MS1Connect: a mass spectrometry run similarity measure

Andy Lin¹, Brooke L. Deatherage Kaiser¹, Janine R. Hutchison¹, Jeffrey A. Bilmes², and William Stafford Noble^{*3,4}

¹Chemical and Biological Signatures, Pacific Northwest National Laboratory, Richland, WA, USA

²Department of Electrical and Computer Engineering, University of Washington, Seattle, WA, USA

³Department of Genome Sciences, University of Washington, Seattle, WA, USA

⁴Paul G. Allen School for Computer Science and Engineering, University of Washington, Seattle, WA, USA

Abstract

Interpretation of newly acquired mass spectrometry data can be improved by identifying, from an online repository, previous mass spectrometry runs that resemble the new data. However, this retrieval task requires computing the similarity between an arbitrary pair of mass spectrometry runs. This is particularly challenging for runs acquired using different experimental protocols. We propose a method, MS1Connect, that calculates the similarity between a pair of runs by examining only the intact peptide (MS1) scans, and we show evidence that the MS1Connect score is accurate. Specifically, we show that MS1Connect outperforms several baseline methods on the task of predicting the species from which a given proteomics sample originated. In addition, we show that MS1Connect scores are highly correlated with similarities computed from fragment (MS2) scans, even though this data is not used by MS1Connect. The MS1Connect software will be made available upon acceptance at <https://github.com/bmx8177/MS1Connect>.

1 Introduction

Over the years a wealth of proteomics data has been deposited into online repositories such as PRIDE¹ and MassIVE.² Researchers may be interested in finding data in these online repositories that are similar to their own data in order to perform some joint analysis. However, it is difficult to identify which repository runs should be analyzed with your data. This is because it is hard to measure the similarity of a pair of runs, especially across different studies. Specifically, biologically irrelevant differences in sample preparation, liquid chromatography, and instrument parameters all affect the resulting data.

Currently, few methods exist for measuring the similarity of a pair of proteomics runs. One metric is to count the number of confidently detected peptides in common.³ Unfortunately, this method requires knowing the species composition of the samples, which is not always known, and requires a database search. Another previously developed method for measuring the similarity between a pair of mass spectrometry runs directly compares the set of peptide fragment spectra (MS2 scans) from each run. Conceptually this method works by calculating the spectra dot product for all pairs of spectra in two runs and then counting the fraction of scores above a threshold.⁴ This approach has been successfully used for species identification,⁵ molecular phylogenetics,⁶ and differentiating between experimental protocols.⁷

Unfortunately these MS2 based methods do not account for differences in how MS2 scans are collected. MS2 spectra collected by data-independent acquisition (DIA), data-dependent acquisition (DDA), and targeted analyses are all different from each other. Therefore, new methods are needed that can measure the similarity between a pair of proteomics runs regardless of how the data was acquired.

*Corresponding author: Box 355065, Foegen Building, S220B, 3720 15th Ave NE, Seattle, WA 98195-5065, william-noble@uw.edu

In this work, we describe a new method, MS1Connect, that only uses intact peptide scans (MS1 scans) to calculate the similarity between a pair of runs. Since MS1 scans are always collected in the same way, this data can be used to compare targeted, DDA, and DIA data. To our knowledge, MS1 data has not been used to measure the overall similarity between a pair of proteomics runs. However, several methods have been developed to align MS1 features maps of a pair of proteomics runs,^{8–10} but these methods do not report an overall similarity score.

Our method, MS1Connect, frames scoring the similarity between a pair of proteomics runs as a maximum bipartite matching problem. A bipartite graph consists of two disjoint sets of vertices and a set of edges that connect the two sets of vertices. In a maximum bipartite matching problem, the goal is to select a subset of edges in a bipartite graph that maximizes some objective value subject to some constraint. In our setting, each of the two disjoint sets of vertices represent the set of MS1 features found in a run, and edges link MS1 features, in different runs, whose masses match within some tolerance. In addition, we require that every MS1 feature be associated with at most one edge in the set of selected edges.

The MS1Connect objective function consists of a weighted combination of three modular terms and a fourth supermodular term. Modular and supermodular functions are both types of set functions. In a modular function the sum is equals to the sum of its parts. More specifically, given two sets of disjoint items X and Y and a scoring function f , $f(X) + f(Y) = f(X \cup Y)$. On the other hand, for a supermodular function, the sum is greater than the sum of its parts. In this case $f(X) + f(Y) \leq f(X \cup Y)$. While there exists a robust set of literature dedicated to the theory of supermodular maximization,^{11–14} these functions have rarely been applied to the analysis of biological data. The one example we are aware of uses a surrogate function to estimate the supermodular relationship between fragment ions in database searching.¹⁵

Each of the MS1Connect terms measures a different aspect of proteomics run similarity. The first modular term (M_1) favors solutions with a large number of edges. The second modular term (M_2) favors selection of edges with high intensity values. The third modular term (M_3) favors edges with small normalized retention time shifts. We define the normalized retention time shift as the difference in normalized retention times between the two MS1 features in an edge. Finally, the fourth supermodular term (M_4) favors solutions with pairs of edges that are similar to each other. Two edges are similar if they have similar normalized retention time shifts and if the normalized retention times of the MS1 features, in the same run, are similar to each other.

We show evidence that the MS1Connect score accurately measures the similarity between two proteomics runs. There are many ways to define whether two runs are similar to each other. For this work we focus on the task of species prediction. We show that MS1Connect scores outperforms baselines for predicting the species a sample originated from. In addition, we show that MS1Connect scores are able to recapitulate similarities based off MS2 spectra. Specifically, we show a high correlation between MS1Connect scores and the Jaccard index between the sets of confidently detected peptides for a pair of runs.

2 Approach

2.1 Representation of a mass spectrometry run

We represented each tandem mass spectrometry run as a bag of MS1 features, where each feature nominally corresponds to a peptide detected in a set of precursor scans. Each MS1 feature is represented as a tuple of four values: m/z , intensity, charge, and retention time (in seconds). All four of these values are reported by pyOpenMS,¹⁶ the tool we use for MS1 feature detection. To generate the input files for pyOpenMS, we used Proteowizard version 3.0¹⁷ to convert Thermo RAW files to .mzML format.¹⁸ We use the N most intense MS1 features to represent each run, where N is a hyperparameter. In Supplemental Section 1.1 we discuss how we normalized the values returned by pyOpenMS.

2.2 Mass spectrometry run matching as a maximum bipartite matching problem

We frame the measurement of the similarity between a pair of mass spectrometry runs as a maximum bipartite matching problem. In this approach, we aim to select a set of edges in a given bipartite graph that achieves a maximum objective value.

A bipartite graph $G = (U, V, E)$ consists of two disjoint sets of vertices, U and V , and a set of edges $E \subseteq U_{r_1} \times V_{r_2}$, where each edge e connects a vertex $u \in U$ to a vertex $v \in V$. For our specific formulation, U and V are the sets of MS1 features from two different mass spectrometry runs, r_1 and r_2 , and the edges E link MS1 features

between U and V (Supplemental Figure 1). For this reason, we have one bipartite graph $G_{r_1, r_2} = (U_{r_1}, V_{r_2}, E_{r_1, r_2})$ associated with every run pair r_1, r_2 , where $E_{r_1, r_2} \subseteq U_{r_1} \times V_{r_2}$. For notational simplicity, we drop the r_1, r_2 subscripts except when they are needed for run-pair disambiguation.

We include in the graph edges between all pairs of MS1 features whose charges match and whose m/z values match within some tolerance δ_1 (in units of ppm):

$$E = \left\{ (u, v) : \left| \frac{m(v) - m(u)}{\frac{1}{2}(m(v) + m(u))} 10^6 \right| \leq \delta_1 \wedge c(v) = c(u) \right\} \subseteq U \times V,$$

where $m(u)$ is the m/z of u , $m(v)$ is the m/z of v , $c(v)$ is the charge of MS1 feature v , and $c(u)$ is the charge of u . By connecting MS1 features with the same charge and similar m/z , these edges attempt to connect the same peptide precursor.

The goal in maximum bipartite matching is to select a set of edges A that achieves a maximum score, as measured by a specified objective function S , subject to some matching constraints. Specifically, because we expect that each peptide precursor will be detected at most once per run, we require that a valid matching connects each feature in U to at most one feature in V and vice versa. More formally, for two runs r_1 and r_2 , we want to choose a subset of edges $A \subseteq E$ that achieves a maximum value of the objective defined below subject to the following constraints: $\forall e \in A, \text{degree}(u(e)) \leq 1$ and $\text{degree}(v(e)) \leq 1$, where $u(e)$ retrieves the relevant MS1 feature from U and $v(e)$ retrieves the corresponding MS1 feature in V . To designate these constraints, we define $\mathcal{E}_U = \{A \subseteq E : \forall e \in A, \text{degree}(u(e)) \leq 1\}$ which is the set of all subsets of edges that satisfy by the required degree constraint of the corresponding nodes on the U side, and we correspondingly define $\mathcal{E}_V = \{A \subseteq E : \forall e \in A, \text{degree}(v(e)) \leq 1\}$ for the V side.

2.3 Scoring a candidate matching

Our handcrafted score function S_{r_1, r_2} uses a maximum matching approach which maximizes an objective that consists of a weighted combination of four terms. The score for two runs r_1, r_2 is defined as follows:

$$S_{r_1, r_2} = \left(\sum_{j=1}^4 \lambda_j M_j(E_{r_1, r_2}) \right) \times \max_{A \subseteq E_{r_1, r_2} : A \in (\mathcal{E}_{U_{r_1}} \cap \mathcal{E}_{V_{r_2}})} \sum_{i=1}^4 \lambda_i \frac{M_i(A)}{M_i(E_{r_1, r_2})}, \quad (1)$$

where M_1 through M_4 are terms that are defined below, λ_1 through λ_4 are convex mixture hyperparameters that weight the relative importance of each term, and E_{r_1, r_2} is the set of valid edges between runs r_1 and r_2 . Each term in the inner maximization is normalized by the score for the full set E with no matching constraints. The inner maximization maximizes over all subsets of edges that satisfy by both degree constraints, and that is indicated via $A \in (\mathcal{E}_{U_{r_1}} \cap \mathcal{E}_{V_{r_2}})$ since if A is a member of both $\mathcal{E}_{U_{r_1}}$ and $\mathcal{E}_{V_{r_2}}$, then no node incident to any edge has more than degree one in the matching. The leading summation serves to un-normalize the objective value via a multiplicative (r_1, r_2) -dependent constant. This two-step normalization and un-normalization process is used to make the λ hyperparameters interpretable (i.e., each λ_i indicates the relative contribution of each term) and also to ensure that the scores are calibrated over multiple distinct pairs of runs. As a result of the normalization process, the four λ values must sum to one and range between zero and one, inclusive. We next describe the M_i terms, each of which measures a different aspect of proteomics run similarity (Figure 1).

The first term M_1 counts the number of edges selected in a solution: $M_1(A) = |A|$. Thus, this term favors solutions that contain a large number of edges. Runs that are similar will share many of the same peptides. As a result, these peptides will be linked by edges, thereby generating many edges.

The second term sums the product of the two intensities associated with each edge found in a selected matching, $M_2(A) = \sum_{e \in A} I(u(e))I(v(e))$, where $I(u)$ is the intensity of feature u and $I(v)$ is the intensity of feature v . This term favors solutions that select edges with high intensity values. The intensity of an MS1 feature produced by a peptide is a function of the sequence and abundance of that peptide. Runs that are similar to each other should have similar peptide expression levels. By multiplying the two intensities together, we focus on the most intense peptide features.

The third term sums the negative exponent of the absolute difference of the normalized retention time shift of all the selected edges: $M_3(A) = \sum_{e \in A} \exp(-\alpha|s(e)|)$, where α is a hyperparameter that determines how quickly the exponential function decreases and $s(e)$ is the normalized retention time shift between u and v : $s(e) = t(v(e)) -$

Figure 1: **Schematic of four MS1Connect terms.** This figure illustrates what each term of the MS1Connect objective function focuses on when scoring a matching. Each subplot shows bipartite graph formed by a pair of mass spectrometry runs. The red and gray lines are edges that link MS1 features with similar m/z . Note that the m/z axis is not shown. In addition, note that, in a specific matching, each MS1 feature is associated with at most one edge. (A) The M_1 term favors solutions that contain a large number of edges. (B) The M_2 term favors solutions that select edges with high intensity values. (C) The M_3 term favors edges that have small normalized retention time shifts. (D) The M_4 term favors solutions that contain many similar pairs of edges. The best solution chooses the three red edges in the middle instead of the gray edge because these three edges are similar to each other. This similarity indicates that the large retention time difference is correct and due to systematic differences in chromatography conditions.

$t(u(e))$, where $t(u)$ and $t(v)$ are the normalized retention times of MS1 features u and v , respectively. This term favors edges that have small normalized retention time shifts. Such edges are better because they are more likely to link the same peptide feature. This is because we expect the time it takes for peptides to elute off the column, in similar chromatography conditions, to be similar from one run to the next.

Finally, the fourth term scores pairs of edges rather than a single edge. This term favors solutions that contain many similar pairs of edges. We say a pair of edges are similar to each other when they have similar normalized retention time shifts and the normalized retention times of the MS1 features, in the same run, are similar to each other. We define the term as

$$M_4(A) = \sum_{e_1, e_2 \in A} \exp(-\beta|s(e_1) - s(e_2)|) \exp(-\gamma|t(v(e_1)) - t(v(e_2))|), \quad (2)$$

where β and γ are hyperparameters. The first exponent is large when the normalized retention time shifts of the two edges are similar. On the other hand, the second exponent is large when pairs of MS1 features in the same run have similar normalized retention times. Altogether, this term scores edge pair similarity while allowing for systematic retention time shifts between two different mass spectrometry runs. Such shifts are common, reflecting differences in chromatography conditions between runs. Note that, for efficiency, we set M_4 to zero when $|s(e_1) - s(e_2)| > 0.01$ or when $|t(v(e_1)) - t(v(e_2))| > 0.01$.

2.4 Selecting the best matching

In order to choose the best matching, we must choose a feasible subset of edges that maximizes our objective function, that is, we must compute

$$\max_{A \subseteq E: A \in (\mathcal{E}_U \cap \mathcal{E}_V)} M(A) \quad (3)$$

where $M(A) = \sum_{i=1}^4 \lambda_i \frac{M_i(A)}{M_i(E)}$. The first thing to note is that $M(A)$ is the summation over entries of a $|U||V| \times |U||V|$ matrix. That is, $M(A) = \sum_{e_1, e_2 \in A} m(e_1, e_2)$ where $m(e_1, e_2) = \sum_{i=1}^4 \lambda_i m_i(e_1, e_2)$ and where $m_i(e_1, e_2)$ is the element of the corresponding sub-objective M_i defined above (that is, $M_i(A) = \sum_{e_1, e_2 \in A} m_i(e_1, e_2)$). Note that for $i \in \{1, 2, 3\}$ we have that $m_i(e_1, e_2) = 0$ whenever $e_1 \neq e_2$. We also note that $m_4(e_1, e_2) = 1$ whenever $e_1 = e_2$. Hence, we can define $m(e_1, e_2)$ as follows:

$$m(e_1, e_2) = \begin{cases} \sum_{i=1}^3 \lambda_i \frac{m_i(e_1, e_2)}{M_i(E)}, & \text{if } e_1 = e_2 \\ \lambda_4 \frac{m_4(e_1, e_2)}{M_4(E)}, & \text{if } e_1 \neq e_2. \end{cases} \quad (4)$$

Ordinarily, computing such a maximization over an exponential number of subsets would be intractable. It turns out, however, that this maximization problem exhibits useful structure that allows an efficient algorithm to be used to obtain

an approximate solution. Specifically we use the greedy algorithm to solve an instance of supermodular maximization subject to matroid constraints. In Supplemental Section 1.2 we discuss our approach to select the best matching as well as the approximation bound of the solution provided by the greedy algorithm.^{11–13,19,20}

Lastly, we note again that the above is described for a generic bipartite graph $G = (U, V, E)$. In practice, we have one distinct bipartite graph for each pair r_1, r_2 of runs; hence, the greedy algorithm is run for each run pair.

3 Methods

Additional methods can be found in the Supplement.^{21–28}

3.1 MS1Connect versions

Since the objective function of MS1Connect contains four terms, we compared five different versions of MS1Connect against each other. Four of the versions used a single term of the objective function while the fifth version used all four terms of the objective function. We denote the version of MS1Connect that only uses the M_1 term as “MS1Connect (M_1 only)”, and we denote the version of MS1Connect that uses all four terms as “MS1Connect ($M_1 - M_4$)”.

3.2 Baseline similarity measures

In order to ensure that our objective obtained sensible results, we compared MS1Connect against several baseline similarity measures. In general, our baseline similarity measures are based on binning the MS1 features and computing scalar products. More specifically, we represent the two runs, X and Y , as two matrices with dimensions m and n , where m is the number of m/z bins and n is the number of retention time bins. The value in each cell is the number of MS1 features whose discretized m/z and retention time fall into that bin. Both δ_2 , the m/z bin width, and n , the number of retention time bins, are hyperparameters.

Our baseline scores are defined as

$$S_{X,Y} = \sum_{i=0}^m \sum_{j=0}^n G(X_{i,j}, Y_{i,j}), \quad (5)$$

where X and Y are the matrix representations of the proteomics runs and G is the product, maximum, or minimum of the two input values. Alternatively, G could be defined as $G(S, T) = \mathbb{1}_{\{S \geq 1\}} \mathbb{1}_{\{T \geq 1\}}$, where $\mathbb{1}$ is the indicator function. Finally, we considered the baselines that do not consider retention time ($n = 1$) separately from baselines that do consider retention time ($n \geq 2$) for a total of eight different baselines.

3.3 Evaluation metrics

We use two different measures to quantify the performance of a given MS1 similarity score at predicting the metadata label of a query run given a repository of runs with known metadata labels.

The first performance measure is the per-query average precision (QAP), defined as the average precision across all queries $q \in R$:

$$\text{QAP} = \frac{1}{|R|} \sum_{q \in R} \sum_{k=1}^C P_k(S, q, R \setminus \{q\}) (R_k(S, q, R \setminus \{q\}) - R_{k-1}(S, q, R \setminus \{q\})), \quad (6)$$

where $P_k(S, q, R)$ and $R_k(S, q, R)$ are the precision and recall, respectively, after k repository runs have been retrieved from R using query q and similarity S . Note that the query run q is not included in the ranking when computing the precision and recall.

The second measure, aggregate average precision (AAP), is similar to QAP except that the average precision is calculated once on an aggregated list of similarity scores. This aggregated list is produced by sorting together all pairs of runs, considering only the upper triangle of the run-by-run matrix. In this work we focused on the QAP because these two performance metrics are highly correlated (Supplemental Figure 2).

method	species training set		species test set	
	QAP	AAP	QAP	AAP
MS1Connect ($M_1 - M_4$)	0.8115	0.7301	0.8483	0.7983
MS1Connect (M_1 only)	0.7873	0.6992	0.8112	0.7672
MS1Connect (M_2 only)	0.7096	0.5801	0.6686	0.5806
MS1Connect (M_3 only)	0.7891	0.7000	0.8130	0.7675
MS1Connect (M_4 only)	0.8076	0.7322	0.8497	0.7993
# m/z bins in common (indicator)	0.7847	0.6978	0.8211	0.7855
# $m/z + RT$ bins in common (indicator)	0.7677	0.6856	0.8083	0.7509
# m/z bins in common (product)	0.7771	0.7062	0.7951	0.7529
# $m/z + RT$ bins in common (product)	0.7591	0.6838	0.8107	0.7532
# m/z bins in common (min)	0.7831	0.6979	0.8225	0.7868
# $m/z + RT$ bins in common (min)	0.7665	0.6857	0.8092	0.7514

Table 1: **QAP and AAP.** A table of the per-query average precision (QAP) and the aggregate average precision (AAP) of the best performing hyperparameter set for each method in the species training and test datasets. Bold values denote the best and second best performance for each column. MS1Connect ($M_1 - M_4$) and MS1Connect (M_4 only) had the overall best performance. Note that the two supermodular methods outperformed all other methods.

4 Results

4.1 MS1Connect scores can be used for species prediction

To determine whether the MS1Connect score successfully measures the similarity of a pair of runs, we investigated whether these scores can be used to predict the species label of a proteomics run. Given a set of runs, we designated a single run as the query run and the remaining runs as the repository. The repository runs were then ordered by their similarity to the query run. Then, the average precision was calculated based on whether the repository runs had the same species label as the query run. We repeated this process for each of the runs in the dataset to calculate our two metrics, QAP and AAP. MS1Connect has nine different hyperparameters (Table 1), and to determine which set of hyperparameters achieved the best performance on the training set, as measured by QAP, we sampled the hyperparameter space using a random grid search.

We found that MS1Connect scores can be successfully used for predicting what species a sample was generated from (Table 1). Comparing the various versions of MS1Connect against each other, we discovered that the two supermodular versions of MS1Connect, MS1Connect ($M_1 - M_4$) and MS1Connect (M_4 only), had the best performance. Specifically, these two versions had the best or second best performance, as measured by QAP or AAP, in both the training and test data. These two versions are supermodular because they incorporate the supermodular M_4 term. The higher performance of these two supermodular versions of MS1Connect, compared to the three modular versions of MS1Connect, suggests that using supermodularity may lead to higher performance. In Supplemental Note 3.1 we discuss a possible reason why supermodular methods outperform modular methods.

Next, we compared the performance of MS1Connect against our baseline similarity methods (Table 1). We found that the supermodular MS1Connect methods outperformed all of the baselines. In addition, we found the performance of MS1Connect (M_1 only) and MS1Connect (M_3 only) to be similar to the performance of the baselines while MS1Connect (M_2 only) under-performed the baselines. Note that Table 1 does not include the performance of the two baselines where G is the maximum of the two input values as these methods exhibited poor performance (≤ 0.13 for both QAP and AAP in the training set).

After we compared MS1Connect against the baselines, we compared the various baselines against each other. We found that performance generally decreased when retention time was considered (Table 1). This result may seem unexpected because retention times, in principle, should improve the specificity of a matching. However, binning retention times could lead to a significant number of edge effects because of systemic retention time shifts that reflect differences in chromatography conditions. As an example, the number of MS1 features in common between two *C. albicans* runs, 1512006-2-TRIS1-10-Calbicans.raw and Control-60min_R3.raw, decreased by 30.8% after increasing the number of retention time bins from one to two.

Following the quantitative comparison, we examined the MS1Connect scores in the species training data to understand whether these scores made qualitative sense. We expected that runs generated by the same experiment should be highly similar to each other. Beyond that, we also expected runs from the same species to also be similar with each other.

Using the hyperparameter set that yielded the best performance, as measured by QAP on the training data, we visualized the MS1Connect scores between all pairwise runs in the species training dataset as a heatmap (Figure 2). In this heatmap the rows and columns are ordered by expected similarity. In general, the resulting heatmap matched our expectation. For example, the strong diagonal component in the heatmap resulted from runs that were generated by

Figure 2: **MS1Connect for species prediction.** A heatmap and dendrogram of MS1Connect scores showing the structure of our species training dataset. Each cell is colored by the MS1Connect score between a pair of runs. The solid white lines denote the border between different species while the dotted gray lines delineate the border between different experiments (PRIDE ID). The dendrogram shown is the phylogenetic tree of the nine species found in the training dataset.

the same experiment. In addition, the block-diagonal structure (delimited in the figure by the solid white lines) showed that runs generated from the same species tend to have high scores. For example, all the runs generated by *S. aureus* and *C. albicans* generally only have high scores with each other. On the other hand, we also found a few examples of runs whose similarity profile did not match our expectation. One prominent example is a set of five *E. coli* runs from the same experiment that only had high MS1Connect scores when comparing the same run to itself. A second example is a set of five *A. thaliana* runs from the same experiment where a run often did not have a high MS1Connect score to itself. This result occurred because these 10 runs have few MS1 features. In general, larger MS1Connect scores can be achieved when a run has a larger number of MS1 features. In this specific situation, these 10 runs have the fewest number of MS1 features in the overall dataset.

One unexpected finding was that MS1Connect scores may be sensitive enough to detect inter-species relationships. Considering the human and mouse runs, the MS1Connect scores indicated that these two species have some degree of similarity. This finding fits with our phylogenetic understanding of these two species, as human and mice are both mammals. These two species are highly similar to each other in the context of our dataset, which also included bacteria, fungi, and plants. Another example is that *B. anthracis* and *B. cereus* runs are similar to each other, which is expected since these two species are in the same genus. In the species test data, we saw comparable results with the two gram-negative bacterial species, *E. coli* and *S. enterica*, being similar to each other (Supplementary Figure 5).

A close inspection of the heatmap showed an unexpectedly high similarity between a group of *S. aureus* runs and a group of *H. sapiens* runs (Figure 2). The *S. aureus* runs originated from a study that used both cultures and clinical samples of *S. aureus*. We hypothesized that these runs originated from clinical samples and therefore contained a large number of human peptides. This type of contamination would cause these bacterial runs to be similar to human runs. We were unable to confirm this information from the PRIDE metadata, because the submission did not track which samples were cultured-based and which samples were clinically-based. However, a database search of the *S. aureus* runs against a concatenated *S. aureus* and *H. sapiens* proteome detected a large number of human peptides, supporting our hypothesis.

4.2 MS1Connect scores replicate database search results

Our analyses so far suggest that MS1Connect scores can successfully measure the similarity between pairs of proteomics runs. Next, we compared whether our MS1-based similarity method replicates similarities generated from MS2 data. We expect that these two methods should generally agree with each other.

To test this hypothesis, we compared the MS1Connect scores against the similarities generated by a MS2-based method on a set of samples with known species composition. For the MS2-based method, we calculated the Jaccard index between the sets of confidently detected peptides at a 1% FDR threshold for a pair of runs. The dataset we used contained a set of 16 samples, labeled “A” through “P”, where each sample contained one, two, or three different organisms (Supplementary Table 6). Each sample, except for the sample labeled “P” was run on a mass spectrometer

Figure 3: **Similarity between MS1Connect scores and MS2-based Jaccard index.** A heatmap of similarities scores where the upper triangle is MS1Connect scores and the lower triangle is the Jaccard index between the two sets of confidently detected peptides at a 1% FDR threshold for the multi-species dataset. These two methods are highly correlated with a Spearman rank correlation of 0.91. The solid white lines denote the border between samples. A description of what species are in each sample can be found in Supplementary Table 6. Within a box delineated by the white lines, the first three rows are the runs from one laboratory while the fourth row is the run from a second laboratory.

four times. Three of the runs were analyzed by a single instrument in one laboratory while the fourth run was analyzed on a different instrument in a second laboratory. The two laboratories are independent from each other, and therefore the samples were run with differing instrument parameters and chromatography conditions.

Our results showed that MS1Connect scores indeed have a high correspondence to similarities measured from MS2 data. A heatmap where the upper triangle is MS1Connect scores and the lower triangle is the Jaccard index showed that these two methods are highly consistent with each other (Figure 3). In addition, we found that these two scores are highly correlated, with a Spearman rank correlation of 0.91. Overall, this result showed that MS1Connect scores can replicate MS2-based analyses. In addition, our result suggested that MS1Connect scores can measure the similarity between runs that contain multiple species in the face of differing chromatography conditions and instrumentation.

A distinct checkerboard pattern in the lower triangle of the heatmap occurred because three out of the four runs from each sample are run on a single instrument while the fourth is run on a second instrument (Figure 3). This result matched our expectation since the three runs from the same instrument should be more similar to each other than the run from the second instrument. This pattern, while not as prominent, also exists on the MS1Connect side of the heatmap and further showed that MS1Connect scores can replicate database search results.

4.3 MS1Connect scores can identify mislabeled runs

Having shown the ability of MS1Connect to accurately measure the similarity between pairs of proteomics runs, we hypothesized that our method could be used to identify runs with mislabeled metadata. To this end, we analyzed data collected from seven different *Bacillus* species.²⁸ Each species was analyzed by mass spectrometry six times. We calculated the MS1Connect scores between all of the runs. In addition, we calculated the Jaccard index between the sets of confidently detected peptides at 1% FDR threshold for all pairs of runs.

We found that MS1Connect scores identify runs whose metadata may have been mislabeled. Specifically, when visualizing a heatmap of the MS1Connect scores between all pairs of runs (upper triangle of Figure 4), we observed that one of the runs labeled as *B. tonyonensis* exhibits low similarity with other *B. tonyonensis* runs but has high similarity with other *B. cytotoxicus* runs (last row/column of the heatmap). A database search also confirmed this result (lower triangle of Figure 4). Together, these results strongly suggest that a single run labeled as *B. tonyonensis* should instead be labeled *B. cytotoxicus*. This mislabelling could have occurred in many ways. For example, during the sample preparation process a vial could have been mislabeled or there could have been an error during some pipetting step. Alternatively, when placing samples into an autosampler, a vial could have been placed incorrectly onto the tray or

Figure 4: **Possible identification of a mislabeled run.** The figure is a heatmap of pairwise similarity scores, where the upper triangle is MSIConnect scores and the lower triangle is the Jaccard index between the two sets of confidently detected peptides at 1% FDR threshold for the *Bacillus* genus dataset. The x- and y-axis show the species labels of each run, as described by the PRIDE repository, with six runs per species. The solid white lines denote the border between species. Both the MSIConnect scores and the Jaccard index indicate that a run labeled as *B. toyonensis* (last row/column of the matrix) was mislabeled and should be labeled *B. cytotoxicus*.

the vial position could have been incorrectly recorded into the instrument.

4.4 Analysis of hyperparameters

After the assessment of our method, we investigated the performance of MS1Connect as a function of the hyperparameters to determine if any of the hyperparameters were correlated with performance. Specifically, we studied the performance of the 2500 different hyperparameterizations of MS1Connect, as described in Supplemental Section 2.1, with respect to each hyperparameter.

This analysis revealed trends in the performance of various hyperparameterizations of MS1Connect as a function of N and δ_1 (Figure 5). Specifically, we found that the best performance, as measured by QAP on the training dataset, occurred when the number of MS1 features used to generate the bipartite graph N was set to 4000 and when the m/z tolerance δ_1 was set to 4 ppm. We note that 4 ppm is well within the range of precursor mass tolerances researchers typically use for database searching. On the other hand, we found no trends in the performance of MS1Connect as a function of the remaining seven hyperparameters (Supplementary Figures 6–7).

Given that best performance occurred when $N = 4000$ and m/z tolerance = 4 ppm, we fixed these two hyperparameters and then asked whether there were any trends in the performance as a function of the remaining hyperparameters. We found that λ_4 was correlated with performance (Supplemental Figure 8A). The performance increased as λ_4 increased from 0.0 to 0.9, then performance decreased as λ_4 approached 1.0. This trend suggested that larger λ_4 may lead to higher performance and hence that supermodularity is important for achieving high performance. This is also shown by the fact that the best performing hyperparameterizations of MS1Connect occurred when $\lambda_4 = 0.9$. In addition to λ_4 , λ_2 was also found to be correlated with performance (Supplemental Figure 8B). In general, the performance MS1Connect decreased as λ_2 increased from 0.0 to 1.0. This result indicated that M_2 , which measures intensity, is not useful for achieving high performance. However, we note that another hyperparameter, N , limits MS1 features in the bipartite graph to high-intensity features. None of the remaining hyperparameters were correlated with performance (Supplementary Figures 9–10).

Figure 5: **Performance as a function of # of MS1 features or m/z tolerance.** These figures plot the performance of 2500 different hyperparameterizations of MS1Connect as a function of (A) # of MS1 features or (B) m/z tolerance. The best performance occurs when m/z tolerance is set to 4 ppm and # of MS1 features is set to 4000. Note that the values near the top are the number of points found in each column.

5 Discussion

In this work, we introduced a new method, MS1Connect, that measures the similarity between pairs of proteomics runs. We showed evidence that MS1Connect successfully measures proteomics run similarity. Specifically, we showed that MS1Connect scores can be used for classifying what species a sample was generated from, and we showed evidence that MS1Connect scores can replicate database search results.

We also considered including in our empirical comparison a method based on dynamic time warping, but this approach turned out to be computationally infeasible. For this approach, we used FastDTW,²⁹ with a cosine distance as the distance metric, to score the warping between two runs. We estimated that running dynamic time warping between all pairs of runs in our training dataset for the largest m/z bin width (0.01 Da) in our hyperparameter range would take approximately three weeks and 300 Gb of memory. The large memory requirement is due to the large matrices being compared to each other. For example, the run with the largest matrix had approximately 30,000 time bins and 700,000 m/z bins (m/z bin width of 0.0035 Da). The long time requirement resulted from having to calculate the warping between all pairs of runs (>13,000 warpings).

Since MS1Connect only uses MS1 data, it is agnostic to acquisition style. Therefore, MS1Connect can calculate the similarity between runs that were collected by DIA, DDA, or by targeted methods. While we did not specifically test this idea, we note that one set of five *S. aureus* runs in the training data was collected by DIA while the remaining runs were collected by DDA. Our results suggested that MS1Connect scores can measure the similarity between DIA and DDA runs. In turn, this ability to measure similarity is the first step for allowing future joint analysis of DDA and DIA data.

While we have shown that MS1Connect scores can be used to predict metadata labels of runs, future works needs to be conducted to determine the limits of our method. For example, none of the data used in this study was isobarically labeled. MS1Connect could be extended to include these types of samples. In addition, MS1Connect has not been tested on a wide variety of samples, such as fractionated samples. Additional work could be pursued to test the applicability of MS1Connect in these scenarios.

Finally, we note that while we developed MS1Connect for use in proteomics our method can be used to analyze data from other mass spectrometry-based fields such as metabolomics and lipidomics. Our method may be especially useful in these fields because MS2-based identifications are frequently missing. Therefore, MS2-based methods may not be as valuable. In the future, MS1Connect could be applied to data generated by metabolomics and lipidomics.

6 Acknowledgments

Some of the work was funded by NIH award R01GM121818. In addition, some of the research described in this paper was conducted under the Laboratory Directed Research and Development Program at Pacific Northwest National

Laboratory, a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy. Andy Lin is grateful for the support of the Linus Pauling Distinguished Postdoctoral Fellowship program. Pacific Northwest National Laboratory is a multiprogram national laboratory operated by Battelle Memorial Institute for the United States Department of Energy under contract DE-AC06-76RLO.

References

- [1] Y. Perez-Riverol, A. Csordas, J. Bai, M. Bernal-Llinares, S. Hewapathirana, D. J. Kundu, A. Inuganti, J. Griss, G. Mayer, M. Eisenacher, E. Pérez, J. Uszkoreit, J. Pfeuffer, T. Sachsenberg, S. Yilmaz, S. Tiwary, J. Cox, E. Audain, M. Walzer, A. F. Jarnuczak, T. Ternent, A. Brazma, and J. A. Vizcaíno. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res*, 47(D1):D442–D450, 01 2019.
- [2] M. Choi, J. Carver, C. Chiva, M. Tzouros, T. Huang, T. H. Tsai, B. Pullman, O. M. Bernhardt, R. Hüttenhain, G. C. Teo, Y. Perez-Riverol, J. Muntel, M. Müller, S. Goetze, M. Pavlou, E. Verschueren, B. Wollscheid, A. I. Nesvizhskii, L. Reiter, T. Dunkley, E. Sabidó, N. Bandeira, and O. Vitek. MassIVE.quant: a community resource of quantitative mass spectrometry-based proteomics datasets. *Nature Methods*, 17(10):981–984, 10 2020.
- [3] D. L. Tabb, L. Vega-Montoto, P. A. Rudnick, A. M. Variyath, A. J. Ham, D. M. Bunk, L. E. Kilpatrick, D. D. Billheimer, R. K. Blackman, H. L. Cardasis, S. A. Carr, K. R. Clauser, J. D. Jaffe, K. A. Kowalski, T. A. Neubert, F. E. Regnier, B. Schilling, T. J. Tegeler, M. Wang, P. Wang, J. R. Whiteaker, L. J. Zimmerman, S. J. Fisher, B. W. Gibson, C. R. Kinsinger, M. Mesri, H. Rodriguez, S. E. Stein, P. Tempst, A. G. Paulovich, D. C. Liebler, and C. Spiegelman. Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *Journal of Proteome Research*, 9(2):761–776, Feb 2010.
- [4] V. Rieder, B. Blank-Landeshammer, M. Stuhr, T. Schell, K. Biß, L. Kollipara, A. Meyer, M. Pfenninger, H. Westphal, A. Sickmann, and J. Rahnenführer. DISMS2: A flexible algorithm for direct proteome-wide distance calculation of LC-MS/MS runs. *BMC Bioinformatics*, 18(1):148, Mar 2017.
- [5] I. Belghit, E. J. Lock, O. Fumière, M. C. Lecrenier, P. Renard, M. Dieu, M. H. G. Berntssen, M. Palmblad, and J. D. Rasinger. Species-Specific Discrimination of Insect Meals for Aquafeeds by Direct Comparison of Tandem Mass Spectra. *Animals (Basel)*, 9(5), May 2019.
- [6] M. Palmblad and A. M. Deelder. Molecular phylogenetics by direct comparison of tandem mass spectra. *Rapid Commun Mass Spectrom*, 26(7):728–732, Apr 2012.
- [7] S. J. van der Plas-Duivesteijn, T. Wulff, O. Klychnikov, D. Ohana, H. Dalebout, P. A. van Veelen, J. de Keijzer, M. A. Nessen, Y. E. van der Burgt, A. M. Deelder, and M. Palmblad. Differentiating samples and experimental protocols by direct comparison of tandem mass spectra. *Rapid Commun Mass Spectrom*, 30(6):731–738, Mar 2016.
- [8] H. L. Röst, T. Sachsenberg, S. Aiche, C. Bielow, H. Weisser, F. Aicheler, S. Andreotti, H. Ehrlich, P. Gutenbrunner, E. Kenar, X. Liang, S. Nahnsen, L. Nilse, J. Pfeuffer, G. Rosenberger, M. Rurik, U. Schmitt, J. Veit, M. Walzer, D. Wojnar, W. E. Wolski, O. Schilling, J. S. Choudhary, L. Malmström, R. Aebersold, K. Reinert, and O. Kohlbacher. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nature Methods*, 13(9):741, 2016.
- [9] J. Cox and M. Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26:1367–1372, 2008.
- [10] X. Wang, S. Shen, S. S. Rasam, and J. Qu. MS1 ion current-based quantitative proteomics: A promising solution for reliable analysis of large biological cohorts. *Mass Spectrom Rev*, 38(6):461–482, 11 2019.
- [11] U. Feige. A threshold of $\ln n$ for approximating set cover. *Journal of the ACM*, 45(4):634–652, 1998.
- [12] H. Lin and J. A. Bilmes. Optimal selection of limited vocabulary speech corpora. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Florence, Italy, August 2011.

- [13] W. Bai and J. Bilmes. Greed is still good: Maximizing monotone Submodular+Supermodular (BP) functions. In *Proceedings of the 35th International Conference on Machine Learning*, pages 304–313, 10–15 Jul 2018.
- [14] S. Ji, D. Xu, M. Li, Y. Wang, and D. Zhang. Stochastic greedy algorithm is still good: Maximizing submodular + supermodular functions. In H. Le Thi, H. Le, and T. Pham Dinh, editors, *Optimization of Complex Systems: Theory, Models, Algorithms and Applications*, pages 488–497, Cham, 2020. Springer International Publishing.
- [15] W. Bai, J. A. Bilmes, and W. S. Noble. Bipartite matching generalizations for peptide identification in tandem mass spectrometry. In *ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 327–336, Seattle, WA, 2016.
- [16] H. L. Röst, U. Schmitt, R. Aebersold, and L. Malmström. pyOpenMS: a Python-based interface to the OpenMS mass-spectrometry algorithm library. *Proteomics*, 14(1):74–77, Jan 2014.
- [17] M. C. Chambers, B. Maclean, R. Burke, D. Amodei, D. L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, K. Hoff, D. Kessner, N. Tasman, N. Shulman, B. Frewen, T. A. Baker, M. Y. Brusniak, C. Paulse, D. Creasy, L. Flashner, K. Kani, C. Moulding, S. L. Seymour, L. M. Nuwaysir, B. Lefebvre, F. Kuhlmann, J. Roark, P. Rainer, S. Detlev, T. Hemenway, A. Huhmer, J. Langridge, B. Connolly, T. Chadick, K. Holly, J. Eckels, E. W. Deutsch, R. L. Moritz, J. E. Katz, D. B. Agus, M. J. MacCoss, D. L. Tabb, and P. Mallick. A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology*, 30(10):918–920, 2012.
- [18] L. Martens, M. Chambers, M. Sturm, D. Kessner, F. Levander, J. Shofstahl, W. H. Tang, A. Römpf, S. Neumann, A. D. Pizarro, L. Montecchi-Palazzi, N. Tasman, M. Coleman, F. Reisinger, P. Souda, H. Hermjakob, P. A. Binz, and E. W. Deutsch. mzML—a community standard for mass spectrometry data. *Mol Cell Proteomics*, 10(1):R110.000133, Jan 2011.
- [19] J. Edmonds. Matroids, submodular functions, and certain polyhedra. *Combinatorial Structures and Their Applications*, pages 69–87, 1970.
- [20] J.G. Oxley. *Matroid Theory: Second Edition*. Oxford University Press, 2011.
- [21] J. K. Cole, J. R. Hutchison, R. S. Renslow, Y. M. Kim, W. B. Chrisler, H. E. Engelmann, A. C. Dohnalkova, D. Hu, T. O. Metz, J. K. Fredrickson, and S. R. Lindemann. Phototrophic biofilm assembly in microbial-mat-derived uncyanobacterial consortia: model systems for the study of autotroph-heterotroph interactions. *Front Microbiol*, 5:109, 2014.
- [22] M. Chatterjee, S. Gupta, A. Bhar, and S. Das. Optimization of an Efficient Protein Extraction Protocol Compatible with Two-Dimensional Electrophoresis and Mass Spectrometry from Recalcitrant Phenolic Rich Roots of Chickpea (*Cicer arietinum* L.). *Int J Proteomics*, 2012:536963, 2012.
- [23] R. T. Kelly, J. S. Page, Q. Luo, R. J. Moore, D. J. Orton, K. Tang, and R. D. Smith. Chemically etched open tubular and monolithic emitters for nanoelectrospray ionization mass spectrometry. *Analytical Chemistry*, 78(22):7796–7801, Nov 2006.
- [24] S. McIlwain, K. Tamura, A. Kertesz-Farkas, C. E. Grant, B. Diamant, B. Frewen, J. J. Howbert, M. R. Hoopmann, L. Käll, J. K. Eng, M. J. MacCoss, and W. S. Noble. Crux: rapid open source protein tandem mass spectrometry analysis. *Journal of Proteome Research*, 13(10):4488–4491, 2014. PMC4184452.
- [25] C. Y. Park, A. A. Klammer, L. Käll, M. P. MacCoss, and W. S. Noble. Rapid and accurate peptide identification from tandem mass spectra. *Journal of Proteome Research*, 7(7):3022–3027, 2008.
- [26] A. Lin, J. J. Howbert, and W. S. Noble. Combining high-resolution and exact calibration to boost statistical power: A well-calibrated score function for high-resolution ms2 data. *Journal of Proteome Research*, 17:3644–3656, 2018.
- [27] The UniProt Consortium. UniProt: a worldwide hub for protein knowledge. *Nucleic Acids Research*, pages D506–D515, 2019.
- [28] S. Pfrunder, J. Grossmann, P. Hunziker, R. Brunisholz, M. Gekenis, and D. Drissner. Bacillus cereus group-type strain-specific diagnostic peptides. *Journal of Proteome Research*, 15(9):3098–3107, September 2016.

- [29] S. Salvador and P. Chan. FastDTW: Toward accurate dynamic time warping in linear time and space. *arXiv*, 2004.