled collection tube and 75 µl of NH_4HCO_3 was added to the filter along with 4 µl of 1 µg/µl molecular grade trypsin (Thermo Fisher, Waltham, MA) then incubated at 37 °C for 3 h. After digestion, 40 µl of NH_4HCO_3 was added to the sample and centrifuged at 14,000 *xg* for 20 min. Another 40 µl of NH_4HCO_3 was added to the top of the filter, mixed and centrifuged again for 10 min. The filter was discarded, and the collected peptides were treated with potassium chloride (KCl) in order to ensure all the SDS was removed[41]. To accomplish this, potassium chloride was added to the peptides in NH_4HCO_3 resulting in a final concentration of 2M KCl, then mixed and allowed to rest for 10 min. at room temperature. To pellet the SDS, the peptide solution containing NH_4HCO_3 and KCl was centrifuged at 14,000 *xg* for 10 min. The supernatant was transferred to a fresh tube without disturbing the SDS pellet and salts removed

using a microspin C18 column according to the manufacturer's instructions (the Nest Group, Inc., Southborough, MA). Peptides from the aliquots of 10 g of soil were combined to generate a single peptide sample. A bicinchoninic acid (BCA) assay (Thermo Fisher Scientific, Waltham, MA) was performed to determine the peptide concentration.

The peptide sample was separated with a commercial Waters (Milford, MA) XBridge 5 µm particle size C18 column, (4.6mm i. d. x 250mm length) with an attached 20mm long x 4.6mm i. d. guard column. Fractionation was performed at 0.5 mL/min using an Agilent 1100 series HPLC system (Agilent Technologies, Santa Clara, CA) with two mobile phases: A) 10mM NH_4HCO_2 (pH 10.0), and B) 10mM NH_4HCO_2 (pH 10.0) with acetonitrile (10:90). A six step gradient was adjusted over 120 min by replacing mobile phase A with B according to: 1) 100%–95% over the first 10 min., 2) 95%–65% from minutes 10 to 70, 3) 65%–30% from minutes 70 to 85, 4) then maintained mobile phase A at 30% from minutes 85 to 95, 5) re-equilibrating with 100% mobile phase A from minute 95 to 105, and 6) holding mobile phase A at 100% until minute 120. Fractions were collected every 1.25 min (96 fractions over the entire gradient) with every 24th fraction combined for a total of 24 final fractions (rows of a 96 well plate were pooled by every other row). All fractions were dried under vacuum and suspended in 25 µl H_2O. A final BCA assay was done on the fractions and each were diluted to 0.1 µg/µl for LC-MS/MS analysis. (see LC-MS/MS data acquisition methods above).

Analyzing 16S rRNA amplicon sequences

16S rRNA gene amplicon sequencing data was downloaded from https://osf.io/4uvj7/, performed using the protocol developed by the Earth Microbiome Project[1]. Please refer to the previous studies[19,42] for 16S rRNA gene amplicon sequencing in detail. The 16S rRNA amplicon sequences were first re-processed by Hundo pipeline[43] (v1.2.8), a command line interface work comprising a set of existing software together with validated custom methods derived from QIIME[44]. In brief, the sequences were first quality filtered to remove the adaptors and contaminated reads from Phix genomes by BBDuk2[45]. The passing reads were merged and checked for chimera, which were subjected to be clustered into OTUs by VSEARCH[46] using the default parameters. The abundance of each OTU was estimated by the read coverage of the OTU representative sequences (VSEARCH). In comparison to the Silva database[47] implemented in Hundo, NCBI database was reported with a higher confidence of lineage assignment to lower taxonomy levels[48]. The de-replicated representative sequences of each OTU were then annotated following the same workflow coded in Hundo with modifications and using NCBI 16S Refseq database (https://www.ncbi.nlm.nih.gov/refseq/targetedloci/16S_process/, accessed on Apr 9th, 2020) instead. The top 25 hits of each OTU representative sequence were kept and screened for ones with percent identity higher than 85% and bit score greater than 125. For OTUs with more than one qualified hits, we will perform Lowest common ancestor (LCA) algorism using a R package, taxize[49]. OTUs with only one qualified hit adopted the lineage of the hits and the rest were left unclassified.

Constructing protein database for metaproteomic data analysis with Kaiko

Raw mass spectrometry files were converted to the PSI open format mzML[36] using msConvert[37], which were converted to MGF files compatible with the Kaiko model. After performing Kaiko prediction, as used for assigning taxa to the unknown samples, we used Kaiko's top 25% scoring peptide sequences predicted from each sample to identify the most likely candidate organisms using DIAMOND over the Uniref100 database. The protein database was constructed by aggregating all the reference sequences associated with top 100 bacterial organisms from the Uniref100 into a single fasta (8.2GB).

Peptide identification and functional analysis with the constructed database

Against the protein database constructed from the Kaiko prediction, MSGF+ was performed to identify peptide sequences with the false discovery rate (FDR) cutoff. The search parameters and values or settings were as follows: PrecursorMassTolerance, 20.0 ppm; IsotopeErrorRange, −1,1; TargetDecoyAnalysis, true; FragmentationMethod, as written in the spectrum; InstrumentID, 0; Enzyme, Tryp; NumTolerableTermini, 2; MinPeptideLength, 6; MaxPeptideLength, 50; MinCharge, 2; MaxCharge, 5; and NumMatchesPerSpec, 1. PSM results from MSGF+ were filtered using MSnID (v1.20.0)[50]. Filters based on the cleavage patterns for the trypsin were applied, e.g., nuIrregCleavages==0 and numMissCleavages<=2. Optimizing the MS/MS filter was applied to achieve the maximum number of identifications within a given FDR upper limit threshold. Nelder-Mead method was employed for parameter optimization (MS-GF:SpecEValue and absParentMassErrorPPM), and for 5% peptide FDR, SpecEValue≤1.0e-11, and 11 ppm mass window with the ppm offset adjustment were determined. For functional annotation for metaproteomics, Unipept[29] (v4.3.5, https://unipept.ugent.be/datasets, accessed on Jun 2nd, 2020) was used with "Equate I and L" and "Filter duplicate peptides" options.

Data Availability

The mass spectrometry proteomics data for this benchmark set are split into two separate depositions, for the training and testing datasets respectively. The training dataset consists of spectra from 51 organisms and has been deposited to the ProteomeXchange Consortium via the PRIDE[25] partner repository with the dataset identifier PXD010000. The testing dataset consists of spectra for 4 organisms and has been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD010613. The metaproteomics dataset has been deposited to the MassIVE Repository with the accession identifier MSV000086336.

References

1. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
2. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **13**, 227–232 (2012).
3. Eng, J. K., McCormack, A. L. & Yates, J. R. 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 (1994).
4. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
5. Mao, D.-P., Zhou, Q., Chen, C.-Y. & Quan, Z.-X. Coverage evaluation of universal bacterial primers using the metagenomic datasets. *BMC Microbiol.* **12**, 66 (2012).
6. Howe, A. C. *et al.* Tackling soil diversity with the assembly of large, complex metagenomes. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 4904–4909 (2014).
7. Rodriguez-R, L. M. & Konstantinidis, K. T. Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinforma. Oxf. Engl.* **30**, 629–635 (2014).
8. Poretsky, R., Rodriguez-R, L. M., Luo, C., Tsementzi, D. & Konstantinidis, K. T. Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. *PloS One* **9**, e93827 (2014).
9. Pereira-Marques, J. *et al.* Impact of Host DNA and Sequencing Depth on the Taxonomic Resolution of Whole Metagenome Sequencing for Microbiome Analysis. *Front. Microbiol.* **10**, 1277 (2019).

10. Rodriguez-R, L. M. & Konstantinidis, K. T. Estimating coverage in metagenomic data sets and why it matters. *ISME J.* **8**, 2349–2351 (2014).

11. Frank, A. & Pevzner, P. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* **77**, 964–973 (2005).

12. Tran, N. H., Zhang, X., Xin, L., Shan, B. & Li, M. De novo peptide sequencing by deep learning. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 8247–8252 (2017).

13. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

14. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).

15. Heyer, R. *et al.* Challenges and perspectives of metaproteomic data analysis. *J. Biotechnol.* **261**, 24–36 (2017).

16. Xiao, J. *et al.* Metagenomic Taxonomy-Guided Database-Searching Strategy for Improving Metaproteomic Analysis. *J. Proteome Res.* **17**, 1596–1605 (2018).

17. Jagtap, P. *et al.* A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies. *Proteomics* **13**, 1352–1357 (2013).

18. McClure, R. S. *et al.* Integrated network modeling approach defines key metabolic responses of soil microbiomes to perturbations. *Sci. Rep.* **10**, 10882 (2020).

19. Roy Chowdhury, T. *et al.* Metaphenomic Responses of a Native Prairie Soil Microbiome to Moisture Perturbations. *mSystems* **4**, (2019).

20. Delgado-Baquerizo, M. *et al.* A global atlas of the dominant bacteria found in soil. *Science* **359**, 320–325 (2018).

21. Starke, R., Jehmlich, N. & Bastida, F. Using proteins to study how microbes contribute to

soil ecosystem services: The current state and future perspectives of soil metaproteomics. *J. Proteomics* **198**, 50–58 (2019).

22. Tedersoo, L. *et al.* Fungal biogeography. Global diversity and geography of soil fungi. *Science* **346**, 1256688 (2014).

23. Becraft, E. D. *et al.* Rokubacteria: Genomic Giants among the Uncultured Bacterial Phyla. *Front. Microbiol.* **8**, 2264 (2017).

24. Wilson, M. C. *et al.* An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature* **506**, 58–62 (2014).

25. Wang, W. *et al.* Soil Property and Plant Diversity Determine Bacterial Turnover and Network Interactions in a Typical Arid Inland River Basin, Northwest China. *Front. Microbiol.* **10**, 2655 (2019).

26. Ogwu, M. C. *et al.* Community Ecology of Deinococcus in Irradiated Soil. *Microb. Ecol.* **78**, 855–872 (2019).

27. Li, H.-Y. *et al.* The chemodiversity of paddy soil dissolved organic matter correlates with microbial community at continental scales. *Microbiome* **6**, 187 (2018).

28. Deng, J., Yin, Y., Zhu, W. & Zhou, Y. Variations in Soil Bacterial Community Diversity and Structures Among Different Revegetation Types in the Baishilazi Nature Reserve. *Front. Microbiol.* **9**, 2874 (2018).

29. Gurdeep Singh, R. *et al.* Unipept 4.0: Functional Analysis of Metaproteome Data. *J. Proteome Res.* **18**, 606–615 (2019).

30. Ma, B. Novor: real-time peptide de novo sequencing software. *J. Am. Soc. Mass Spectrom.* **26**, 1885–1894 (2015).

31. Crits-Christoph, A., Diamond, S., Butterfield, C. N., Thomas, B. C. & Banfield, J. F. Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. *Nature* **558**,

440–444 (2018).

32. Hug, L. A. *et al.* Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages. *Environ. Microbiol.* **18**, 159–173 (2016).

33. Nakayasu, E. S. *et al.* Ancient Regulatory Role of Lysine Acetylation in Central Metabolism. *mBio* **8**, e01894-17, /mbio/8/6/mBio.01894-17.atom (2017).

34. Kelly, R. T. *et al.* Chemically etched open tubular and monolithic emitters for nanoelectrospray ionization mass spectrometry. *Anal. Chem.* **78**, 7796–7801 (2006).

35. Ma, B. *et al.* PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom. RCM* **17**, 2337–2342 (2003).

36. Deutsch, E. mzML: a single, unifying data format for mass spectrometer output. *Proteomics* **8**, 2776–2777 (2008).

37. Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinforma. Oxf. Engl.* **24**, 2534–2536 (2008).

38. Folch, J., Lees, M. & Sloane Stanley, G. H. A simple method for the isolation and purification of total lipides from animal tissues. *J. Biol. Chem.* **226**, 497–509 (1957).

39. Nicora, C. D. *et al.* The MPLEx Protocol for Multi-omic Analyses of Soil Samples. *J. Vis. Exp.* 57343 (2018) doi:10.3791/57343.

40. Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* **6**, 359–362 (2009).

41. Zhou, J.-Y. *et al.* Simple sodium dodecyl sulfate-assisted sample preparation method for LC-MS-based proteomics applications. *Anal. Chem.* **84**, 2862–2867 (2012).

42. White, R. A. *et al.* Moleculo Long-Read Sequencing Facilitates Assembly and Genomic Binning from Complex Soil Metagenomes. *mSystems* **1**, (2016).

43. Brown, J., Zavoshy, N., Brislawn, C. J. & McCue, L. A. *Hundo: a Snakemake workflow for microbial community sequence data*. https://peerj.com/preprints/27272v1 (2018) doi:10.7287/peerj.preprints.27272v1.

44. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336 (2010).

45. Bushnell, B. *BBMap: A Fast, Accurate, Splice-Aware Aligner*. https://www.osti.gov/biblio/1241166 (2014).

46. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).

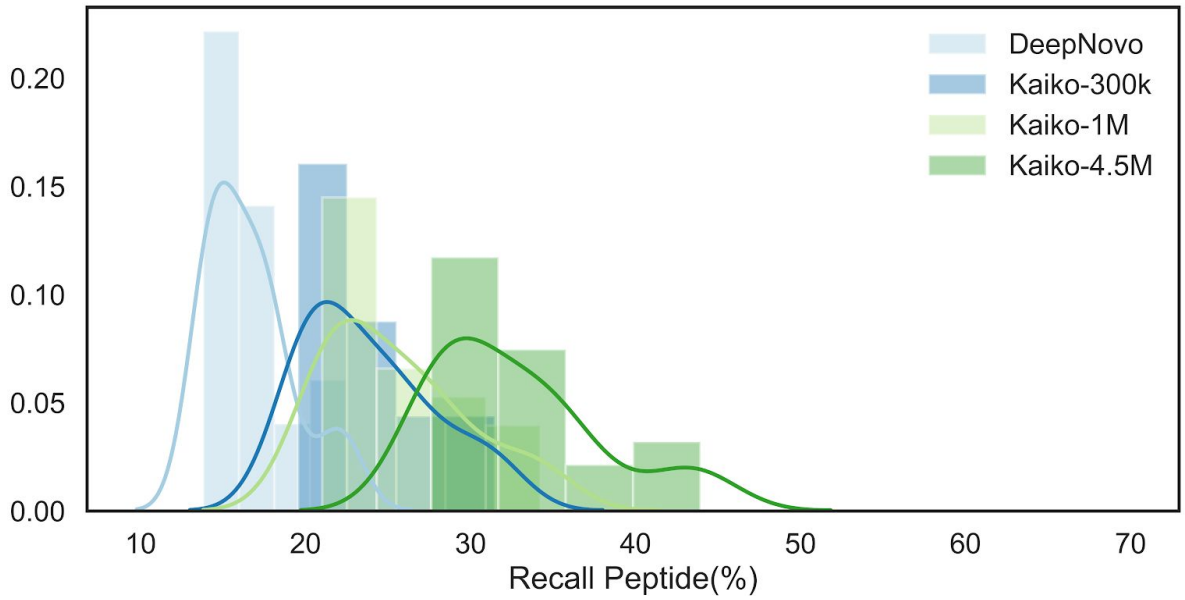
47. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590-596 (2013).

48. Balvočiūtė, M. & Huson, D. H. SILVA, RDP, Greengenes, NCBI and OTT - how do these taxonomies compare? *BMC Genomics* **18**, 114 (2017).

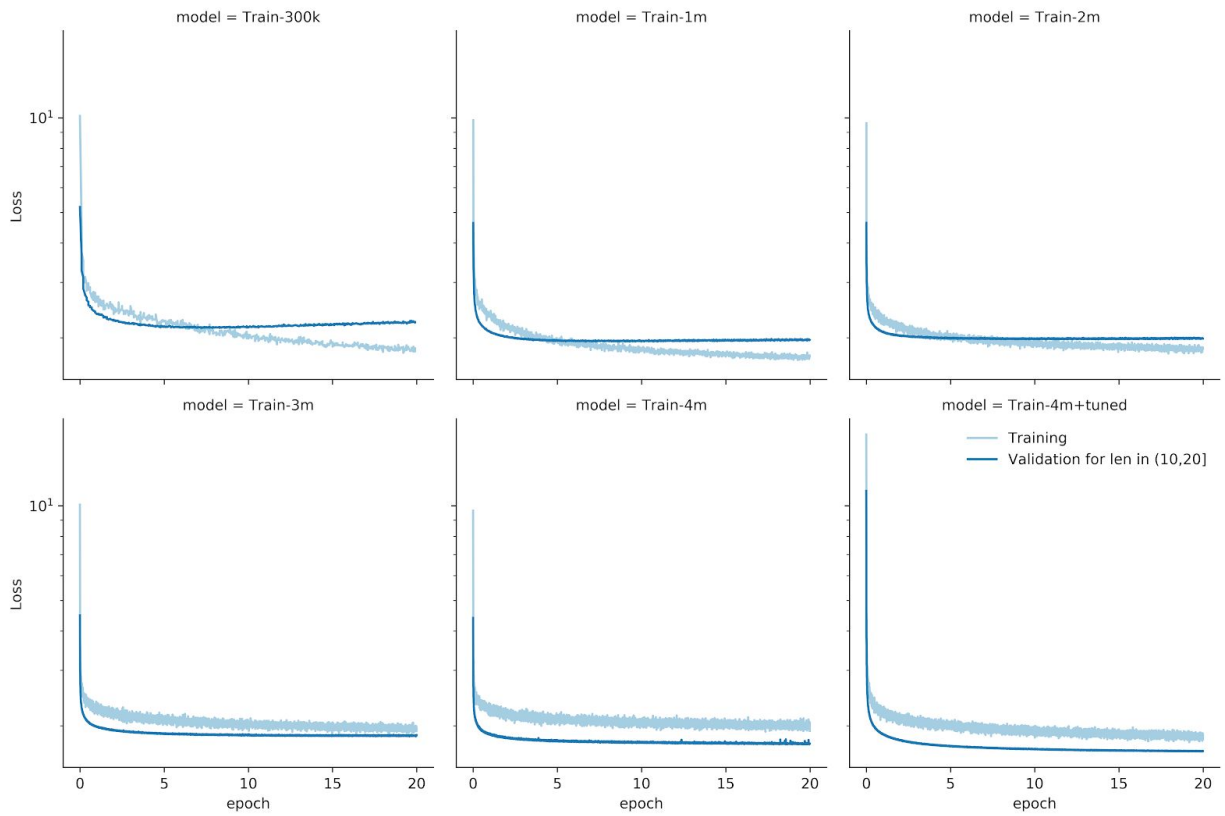
49. Chamberlain, S. A. & Szöcs, E. taxize: taxonomic search and retrieval in R. *F1000Research* **2**, 191 (2013).

50. *MSnID: Utilities for Exploration and Assessment of Confidence of LC-MSn Proteomics Identifications*. (Bioconductor, 2020). doi:10.18129/B9.BIOC.MSNID.

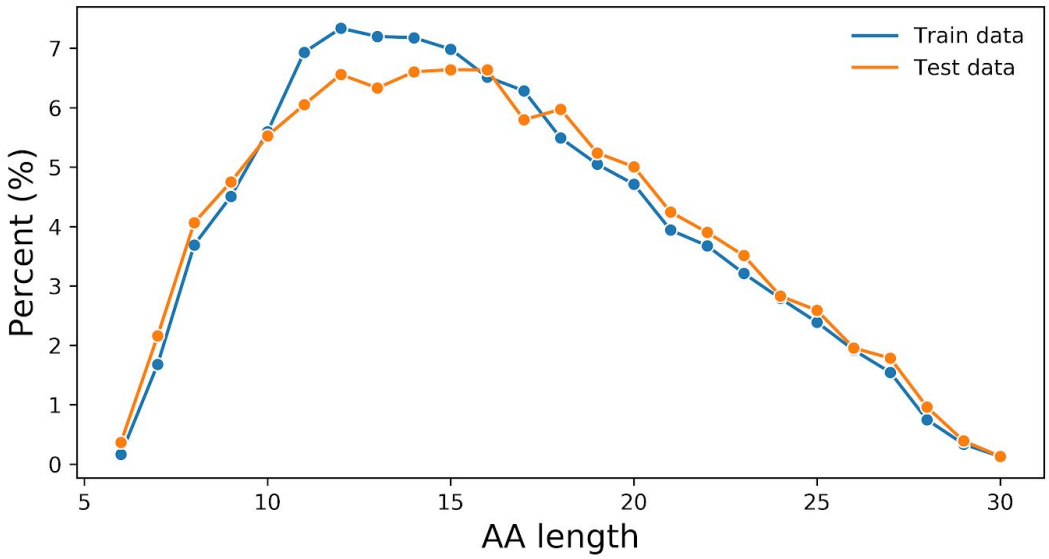
Supplementary Figures



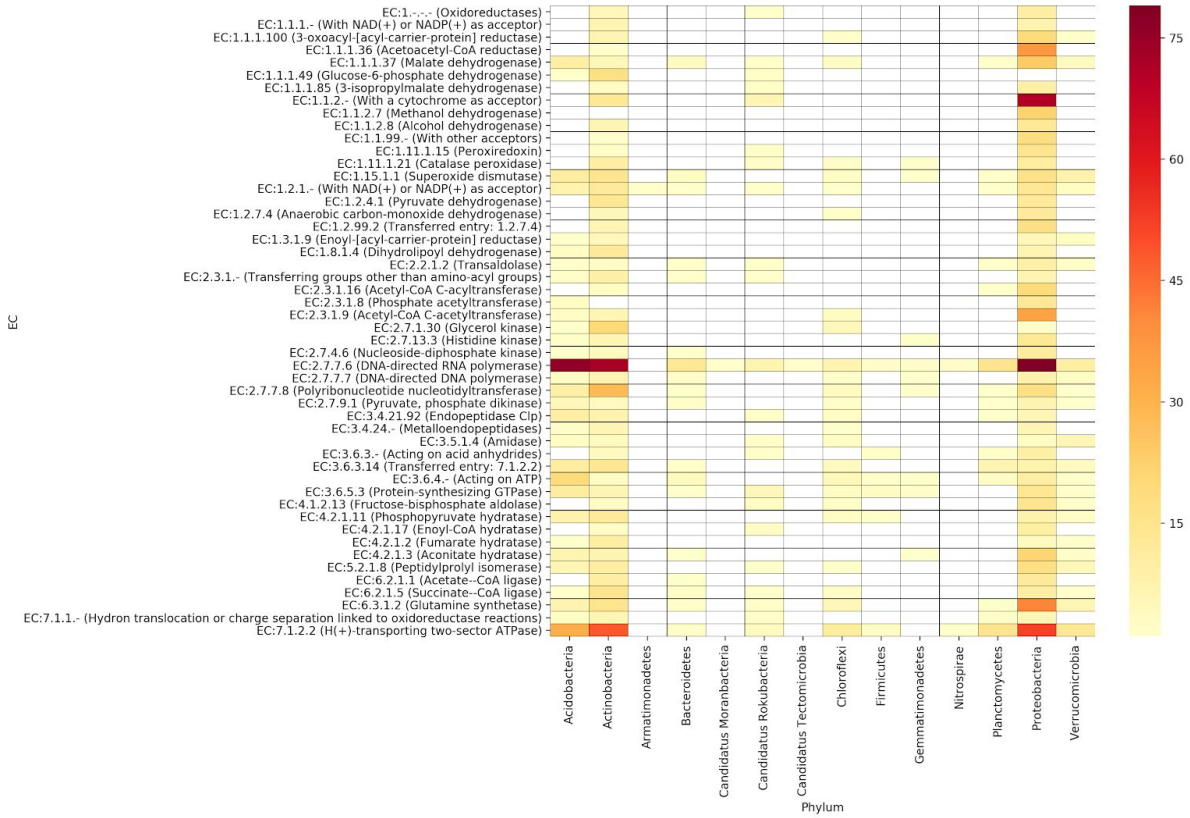
Supplementary Figure 1. Improving deep neural networks with more training data. The accuracy of peptide/spectrum matching is shown for four deep neural network models. DeepNovo is a pre-trained publicly available model trained on 50,000 spectra. Kaiko was trained with varying numbers of spectra. The final model was trained with 4.5 million spectra. A significant improvement is seen in model performance with increased training data.



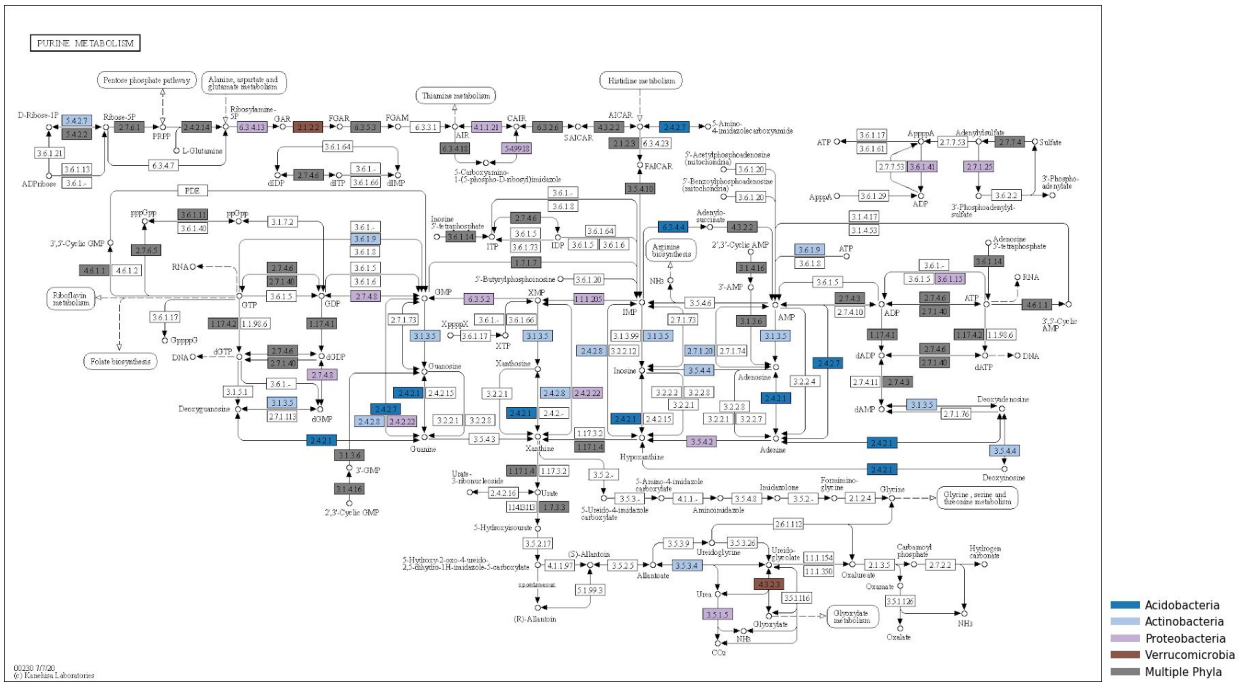
Supplementary Figure 2 - Training and validation error. During the epochs of learning for the deep neural network, progress is measured by evaluating the accuracy of spectrum annotation. We employ a cross-entropy loss function, which represents how well the algorithm is being trained, with small numbers being better. The light blue line represents the error on batches of training data. The dark blue lines represent the error on the random samples of the validation data. When the training error improves beyond the validation error, the model is likely to be overfitting.



Supplementary Figure 3 - Distribution of peptide lengths used for training and testing the Kaiko model.



Supplementary Figure 4. Heatmap of the peptide counts for the most common functions over the diverse phyla. Columns and rows in the heatmap represent the phyla and EC numbers, respectively. Cell colors indicate the number of phyla-affiliated peptides corresponding to a specific phylum and function.



Supplementary Figure 5. Taxa-specific peptides for enzymes in purine metabolism at the phylum level. Many EC numbers were uniquely detected in four phyla (Acidobacteria, Actinobacteria, Proteobacteria, and Verrucomicrobia).

Supplementary Tables

Supplementary Table 1. LC/MS data files for training and testing the Kaiko model.

Index	Files	Species	# PSM	# Cumulative PSM
0	Biodiversity_A_cryptum_FeTSB_anaerobic_1_01Jun16_Pippin_16-03-39	Acidiphilium_cryptum_JF-5	6659	6659
1	Biodiversity_A_cryptum_FeTSB_anaerobic_2_01Jun16_Pippin_16-03-39	Acidiphilium_cryptum_JF-5	8532	15191
2	Biodiversity_A_cryptum_FeTSB_anaerobic_3_01Jun16_Pippin_16-03-39	Acidiphilium_cryptum_JF-5	7379	22570
3	Biodiversity_A_faecalis_LB_aerobic_01_26Feb16_Arwen_16-01-01	Alcaligenes_faecalis	15496	38066
4	Biodiversity_A_faecalis_LB_aerobic_02_26Feb16_Arwen_16-01-01	Alcaligenes_faecalis	15367	53433
5	Biodiversity_A_faecalis_LB_aerobic_03_26Feb16_Arwen_16-01-01	Alcaligenes_faecalis	15035	68468
6	Biodiversity_A_tumefaciens_R2A_aerobic_1_23Nov16_Pippin_16-09-11	Agrobacterium_tumefaciens_IAM_12048	12994	81462
7	Biodiversity_A_tumefaciens_R2A_aerobic_2_23Nov16_Pippin_16-09-11	Agrobacterium_tumefaciens_IAM_12048	12442	93904
8	Biodiversity_A_tumefaciens_R2A_aerobic_3_23Nov16_Pippin_16-09-11	Agrobacterium_tumefaciens_IAM_12048	11916	105820
9	Biodiversity_B_bifidum_CMcarb_anaerobic_01_26Feb16_Arwen_16-01-01	Bifidobacterium_bifidum_ATCC29521	14409	120229
10	Biodiversity_B_bifidum_CMcarb_anaerobic_02_26Feb16_Arwen_16-01-01	Bifidobacterium_bifidum_ATCC29521	13731	133960
11	Biodiversity_B_bifidum_CMcarb_anaerobic_03_26Feb16_Arwen_16-01-01	Bifidobacterium_bifidum_ATCC29521	13854	147814
12	Biodiversity_B_cereus_ATCC14579_LB_aerobic_1_17July16_Samwise_16-04-10	Bacillus_cereus_ATCC14579	23828	171642
13	Biodiversity_B_cereus_ATCC14579_LB_aerobic_2_17July16_Samwise_16-04-10	Bacillus_cereus_ATCC14579	23693	195335
14	Biodiversity_B_cereus_ATCC14579_LB_aerobic_3_17July16_Samwise_16-04-10	Bacillus_cereus_ATCC14579	22460	217795
15	Biodiversity_B_cereus_PN_L_CL_1_09Oct16_Pippin_16-05-06	Bacillus_cereus_ATCC14579	22349	240144
16	Biodiversity_B_cereus_PN_L_CL_2_09Oct16_Pippin_16-05-06	Bacillus_cereus_ATCC14579	22572	262716
17	Biodiversity_B_cereus_PN_L_CL_3_09Oct16_Pippin_16-05-06	Bacillus_cereus_ATCC14579	23153	285869

18	Biodiversity_B_fragilis_01_28Jul15_Arwen_14-12-03	Bacteroides_fragilis_638R	19454	305323
19	Biodiversity_B_fragilis_Carb_01_28Oct15_Arwen_15-07-13	Bacteroides_fragilis_638R	17656	322979
20	Biodiversity_B_fragilis_CMcarb_anaerobic_01_01Feb16_Arwen_15-07-13	Bacteroides_fragilis_638R	21410	344389
21	Biodiversity_B_fragilis_CMcarb_anaerobic_02_01Feb16_Arwen_15-07-13	Bacteroides_fragilis_638R	21703	366092
22	Biodiversity_B_fragilis_CMcarb_anaerobic_03_01Feb16_Arwen_15-07-13	Bacteroides_fragilis_638R	22366	388458
23	Biodiversity_B_fragilis_CMgluc_anaerobic_01_01Feb16_Arwen_15-07-13	Bacteroides_fragilis_638R	19770	408228
24	Biodiversity_B_fragilis_CMgluc_anaerobic_02_01Feb16_Arwen_15-07-13	Bacteroides_fragilis_638R	20803	429031
25	Biodiversity_B_fragilis_CMgluc_anaerobic_03_01Feb16_Arwen_15-07-13	Bacteroides_fragilis_638R	20515	449546
26	Biodiversity_B_fragilis_LB_anaerobic_01_01Feb16_Arwen_15-07-13	Bacteroides_fragilis_638R	21122	470668
27	Biodiversity_B_fragilis_LB_anaerobic_02_01Feb16_Arwen_15-07-13	Bacteroides_fragilis_638R	21756	492424
28	Biodiversity_B_fragilis_LB_anaerobic_03_01Feb16_Arwen_15-07-13	Bacteroides_fragilis_638R	22228	514652
29	Biodiversity_B_fragilis_LIB_aerobic_01_01Feb16_Arwen_15-07-13	Bacteroides_fragilis_638R	20523	535175
30	Biodiversity_B_fragilis_LIB_aerobic_02_01Feb16_Arwen_15-07-13	Bacteroides_fragilis_638R	20037	555212
31	Biodiversity_B_fragilis_LIB_aerobic_03_01Feb16_Arwen_15-07-13	Bacteroides_fragilis_638R	21645	576857
32	Biodiversity_B_fragilis_LIB_anaerobic_01_08Feb16_Arwen_15-07-13	Bacteroides_fragilis_638R	1055	577912
33	Biodiversity_B_fragilis_LIB_anaerobic_02_01Feb16_Arwen_15-07-13	Bacteroides_fragilis_638R	22114	600026
34	Biodiversity_B_fragilis_LIB_anaerobic_03_01Feb16_Arwen_15-07-13	Bacteroides_fragilis_638R	21766	621792
35	Biodiversity_B_infantis_CMcarb_anaerobic_01_26Feb16_Arwen_16-01-01	Bifidobacterium_longum_infantis_ATCC15697	11900	633692
36	Biodiversity_B_infantis_CMcarb_anaerobic_02_26Feb16_Arwen_16-01-01	Bifidobacterium_longum_infantis_ATCC15697	12737	646429
37	Biodiversity_B_infantis_CMcarb_anaerobic_03_26Feb16_Arwen_16-01-01	Bifidobacterium_longum_infantis_ATCC15697	11620	658049
38	Biodiversity_B_subtilis_NCIB3610_24h_plates_1_13Jun16_Pip	Bacillus_subtilis_NCIB3610	14159	672208

	pin_16-03-39			
39	Biodiversity_B_subtilis_NCIB3610_24h_plates_2_13Jun16_Pippin_16-03-39	Bacillus_subtilis_NCIB3610	12880	685088
40	Biodiversity_B_subtilis_NCIB3610_24h_plates_3_13Jun16_Pippin_16-03-39	Bacillus_subtilis_NCIB3610	12518	697606
41	Biodiversity_B_subtilis_NCIB3610_48h_plates_1_13Jun16_Pippin_16-03-39	Bacillus_subtilis_NCIB3610	9087	706693
42	Biodiversity_B_subtilis_NCIB3610_48h_plates_2_13Jun16_Pippin_16-03-39	Bacillus_subtilis_NCIB3610	8258	714951
43	Biodiversity_B_subtilis_NCIB3610_48h_plates_3_13Jun16_Pippin_16-03-39	Bacillus_subtilis_NCIB3610	5163	720114
44	Biodiversity_B_subtilis_NCIB3610_pellet_1_03May16_Samwise_16-03-32	Bacillus_subtilis_NCIB3610	20922	741036
45	Biodiversity_B_subtilis_NCIB3610_pellet_2_03May16_Samwise_16-03-32	Bacillus_subtilis_NCIB3610	21034	762070
46	Biodiversity_B_subtilis_NCIB3610_plates_1_03May16_Samwise_16-03-32	Bacillus_subtilis_NCIB3610	12240	774310
47	Biodiversity_B_subtilis_NCIB3610_plates_2_03May16_Samwise_16-03-32	Bacillus_subtilis_NCIB3610	13306	787616
48	Biodiversity_B_subtilis_pellet_set2_1_13Jun16_Pippin_16-03-39	Bacillus_subtilis_NCIB3610	17709	805325
49	Biodiversity_B_subtilis_pellet_set2_2_13Jun16_Pippin_16-03-39	Bacillus_subtilis_NCIB3610	17532	822857
50	Biodiversity_B_subtilis_pellet_set2_3_13Jun16_Pippin_16-03-39	Bacillus_subtilis_NCIB3610	18214	841071
51	Biodiversity_B_thet_CMcarb_anaerobic_01_01Feb16_Arwen_15-07-13	Bacteroides_thetaiotaomicron_VPI-5482	22586	863657
52	Biodiversity_B_thet_CMcarb_anaerobic_02_01Feb16_Arwen_15-07-13	Bacteroides_thetaiotaomicron_VPI-5482	25220	888877
53	Biodiversity_B_thet_CMcarb_anaerobic_03_01Feb16_Arwen_15-07-13	Bacteroides_thetaiotaomicron_VPI-5482	22535	911412
54	Biodiversity_B_thet_CMgluc_anaerobic_01_01Feb16_Arwen_15-07-13	Bacteroides_thetaiotaomicron_VPI-5482	20596	932008
55	Biodiversity_B_thet_CMgluc_anaerobic_02_01Feb16_Arwen_15-07-13	Bacteroides_thetaiotaomicron_VPI-5482	20725	952733
56	Biodiversity_B_thet_CMgluc_anaerobic_03_01Feb16_Arwen_15-07-13	Bacteroides_thetaiotaomicron_VPI-5482	20639	973372
57	Biodiversity_B_thet_LB_anaerobic_01_01Feb16_Arwen_15-07-13	Bacteroides_thetaiotaomicron_VPI-5482	22310	995682

58	Biodiversity_B_thet_LB_anaerobic_02_01Feb16_Arwen_15-07-13	Bacteroides_thetaiotaomicron_VPI-5482	20736	1016418
59	Biodiversity_B_thet_LB_anaerobic_03_01Feb16_Arwen_15-07-13	Bacteroides_thetaiotaomicron_VPI-5482	17178	1033596
60	Biodiversity_B_thet_LIB_anaerobic_01_01Feb16_Arwen_15-07-13	Bacteroides_thetaiotaomicron_VPI-5482	23175	1056771
61	Biodiversity_B_thet_LIB_anaerobic_02_01Feb16_Arwen_15-07-13	Bacteroides_thetaiotaomicron_VPI-5482	21920	1078691
62	Biodiversity_B_thet_LIB_anaerobic_03_01Feb16_Arwen_15-07-13	Bacteroides_thetaiotaomicron_VPI-5482	22215	1100906
63	Biodiversity_B_thetaiotaomicron_Carb_01_26Aug15_Arwen_15-07-13	Bacteroides_thetaiotaomicron_VPI-5482	22781	1123687
64	Biodiversity_B_thetaiotaomicron_Glc_01_26Aug15_Arwen_15-07-13	Bacteroides_thetaiotaomicron_VPI-5482	25625	1149312
65	Biodiversity_Bacillus_subtilis_LB_01_27Dec15_Arwen_15-07-13	Bacillus_subtilis_168	23891	1173203
66	Biodiversity_Bacillus_subtilis_LB_02_27Dec15_Arwen_15-07-13	Bacillus_subtilis_168	23513	1196716
67	Biodiversity_Bacillus_subtilis_LB_03_27Dec15_Arwen_15-07-13	Bacillus_subtilis_168	25596	1222312
68	Biodiversity_C_Baltica_T240_R1_C_27Jan16_Arwen_15-07-13	Cellulophaga_baltica_18	23983	1246295
69	Biodiversity_C_Baltica_T240_R1_Inf_27Jan16_Arwen_15-07-13	Cellulophaga_baltica_18	26844	1273139
70	Biodiversity_C_Baltica_T240_R2_C_27Jan16_Arwen_15-07-13	Cellulophaga_baltica_18	26240	1299379
71	Biodiversity_C_Baltica_T240_R2_Inf_27Jan16_Arwen_15-07-13	Cellulophaga_baltica_18	26536	1325915
72	Biodiversity_C_Baltica_T240_R3_C_27Jan16_Arwen_15-07-13	Cellulophaga_baltica_18	27084	1352999
73	Biodiversity_C_Baltica_T240_R3_Inf_27Jan16_Arwen_15-07-13	Cellulophaga_baltica_18	27658	1380657
74	Biodiversity_C_comes_Carb_01_14Sep15_Arwen_15-07-13	Coprococcus_comes_ATCC27758	20528	1401185
75	Biodiversity_C_comes_Glc_01_28Oct15_Arwen_15-07-13	Coprococcus_comes_ATCC27758	21095	1422280
76	Biodiversity_C_comes_LIB_01_28Oct15_Arwen_15-07-13	Coprococcus_comes_ATCC27758	21532	1443812
77	Biodiversity_C_freundii_LB_01_14Sep15_Arwen_15-07-13	Citrobacter_freundii	23455	1467267
78	Biodiversity_C_freundii_LIB_01_28Oct15_Arwen_15-07-13	Citrobacter_freundii	22562	1489829
79	Biodiversity_C_gilvus_GS2_anaerobic_01_01Feb16_Arwen_15	Cellulomonas_gilvus_ATCC1312	25544	1515373

	-07-13	7		
80	Biodiversity_C_gilvus_GS2_anaerobic_02_01Feb16_Arwen_15-07-13	Cellulomonas_gilvus_ATCC13127	24443	1539816
81	Biodiversity_C_gilvus_GS2_anaerobic_03_01Feb16_Arwen_15-07-13	Cellulomonas_gilvus_ATCC13127	24651	1564467
82	Biodiversity_C_indologenes_LIB_aerobic_01_03May16_Samwise_16-03-32	Chryseobacterium_indologenes	12314	1576781
83	Biodiversity_C_indologenes_LIB_aerobic_02_03May16_Samwise_16-03-32	Chryseobacterium_indologenes	12289	1589070
84	Biodiversity_C_indologenes_LIB_aerobic_03_03May16_Samwise_16-03-32	Chryseobacterium_indologenes	12315	1601385
85	Biodiversity_C_ljungdahlii_CO_anaerobic_1_04Oct16_Pippin_16-05-06	Clostridium_ljungdahlii_DMS_13528	20363	1621748
86	Biodiversity_C_ljungdahlii_CO_anaerobic_2_04Oct16_Pippin_16-05-06	Clostridium_ljungdahlii_DMS_13528	21268	1643016
87	Biodiversity_C_ljungdahlii_CO_anaerobic_3_04Oct16_Pippin_16-05-06	Clostridium_ljungdahlii_DMS_13528	20785	1663801
88	Biodiversity_C_ljungdahlii_Fructose_anaerobic_1_04Oct16_Pippin_16-05-06	Clostridium_ljungdahlii_DMS_13528	21315	1685116
89	Biodiversity_C_ljungdahlii_Fructose_anaerobic_2_04Oct16_Pippin_16-05-06	Clostridium_ljungdahlii_DMS_13528	21903	1707019
90	Biodiversity_C_ljungdahlii_Fructose_anaerobic_3_04Oct16_Pippin_16-05-06	Clostridium_ljungdahlii_DMS_13528	22204	1729223
91	Biodiversity_C_necator_R2A_aerobic_1_23Nov16_Pippin_16-09-11	Cupriavidus_necator_N-1	17734	1746957
92	Biodiversity_C_necator_R2A_aerobic_2_23Nov16_Pippin_16-09-11	Cupriavidus_necator_N-1	16630	1763587
93	Biodiversity_C_necator_R2A_aerobic_3_23Nov16_Pippin_16-09-11	Cupriavidus_necator_N-1	16005	1779592
94	Biodiversity_Cellulomonas_gilvus_GS2_01_27Dec15_Arwen_15-07-13	Cellulomonas_gilvus_ATCC13127	24408	1804000
95	Biodiversity_Cellulomonas_gilvus_GS2_02_27Dec15_Arwen_15-07-13	Cellulomonas_gilvus_ATCC13127	24056	1828056
96	Biodiversity_Cellulomonas_gilvus_GS2_03_27Dec15_Arwen_15-07-13	Cellulomonas_gilvus_ATCC13127	23653	1851709
97	Biodiversity_Cibrobacter_freundii_LB_aerobic_01_01Feb16_Arwen_15-07-13	Citrobacter_freundii	23581	1875290
98	Biodiversity_Cibrobacter_freundii_LB_aerobic_02_01Feb16_Arwen_15-07-13	Citrobacter_freundii	22579	1897869

99	Biodiversity_Cibrobacter_freundii_LB_aerobic_03_01Feb16_Arwen_15-07-13	Citrobacter_freundii	23273	1921142
100	Biodiversity_D_acidovorans_TGY_aerobic_01_29Apr16_Samwise_16-03-32_renamed	Delftia_acidovorans_SPH1	21871	1943013
101	Biodiversity_D_acidovorans_TGY_aerobic_02_29Apr16_Samwise_16-03-32_renamed	Delftia_acidovorans_SPH1	20549	1963562
102	Biodiversity_D_acidovorans_TGY_aerobic_03_29Apr16_Samwise_16-03-32_renamed	Delftia_acidovorans_SPH1	19125	1982687
103	Biodiversity_D_longicatena_CarbI_01_26Aug15_Arwen_15-07-13	Dorea_longicatena_DSM13814	23197	2005884
104	Biodiversity_D_longicatena_CarbII_01_26Aug15_Arwen_15-07-13	Dorea_longicatena_DSM13814	22441	2028325
105	Biodiversity_D_longicatena_Glc_01_28Oct15_Arwen_15-07-13	Dorea_longicatena_DSM13814	19551	2047876
106	Biodiversity_F_novicida_TSB_aerobic_01_01Feb16_Arwen_15-07-13	Francisella_novicida_U112	25900	2073776
107	Biodiversity_F_novicida_TSB_aerobic_02_01Feb16_Arwen_15-07-13	Francisella_novicida_U112	23556	2097332
108	Biodiversity_F_novicida_TSB_aerobic_03_01Feb16_Arwen_15-07-13	Francisella_novicida_U112	23189	2120521
109	Biodiversity_F_prausnitzii_Carb_01_28Oct15_Arwen_15-07-13	Faecalibacterium_prausnitzii	11204	2131725
110	Biodiversity_F_prausnitzii_Glc_01_28Oct15_Arwen_15-07-13	Faecalibacterium_prausnitzii	13901	2145626
111	Biodiversity_F_prausnitzii_LIB_01_28Oct15_Arwen_15-07-13	Faecalibacterium_prausnitzii	12858	2158484
112	Biodiversity_F_succinogenes_MDM_01_27Dec15_Arwen_15-07-13	Fibrobacter_succinogenes_S85	20923	2179407
113	Biodiversity_F_succinogenes_MDM_02_27Dec15_Arwen_15-07-13	Fibrobacter_succinogenes_S85	23278	2202685
114	Biodiversity_F_succinogenes_MDM_03_27Dec15_Arwen_15-07-13	Fibrobacter_succinogenes_S85	21665	2224350
115	Biodiversity_HL111_HLHglutamate_aerobic_1_14July16_Pippin_16-05-01	Erythrobacter_HL-111	16707	2241057
116	Biodiversity_HL111_HLHglutamate_aerobic_2_14July16_Pippin_16-05-01	Erythrobacter_HL-111	15709	2256766
117	Biodiversity_HL111_HLHglutamate_aerobic_3_14July16_Pippin_16-05-01	Erythrobacter_HL-111	17636	2274402
118	Biodiversity_HL48_HLHxylose_aerobic_1_09Jun16_Pippin_16-03-39	Halomonas_HL-48	19564	2293966
119	Biodiversity_HL48_HLHxylose_aerobic_2_09Jun16_Pippin_16-03-39	Halomonas_HL-48	11419	2305385

120	Biodiversity_HL48_HLHxylose_aerobic_3_09Jun16_Pippin_16-03-39	Halomonas_HL-48	19558	2324943
121	Biodiversity_HL49_HLHYE_aerobic_1_05Oct16_Pippin_16-05-06	Algoriphagus_marincola_HL-49	20745	2345688
122	Biodiversity_HL49_HLHYE_aerobic_2_05Oct16_Pippin_16-05-06	Algoriphagus_marincola_HL-49	25153	2370841
123	Biodiversity_HL49_HLHYE_aerobic_3_05Oct16_Pippin_16-05-06	Algoriphagus_marincola_HL-49	24520	2395361
124	Biodiversity_HL69_HLA_aerobic_1_05Oct16_Pippin_16-05-06	Cyanobacterium_stanieri	15059	2410420
125	Biodiversity_HL69_HLA_aerobic_2_05Oct16_Pippin_16-05-06	Cyanobacterium_stanieri	14529	2424949
126	Biodiversity_HL69_HLA_aerobic_3_05Oct16_Pippin_16-05-06	Cyanobacterium_stanieri	16531	2441480
127	Biodiversity_HL91_HLHsucrose_aerobic_1_09Jun16_Pippin_16-03-39	Rhodobacteraceae_bacterium_HL-91	18271	2459751
128	Biodiversity_HL91_HLHsucrose_aerobic_2_09Jun16_Pippin_16-03-39	Rhodobacteraceae_bacterium_HL-91	18280	2478031
129	Biodiversity_HL91_HLHsucrose_aerobic_3_09Jun16_Pippin_16-03-39	Rhodobacteraceae_bacterium_HL-91	19857	2497888
130	Biodiversity_HL93_HLHfructose_aerobic_1_09Jun16_Pippin_16-03-39	Halomonas_HL-93	9976	2507864
131	Biodiversity_HL93_HLHfructose_aerobic_2_09Jun16_Pippin_16-03-39	Halomonas_HL-93	9277	2517141
132	Biodiversity_HL93_HLHfructose_aerobic_3_09Jun16_Pippin_16-03-39	Halomonas_HL-93	8961	2526102
133	Biodiversity_L_monocytogenes_BHI_aerobic_01_27Feb17_Pippin_16-11-03	Listeria_monocytogenes_10403S	27172	2553274
134	Biodiversity_L_monocytogenes_BHI_aerobic_02_27Feb17_Pippin_16-11-03	Listeria_monocytogenes_10403S	25972	2579246
135	Biodiversity_L_monocytogenes_BHI_aerobic_03_27Feb17_Pippin_16-11-03	Listeria_monocytogenes_10403S	26591	2605837
136	Biodiversity_Lactobacillus_casei_MRS_01_27Dec15_Arwen_15-07-13	Lactobacillales_casei	11555	2617392
137	Biodiversity_Lactobacillus_casei_MRS_02_27Dec15_Arwen_15-07-13	Lactobacillales_casei	10984	2628376
138	Biodiversity_Lactobacillus_casei_MRS_03_27Dec15_Arwen_15-07-13	Lactobacillales_casei	12133	2640509
139	Biodiversity_M_luteus_LIB_aerobic_01_26Feb16_Arwen_16-01-01	Micrococcus_luteus	14623	2655132
140	Biodiversity_M_luteus_LIB_aerobic_02_26Feb16_Arwen_16-01-01	Micrococcus_luteus	14694	2669826

141	Biodiversity_M_luteus_LIB_aerobic_03_26Feb16_Arwen_16-01-01	Micrococcus_luteus	14865	2684691
142	Biodiversity_M_smegmatis_BHI_aerobic_1_05Oct16_Pippin_16-05-06	Mycobacterium_smegmatis	22302	2706993
143	Biodiversity_M_smegmatis_BHI_aerobic_2_05Oct16_Pippin_16-05-06	Mycobacterium_smegmatis	23937	2730930
144	Biodiversity_M_smegmatis_BHI_aerobic_3_05Oct16_Pippin_16-05-06	Mycobacterium_smegmatis	23123	2754053
145	Biodiversity_M_xanthus_DZ2_24h_plates_1_13Jun16_Pippin_16-03-39	Myxococcus_xanthus_DZ2	17676	2771729
146	Biodiversity_M_xanthus_DZ2_24h_plates_2_13Jun16_Pippin_16-03-39	Myxococcus_xanthus_DZ2	18291	2790020
147	Biodiversity_M_xanthus_DZ2_24h_plates_3_13Jun16_Pippin_16-03-39	Myxococcus_xanthus_DZ2	17586	2807606
148	Biodiversity_M_xanthus_DZ2_48h_plates_1_13Jun16_Pippin_16-03-39	Myxococcus_xanthus_DZ2	20435	2828041
149	Biodiversity_M_xanthus_DZ2_48h_plates_2_13Jun16_Pippin_16-03-39	Myxococcus_xanthus_DZ2	18715	2846756
150	Biodiversity_M_xanthus_DZ2_48h_plates_3_13Jun16_Pippin_16-03-39	Myxococcus_xanthus_DZ2	19998	2866754
151	Biodiversity_M_xanthus_DZ2_pellet_1_03May16_Samwise_16-03-32	Myxococcus_xanthus_DZ2	24459	2891213
152	Biodiversity_M_xanthus_DZ2_pellet_2_03May16_Samwise_16-03-32	Myxococcus_xanthus_DZ2	24181	2915394
153	Biodiversity_M_xanthus_DZ2_plates_1_03May16_Samwise_16-03-32	Myxococcus_xanthus_DZ2	18520	2933914
154	Biodiversity_M_xanthus_DZ2_plates_2_03May16_Samwise_16-03-32	Myxococcus_xanthus_DZ2	17909	2951823
155	Biodiversity_M_xanthus_pellet_set2_1_13Jun16_Pippin_16-03-39	Myxococcus_xanthus_DZ2	23354	2975177
156	Biodiversity_M_xanthus_pellet_set2_2_13Jun16_Pippin_16-03-39	Myxococcus_xanthus_DZ2	23884	2999061
157	Biodiversity_M_xanthus_pellet_set2_3_13Jun16_Pippin_16-03-39	Myxococcus_xanthus_DZ2	22772	3021833
158	Biodiversity_P_denitrificans_LIB_aerobic_01_29Apr16_Samwise_16-03-32_renamed	Paracoccus_denitrificans	22165	3043998
159	Biodiversity_P_denitrificans_LIB_aerobic_02_29Apr16_Samwise_16-03-32_renamed	Paracoccus_denitrificans	20888	3064886
160	Biodiversity_P_denitrificans_LIB_aerobic_03_29Apr16_Samwise_16-03-32_renamed	Paracoccus_denitrificans	23115	3088001

161	Biodiversity_P_hydrogenalis_01_28Jul15_Arwen_14-12-03	Anaerococcus_hydrogenalis_DSM_7454	18519	3106520
162	Biodiversity_P_hydrogenalis_CMgluc_anaerobic_01_26Feb16_Arwen_16-01-01	Anaerococcus_hydrogenalis_DSM_7454	12813	3119333
163	Biodiversity_P_hydrogenalis_CMgluc_anaerobic_02_26Feb16_Arwen_16-01-01	Anaerococcus_hydrogenalis_DSM_7454	13371	3132704
164	Biodiversity_P_hydrogenalis_CMgluc_anaerobic_03_26Feb16_Arwen_16-01-01	Anaerococcus_hydrogenalis_DSM_7454	12649	3145353
165	Biodiversity_P_polymyxa_TBS_aerobic_1_17July16_Samwise_16-04-10	Paenibacillus_polymyxa_ATCC842	25623	3170976
166	Biodiversity_P_polymyxa_TBS_aerobic_2_17July16_Samwise_16-04-10	Paenibacillus_polymyxa_ATCC842	25057	3196033
167	Biodiversity_P_polymyxa_TBS_aerobic_3_17July16_Samwise_16-04-10	Paenibacillus_polymyxa_ATCC842	24268	3220301
168	Biodiversity_P_ruminicola_MDM_anaerobic_1_09Jun16_Pippin_16-03-39	Prevotella_ruminicola_23_ATCC_19189	17277	3237578
169	Biodiversity_P_ruminicola_MDM_anaerobic_2_09Jun16_Pippin_16-03-39	Prevotella_ruminicola_23_ATCC_19189	17543	3255121
170	Biodiversity_R_gnavus_01_28Jul15_Arwen_14-12-03	Ruminococcus_gnavus	20132	3275253
171	Biodiversity_R_gnavus_Carb_01_28Oct15_Arwen_15-07-13	Ruminococcus_gnavus	22004	3297257
172	Biodiversity_R_jostii_R2A_aerobic_1_23Nov16_Pippin_16-09-11	Rhodococcus_jostii_RHA1	24374	3321631
173	Biodiversity_R_jostii_R2A_aerobic_2_23Nov16_Pippin_16-09-11	Rhodococcus_jostii_RHA1	23736	3345367
174	Biodiversity_R_jostii_R2A_aerobic_3_23Nov16_Pippin_16-09-11	Rhodococcus_jostii_RHA1	22296	3367663
175	Biodiversity_R_palustris_PM_aerobic_1_01Jun16_Pippin_16-03-39	Rhodopseudomonas_palustris	21988	3389651
176	Biodiversity_R_palustris_PM_aerobic_2_01Jun16_Pippin_16-03-39	Rhodopseudomonas_palustris	21998	3411649
177	Biodiversity_R_palustris_PM_aerobic_3_01Jun16_Pippin_16-03-39	Rhodopseudomonas_palustris	20740	3432389
178	Biodiversity_R_palustris_PMnitro_anaerobic_1_01Jun16_Pippin_16-03-39	Rhodopseudomonas_palustris	20820	3453209
179	Biodiversity_R_palustris_PMnitro_anaerobic_2_01Jun16_Pippin_16-03-39	Rhodopseudomonas_palustris	20800	3474009
180	Biodiversity_R_palustris_PMnitro_anaerobic_3_01Jun16_Pippin_16-03-39	Rhodopseudomonas_palustris	19401	3493410
181	Biodiversity_R_palustris_PMnonnitro_anaerobic_1_01Jun16_P	Rhodopseudomonas_palustris	21220	3514630

	ippin_16-03-39			
182	Biodiversity_R_palustris_PMnonnitro_anaerobic_2_01Jun16_Pippin_16-03-39	Rhodopseudomonas_palustris	21947	3536577
183	Biodiversity_R_palustris_PMnonnitro_anaerobic_3_01Jun16_Pippin_16-03-39	Rhodopseudomonas_palustris	20793	3557370
184	Biodiversity_S_agalactiae_LIB_aerobic_01_26Feb16_Arwen_16-01-01	Streptococcus_agalactiae	12558	3569928
185	Biodiversity_S_agalactiae_LIB_aerobic_02_26Feb16_Arwen_16-01-01	Streptococcus_agalactiae	11366	3581294
186	Biodiversity_S_agalactiae_LIB_aerobic_03_26Feb16_Arwen_16-01-01	Streptococcus_agalactiae	11845	3593139
187	Biodiversity_S_aurantiaca_CYE_aerobic_1_17July16_Samwise_16-04-10	Stigmatella_aurantiaca_DW431	26687	3619826
188	Biodiversity_S_aurantiaca_CYE_aerobic_2_17July16_Samwise_16-04-10	Stigmatella_aurantiaca_DW431	25198	3645024
189	Biodiversity_S_aurantiaca_CYE_aerobic_3_17July16_Samwise_16-04-10	Stigmatella_aurantiaca_DW431	28243	3673267
190	Biodiversity_S_elongatus_BG11_aerobic_1_14July16_Pippin_16-05-01	Synechococcus_elongatus_PCC 7942	16601	3689868
191	Biodiversity_S_elongatus_BG11_aerobic_2_14July16_Pippin_16-05-01	Synechococcus_elongatus_PCC 7942	16512	3706380
192	Biodiversity_S_elongatus_BG11_aerobic_3_14July16_Pippin_16-05-01	Synechococcus_elongatus_PCC 7942	16383	3722763
193	Biodiversity_S_elongatus_BG11NaCl_aerobic_1_05Oct16_Pippin_16-05-06	Synechococcus_elongatus_PCC 7942	18618	3741381
194	Biodiversity_S_elongatus_BG11NaCl_aerobic_2_05Oct16_Pippin_16-05-06	Synechococcus_elongatus_PCC 7942	19025	3760406
195	Biodiversity_S_elongatus_BG11NaCl_aerobic_3_05Oct16_Pippin_16-05-06	Synechococcus_elongatus_PCC 7942	18419	3778825
196	Biodiversity_S_griseorubens_HSM_aerobic_1_23Nov16_Pippin_16-09-11	Streptomyces_griseorubens	7798	3786623
197	Biodiversity_S_griseorubens_HSM_aerobic_2_23Nov16_Pippin_16-09-11	Streptomyces_griseorubens	7869	3794492
198	Biodiversity_S_griseorubens_HSM_aerobic_3_23Nov16_Pippin_16-09-11	Streptomyces_griseorubens	5960	3800452
199	Biodiversity_S_thermosulf_FeYE_anaerobic_1_01Jun16_Pippin_16-03-39	Sulfobacillus_thermosulfidooxidans	14607	3815059
200	Biodiversity_S_thermosulf_FeYE_anaerobic_2_01Jun16_Pippin_16-03-39	Sulfobacillus_thermosulfidooxidans	14762	3829821

201	Biodiversity_S_thermosulf_FeYE_anaerobic_3_01Jun16_Pippin_16-03-39	Sulfobacillus_thermosulfidooxidans	15862	3845683
202	Cj_media_MH_R1_23Feb15_Arwen_14-12-03	Campylobacter_jejuni	24941	3870624
203	Cj_media_MH_R2_23Feb15_Arwen_14-12-03	Campylobacter_jejuni	24931	3895555
204	Cj_media_MH_R3_23Feb15_Arwen_14-12-03	Campylobacter_jejuni	21400	3916955
205	Cj_media_MH_R4_23Feb15_Arwen_14-12-03	Campylobacter_jejuni	25037	3941992
206	Cj_media_MH_R5_23Feb15_Arwen_14-12-03	Campylobacter_jejuni	20168	3962160
207	LP_LS_Phi_Stat_R1_30Sep14_Pippin_13-04-12	Legionella_pneumophila	26779	3988939
208	LP_LS_Phi_Stat_R2_30Sep14_Pippin_13-04-12	Legionella_pneumophila	27771	4016710
209	LP_LS_Phi_Stat_R3_30Sep14_Pippin_13-04-12	Legionella_pneumophila	25938	4042648
210	P_putida_01Dec15_1_21Mar16_Arwen_16-01-03	Pseudomonas_putida_KT2440	23697	4066345
211	P_putida_01Dec15_2_21Mar16_Arwen_16-01-03	Pseudomonas_putida_KT2440	24322	4090667
212	P_putida_17Nov15_1_21Mar16_Arwen_16-01-03	Pseudomonas_putida_KT2440	22904	4113571
213	P_putida_17Nov15_2_21Mar16_Arwen_16-01-03	Pseudomonas_putida_KT2440	22395	4135966
214	P_putida_18Nov15_1_21Mar16_Arwen_16-01-03	Pseudomonas_putida_KT2440	22246	4158212
215	P_putida_18Nov15_2_21Mar16_Arwen_16-01-03	Pseudomonas_putida_KT2440	22937	4181149
216	S_venezuelae_GYM_1_21Mar16_Arwen_16-01-03	Streptomyces_venezuelae	12030	4193179
217	S_venezuelae_GYM_2_21Mar16_Arwen_16-01-03	Streptomyces_venezuelae	11099	4204278
218	S_venezuelae_MYM_1_21Mar16_Arwen_16-01-03	Streptomyces_venezuelae	14276	4218554
219	S_venezuelae_MYM_2_21Mar16_Arwen_16-01-03	Streptomyces_venezuelae	14322	4232876
220	QC_Shew_13_05_500ng_2_100uL_5hr_30Mar14_Samwise_13-07-17	Shewanella_oneidensis_MR-1	49123	4281999
221	QC_Shew_13_05_500ng_2_5hr_19Mar14_Samwise_13-07-17	Shewanella_oneidensis_MR-1	50274	4332273
222	QC_Shew_13_05_500ng_2_5hr_24Mar14_Samwise_13-07-17	Shewanella_oneidensis_MR-1	50273	4382546
223	M_alcali_copp_CH4_B1_T1_07_QE_23Mar18_Oak_18-01-07	Methylomicrobium_alcaliphilum	21112	4403658
224	M_alcali_copp_CH4_B1_T2_08_QE_23Mar18_Oak_18-01-07	Methylomicrobium_alcaliphilum	21074	4424732
225	M_alcali_copp_CH4_B2_T1_09_QE_23Mar18_Oak_18-01-07	Methylomicrobium_alcaliphilum	18470	4443202
226	M_alcali_copp_CH4_B2_T2_10_QE_23Mar18_Oak_18-01-07	Methylomicrobium_alcaliphilum	18257	4461459

227	M_alcali_copp_CH4_B3_T1_11_QE_23Mar18_Oak_18-01-07	Methylomicrobium_alcaliphilum	20464	4481923
228	M_alcali_copp_CH4_B3_T2_12_QE_23Mar18_Oak_18-01-07	Methylomicrobium_alcaliphilum	20368	4502291
229	M_alcali_copp_MeOH_B1_T1_01_QE_23Mar18_Oak_18-01-07	Methylomicrobium_alcaliphilum	18067	4520358
230	M_alcali_copp_MeOH_B1_T2_02_QE_23Mar18_Oak_18-01-07	Methylomicrobium_alcaliphilum	18229	4538587
231	M_alcali_copp_MeOH_B2_T1_03_QE_23Mar18_Oak_18-01-07	Methylomicrobium_alcaliphilum	17445	4556032
232	M_alcali_copp_MeOH_B2_T2_04_QE_23Mar18_Oak_18-01-07	Methylomicrobium_alcaliphilum	18191	4574223
233	M_alcali_copp_MeOH_B3_T1_05_QE_23Mar18_Oak_18-01-07	Methylomicrobium_alcaliphilum	15781	4590004
234	M_alcali_copp_MeOH_B3_T2_06_QE_23Mar18_Oak_18-01-07	Methylomicrobium_alcaliphilum	14536	4604540
235	Alverdy_Efae_1A_lys_13Jul13_Pippin_12-12-39	Enterococcus_faecalis	16181	4620721
236	Alverdy_Efae_1B_lys_13Jul13_Pippin_12-12-39	Enterococcus_faecalis	16055	4636776
237	Alverdy_Efae_1C_lys_13Jul13_Pippin_12-12-39	Enterococcus_faecalis	15814	4652590
238	Biodiversity_A_muciniphila_test_27Feb17_Pippin_16-11-03	Akkermansia_muciniphila_ATCC_BAA-835	21214	4673804
239	Ha_150NaCl_1_13_QE_12Aug15_Arwen_14-12-03	Halanaerobium_congolense	27902	4701706
240	Ha_150NaCl_2_14_QE_12Aug15_Arwen_14-12-03	Halanaerobium_congolense	29961	4731667
241	Ha_150NaCl_3_15_QE_12Aug15_Arwen_14-12-03	Halanaerobium_congolense	27892	4759559
242	Ha_200NaCl_1_22_QE_21Jan16_Arwen_15-07-13	Halanaerobium_congolense	29705	4789264
243	Ha_200NaCl_2_23_QE_21Jan16_Arwen_15-07-13	Halanaerobium_congolense	28807	4818071
244	Ha_200NaCl_3_24_QE_21Jan16_Arwen_15-07-13	Halanaerobium_congolense	29544	4847615
245	Ha_250NaCl_1_16_QE_12Aug15_Arwen_14-12-03	Halanaerobium_congolense	15326	4862941
246	Ha_250NaCl_2_17_QE_12Aug15_Arwen_14-12-03	Halanaerobium_congolense	19355	4882296
247	Ha_250NaCl_3_18_QE_12Aug15_Arwen_14-12-03	Halanaerobium_congolense	22461	4904757
248	YJ_Cc_WT1_C_P_9Jan17_Pippin_16-09-11	Caulobacter_crescentus_NA1000	22820	4927577
249	YJ_Cc_WT1_IM_P_9Jan17_Pippin_16-09-11	Caulobacter_crescentus_NA1000	19081	4946658
250	YJ_Cc_WT1_OM_P_9Jan17_Pippin_16-09-11	Caulobacter_crescentus_NA10	26060	4972718

		00		
251	YJ_Cc_WT1_P_Prot_9Jan17_Pippin_16-09-11	Caulobacter_crescentus_NA1000	21323	4994041
252	YJ_Cc_WT1_WC_P_9Jan17_Pippin_16-09-11	Caulobacter_crescentus_NA1000	17975	5012016
253	YJ_Cc_WT2_C_P_9Jan17_Pippin_16-09-11	Caulobacter_crescentus_NA1000	22620	5034636
254	YJ_Cc_WT2_IM_P_9Jan17_Pippin_16-09-11	Caulobacter_crescentus_NA1000	19881	5054517
255	YJ_Cc_WT2_OM_P_9Jan17_Pippin_16-09-11	Caulobacter_crescentus_NA1000	20632	5075149
256	YJ_Cc_WT2_P_Prot_9Jan17_Pippin_16-09-11	Caulobacter_crescentus_NA1000	24577	5099726
257	YJ_Cc_WT2_WC_P_9Jan17_Pippin_16-09-11	Caulobacter_crescentus_NA1000	16579	5116305

Supplementary Table 2. Top 20 of EC numbers most frequently matched from the unique peptides using Unipept 4.3 with the identified peptide sequences.

EC number	Name	PepCounts
EC:2.7.7.6	DNA-directed RNA polymerase	755
EC:7.1.2.2	H(+)-transporting two-sector ATPase	688
EC:3.6.5.3	Protein-synthesizing GTPase	269
EC:2.7.13.3	Histidine kinase	263
EC:3.6.4.12	DNA helicase	230
EC:3.6.4.-	Acting on ATP; involved in cellular and subcellular movement	208
EC:5.2.1.8	Peptidylprolyl isomerase	208
EC:6.3.1.2	Glutamine synthetase	200
EC:2.7.7.7	DNA-directed DNA polymerase	198
EC:1.2.1.-	With NAD(+) or NADP(+) as acceptor	194
EC:3.6.3.-	Acting on acid anhydrides; catalyzing transmembrane movement of substances	169
EC:2.7.7.8	Polyribonucleotide nucleotidyltransferase	158
EC:3.1.-.-	Acting on ester bonds	153

EC:1.1.2.-	With a cytochrome as acceptor	142
EC:3.6.3.14	Transferred entry: 7.1.2.2	140
EC:1.1.1.100	3-oxoacyl-[acyl-carrier-protein] reductase	136
EC:4.2.1.3	Aconitate hydratase	133
EC:1.-.-.-	Oxidoreductases	130
EC:1.1.1.37	Malate dehydrogenase	122
EC:2.3.1.9	Acetyl-CoA C-acetyltransferase	116

Supplementary Table 3. Functional distribution of the unique peptides across the phyla taxa. Each row and column represent the different EC numbers and phyla, respectively. The number in each cell indicates the number of unique peptides annotated by Unipept. This table was transformed from the original output file provided by Unipept.
(It's too big to add here so it will be attached as an Excel file)