Supplement to "A learned embedding for efficient joint analysis of millions of mass spectra"

Damon H. May[1], Jeffrey Bilmes[1,2], and William S. Noble[1,2]

[1]Department of Genome Sciences, University of Washington
[2]Paul G. Allen School of Computer Science and Engineering, University of Washington

November 29, 2018

Contents

1 Supplementary Note 1: Exploring hyperparameter space and training data structure

To arrive at the model used in GLEAMS, we explored the space of model structure and hyperparameter settings extensively but not exhaustively. To compare trained models, we assessed their performance by the area under the concentrated receiver operator characteristic (CROC) curve [1] on held-out same-label and different-label pairs of spectra, as described in Methods. Most hyperparameters were selected based on the results of training on 100,000 pairs of spectra. However, the network structure (the numbers, types and sizes of layers used for each type of input feature, and the number of fully-connected layers after concatenation) was selected from a wide variety of structures all trained on one million pairs of spectra. The model presented had the highest CROC of all models considered. Below, we describe the various models and hyperparameter settings that we explored.

We considered different types and encodings of input features. The model using all three feature types (precursor, binned fragment and reference spectrum similarity) outperformed models using one or two of those feature types. We considered encoding precursor mass and m/z as single, real-value features (with or without scaling) and with an arbitrary binary encoding lacking the locality benefits of Gray Code. We also considered binning fragment features at 0.02 Da (convolving peak intensities with a Gaussian representing estimated fragment measurement error); the resulting enormous number of features was a great impediment to training. We further considered using the hashing trick (defining a hash function to map large numbers of features to a smaller number of features, with collisions) to reduce this dimensionality to 2,000, 4,000 or 6,000 features, which improved performance but still lagged behind the 1 m/z binning. We considered using 500 and 1000 reference spectrum similarity features.

We considered many different structures for the embedder model. For each input type (precursor, fragment, reference spectrum) we considered one to three fully-connected layers of various sizes, one to three convolutional layers followed by max pooling, a recurrent neural network, a dense layer followed by convolutional layers, and long short-term memory (LSTM) layers (single-directional and bidirectional). We also considered one to four fully-connected layers after concatenation of the network outputs from the three input types. Surprisingly, deeper networks generally trained more slowly and also reached lower final AUCROCs: the only input type that benefited from more than a single layer was the precursor input type.

We considered several values for the hyperparameters associated with the convolutional neural networks (CNNs): number of filters (20, 30 or 50), kernel size (2,3,4), stride length (1,2), pooling kernel size (1,2) and pooling stride length (1,2). Of particular note, we discovered that the size of the last layer on the precursor features needed to be small (we settled on five) compared to the number of filters (we settled on 30) used in the CNNs on the binned fragment and reference spectrum features.

We considered several different nonlinearities for all network layers: ELU, ReLU, SELU, PReLU and sigmoid, as well as linear activation. We considered training with a fixed learning rate, as well as with Adam, RMSprop, and Adagrad using several learning rates.

We considered batch normalization and dropout with proportion 0.0005 to 0.2, and L1 and L2 regularization. All decreased performance.

We considered several sizes for the embedded dimension: 8, 16, 24, 32, 64 and 128. Higher dimensionality gave monotonically higher CROC but slowed down operations on the embedded spectra such as k-nearest-neighbor search. The improvement from 24 to 32 was substantial, and the improvement from 32 to 64 was minimal.

We considered four approaches to training and validation data set construction before settling on the combination of observed and theoretical spectra described in Methods. First, we used only pairs of observed spectra with the same or different peptide labels, with no further restrictions. Second, we imposed a 3 Da maximum on the difference between the two precursor masses. The second approach led to higher AUCROC than the first approach, even when using a validation set without the precursor mass restriction. We suspect this improvement arose because, without the restriction, too many of the different-label pairs were "too easy," having many differences between the spectra and insufficiently representing the difficult task of discriminating pairs of spectra that share more characteristics. This observation led us to our third approach, in which we used same-label pairs of real spectra with precursor masses within 0.2 Da and different-label pairs between observed spectra and theoretical spectra generated by MS2PIP [2;3] representing decoy peptides from the top five search results from Comet search. With this approach the network learned how to separate real from

theoretical spectra but did not learn as well to separate positive- and negative-label pairs of real spectra. To address this issue, we developed our fourth and final method, described in Methods, in which we added to the third approach different-labeled pairs of real spectra and different-labeled pairs of theoretical spectra.

2 Supplementary Table

Experiment	Instrument	Organism	Additional Search Parameters
Training Datasets			
2013poulsen-PXD000307	TripleTOF	human	
2014kim-kidney[4]	Orbitrap Velos	human	
2014kim-lung[4]	Orbitrap Elite	human	
2014kim-adrenalgland[4]	Orbitrap Velos	human	
2014kim-monocytes[4]	Orbitrap Velos	human	
2014kim-rectum[4]	Orbitrap Velos	human	
2014kim-gut[4]	Orbitrap Velos	human	
2014kim-fetalovary[4]	Orbitrap Elite	human	
2014kim-fetalplacenta[4]	Orbitrap Elite	human	
2015clark-redefining[5]	LTQ Orbitrap	human	TMT 6-plex (229.1629 to K, N-terminus)
2015tanca-impact[6]	Orbitrap Velos	human gut microbiome	
2015uszkoreit-intuitive[7]	Orbitrap Elite	mouse	
2016mann-unpublished	QExactive	human	
2016may-metapeptides[8]	QExactive	ocean microbiome	
2016saraf-dynamic[9]	LTQ-Orbitrap	human	
2016zhong-quantitative[10]	Orbitrap Velos	human	
Test Datasets			
2014kim-cd4tcell[4]	Orbitrap Elite	human	
2014kim-adultovary[4]	Orbitrap Elite	human	
2014kim-eart[4]	Orbitrap Elite	human	
2015radoshevich-isg15[11]	QExactive	human	
2016audain-in-depth[12]	LTQ Orbitrap	yeast	
2016schittmayer-cleaning[13]	Orbitrap Velos	yeast	

Table 1: **Experiments used in the training and validation of the embedder network.**

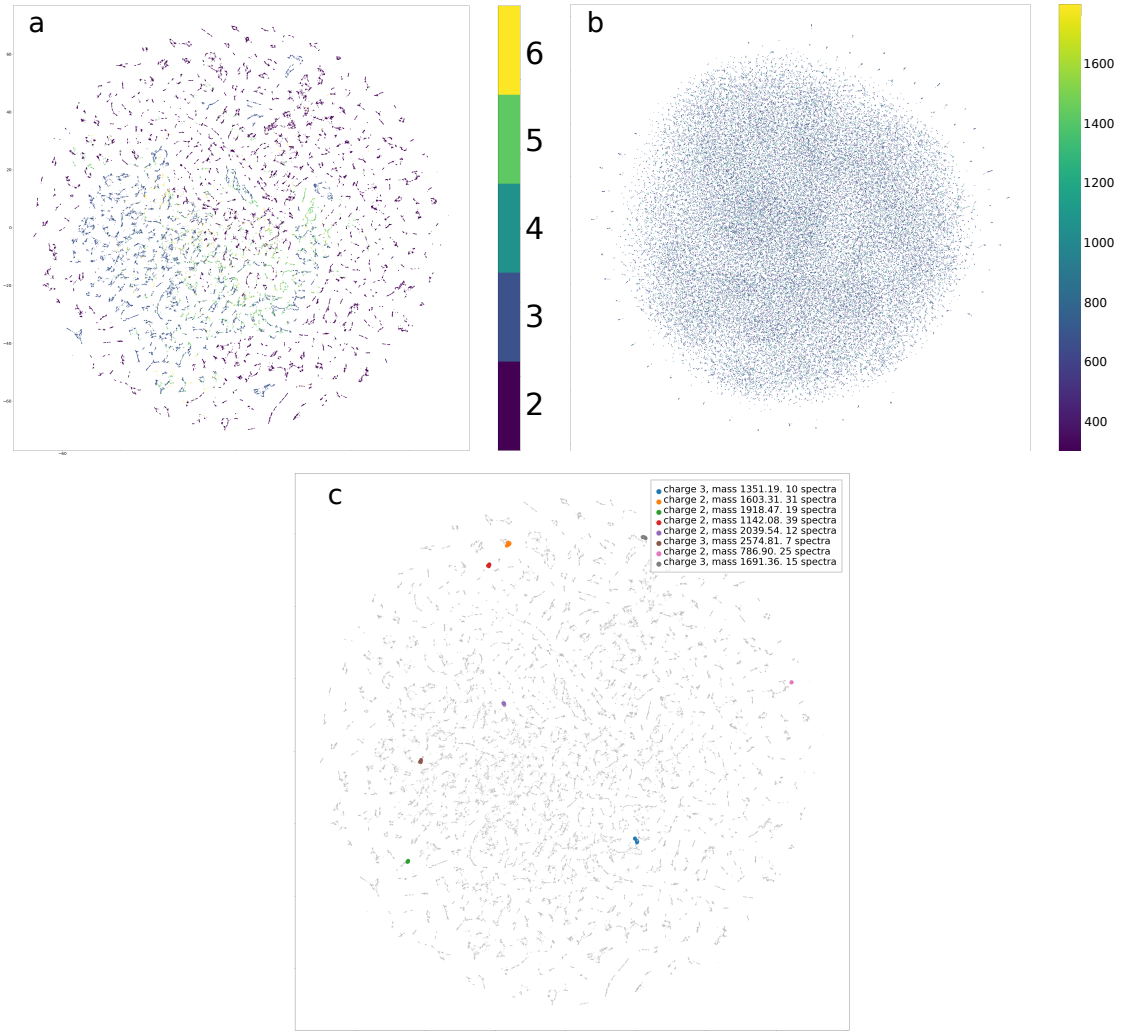
3 Supplementary Algorithm

Supplementary Algorithm 1 Hub-and-spoke spectrum community detection. First, 'hub' spectra are associated with their 'spoke' neighbors (based on k-nearest neighbors search) within distance threshold τ in a greedy fashion, with the most-connected hubs chosen first. Then adjacent hub-and-spoke communities are combined if their hubs are within τ, leaving out any spokes not within τ of the hub of the larger community. The second step is necessary because the first pass is based on a limited value of k (1000), much less than the number of nodes. In the below, D_N is a mapping from spectra to neighbors; thus for a given spectrum d, $D_N(d)$ are the neighbors (within distance τ) of d.

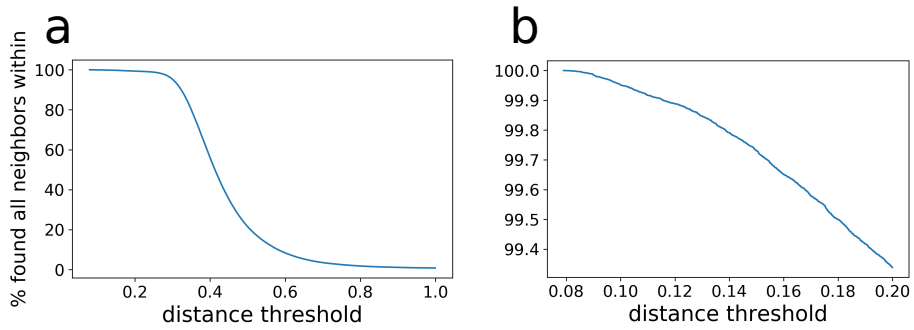
Input: Dictionary D_N mapping each spectrum to its neighbors within distance τ.

$S \leftarrow \emptyset$ ▷ S accumulates a set of assigned 'spoke' spectra.
$D_{HS} \leftarrow \emptyset$ ▷ D_{HS} is a dictionary mapping each hub to a set of spokes.
Sort keys of D_N first by (1) neighbor count, and then by (2) mean neighbor distance (both descending).
for $d \in$ keys of D_N in order **do**
    if $d \notin S$ **then**
        $S_d \leftarrow \{n : n \in D_N(d),\ n \notin S, \text{ and } n \notin D_{HS}\}$
        if $S_d \neq \emptyset$ **then**
            $D_{HS}(d) = S_d$
            $S \leftarrow S \cup S_d$
$H \leftarrow$ keys of D_{HS} ▷ H is the list of hub spectra
Sort H by number of spokes per hub (ascending).
for $h_1 \in H$ in order **do**
    $N_H \leftarrow \{n : n \in D_{HS}(h_1), n \in H \text{ and } |D_{HS}(n)| \geq |D_{HS}(h_1)|\}$
    if $N_H \neq \emptyset$ **then**
        $h_2 = \text{argmin}_{x \in N_H} ||h_1 - x||$ ▷ h_2 is the closest hub neighbor of h_1 with at least as many spokes
        $D_{HS}(h_2) \leftarrow D_{HS}(h_2) \cup \{h_1\}$
        for $s \in D_{HS}(h_1)$ **do**
            if $||h_2 - s|| < \tau$ **then**
                $D_{HS}(h_2) \leftarrow D_{HS}(h_2) \cup \{s\}$
        remove h_1 from D_{HS}
return D_{HS}

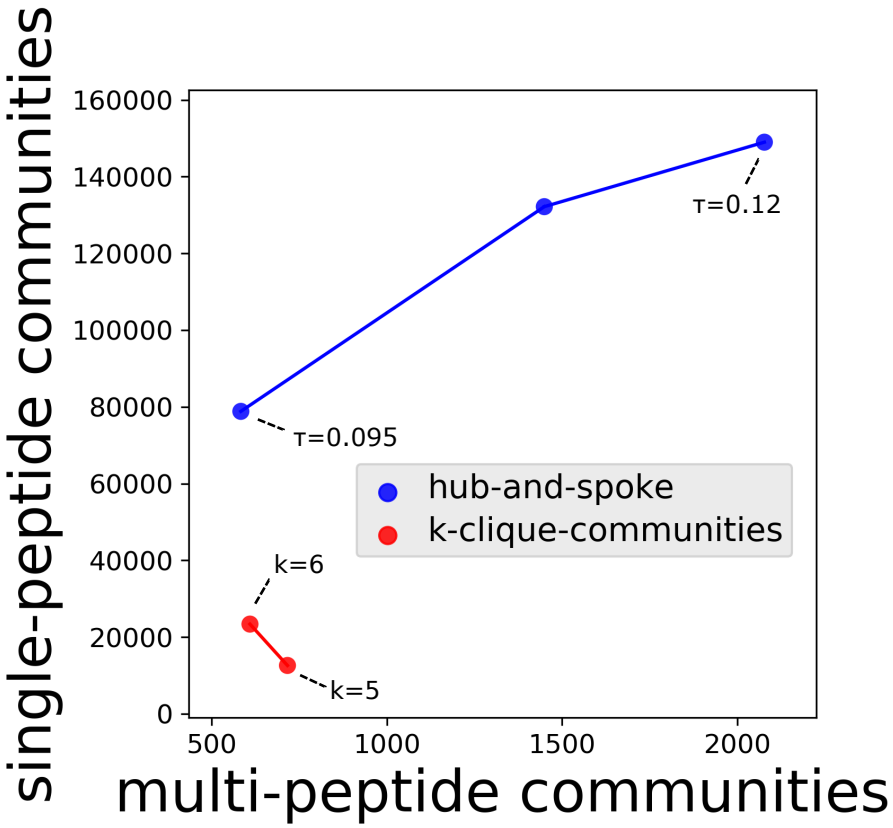
4 Supplementary figures



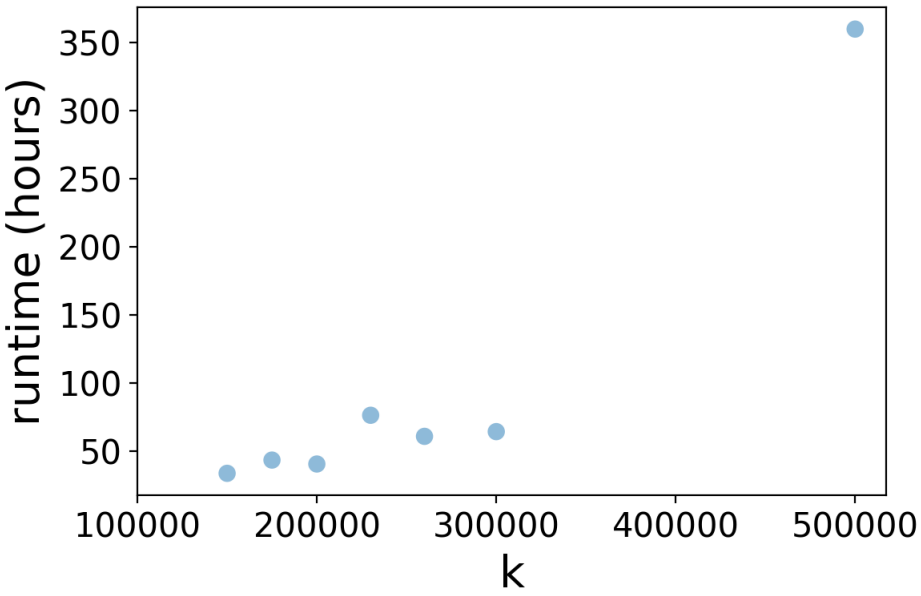
Supplementary Figure 1: **Additional t-SNE projections.** Each point represents a spectrum. (A) colored by charge state. (B) With the per-spectrum values for each of the 32 dimensions of the embedded spectrum matrix shuffled prior to running t-SNE. The lack of "clumpy" structure in this plot demonstrates that the structure observed in the unpermuted plots is not an artifact of the t-SNE algorithm. (C) Unpermuted spectra, with all spectra within a single randomly chosen charge state and 1.000507 Da mass bin (158 mass spectra across all eight bins) each given a different color and larger dot size. The spectra from each mass bin all occur within the same globular structure. Legend indicates charges and centers of each mass bin.



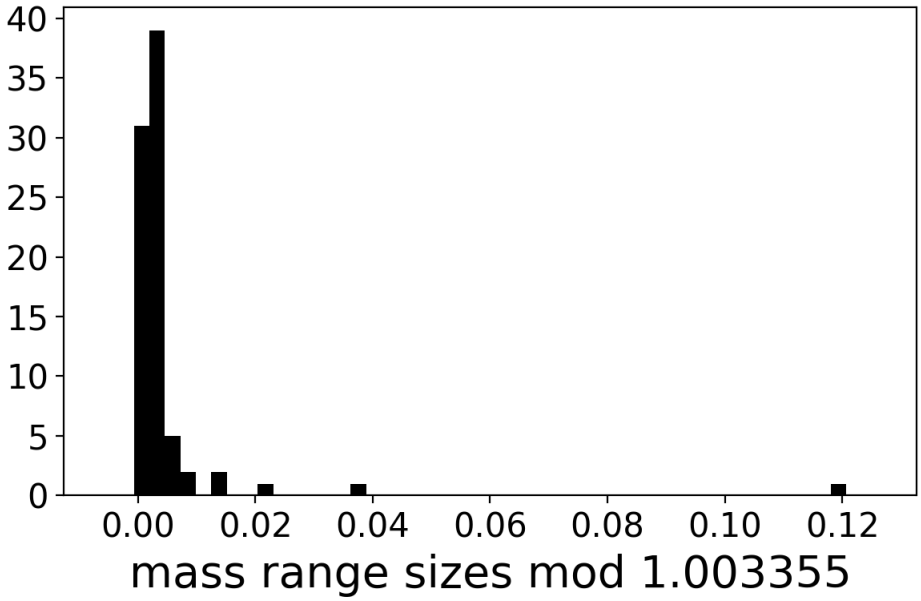
Supplementary Figure 2: **Percentage of spectra with 1000 nearest neighbors within distance thresholds.** The percentage of embedded spectra (vertical axis) having all 1000 of their nearest 1000 neighbors within a given Euclidean distance threshold (horizontal axis). (A) Distance thresholds < 1 (B) Distance thresholds < 0.2



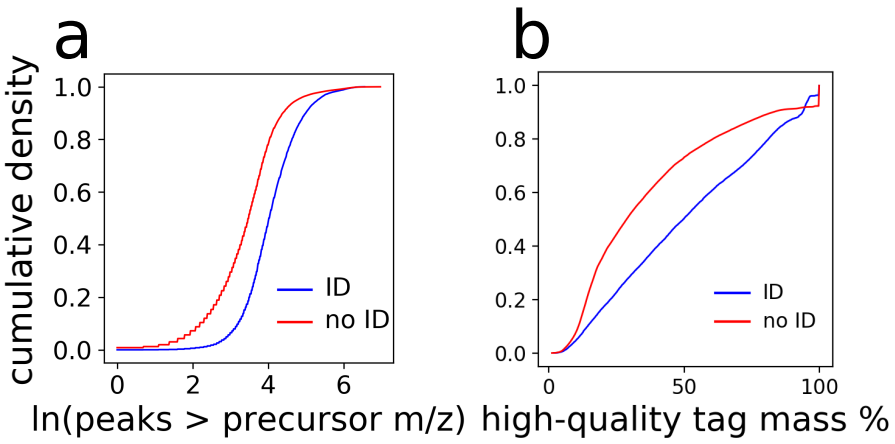
Supplementary Figure 3: **Comparing hub-and-spoke and k-clique-communities methods for community detection.** Comparisons of the numbers of single-peptide and multi-peptide communities detected among the 3,390,759 charge-2 repository spectra by the hub-and-spoke method with τ ranging from 0.095 to 0.12 and the k-clique-communities method with $k = 5$ and $k = 6$.



Supplementary Figure 4: **Running time for k-means clustering on the charge-2 spectra as a function of k.** k-means clustering was performed with an Intel Xeon(R) E5-2650 CPU and 90GB memory available.



Supplementary Figure 5: **Communities with single amino acid substitutions appear to be generated by a single peptide.** Mass ranges, modulo 1.003355 (the mass difference between ^{13}C and ^{12}C), of all 82 spectrum communities containing only spectrum identifications representing a single E-to-K amino acid substitution.



Supplementary Figure 6: **Quality of identified and unidentified spectra.** Cumulative density functions for two proxies for spectrum quality, for spectra that were identified (blue line) or were not identified (red line) by database search. (A) The natural log of the one-padded count of fragment peaks higher than the precursor m/z. (B) The mass of the longest high quality sequence tag found by Novor as a percentage of the mass of the precursor ion.

References

[1] S Joshua Swamidass, Chloé-agathe Azencott, Kenny Daily, and Pierre Baldi. A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval. *Bioinformatics*, 26(10):1348–1356, 2010.

[2] Sven Degroeve, Lennart Martens, and Igor Jurisica. MS2PIP: A tool for MS/MS peak intensity prediction. *Bioinformatics*, 29(24):3199–3203, 2013.

[3] Sven Degroeve, Davy Maddelein, and Lennart Martens. MS2PIP prediction server: Compute and visualize MS2peak intensity predictions for CID and HCD fragmentation. *Nucleic Acids Research*, 43(W1):W326–W330, 2015.

[4] Min-Sik Kim, Sneha M. Pinto, Derese Getnet, Raja Sekhar Nirujogi, Srikanth S. Manda, Raghothama Chaerkady, Anil K. Madugundu, Dhanashree S. Kelkar, Ruth Isserlin, Shobhit Jain, Joji K. Thomas, Babylakshmi Muthusamy, Pamela Leal-Rojas, Praveen Kumar, Nandini A. Sahasrabuddhe, Lavanya Balakrishnan, Jayshree Advani, Bijesh George, Santosh Renuse, Lakshmi Dhevi N. Selvan, Arun H. Patil, Vishalakshi Nanjappa, Aneesha Radhakrishnan, Samarjeet Prasad, Tejaswini Subbannayya, Rajesh Raju, Manish Kumar, Sreelakshmi K. Sreenivasamurthy, Arivusudar Marimuthu, Gajanan J. Sathe, Sandip Chavan, Keshava K. Datta, Yashwanth Subbannayya, Apeksha Sahu, Soujanya D. Yelamanchi, Savita Jayaram, Pavithra Rajagopalan, Jyoti Sharma, Krishna R. Murthy, Nazia Syed, Renu Goel, Aafaque A. Khan, Sartaj Ahmad, Gourav Dey, Keshav Mudgal, Aditi Chatterjee, Tai-Chung Huang, Jun Zhong, Xinyan Wu, Patrick G. Shaw, Donald Freed, Muhammad S. Zahari, Kanchan K. Mukherjee, Subramanian Shankar, Anita Mahadevan, Henry Lam, Christopher J. Mitchell, Susarla Krishna Shankar, Parthasarathy Satishchandra, John T. Schroeder, Ravi Sirdeshmukh, Anirban Maitra, Steven D. Leach, Charles G. Drake, Marc K. Halushka, Keshava Prasad, Ralph H. Hruban, Candace L. Kerr, Gary D. Bader, Christine A. Iacobuzio-Donahue, Harsha Gowda, and Akhilesh Pandey. A draft map of the human proteome A. *Nature*, 509(7502):575–581, 2014.

[5] David J. Clark, William E. Fondrie, Zhongping Liao, Phyllis I. Hanson, Amy Fulton, Li Mao, and Austin J. Yang. Redefining the Breast Cancer Exosome Proteome by Tandem Mass Tag Quantitative Proteomics and Multivariate Cluster Analysis. *Analytical Chemistry*, 87(20):10462–10469, 2015.

[6] Alessandro Tanca, Antonio Palomba, Salvatore Pisanu, Maria Filippa Addis, and Sergio Uzzau. A human gut metaproteomic dataset from stool samples pretreated or not by differential centrifugation. *Data in Brief*, 4:559–562, 2015.

[7] Julian Uszkoreit, Alexandra Maerkens, Yasset Perez-Riverol, Helmut E. Meyer, Katrin Marcus, Christian Stephan, Oliver Kohlbacher, and Martin Eisenacher. PIA: An Intuitive Protein Inference Engine with a Web-Based User Interface. *Journal of Proteome Research*, 14(7):2988–2997, 2015.

[8] Damon H. May, Emma Timmins-Schiffman, Molly P. Mikan, H. Rodger Harvey, Elhanan Borenstein, Brook L. Nunn, and William Stafford Noble. An alignment-free 'metapeptide' strategy for metaproteomic characterization of microbiome samples using shotgun metagenomic sequencing. *Journal of Proteome Research*, 15(8):acs.jproteome.6b00239, 2016.

[9] Anita Saraf, Serena Cervantes, Evelien M. Bunnik, Nadia Ponts, Mihaela E. Sardiu, Duk Won D. Chung, Jacques Prudhomme, Joseph M. Varberg, Zhihui Wen, Michael P. Washburn, Laurence Florens, and Karine G. Le Roch. Dynamic and combinatorial landscape of histone modifications during the intraerythrocytic developmental cycle of the malaria parasite. *Journal of Proteome Research*, 15(8):2787–2801, 2016.

[10] Lijun Zhong, Juntuo Zhou, Xi Chen, Yaxin Lou, Dan Liu, Xiajuan Zou, Bin Yang, Yuxin Yin, and Yan Pan. Quantitative proteomics study of the neuroprotective effects of B12 on hydrogen peroxide-induced apoptosis in SH-SY5Y cells. *Scientific Reports*, 6(February 2015):22635, 2016.

[11] Lilliana Radoshevich, Francis Impens, David Ribet, Juan J. Quereda, To Nam Tham, Marie Anne Nahori, Helene Bierne, Olivier Dussurget, Javier Pizarro-Cerda, Klaus Peter Knobeloch, and Pascale Cossart. ISG15 counteracts Listeria monocytogenes infection. *eLife*, 4(AUGUST2015):1–23, 2015.

[12] Enrique Audain, Julian Uszkoreit, Timo Sachsenberg, Julianus Pfeuffer, Xiao Liang, Henning Hermjakob, Aniel Sanchez, Martin Eisenacher, Knut Reinert, David L. Tabb, Oliver Kohlbacher, and Yasset Perez-Riverol. In-depth analysis of protein inference algorithms using multiple search engines and well-defined metrics. *Journal of Proteomics*, 150:170–182, 2016.

[13] Matthias Schittmayer, Katarina Fritz, Laura Liesinger, Johannes Griss, and Ruth Birner-Gruenberger. Cleaning out the Litterbox of Proteomic Scientists Favorite Pet: Optimized Data Analysis Avoiding Trypsin Artifacts. *Journal of Proteome Research*, 15(4):1222–1229, 2016.