Proteomics of natural bacterial isolates powered by deep learning-based *de novo* identification.

Joon-Yong Lee[1], Hugh D. Mitchell[1], Meagan C. Burnet[1], Sarah C. Jenson[2], Eric D. Merkley[2], Anil K. Shukla[1], Ernesto S. Nakayasu[1], Samuel H. Payne[3*]

1. Biological Sciences Division, Pacific Northwest National Laboratory, Richland WA
2. Signature Sciences and Technology Division, Pacific Northwest National Laboratory, Richland WA
3. Biology Department, Brigham Young University, Provo UT
* contact: sam_payne@byu.edu

Abstract

The fundamental task in proteomic mass spectrometry is identifying peptides from their observed spectra. Where protein sequences are known, standard algorithms utilize these to narrow the list of peptide candidates. If protein sequences are unknown, a distinct class of algorithms must interpret spectra *de novo.* Despite decades of effort on algorithmic constructs and machine learning methods, *de novo* software tools remain inaccurate when used on environmentally diverse samples. Here we train a deep neural network on 5 million spectra from 55 phylogenetically diverse bacteria. This new model outperforms current methods by 25-100%. The diversity of organisms used for training also improves the generality of the model, and ensures reliable performance regardless of where the sample comes from. Significantly, it also achieves a high accuracy in long peptides which assist in identifying taxa from samples of unknown origin. With the new tool, called Kaiko, we analyze proteomics data from six natural soil isolates for which a proteome database did not exist. Without any sequence information, we correctly identify the taxonomy of these soil microbes as well as annotate thousands of peptide spectra.

Background and Summary

Machine learning is an important tool for developing the complex mathematical and statistical models for proteomics data analysis. The most common application of machine learning within computational proteomics is the design and implementation of a scoring metric that determines the quality of a peptide/spectrum match, a topic that has been rigorously studied for more than 25 years. The primary task of these tools is to rank candidate peptides as to how likely they are to have generated an observed spectrum, and then to identify the peptide/spectrum matches considered to be correct. A wide variety of machine learning methods have been explored to help parameterize these models, including: support vector machines[1], Bayesian networks[2], rank-based scoring[3], semi-supervised learning[4] and many other techniques[5–8].

The most successful proteomics algorithms also rely on a protein sequence database, which limits the peptide candidates for scoring[9]. Historically, these algorithms have been much more accurate than methods which do not use a sequence database, called *de novo* algorithms[10,11] . This reliance on a database is entirely appropriate in the post-genomic era where an increasing number of organisms have an annotated genome. However, there remain several biologically and environmentally essential research areas where a complete and accurate protein database is still unlikely, including: antibodies produced by programmed genomic hypermutation, environmental samples, forensics, and natural isolates. For situations such as these, the proteomics community still relies on alternative computational methods like *de novo* spectrum annotation[12,13].

Recent advances in both algorithms and computational infrastructure have enabled a breakthrough in deep neural network-based machine learning, often simply called deep learning[14]. These breakthroughs have revolutionized capabilities in speech[15] and image recognition[16], language translation[17], and many other challenging computational problems. Deep learning has also recently been used in proteomics[18,19].

One significant challenge with deep learning is that the immense number of parameters in deep neural networks requires very large training data to avoid overfitting and to make a model which generalizes well to unseen data. The ProteomeXchange consortium[20] hosts data for the proteomics community, however, mining these repositories to gather sufficiently large training and testing data is hampered by two factors. First, the diversity of data, specifically the number of distinct peptides, is low compared to the overall volume of data. This is due to the fact that many datasets are deposited on the same model systems (e.g. *Homo sapiens*) and therefore identify many of the same peptides. Second, the datasets are submitted by a wide variety of labs with an equally large variety in methodology, instrumentation, data quality and meta-data. Thus, aggregating datasets to amass an appropriate number of distinct peptides is a time-consuming and laborious adventure.

We present both a sufficiently large and diverse benchmark dataset for deep learning, and our *de novo* algorithm trained on these data. To address the challenge of dataset size, we created bottom-up proteomics data from 55 phylogenetically diverse bacteria, which produced over 1

million confidently identified peptides from more than 5 million spectra. Using these data, we train a deep neural network, called Kaiko, and achieve a significant improvement in spectrum annotation accuracy relative to the most recent and best performing *de novo* tools. Finally, we use the tool to analyze proteomics data and correctly identify the taxonomy of six natural soil isolate bacteria without using any sequence information.

Methods

Data generation

Cell culture and sample preparation. The growth, sample preparation and data collection was reported previously[21]. 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 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[21] 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[22]. 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. 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[21]. 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[18]. 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.

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 Supplemental Figure 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[18], PEAKS[11] and Novor[23]. 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. 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 MGF format using MSConvert. 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[24] 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.

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.

Code Availability

All software used in this project is open source under the BSD license and available on GitHub. The Kaiko tool is available at https://github.com/PNNL-Comp-Mass-Spec/Kaiko. The Jupyter notebooks and R code used to analyze the results and create figures are available at https://github.com/PNNL-Comp-Mass-Spec/Kaiko_Publication.

Results

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. Unlike many of the established tools for *de novo* identification, our tool is not a dynamic programming[10,11,26] or decision tree approach[23]. Instead, following DeepNovo[18], we have used a deep neural network to let the algorithm learn about the data and generate a scoring method on its own. Using these learned parameters, the neural network compares peptide sequences against the observed spectrum. Candidate peptide sequences are not limited by a database or any other outside information; they are simply chosen from the combinatorial set of all 20^N possible amino acid sequences. The top scoring candidate is reported as the best peptide/spectrum match. Our newly trained deep neural network model is called Kaiko, after the Japanese deep ocean submersible used to explore the Marianas Trench.

The structure of our neural network began with the recently published and open source DeepNovo method[18]; however, adaptations have been made to improve model performance (see Methods). Importantly, we have amassed a dramatically larger dataset, which is necessary for meaningful training of deep neural networks. For training and validation, we use 4,604,540 spectra and 927,316 peptides from 51 distinct bacteria (Figure 1, Supplemental Table 1). Data from four additional organisms were held out for testing of the final model (511,765 spectra and 90,048 peptides). As Kaiko is a deep neural network, it is essential to have sufficient training data for parameter optimization. Training a deep neural network with insufficient data leads to poor overall performance (Supplemental Figure 1). For our neural network architecture, training events with less than 3 million spectra resulted in severely overfit models (Supplemental Figure 2).

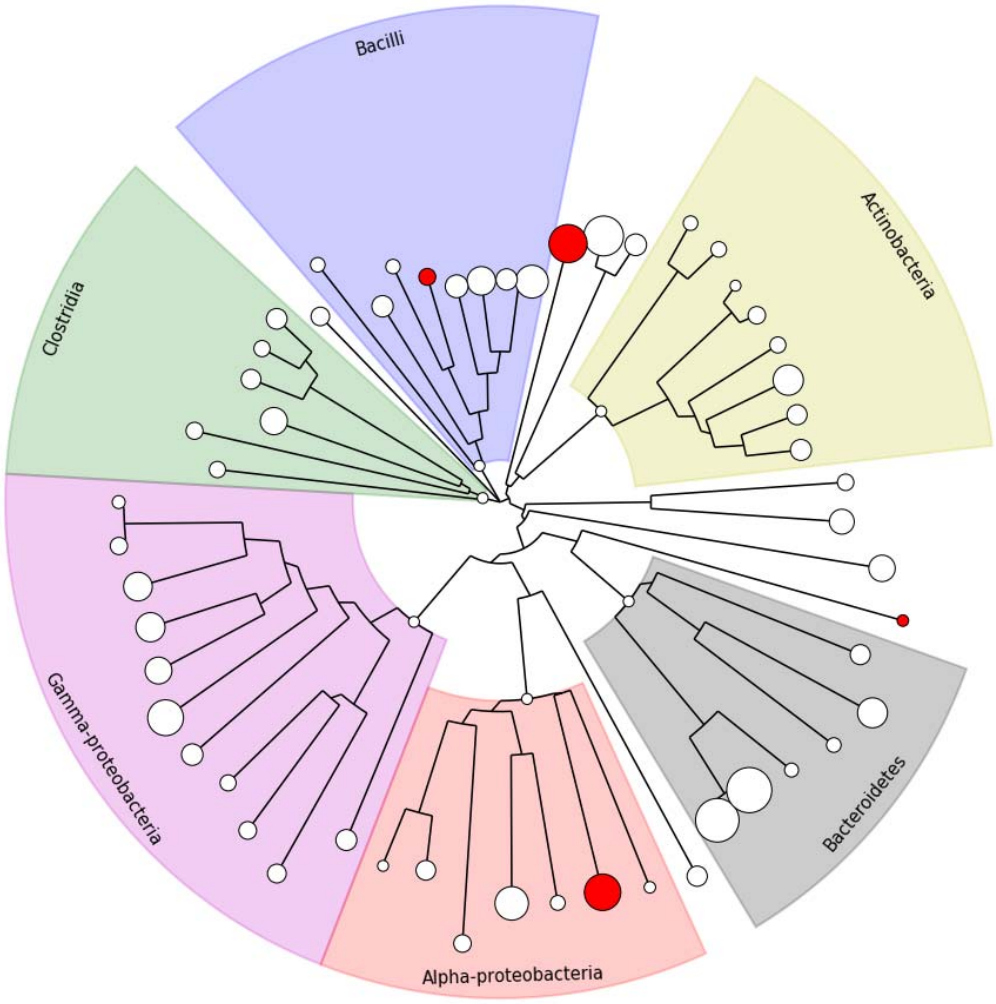

Figure 1 - Bacteria represented in training and testing data. A phylogenetic tree built from the multiple sequence alignment of rplB is shown for all organisms in the training (white nodes) and testing datasets (red nodes). The size of the node is scaled to represent the number of spectra used. Large taxonomic divisions as defined by NCBI's taxonomy are colored for convenience.

Predicting correct peptide sequences

The primary metric for spectrum identification is whether the algorithm scores the correct peptide sequence as the best peptide/spectrum match among the many possible candidates. To evaluate the accuracy of Kaiko and other *de novo* algorithms, we compare their top result for each spectrum in the test dataset to the answer derived by MSGF+[27], a database search scoring algorithm (see Methods). Test spectra are from four organisms, and are represented by multiple mass spectrometry analyses from replicates and/or different experimental conditions. Each spectral file was analyzed by Kaiko, PEAKS[11], Novor[23] and DeepNovo[18]. For each analysis, we simply compute the percent of spectra that are correctly annotated. Kaiko achieved an average accuracy of 33% over all testing files and organisms (Figure 2). When considering the top five spectrum annotations, average accuracy exceeded 41%.

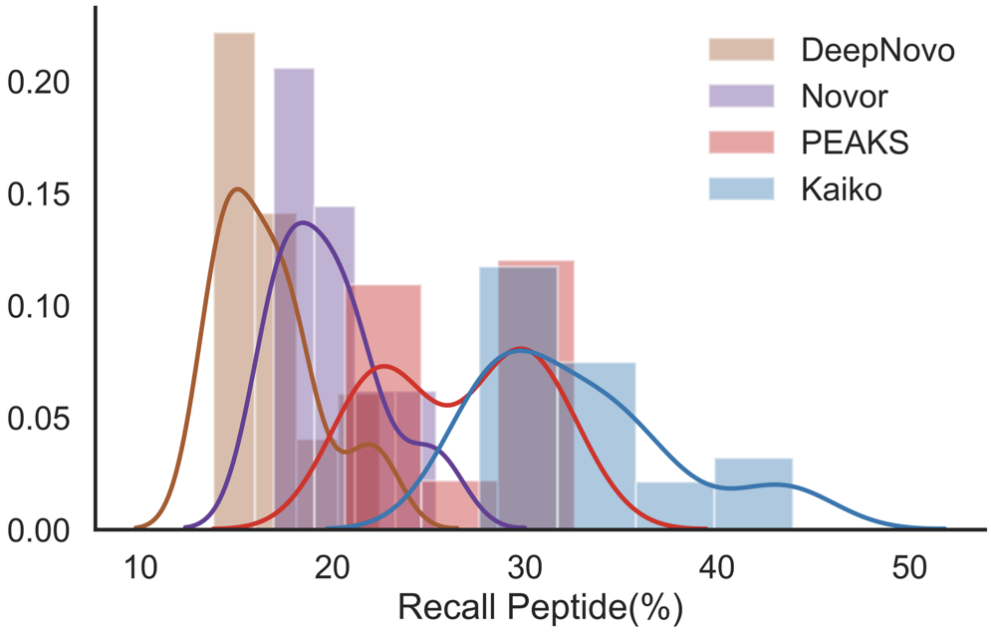

Figure 2 - Accuracy of spectrum annotation. Four *de novo* spectrum annotation tools were benchmarked against the testing data, which was bottom-up proteomics data from four diverse bacterial species. The accuracy of each program is shown in the bar histogram. The line series represents an approximated continuous distribution of the histogram.

We next looked at model performance as a function of peptide length (Figure 3). Most algorithms performed well with short peptides, length < 8. Unfortunately, these peptides are infrequent in bottom-up proteomics data samples (Supplementary Figure 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.

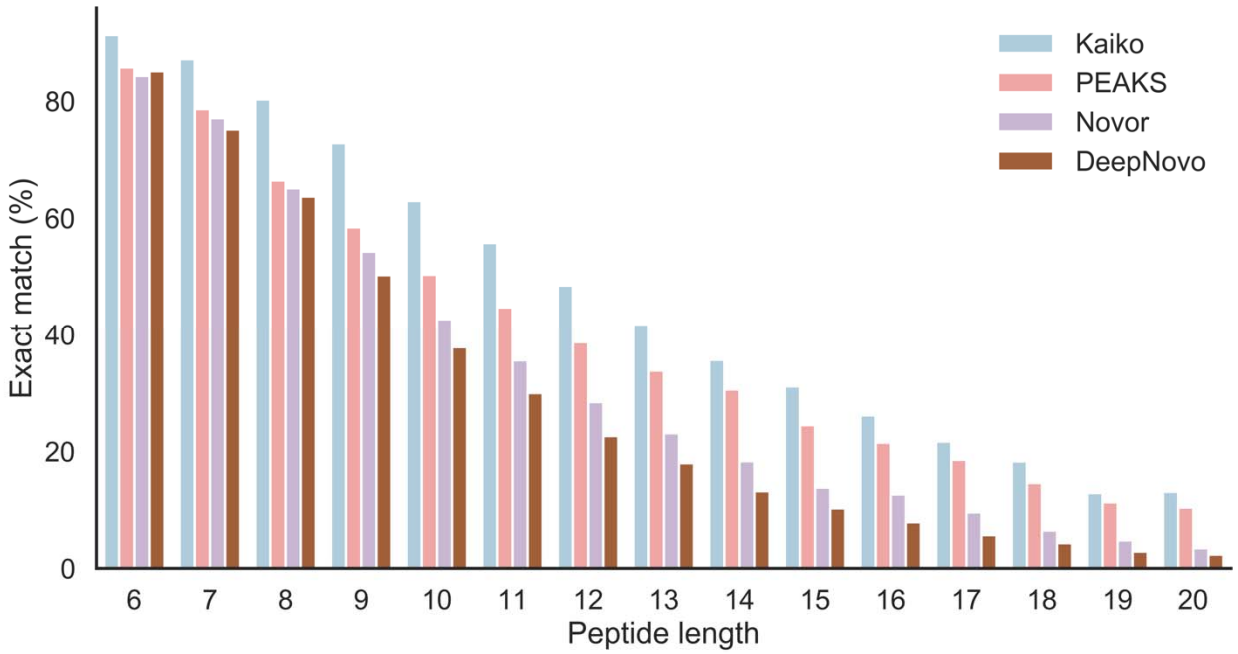

Figure 3 - Sequence prediction accuracy by peptide length. For each peptide sequence length, the accuracy of spectrum annotation is shown for each of the four algorithms.

Performance differences on diverse species

One important aspect of Kaiko's model is its generality, meaning the ability to perform well on diverse samples. As a measure of how representative the test data is of Kaiko's future use on any given organisms, we looked at the relatedness of the test set compared to the organisms within the training set. For *Enterococcus faecalis*, its closest relative within the training set is *Streptococcus agalactiae*; both are from the order Lactobacillales. These two organisms diverged ~1 billion years ago[28]. (For reference, humans and the plant model organism *Arabidopsis* diverged 1.6 billion years ago[29]). This is the closest pair of organisms between the training and testing datasets. The proteomics data from these two organisms share 306 peptides, or 1.3%. At the other extreme, the test set contains *Akkermansia muciniphila*, which is from the bacterial phylum Verrucomicrobia. There are no members of this phylum in the training set. This means that the closest relative within our training set for *A. muciniphila* diverged over 3 billion years ago, near the bacterial phyla radiation. Although Kaiko was more accurate for *E. faecalis* than *A. muciniphila* (42% accuracy compared to 30% respectively), it should be noted that it performed equally well for *A. muciniphila* as for *Caulobacter crescentus* and *Halanaerobium congolense* (30%, 28% and 34%, respectively). Both *C. crescentus* and *H. congolense* are related to organisms within the training set at the level of class. Therefore, we believe that Kaiko will perform consistently well regardless of the organism and for peptides not present in the training set.

Predicting correct substrings

The imperfect fragmentation of peptides in the mass spectrometer leads to the incomplete observation of b/y fragment ions, a situation that becomes more likely as peptide length

increases. Spectral interpretation can also be complicated by co-fragmenting peptides[30,31]. Therefore, for some spectra it is not a realistic expectation to identify the full and complete sequence. In this situation, a long substring that is correct is an important goal, as it can be used for partial database matching, sequence tag searching, or blast-like searches of similar sequences [32–35].

For each spectrum, we recorded the longest correct substring from the *de novo* prediction (Figure 4). For length 9 peptides, all algorithms annotated at least 50% of the spectra correctly. Kaiko was significantly better than other algorithms, annotating 72% of spectra correctly; the next best algorithms, PEAKS, annotated 58% of spectra correctly. As peptide length increased, accuracy decreased across all algorithms. However, Kaiko was the best performer at peptides of any length. On average Kaiko correctly annotated 25% more spectra than PEAKS, and 50-100% more than Novor or DeepNovo.

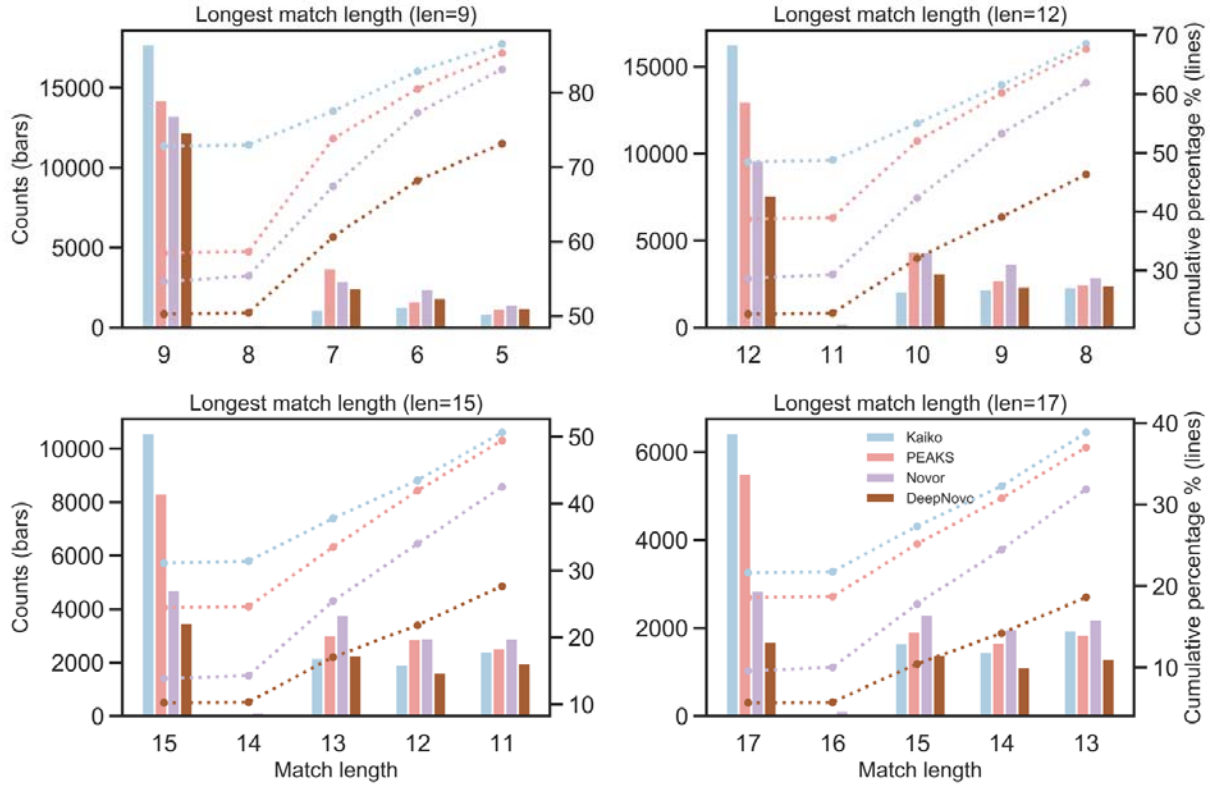

Figure 4 - Longest correct substring. Algorithm performance was assessed by identifying the longest correct substring within the predicted peptide sequence. The bar charts (y-axis legend on the left) shows the raw number of identifications with a correct match length. The dashed line charts (y-axis legend on the right) shows the cumulative percentage of identifications relative to all peptides of a given length.

Proteomics of Natural Isolates

Proteomics analysis of natural bacterial isolates often requires *de novo* spectrum annotation. Although genome and metagenome sequencing has greatly expanded the number of species that contain a complete proteome database, 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. To show the ability of our deep learning-based algorithm to annotate spectra from an 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 spectra with Kaiko and used DIAMOND to identify the closest sequences in the Uniprot database (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 (Figure 5). 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 is no matched proteome and therefore, taxonomic placement through BLAST-like sequence matching is more complicated to interpret. Isolate02 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, Isolate01, is suggested to be a *Gemmobacter* 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[36], this is still a close match.

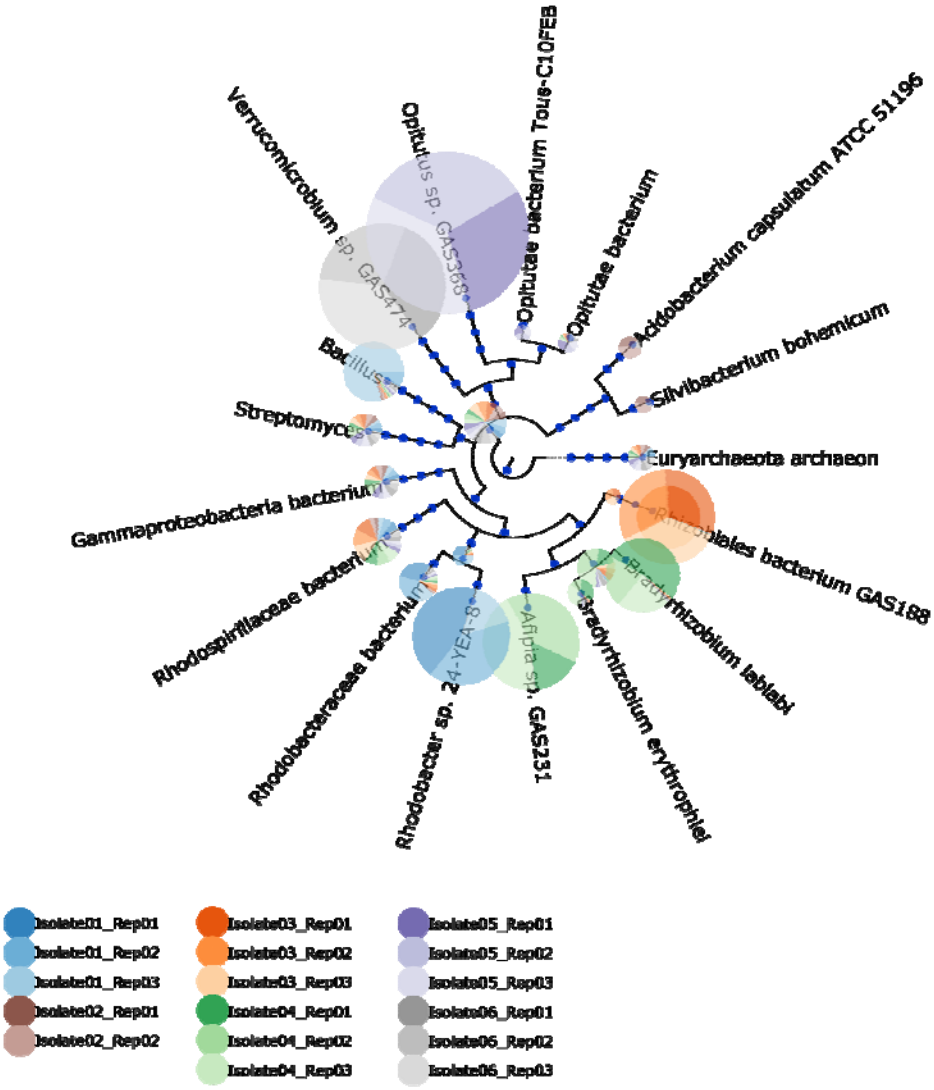

Figure 5 - Taxonomic classification of natural isolates via proteomics. For each of the six natural isolates, replicate proteomics data was annotated with Kaiko. Peptides were matched to sequences in Uniprot and visualized on the phylogenetic tree. Nodes within the tree are from NCBI's Linnaean taxonomy definition of each organism. The size of 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.

Conclusions

We present here a *de novo* spectrum annotation tool, Kaiko, which has significantly improved accuracy compared to other *de novo* algorithms. An important feature of the tool is that it was trained on spectra from a very diverse set of organisms and therefore performs well on the diverse set of species found in natural environments. This is an important step forward for meta-

proteomics, which has long struggled to identify spectra from microbial communities. To demonstrate the application of *de novo* sequencing on data from natural sources, we used Kaiko to identify the taxon of samples from bacterial soil isolates. Kaiko succeeded at scenarios which are expected in natural samples, i.e. samples from phyla where no training data existed and samples from organisms whose taxonomy is currently ambiguous and/or have no sequence representation in public databases.

As machine learning continues to gain momentum in the life sciences and specifically within bioinformatics, it is important to have readily available benchmark datasets so that new tools and computational methodologies can compare themselves to earlier work. Moreover, it is important that the benchmark data be sufficiently large to properly train models and that it represents the diversity of data which could be collected. In this sense, the training and testing spectra presented here, which are publicly available through PRIDE, represents an ideal for mass spectrometry proteomics. Spectra come from 55 bacteria which span greater than 3 billion years of protein sequence evolution. The 5 million spectra represent over 1 million peptide sequences and are all acquired on modern high-resolution instrumentation. We also note that such a highly curated and diverse dataset can be used to improve performance of a wide variety of computational methods beyond *de novo* spectrum identification[37].

Acknowledgments

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 PNNL's Deep Learning for Scientific Discovery initiative and the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Early Career Research Program. 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.

Author contributions

J-Y. L. and S.H.P. conceived and designed the experiment. J-Y.L. wrote software and trained the deep learning model. M.C.B., A.K.S. and E.S.N. prepared samples and acquired mass spectrometry data. J-Y.L., S.C.J. and E.D.M. profiled 3rd party algorithms. J-Y.L., H.D.M., E.S.N. and S.H.P. analyzed data. J-Y.L., E.S.N. and S.H.P. wrote the manuscript with input from all authors.

Conflict of Interest

The authors declare no competing interests.

References

1. Anderson, D. C., Li, W., Payan, D. G. & Noble, W. S. A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *J. Proteome Res.* **2,** 137–146 (2003).
2. Payne, S. H. *et al.* Phosphorylation-specific MS/MS scoring for rapid and accurate phosphoproteome analysis. *J. Proteome Res.* **7,** 3373–3381 (2008).
3. Frank, A. M. A ranking-based scoring function for peptide-spectrum matches. *J. Proteome Res.* **8,** 2241–2252 (2009).
4. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4,** 923–925 (2007).
5. Elias, J. E., Gibbons, F. D., King, O. D., Roth, F. P. & Gygi, S. P. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat. Biotechnol.* **22,** 214–219 (2004).
6. Timm, W., Scherbart, A., Böcker, S., Kohlbacher, O. & Nattkemper, T. W. Peak intensity prediction in MALDI-TOF mass spectrometry: a machine learning study to support quantitative proteomics. *BMC Bioinformatics* **9,** 443 (2008).
7. Edwards, N. J. PepArML: A Meta-Search Peptide Identification Platform for Tandem Mass Spectra. *Curr. Protoc. Bioinforma.* **44,** 13.23.1-23 (2013).
8. Ulintz, P. J., Zhu, J., Qin, Z. S. & Andrews, P. C. Improved classification of mass spectrometry database search results using newer machine learning approaches. *Mol. Cell. Proteomics MCP* **5,** 497–509 (2006).
9. 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).

10. Dancík, V., Addona, T. A., Clauser, K. R., Vath, J. E. & Pevzner, P. A. De novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **6,** 327–342 (1999).
11. 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).
12. Waridel, P. *et al.* Sequence similarity-driven proteomics in organisms with unknown genomes by LC-MS/MS and automated de novo sequencing. *Proteomics* **7,** 2318–2329 (2007).
13. Guthals, A. *et al.* De Novo MS/MS Sequencing of Native Human Antibodies. *J. Proteome Res.* **16,** 45–54 (2017).
14. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521,** 436–444 (2015).
15. Hinton, G. *et al.* Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Process. Mag.* **29,** 82–97 (2012).
16. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv14091556 Cs* (2014).
17. Firat, O., Cho, K. & Bengio, Y. Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism. *ArXiv160101073 Cs Stat* (2016).
18. 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.* (2017). doi:10.1073/pnas.1705691114
19. Kim, M., Eetemadi, A. & Tagkopoulos, I. DeepPep: Deep proteome inference from peptide profiles. *PLoS Comput. Biol.* **13,** e1005661 (2017).
20. Deutsch, E. W. *et al.* The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res.* **45,** D1100–D1106 (2017).
21. Nakayasu, E. S. *et al.* Ancient Regulatory Role of Lysine Acetylation in Central Metabolism. *mBio* **8,** (2017).

22. Kelly, R. T. *et al.* Chemically etched open tubular and monolithic emitters for nanoelectrospray ionization mass spectrometry. *Anal. Chem.* **78,** 7796–7801 (2006).
23. Ma, B. Novor: real-time peptide de novo sequencing software. *J. Am. Soc. Mass Spectrom.* **26,** 1885–1894 (2015).
24. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12,** 59–60 (2015).
25. Vizcaíno, J. A. *et al.* 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44,** D447-456 (2016).
26. Mo, L., Dutta, D., Wan, Y. & Chen, T. MSNovo: a dynamic programming algorithm for de novo peptide sequencing via tandem mass spectrometry. *Anal. Chem.* **79,** 4870–4878 (2007).
27. Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **5,** 5277 (2014).
28. Marin, J., Battistuzzi, F. U., Brown, A. C. & Hedges, S. B. The Timetree of Prokaryotes: New Insights into Their Evolution and Speciation. *Mol. Biol. Evol.* **34,** 437–446 (2017).
29. Hedges, S. B., Blair, J. E., Venturi, M. L. & Shoe, J. L. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evol. Biol.* **4,** 2 (2004).
30. Wang, J., Pérez-Santiago, J., Katz, J. E., Mallick, P. & Bandeira, N. Peptide identification from mixture tandem mass spectra. *Mol. Cell. Proteomics MCP* **9,** 1476–1485 (2010).
31. Kryuchkov, F., Verano-Braga, T., Hansen, T. A., Sprenger, R. R. & Kjeldsen, F. Deconvolution of mixture spectra and increased throughput of peptide identification by utilization of intensified complementary ions formed in tandem mass spectrometry. *J. Proteome Res.* **12,** 3362–3371 (2013).
32. Mann, M. & Wilm, M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **66,** 4390–4399 (1994).

33. Shevchenko, A. *et al.* Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal. Chem.* **73,** 1917–1926 (2001).

34. Wielsch, N. *et al.* Rapid validation of protein identifications with the borderline statistical confidence via de novo sequencing and MS BLAST searches. *J. Proteome Res.* **5,** 2448–2456 (2006).

35. Ma, B. & Johnson, R. De novo sequencing and homology searching. *Mol. Cell. Proteomics MCP* **11,** O111.014902 (2012).

36. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* (2018). doi:10.1038/nbt.4229

37. Payne, S. H. *et al.* The Pacific Northwest National Laboratory library of bacterial and archaeal proteomic biodiversity. *Sci. Data* **2,** 150041 (2015).