Bridging the utility gap between MALDI-TOF and WGS for affordable outbreak cluster detection

Chang Liu CHANGL8@ANDREW.CMU.EDU
Auton Lab, Carnegie Mellon University, Pittsburgh, PA, USA

Jieshi Chen JIESHIC@ANDREW.CMU.EDU
Auton Lab, Carnegie Mellon University, Pittsburgh, PA, USA

Alexander J. Sundermann ALS412@PITT.EDU
Center for Genomic Epidemiology, University of Pittsburgh School of Medicine and Public Health, Pittsburgh, PA, USA

Lee H. Harrison LHARRISO@EDC.PITT.EDU
Center for Genomic Epidemiology, University of Pittsburgh School of Medicine and Public Health, Pittsburgh, PA, USA

Artur Dubrawski AWD@ANDREW.CMU.EDU
Auton Lab, Carnegie Mellon University, Pittsburgh, PA, USA

Abstract

Rapid and accurate detection of emerging outbreak clusters can help contain the spread of diseases with epidemic potential. Among the available pathogen matching methods that can be used to support the task, whole genome sequencing (WGS) offers the highest discriminatory power but is expensive and time-consuming. On the other hand, Matrix-Assisted Laser Desorption Ionization–Time of Flight (MALDI-TOF) mass spectrometry is gaining attention for being a rapid and cost-effective, albeit less precise, alternative. In order to combine the strengths of both MALDI-TOF and WGS, we present **MSMAP**, the first machine learning framework that establishes a mapping between MALDI-TOF mass spectra and the single nucleotide polymorphism (SNP) distances obtained from WGS analysis. We demonstrate the effectiveness of MSMAP in retrieving WGS-defined outbreak clusters on synthetic mass spectrum data and on proprietary data with paired MALDI-TOF and SNP information. The results show that MSMAP augments MALDI-TOF with the discriminatory power of WGS, thus bridging their utility gap and paving the way toward fast, accurate, and cost-effective outbreak cluster detection.

Data and Code Availability The synthetic data and code for this study are available at https://github.com/ChangLiu-DrPatient/MSMAP. The proprietary dataset is de-identified and obtained from a large non-profit research and academic hospital with sharing restrictions due to data use agreements.

Institutional Review Board (IRB) This research does not require IRB approval.

1. Introduction

Healthcare-associated infections may result in outbreaks through the unchecked transmission of pathogens between patients, equipment, and healthcare personnel (Maragakis et al., 2004; Epstein et al., 2014). To help prevent mortality and reduce the resources needed to deal with a massive outbreak, detecting potential outbreaks in their early stages is critical (Palmore and Henderson, 2013). To this end, institutions collect epidemiological data, including patient and environmental samples, to identify outbreak clusters (Archibald and Jarvis, 2011). The clusters contain samples that are considered related to recent transmission events (Poon, 2016), providing a means of tracking the spread of pathogens (Foxman and Riley, 2001).

Among the methods of assessing pathogen similarity needed to detect and attribute outbreak clusters, whole genome sequencing (WGS) provides the highest discriminatory power (Mellmann et al., 2016). Specifically, the number of single-nucleotide polymorphisms (SNP) that differ between two sequences is of-

ten used as a dissimilarity measure ("SNP distance"). Two sequences are assigned to the same cluster if their SNP distance is below a set threshold (Hatherell et al., 2016). However, the labor, cost, and expertise required for the WGS analysis prohibit its broad deployment (Rossen et al., 2018).

On the other hand, due to its low cost and short time to generate results, Matrix-Assisted Laser Desorption Ionization–Time of Flight (MALDI-TOF) mass spectrometry, a standard tool for species identification in clinical microbiology (Croxatto et al., 2012; Clark et al., 2013), is gaining traction as a possible alternative to WGS for outbreak cluster detection (Griffin et al., 2012). MALDI-TOF generates intensity spectra against mass-to-charge ratios (m/z) for microbial proteins, providing a unique "fingerprint" of the microorganism. Cluster analysis of the spectra is then performed to infer outbreak clusters (Rödel et al., 2019; Giraud-Gatineau et al., 2021).

However, for microorganisms with noticeable intraspecies genetic variability, MALDI-TOF can yield highly similar spectra (Veenemans et al., 2016), limiting its discriminative power (Sandrin et al., 2013) and its utility for source attribution in outbreak analysis (Murray, 2010). In fact, several studies comparing MALDI-TOF with WGS found a discrepancy between their detected outbreak clusters (Schlebusch et al., 2017; Dinkelacker et al., 2018), challenging the utility of MALDI-TOF in outbreak investigations.

Given this perceived deficiency of the MALDI-TOF approach, we aim to augment its utility in outbreak cluster detection. Specifically, we develop **MSMAP**, a machine learning framework to learn representations of MALDI-TOF spectra that respect the clustering structure of the corresponding WGS results. Moving beyond research that merely compares the utility of MALDI-TOF and WGS, MSMAP is the *first* attempt to leverage structural information of genomic similarity conveyed by SNP obtained through WGS to inform outbreak cluster analysis driven by MALDI-TOF. We demonstrate the effectiveness of MSMAP in recovering WGS clusters on synthetic mass spectrum data and on proprietary data with paired MALDI-TOF and SNP distance information. Our results show that MSMAP can effectively use Machine Learning (ML) to bridge the modality and capability gaps between MALDI-TOF and WGS in a step towards fast, accurate, yet more cost-effective outbreak cluster detection.

2. Related work

MALDI-TOF vs. WGS for outbreak cluster detection. The use of genetic data to detect and attribute outbreak clusters dates back to 1978 when researchers used restriction fragment length polymorphisms (RFLPs) to cluster isolates from a nosocomial outbreak of herpes simplex virus type 1 (HSV-1) (Buchman et al., 1978). This method has evolved to adopt whole genome sequencing (WGS) to better understand transmission events of pathogens, e.g., *Pseudomonas aeruginosa* (PSA) (Quick et al., 2014; Sundermann et al., 2021) and vancomycin-resistant *Enterococcus faecium* (VRE) (Abdelbary et al., 2019; Sundermann et al., 2020).

MALDI-TOF mass spectrometry has also been applied to outbreak cluster detection in several studies (Rödel et al., 2019; Bar-Meir et al., 2020; Giraud-Gatineau et al., 2021). Additionally, Mohammad et al. (2023) utilized a convolutional neural network (CNN) to encode MALDI-TOF spectra yet simplified the outbreak cluster detection problem to binary classification, i.e., whether an isolate belongs to a fluconazole-resistant clonal subset or not.

Despite the advances in using MALDI-TOF for outbreak cluster detection, several studies pointed out its lack of discriminatory power compared to WGS and suggested caution in interpreting MALDI-TOF clustering results. Schlebusch et al. (2017) applied principal component analysis (PCA) to cluster MALDI-TOF spectra from a VRE outbreak and found vast inconsistencies with WGS-defined clusters; Dinkelacker et al. (2018) used the UPGMA algorithm to cluster MALDI-TOF spectra of *Klebsiella* isolates and found low congruency with the WGS reference.

Departing from the studies that compare MALDI-TOF and WGS, our approach presents the first attempt at explicitly bridging their utility gap for outbreak cluster detection. In addition, we expand the scope of these studies by simultaneously investigating the utility of MALDI-TOF on multiple species.

Machine learning for MALDI-TOF spectrometry. Machine learning is widely applied to MALDI-TOF spectra analysis besides outbreak cluster detection. Weis et al. (2022) used a multilayer perceptron (MLP) to encode MALDI spectra of bacterial strains and predict their antimicrobial resistance (AMR) to a range of drugs. De Waele et al. (2024) developed a dual branch network that encodes MALDI-TOF spectra and drugs to predict AMR profiles. De Waele et al. (2025) pretrained a transformer

model from MALDI-TOF spectra and deployed it for predicting species and AMR. Abdelmoula et al. (2021) employed a variational autoencoder to learn low-dimensional latent features of MALDI-TOF spectra that reveal biologically relevant clusters of tumor regions. These methods corroborate the vast potential of using machine learning to capture MALDI-TOF features relevant to downstream tasks, inspiring us to develop a machine learning framework for predicting WGS clusters from MALDI-TOF spectra.

3. Methods

The MSMAP framework. Our primary goal is to replace expensive WGS with affordable MALDI-TOF in the analysis of microbial isolates to identify clusters of outbreaks of diseases. We propose to accomplish this by training a machine learning model that can retain the structure of microbial similarities measured with WGS-based SNP distances in the equivalent similarity model computed using MALDI-TOF spectrometry.

We train our model using pairwise SNP distances from WGS alongside corresponding MALDI-TOF characteristics from microbial isolates. Once trained, the model will estimate isolate similarities using only MALDI-TOF data without requiring SNP information.

Our approach involves training an autoencoder to learn compact MALDI-TOF embeddings, enhanced by an auxiliary species identification task. This ensures the embeddings are meaningful for an established application of MALDI-TOF spectrometry.

In summary, as presented in Figure 1, training of the MSMAP framework entails three objectives: (1) reconstruction of the input MALDI-TOF spectra through an autoencoder for feature extraction, (2) species identification from MALDI-TOF spectra to retain its inherent utility, and (3) learning a mapping from the distances between MALDI-TOF representations and the corresponding SNP distances.

Formally, we first preprocess the raw spectra, which contain pairs of mass-to-charge ratios (m/z) with intensities, by selecting pairs with m/z values between 2000 and 20000 Daltons and binning them into 6000 equally-spaced bins. The binned spectra are then padded to length 6016 to enable further encoding. With preprocessed spectra $\boldsymbol{x}$ of batch size N as input, an autoencoder consisting of an encoder $\mathrm{Enc}_x(\cdot)$ and a decoder $\mathrm{Dec}(\cdot)$ is trained to embed $\boldsymbol{x}$ onto a bottleneck representation $\boldsymbol{b}$ and uses it to obtain the reconstructed input spectra $\hat{\boldsymbol{x}}$. A further encoding step $\mathrm{Enc}_h(\cdot)$ prepares $\boldsymbol{b}$ for species identification, performed by the classification head $\mathrm{CLS}(\cdot)$, and for alignment with the SNP distances. In all our experiments, the autoencoder adopts the U-Net (Ronneberger et al., 2015) architecture on 1-D data. The mathematical formulation for the MSMAP framework is as follows:

$$\boldsymbol{b} = \mathrm{Enc}_x(\boldsymbol{x}), \tag{1}$$
$$\hat{\boldsymbol{x}} = \mathrm{Dec}(\boldsymbol{b}), \tag{2}$$
$$\boldsymbol{h} = \mathrm{Enc}_h(\boldsymbol{b}), \tag{3}$$
$$\boldsymbol{z} = \mathrm{CLS}(\boldsymbol{h}). \tag{4}$$

We then define the loss functions corresponding to the three learning objectives:

$$\mathcal{L}_{\text{recon}} = \frac{1}{N}\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2, \tag{5}$$
$$\mathcal{L}_{\text{species}} = \mathrm{CE}(\boldsymbol{z}, \boldsymbol{y}), \tag{6}$$
$$\mathcal{L}_{\text{SNP}} = f_{\text{SNP}}(\mathrm{pdist}(\boldsymbol{h}), \boldsymbol{d}). \tag{7}$$

Here, $\boldsymbol{y}$ represents the labels for species identification, $\mathrm{CE}(\cdot)$ stands for the cross-entropy loss, $\mathrm{pdist}(\boldsymbol{h})$ computes the ℓ_2 distance between each pair of the learned MALDI-TOF representations $\boldsymbol{h}$, and $\boldsymbol{d}$ refers to the SNP distance matrix. The function f_{SNP} computes the dissimilarity between these two distances (e.g., mean squared difference).

The loss for MSMAP is a weighted sum of these three losses:

$$\mathcal{L}_{\text{MSMAP}} = \mathcal{L}_{\text{recon}} + \lambda_0 \mathcal{L}_{\text{species}} + \lambda_1 \mathcal{L}_{\text{SNP}}, \tag{8}$$

where λ_0 and λ_1 are hyperparameters controlling the relative importance of the component losses.

During training, we search the combination of λ_0 and λ_1 among the grid $[1, 10, 100] \times [1, 10, 100]$. The choice of $(\lambda_0, \lambda_1) = (10, 1)$ yields the lowest SNP distance mapping loss while preserving the species classification accuracy. We also search for the dimension of the representation $\boldsymbol{h}$ among the grid $[32, 64, 128, 256, 512, 1024]$, with 64 and 512 being optimal for the synthetic spectrum dataset and the proprietary dataset, respectively. We train the model on the proprietary dataset using the Adam optimizer and a cosine with a learning rate of 5e-5 and a weight decay of 1e-5 for 128 epochs. For synthetic spectrum dataset, we use a learning rate of 1e-3, a weight decay of 1e-5, and train for 250 epochs. In both cases, we apply a cosine annealing learning rate scheduler

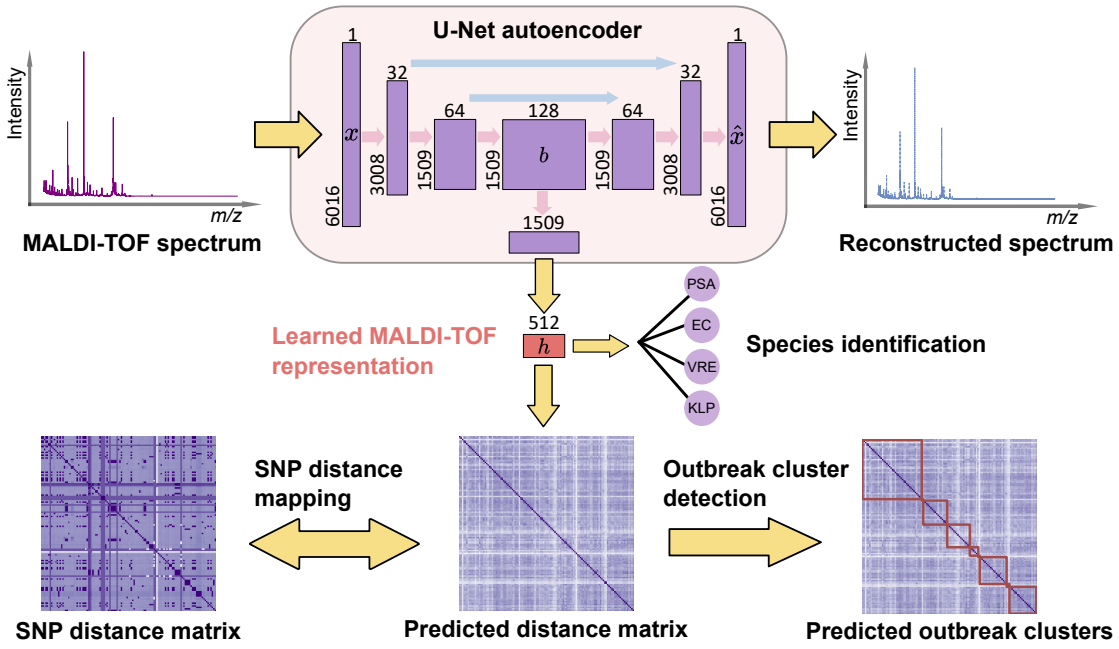

Figure 1: Overview of the MSMAP framework. MSMAP takes MALDI-TOF spectra as input to a U-Net autoencoder, which encodes the spectra to perform spectra reconstruction, species identification, and SNP distance mapping.

whose maximum iterations is set to the number of training epochs.

Through MSMAP, we aim to learn MALDI-TOF representations $\boldsymbol{h}$ that retain discriminatory power on species identification while preserving the clustering structure characterized by SNP distances $\boldsymbol{d}$ from WGS. After training, MSMAP can predict SNP distances from MALDI-TOF spectra obtained for pairs of microbial isolates, thereby bridging the utility gap between MALDI-TOF and WGS.

Baseline methods. To assess MSMAP, we develop three baseline methods. The first, named "clusCLS," borrows from Mohammad et al. (2023), formulating outbreak cluster detection as the classification of cluster labels. Specifically, the ground truth cluster labels $\mathcal{C}_T$ are derived beforehand using the SNP distance matrix $\boldsymbol{d}$ on the full dataset. Then, an additional classification head $\text{CLS}_C(\cdot)$is used to classify these cluster labels:

$$\boldsymbol{z}_c = \text{CLS}_C(\boldsymbol{h} \| \boldsymbol{z}), \tag{9}$$

where "$\|$" stands for the concatenation operation, and $\boldsymbol{h}$ and $\boldsymbol{z}$ are the learned MALDI-TOF representations and species classification logits as computed in Equations 3 and 4, respectively. clusCLS also replaces $\mathcal{L}_{\text{SNP}}$ with the following cross-entropy loss:

$$\mathcal{L}_{\text{clus}} = \text{CE}(\boldsymbol{z}_c, \mathcal{C}_T). \tag{10}$$

The second method, named "onlyCLS," constructs an ablation study by removing the SNP distance mapping objective $\mathcal{L}_{\text{SNP}}$ from the MSMAP loss while preserving all model hyperparameters of MSMAP. Hence, onlyCLS can only rely on the feature extraction capabilities of the autoencoder and the species identification task to implicitly recover the outbreak clusters.

The third method, denoted "rawClus," follows previous comparative studies of MALDI-TOF and WGS (Schlebusch et al., 2017; Dinkelacker et al., 2018) by applying the hierarchical clustering algorithm directly to the pre-processed MALDI-TOF spectra $\boldsymbol{x}$ in hopes of recovering the SNP dissimilarity structure directly.

Obtaining outbreak clusters. We use the trained models to obtain MALDI-TOF representations $\boldsymbol{h}$, which are clustered using the hierarchical clustering algorithm with a distance threshold. We

select the threshold that optimizes the similarity between the resulting MALDI-TOF representation clusters in the training set and the corresponding "ground truth" clusters defined by the SNP distances. In the context of outbreak cluster detection, we propose a novel metric called the *cluster F1 score* to measure the similarity between cluster assignments.

As a preliminary, we first define the *purity* of a cluster assignment with respect to ground truth cluster labels: given a dataset X, a set of clusters $\{C_1, \cdots, C_p\}$, and a ground truth cluster label set Y on X, the purity of the j-th cluster is defined as

$$\text{Purity}(C_j, Y) = \frac{1}{|C_j|} \max_{k \in Y} |C_j \cap T_k|, \tag{11}$$

where T_k is the set of data points in X with label k.

We now define the *cluster F1 score* as follows. Specifically, given a dataset X, predicted clusters $\mathcal{C}_S = \{C_S^1, \cdots, C_S^p\}$ derived from the MALDI-TOF representations $\boldsymbol{h}$, and "ground truth" clusters $\mathcal{C}_T = \{C_T^1, \cdots, C_T^q\}$ derived from the SNP distances $\boldsymbol{d}$, we define the cluster *precision*, *recall*, and *F1 score* of the predicted $\mathcal{C}_S$ with respect to $\mathcal{C}_T$ as follows:

$$\text{Prec}(\mathcal{C}_S, \mathcal{C}_T) = \frac{1}{p} \sum_{i=1}^{p} \text{Purity}(C_S^i, \mathcal{C}_T), \tag{12}$$

$$\text{Rec}(\mathcal{C}_S, \mathcal{C}_T) = \frac{1}{q} \sum_{j=1}^{q} \text{Purity}(C_T^j, \mathcal{C}_S), \tag{13}$$

$$\text{F1}(\mathcal{C}_S, \mathcal{C}_T) = 2 \cdot \frac{\text{Prec}(\mathcal{C}_S, \mathcal{C}_T) \cdot \text{Rec}(\mathcal{C}_S, \mathcal{C}_T)}{\text{Prec}(\mathcal{C}_S, \mathcal{C}_T) + \text{Rec}(\mathcal{C}_S, \mathcal{C}_T)}. \tag{14}$$

To reach high precision, each predicted cluster must contain as few distinct ground truth cluster labels as possible (i.e., be pure), attaining the maximum value of 1 when $p = |X|$ and $|C_S^i| = 1$ for all i. To achieve high recall, the data points with the same ground truth label should be clustered together in the predictions, reaching the maximum value of 1 when $p = 1$ and $|C_S^1| = |X|$. In contrast, a cluster assignment with a high F1 score does not fall into either extreme. Nevertheless, to avoid the impact of singleton ground truth clusters ($|C_T^j| = 1$) on the above metrics, we only evaluate them on the subset of data where the ground truth cluster has more than one element, i.e., $\cup_{j:|C_T^j|\geq 2} C_T^j \subseteq X$.

We argue that thusly proposed *cluster F1 score* has clinical relevance in outbreak cluster detection. To capture most of the isolates with genetic similarity, the predicted outbreak clusters should yield high recall. However, to avoid false alarms that exhaust resources in infection investigations, precision should be high. Being the harmonic mean of the recall and precision, the *cluster F1 score* aims to reflect model performance against both of those objectives in support of outbreak detection and prevention practices.

Note that our approach also allows the use of other standard cluster similarity metrics, such as the Normalized Mutual Information (NMI) and the adjusted Rand Index (ARI). These two metrics, together with the cluster F1 score, are reported when evaluating the outbreak clusters on the validation data.

4. Datasets

To demonstrate the ability of MSMAP to learn MALDI-TOF representations that preserve clustering structure from WGS, we developed a synthetic mass spectrum dataset (SynthSpec) and used a proprietary dataset of MALDI spectra with paired whole genome sequencing SNP distance profiles.

SynthSpec. We constructed a synthetic mass spectrum dataset with defined latent clusters that aims to mimic the relationship between MALDI-TOF and WGS in real-world outbreak investigations. Each spectrum in SynthSpec contains two peaks whose locations and intensities define the species and outbreak clusters (for each species), respectively. Specifically, as illustrated in Figure 2, peak locations are sampled from 1-D truncated normal distributions centered on fixed points for each species, while the intensities are sampled from 2-D Gaussian distributions whose distances exceed a "ground truth" threshold.

With a fixed population size of 6400 spectra (3200 per species), we constructed three versions of SynthSpec with increasing numbers of clusters: SynthSpec(8), SytnhSpec(16), and SynthSpec(32), which contain 8, 16, and 32 clusters per species, respectively. For each dataset, we keep an identical number of spectra per cluster, and the SNP mapping loss in Equation 7 is defined as the mean-squared difference between pdist($\boldsymbol{h}$) and the pairwise distances $\boldsymbol{d}$ between the sampled intensities. More details on the construction of SynthSpec can be found in Appendix A.

Proprietary dataset. The proprietary dataset consists of 8382 bacterial samples with MALDI spectra spanning 17 species of bacteria with corresponding WGS information obtained from a single hospital. A more detailed description of the dataset characteristics can be found in Appendix B. Though the raw

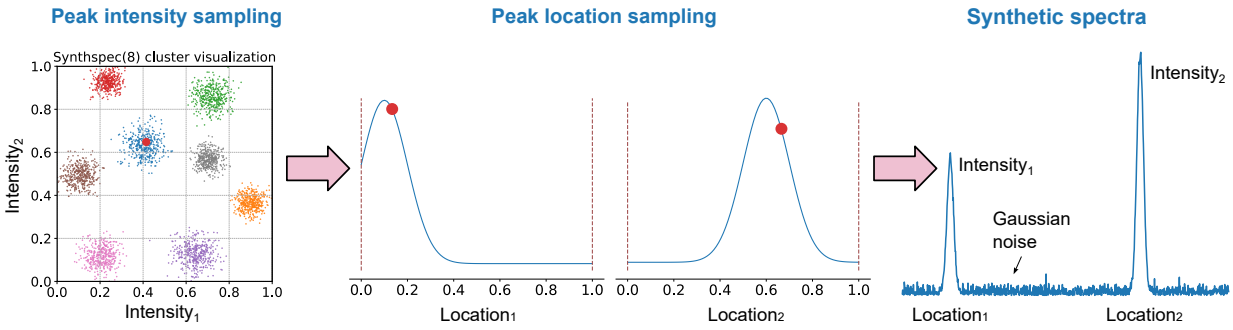

Figure 2: Synthetic mass spectra construction in SynthSpec.

WGS data were unavailable, we have access to their SNP distances. To derive the ground truth outbreak clusters from the SNP distances, we apply hierarchical clustering with complete linkage on the full SNP distance matrix, using 15 as the distance threshold. This value is used operationally by the infection control team in the hospital, from which we obtained proprietary data, as a common reference in their monitoring of the emergence of disease outbreak clusters.

Although small SNP distances are crucial to outbreak cluster detection, actual distances can vary drastically in scale, ranging from less than 10 to greater than 10^5. Here, we develop a custom loss function for Equation 7 to avoid potential overfitting to large SNP distances and nonlinearly focus our model on capturing the impact of SNP distance variability in the lower range of its values:

$$\mathcal{L}_{\text{SNP}} = \frac{1}{N^2} \sum_{i,j \in [N]^2} f_{\text{SNP}}(\text{pdist}(\boldsymbol{h})_{ij}, \boldsymbol{d}_{ij}, t), \quad (15)$$

$$f_{\text{SNP}}(x, y, t) = \begin{cases} (x-y)^2 & y \le t, \\ (\max\{0, t-x\})^2 & y > t, \end{cases} \quad (16)$$

where $\text{pdist}(\boldsymbol{h})$ and $\boldsymbol{d}$ are the MALDI-TOF representation distance and SNP distance matrices, respectively, N is the batch size, and t is the chosen SNP threshold (15 in our case). Under this custom loss function, no penalty is imposed when both the feature distance and SNP distance exceed the SNP threshold, as they have no impact on the results of outbreak cluster detection.

5. Results

For all our experiments, we perform 4-fold cross-validation tests, with each test repeated over five random trials, using different random seeds for data splitting. We report the ARI, NMI, cluster precision, recall, and cluster F1 score for these experiments. Specifically, we first average each metric across the validation folds, then report the mean and the 95% confidence interval (using the t-distribution) of the averaged metric across the five random trials.

5.1. MSMAP yields superior overall clustering performance

As shown in Table 1, while MSMAP falls short of clusCLS in on SynthSpec(8), it outperforms clusCLS across *all* clustering metrics on SynthSpec(16) and SynthSpec(32), suggesting that the distance mapping approach in MSMAP has a marked advantage over the classification approach with a larger number of clusters (and fewer samples per cluster), which is of practical relevance. Additionally, MSMAP consistently outperforms onlyCLS and rawClus on all three versions of SynthSpec datasets, further underscoring the importance of distance mapping objective in $\mathcal{L}_{\text{SNP}}$ in helping MSMAP uncover latent clustering structures. For the proprietary dataset, though the gap in performance between the methods narrows, MSMAP continues to lead in all clustering metrics.

5.2. MSMAP augments MALDI-TOF utility in species-specific outbreak cluster detection

After evaluating the overall clustering performance on the full MALDI-TOF dataset, we now investigate the fine-grained performance for individual species. The correspondence between species labels and names is described in Appendix B. For the MALDI-TOF spectra of each species, we calculate the *lifts* in NMI, ARI, and cluster F1 score between

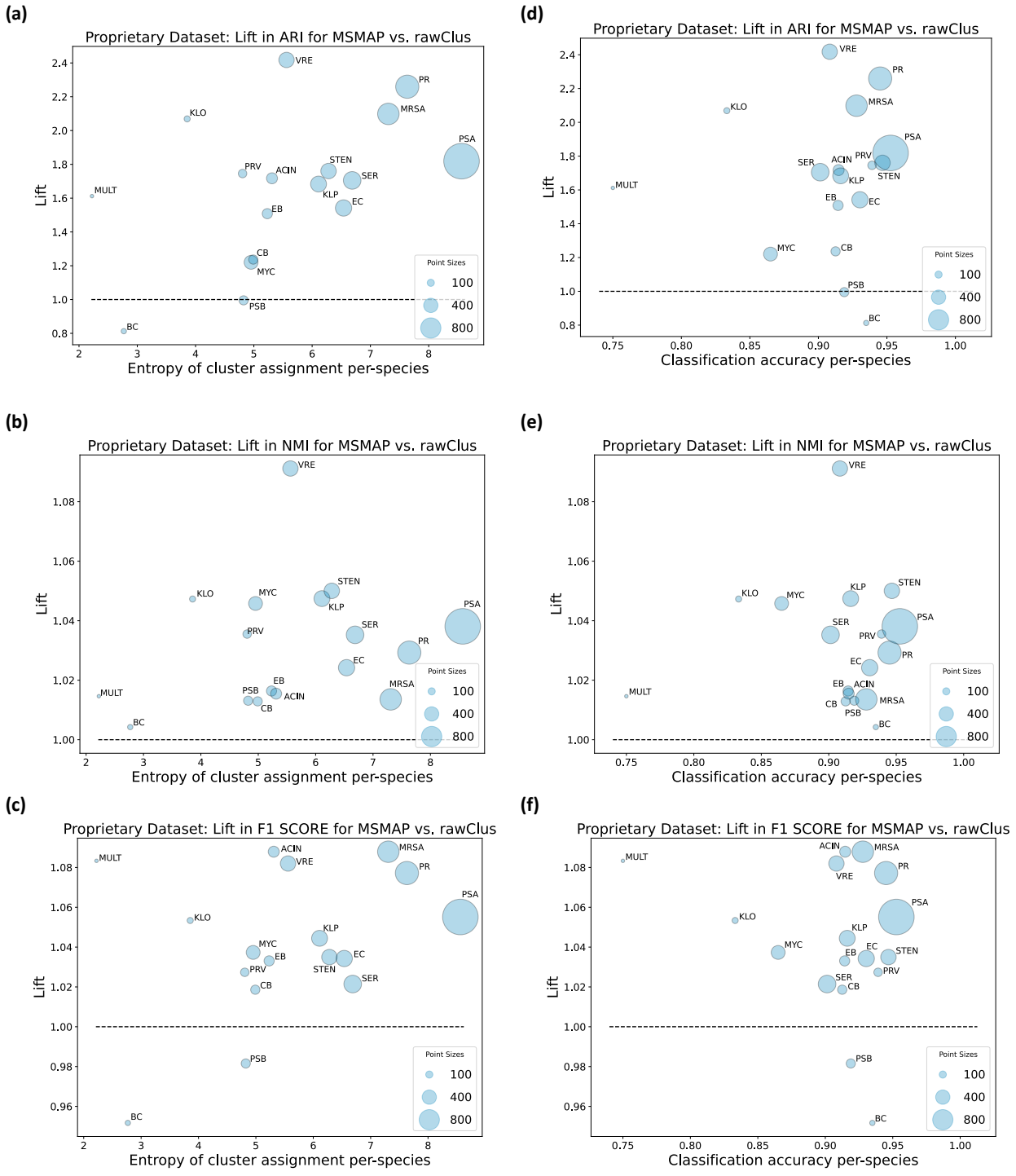

Figure 3: Lifts in ARI, NMI, and cluster F1 score between MSMAP and rawClus for different species in the proprietary dataset, sorted by the entropy of ground truth outbreak clusters or the species classification accuracy. The dot size reflects the number of MALDI-TOF samples of particular species. Species label descriptions are listed in Appendix B.

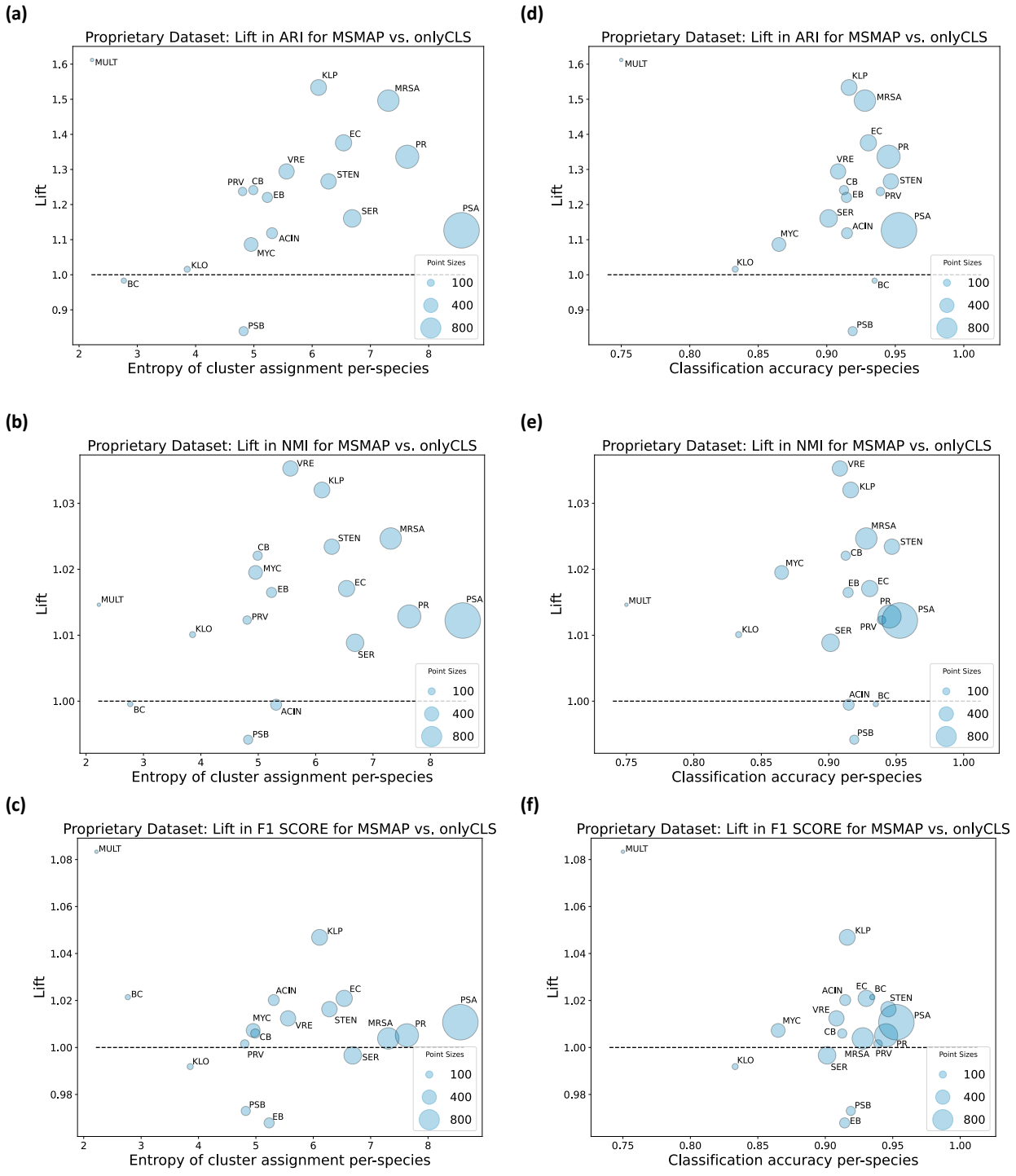

Figure 4: Lifts in ARI, NMI, and cluster F1 score between MSMAP and onlyCLS for different species in the proprietary dataset, sorted by the entropy of ground truth outbreak clusters or the species classification accuracy. The dot size reflects the number of MALDI-TOF samples in that species. Species label descriptions are listed in Appendix B.

Dataset	Model	ARI	NMI	Precision	Recall	F1 Score	Species Accuracy
SynthSpec(8)	MSMAP	0.613 ± 0.019	0.799 ± 0.009	0.855 ± 0.016	0.695 ± 0.043	0.767 ± 0.029	0.879 ± 0.009
	clusCLS	$\mathbf{0.748 \pm 0.019}$	$\mathbf{0.817 \pm 0.011}$	$\mathbf{0.87 \pm 0.011}$	$\mathbf{0.869 \pm 0.013}$	$\mathbf{0.869 \pm 0.012}$	0.916 ± 0.003
	onlyCLS	0.028 ± 0.026	0.298 ± 0.04	0.334 ± 0.049	0.07 ± 0.03	0.114 ± 0.042	$\mathbf{0.953 \pm 0.004}$
	rawClus	0.033 ± 0.003	0.334 ± 0.006	0.401 ± 0.012	0.075 ± 0.006	0.127 ± 0.008	N/A
SynthSpec(16)	MSMAP	$\mathbf{0.596 \pm 0.03}$	$\mathbf{0.809 \pm 0.013}$	$\mathbf{0.754 \pm 0.024}$	$\mathbf{0.759 \pm 0.018}$	$\mathbf{0.756 \pm 0.014}$	0.867 ± 0.022
	clusCLS	0.570 ± 0.013	0.75 ± 0.008	0.749 ± 0.008	0.747 ± 0.008	0.748 ± 0.008	0.902 ± 0.002
	onlyCLS	0.01 ± 0.006	0.345 ± 0.052	0.238 ± 0.047	0.067 ± 0.019	0.103 ± 0.015	$\mathbf{0.961 \pm 0.005}$
	rawClus	0.028 ± 0.003	0.358 ± 0.005	0.274 ± 0.007	0.092 ± 0.004	0.138 ± 0.005	N/A
SynthSpec(32)	MSMAP	$\mathbf{0.389 \pm 0.044}$	$\mathbf{0.739 \pm 0.022}$	$\mathbf{0.506 \pm 0.07}$	$\mathbf{0.727 \pm 0.018}$	$\mathbf{0.595 \pm 0.048}$	0.873 ± 0.009
	clusCLS	0.259 ± 0.009	0.632 ± 0.005	0.443 ± 0.01	0.451 ± 0.009	0.447 ± 0.009	0.894 ± 0.013
	onlyCLS	0.014 ± 0.003	0.37 ± 0.04	0.156 ± 0.032	0.112 ± 0.008	0.129 ± 0.009	$\mathbf{0.954 \pm 0.005}$
	rawClus	0.022 ± 0.002	0.42 ± 0.008	0.211 ± 0.012	0.12 ± 0.004	0.153 ± 0.005	N/A
Proprietary	MSMAP	$\mathbf{0.159 \pm 0.02}$	$\mathbf{0.952 \pm 0.012}$	$\mathbf{0.960 \pm 0.026}$	$\mathbf{0.602 \pm 0.018}$	$\mathbf{0.740 \pm 0.012}$	0.872 ± 0.034
	clusCLS	0.127 ± 0.027	0.918 ± 0.006	0.890 ± 0.007	0.59 ± 0.008	0.709 ± 0.007	0.894 ± 0.003
	onlyCLS	0.132 ± 0.016	0.942 ± 0.006	0.943 ± 0.014	0.599 ± 0.013	0.732 ± 0.01	$\mathbf{0.93 \pm 0.002}$
	rawClus	0.092 ± 0.011	0.928 ± 0.007	0.917 ± 0.02	0.568 ± 0.022	0.701 ± 0.013	N/A

Table 1: Model performance on SynthSpec and the proprietary dataset.

MSMAP and those of direct MALDI-TOF clustering (rawClus) and the ablation model without using SNP information (onlyCLS), respectively. The lift is defined as the ratio of MSMAP's performance to that of rawClus or onlyCLS, measuring the species-specific gain in MALDI-TOF utility for outbreak cluster detection brought by MSMAP.

As presented in Figure 3, MSMAP outperforms rawClus (i.e., lift > 1) in NMI for *all* species. MSMAP also achieves superior ARI and cluster F1 scores for all species except *Pseudomonas* (PSB) and *Burkholderia cepaciae* (BC). These results demonstrate MSMAP's effectiveness in improving MALDI-TOF utility in detecting species-specific outbreak clusters vs. directly applying MALDI-TOF similarity for the same task.

As shown in Figure 4, compared to onlyCLS, MSMAP yields superior ARI for all species except PSB and BC, achieves higher NMI for all species except PSB, BC, and *Acinetobacter baumannii* (ACIN), and leads in cluster F1 score for all species besides PSB, *Klebsiella oxytoca* (KLO), *Enterobacter cloacae* (EB), and *Serratia marcescens* (SER). These results corroborate the significance of the SNP distance mapping objective in MSMAP.

5.3. MSMAP is robust to outbreak cluster diversity

We investigate how such improvements are affected by the diversity of the species' outbreak clusters. Here, we measure diversity using the mean Shannon entropy of the ground truth outbreak cluster assignment on its isolates across the validation sets.

Compared with rawClus, we found that the diversity of outbreak clusters does not negatively impact the lifts in NMI and ARI, and there is a mildly positive impact on the lift in the F1 score (Figure 3(a-c)). For MSMAP vs. onlyCLS, we observe no negative trend for NMI and the F1 scores over cluster diversity, while there is a mildly positive trend for ARI (Figure 4(a-c)). These results show MSMAP's robustness to outbreak cluster diversity and the potential to deal with complex outbreaks.

5.4. MSMAP is agnostic to the difficulty of species identification

Here, we investigate how MSMAP's improvements may be affected by the difficulty of correctly classifying this species itself. As is shown in Figure 3(d-f)) and Figure 4(d-f)), all lifts in performance are not impacted by the species classification accuracy. These results show that MSMAP's ability to augment MALDI-TOF utility in outbreak cluster detection is not affected by the difficulty in classifying species, further substantiating MSMAP's potential to be adopted in various outbreak investigation scenarios.

6. Discussion

We introduced MSMAP, the first machine learning framework designed to bridge the utility gap be-

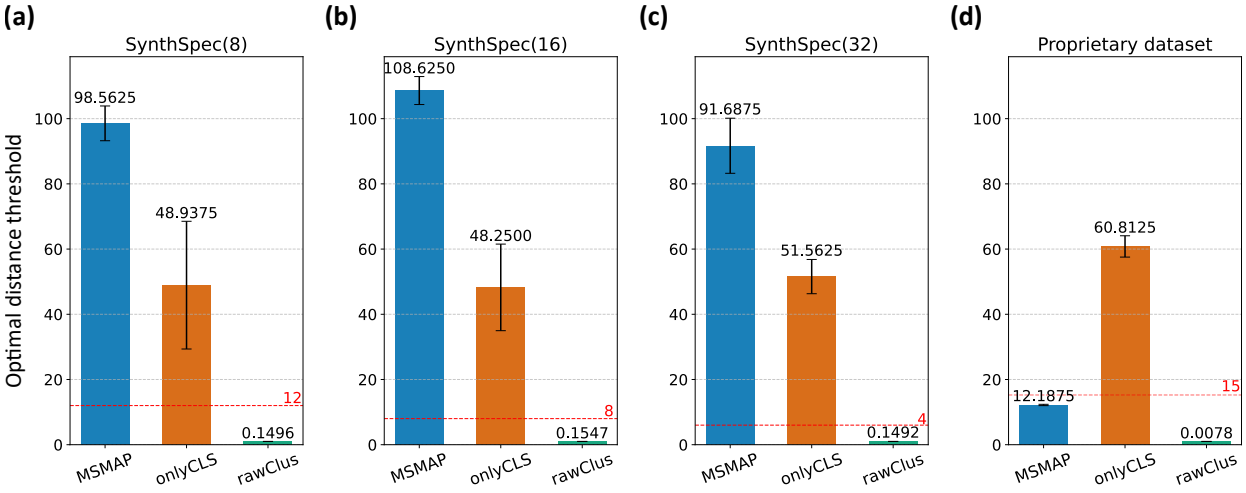

Figure 5: Optimal distance thresholds for clustering MALDI-TOF representations yielded by different methods on SynthSpec and the proprietary dataset. The red dashed line represents the ground truth distance threshold used to derive the ground truth clusters.

tween MALDI-TOF and WGS in the detection of outbreak clusters. Through empirical evaluation, we demonstrated that MSMAP can effectively augment MALDI-TOF utility by capturing and preserving the clustering structures defined by SNP distance. Furthermore, we observed two additional merits of MSMAP: its performance is not negatively affected by (i) the diversity of outbreak clusters and (ii) the difficulty of species classification, suggesting that MSMAP has the potential to be applied to a wide range of outbreak scenarios, in particular wherever access to whole genome sequencing based species similarity analysis is limited or economically infeasible, but MALDI-TOF capacity is available. MSMAP can be integrated into infection control and outbreak detection workflows as a routine and rapid inspection of collected isolates. It reduces the need of performing WGS for outbreak detection: for the species where our proposed modeling approach yields good clustering performance during training, MALDI-TOF can become a viable and cost-effective alternative to WGS. After obtaining the clusters, the infection prevention team will investigate the potential contaminated root causes, including unit, room, procedures, and providers, and perform corresponding interventions to stop the transmissions of the pathogens present in the cluster.

However, while the distance mapping objective (Equation 7) helps MSMAP outperform other methods, we found that the relationship between the optimal distance threshold for clustering representations learned by MSMAP and the ground truth is not consistent across datasets. For SynthSpec, we observe that MSMAP yields optimal distance thresholds that far exceed the ground truth (Figure 5(a-c))). On the other hand, we found that the optimal distance threshold for MSMAP on the proprietary dataset is in close proximity to the ground truth (Figure 5(d))), suggesting that MSMAP learns MALDI-TOF representations whose distances closely mimic that of WGS. We conjecture that this discrepancy results from SynthSpec being an imperfect reflection of real-world MALDI-TOF samples obtained in outbreak investigations, but this behavior calls for further study.

In our evaluations, we noticed that the species with a low lift in clustering performance, e.g., PSB and BC, have small sample sizes (Figure 3 and Figure 4). This indicates a potential limitation when applying the current MSMAP framework to scenarios with limited amounts of data. We plan to address this limitation in future work by incorporating additional supervision that encourages MSMAP to accurately estimate the number of ground truth clusters and their distribution.

Another future research direction aims to further increase the utility of MALDI-TOF in outbreak cluster detection by integrating additional data modalities commonly available in practice, such as an-

timicrobial resistance (AMR) profiles and electronic health records (EHR), taking a further step toward rapid, accurate, and cost-effective outbreak cluster detection.

Acknowledgments

This work has been partially supported by the NIH award 5R01AIR27472-09 and NSF awards 2406231 and 2427948. The authors thank Miss Jiayi Li, Dr. Nicholas Gisolfi, and Dr. Kyle Miller for helpful discussions.

References

Mohamed HH Abdelbary, Laurence Senn, Gilbert Greub, Gregory Chaillou, Estelle Moulin, and Dominique S Blanc. Whole-genome sequencing revealed independent emergence of vancomycin-resistant enterococcus faecium causing sequential outbreaks over 3 years in a tertiary care hospital. *European Journal of Clinical Microbiology & Infectious Diseases*, 38:1163–1170, 2019.

Walid M Abdelmoula, Begona Gimenez-Cassina Lopez, Elizabeth C Randall, Tina Kapur, Jann N Sarkaria, Forest M White, Jeffrey N Agar, William M Wells, and Nathalie YR Agar. Peak learning of mass spectrometry imaging data using artificial neural networks. *Nature communications*, 12(1):5544, 2021.

Lennox K Archibald and William R Jarvis. Health care–associated infection outbreak investigations by the centers for disease control and prevention, 1946–2005. *American journal of epidemiology*, 174 (suppl_11):S47–S64, 2011.

Maskit Bar-Meir, Elihay Berliner, Livnat Kashat, David A Zeevi, and Marc V Assous. The utility of maldi-tof ms for outbreak investigation in the neonatal intensive care unit. *European Journal of Pediatrics*, 179:1843–1849, 2020.

Timothy G Buchman, Bernard Roizman, Garrett Adams, and Beth Hewitt Stover. Restriction endonuclease fingerprinting of herpes simplex virus dna: a novel epidemiological tool applied to a nosocomial outbreak. *Journal of Infectious Diseases*, 138(4):488–498, 1978.

Andrew E Clark, Erin J Kaleta, Amit Arora, and Donna M Wolk. Matrix-assisted laser desorption ionization–time of flight mass spectrometry: a fundamental shift in the routine practice of clinical microbiology. *Clinical microbiology reviews*, 26(3): 547–603, 2013.

Antony Croxatto, Guy Prod'hom, and Gilbert Greub. Applications of maldi-tof mass spectrometry in clinical diagnostic microbiology. *FEMS microbiology reviews*, 36(2):380–407, 2012.

Gaetan De Waele, Gerben Menschaert, and Willem Waegeman. An antimicrobial drug recommender system using maldi-tof ms and dual-branch neural networks. *eLife*, 13:RP93242, 2024.

Gaetan De Waele, Gerben Menschaert, Peter Vandamme, and Willem Waegeman. Pre-trained maldi transformers improve maldi-tof ms-based prediction. *Computers in Biology and Medicine*, 186: 109695, 2025.

Ariane G Dinkelacker, Sophia Vogt, Philipp Oberhettinger, Norman Mauder, Jörg Rau, Markus Kostrzewa, John WA Rossen, Ingo B Autenrieth, Silke Peter, and Jan Liese. Typing and species identification of clinical klebsiella isolates by fourier transform infrared spectroscopy and matrix-assisted laser desorption ionization–time of flight mass spectrometry. *Journal of clinical microbiology*, 56(11):10–1128, 2018.

Lauren Epstein, Jennifer C Hunter, M Allison Arwady, Victoria Tsai, Linda Stein, Marguerite Gribogiannis, Mabel Frias, Alice Y Guh, Alison S Laufer, Stephanie Black, et al. New delhi metallo-β-lactamase–producing carbapenem-resistant escherichia coli associated with exposure to duodenoscopes. *Jama*, 312(14):1447–1455, 2014.

Betsy Foxman and Lee Riley. Molecular epidemiology: focus on infection. *American journal of epidemiology*, 153(12):1135–1141, 2001.

Audrey Giraud-Gatineau, Gaetan Texier, Pierre-Edouard Fournier, Didier Raoult, and Hervé Chaudet. Using maldi-tof spectra in epidemiological surveillance for the detection of bacterial subgroups with a possible epidemic potential. *BMC Infectious Diseases*, 21:1–10, 2021.

Paul M Griffin, Gareth R Price, Jacqueline M Schooneveldt, Sanmarié Schlebusch, Martyn H Tilse, Tess Urbanski, Brett Hamilton, and Deon Venter. Use of matrix-assisted laser desorption

ionization–time of flight mass spectrometry to identify vancomycin-resistant enterococci and investigate the epidemiology of an outbreak. *Journal of clinical microbiology*, 50(9):2918–2931, 2012.

Hollie-Ann Hatherell, Caroline Colijn, Helen R Stagg, Charlotte Jackson, Joanne R Winter, and Ibrahim Abubakar. Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review. *BMC medicine*, 14:1–13, 2016.

Lisa L Maragakis, Sara E Cosgrove, Xiaoyan Song, Denny Kim, Patricia Rosenbaum, Nancy Ciesla, Arjun Srinivasan, Tracy Ross, Karen Carroll, and Trish M Perl. An outbreak of multidrug-resistant acinetobacter baumannii associated with pulsatile lavage wound treatment. *Jama*, 292(24):3006–3011, 2004.

Alexander Mellmann, Stefan Bletz, Thomas Böking, Frank Kipp, Karsten Becker, Anja Schultes, Karola Prior, and Dag Harmsen. Real-time genome sequencing of resistant bacteria provides precision infection control in an institutional setting. *Journal of clinical microbiology*, 54(12):2874–2881, 2016.

Noshine Mohammad, Anne-Cécile Normand, Cécile Nabet, Alexandre Godmer, Jean-Yves Brossas, Marion Blaize, Christine Bonnal, Arnaud Fekkar, Sébastien Imbert, Xavier Tannier, et al. Improving the detection of epidemic clones in candida parapsilosis outbreaks by combining maldi-tof mass spectrometry and deep learning approaches. *Microorganisms*, 11(4):1071, 2023.

PR Murray. Matrix-assisted laser desorption ionization time-of-flight mass spectrometry: usefulness for taxonomy and epidemiology. *Clinical Microbiology and Infection*, 16(11):1626–1630, 2010.

Tara N Palmore and David K Henderson. Managing transmission of carbapenem-resistant enterobacteriaceae in healthcare settings: a view from the trenches. *Clinical Infectious Diseases*, 57(11):1593–1599, 2013.

Art FY Poon. Impacts and shortcomings of genetic clustering methods for infectious disease outbreaks. *Virus evolution*, 2(2):vew031, 2016.

Joshua Quick, Nicola Cumley, Christopher M Wearn, Marc Niebel, Chrystala Constantinidou, Chris M Thomas, Mark J Pallen, Naiem S Moiemen, Amy Bamford, Beryl Oppenheim, et al. Seeking the source of pseudomonas aeruginosa infections in a recently opened hospital: an observational study using whole-genome sequencing. *BMJ open*, 4(11): e006278, 2014.

Jürgen Rödel, Alexander Mellmann, Claudia Stein, Monika Alexi, Frank Kipp, Birgit Edel, Kristin Dawczynski, Christian Brandt, Lothar Seidel, Wolfgang Pfister, et al. Use of maldi-tof mass spectrometry to detect nosocomial outbreaks of serratia marcescens and citrobacter freundii. *European Journal of Clinical Microbiology & Infectious Diseases*, 38:581–591, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

John WA Rossen, Alexander W Friedrich, Jacob Moran-Gilad, et al. Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology. *Clinical microbiology and infection*, 24(4):355–360, 2018.

Todd R Sandrin, Jason E Goldstein, and Stephanie Schumaker. Maldi tof ms profiling of bacteria at the strain level: a review. *Mass spectrometry reviews*, 32(3):188–217, 2013.

Sanmarie Schlebusch, Gareth R Price, Renee L Gallagher, V Horton-Szar, LDH Elbourne, P Griffin, Deon J Venter, Slade O Jensen, and SJ Van Hal. Maldi-tof ms meets wgs in a vre outbreak investigation. *European Journal of Clinical Microbiology & Infectious Diseases*, 36:495–499, 2017.

Alexander J Sundermann, Ahmed Babiker, Jane W Marsh, Kathleen A Shutt, Mustapha M Mustapha, Anthony W Pasculle, Chinelo Ezeonwuka, Melissa I Saul, Marissa P Pacey, Daria Van Tyne, et al. Outbreak of vancomycin-resistant enterococcus faecium in interventional radiology: detection through whole-genome sequencing-based surveillance. *Clinical Infectious Diseases*, 70(11): 2336–2343, 2020.

Alexander J Sundermann, Jieshi Chen, James K Miller, Melissa I Saul, Kathleen A Shutt, Marissa P

Griffith, Mustapha M Mustapha, Chinelo Ezeonwuka, Kady Waggle, Vatsala Srinivasa, et al. Outbreak of pseudomonas aeruginosa infections from a contaminated gastroscope detected by whole genome sequencing surveillance. *Clinical Infectious Diseases*, 73(3):e638–e642, 2021.

J Veenemans, M Welker, A Van Belkum, MC Saccomani, V Girard, A Pettersson, C Verhulst, M Kluytmans-Vandenbergh, and J Kluytmans. Comparison of maldi-tof ms and aflp for strain typing of esbl-producing escherichia coli. *European Journal of Clinical Microbiology & Infectious Diseases*, 35:829–838, 2016.

Caroline Weis, Aline Cuénod, Bastian Rieck, Olivier Dubuis, Susanne Graf, Claudia Lang, Michael Oberle, Maximilian Brackmann, Kirstine K Søgaard, Michael Osthoff, et al. Direct antimicrobial resistance prediction from clinical maldi-tof mass spectra using machine learning. *Nature Medicine*, 28(1):164–174, 2022.

Appendix A. Construction of the SynthSpec dataset

In SynthSpec, we construct synthetic spectra with mass-to-charge ratios (x-axis) ranging from 0 to 128 under the resolution of 0.1. Each spectrum has two peaks of maximum intensity 100. Below, we discuss locations and intensities relative to their maximum values, i.e., on a scale of 0 to 1.

The locations of the peaks determine the species. Specifically, the peak locations are sampled from independent 1-D Gaussian distributions with a standard deviation of 0.1 and truncated to the range of $(0, 1)$. For the 2 species in SynthSpec, we set the mean of the two peaks to be $[0.1, 0.6]$ and $[0.4, 0.8]$, respectively.

The intensities of the peaks determine the outbreak clusters. For simplicity, we use the same set of sample intensities for both species since their distance mapping losses are calculated separately. Specifically, we first generate k 2-D Gaussians ($k = 8, 16, 32$ for SynthSpec(8), SynthSpec(16), and SynthSpec(32), respectively) whose x and y directions are independent and share a common standard deviation. The standard deviations are sampled from i.i.d. uniform distributions. The range of the uniform distribution $(\sigma_{\min}, \sigma_{\max})$ is $(0.03, 0.06)$, $(0.02, 0.04)$, and $(0.01, 0.02)$ for $k = 8, 16, 32$, respectively. We impose the following constraint for each pair of Gaussians:

$$\|\mu_1 - \mu_2\|_2 \geq 2(\sigma_1 + \sigma_2) + t + \alpha, \tag{17}$$

where $\mu_1, \mu_2 \in (0, 1)^2$ and $\sigma_1, \sigma_2 \in (\sigma_{\min}, \sigma_{\max})$ are the mean and the standard deviation of the two Gaussians, respectively. $t = 2\sigma_{\max}$ is the set "ground truth" threshold and $\alpha = 0.005$ is a margin. The assumption behind this constraint is that the points sampled from the 2-D Gaussian are within 2 standard deviations from the mean with high probability. Hence, with high probability, points sampled from different Gaussians will be at least $t + \alpha$ apart, while points sampled from the same Gaussian will be at most t apart. Figures 2 and 6 contain visualization of the Gaussian samples for SynthSpec(8), SynthSpec(16), and SynthSpec(32).

After the Gaussians are generated, we repeatedly sample $3200/k$ points per Gaussian to ensure they are within the range of $(0, 1)^2$. The x and y coordinates of each sampled point are set to the two peak intensities relative to the maximum value 100. Finally, independent unit Gaussian noise is added to every location of the spectrum.

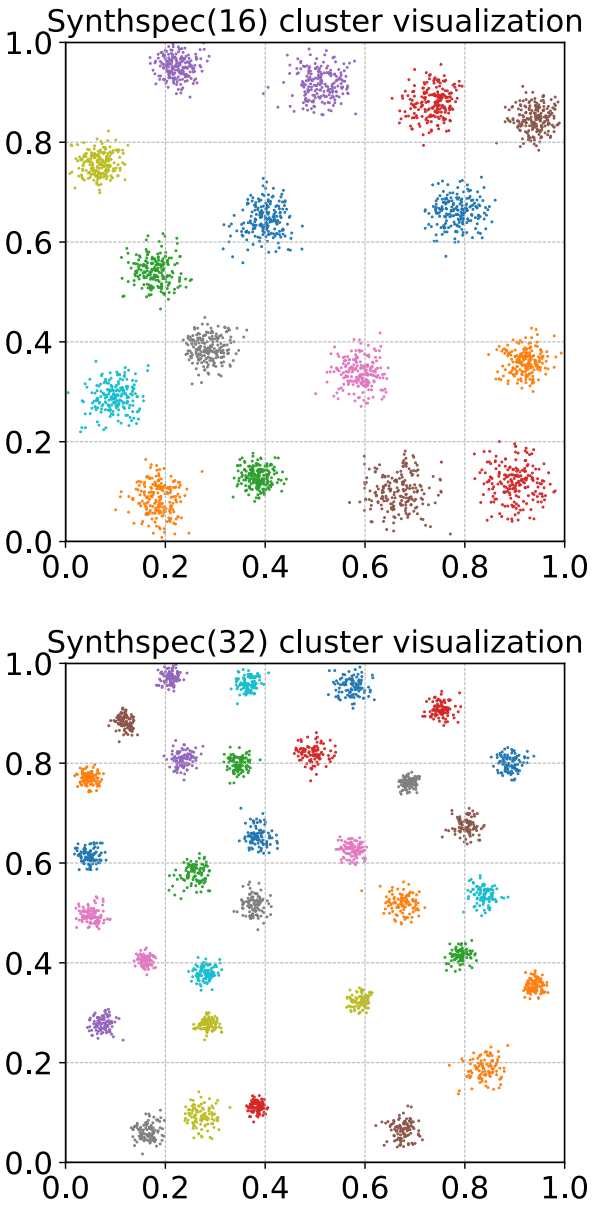

Figure 6: Visualization of clusters in Synthspec(16) and Synthspec(32).

Appendix B. Characteristics of the proprietary dataset

Species	**Label**	**# Samples**	**# Clusters**	**# Singletons**
Acinetobacter baumannii	ACIN	**237**	**127**	**79**
Burkholderia cepaciae	BC	**55**	**17**	**7**
Citrobacter	CB	**167**	**96**	**63**
Enterobacter cloacae	EB	**203**	**97**	**41**
Escherichia coli	EC	**517**	**267**	**154**
Klebsiella oxytoca	KLO	**73**	**42**	**25**
Klebsiella pneumoniae	KLP	**492**	**180**	**76**
Staphylococcus aureus	MRSA	**898**	**484**	**287**
Multiple species (others)	MULT	**24**	**11**	**3**
Mycobacterium	MYC	**375**	**84**	**40**
Proteus mirabilis	PR	**1045**	**602**	**385**
Providencia	PRV	**141**	**78**	**46**
Pseudomonas aeruginosa	PSA	**2460**	**1202**	**704**
Pseudomonas	PSB	**165**	**68**	**26**
Serratia marcescens	SER	**604**	**305**	**179**
Stenotrophomonas maltophilia	STEN	**466**	**228**	**118**
Vancomycin-resistant Enterococcus	VRE	**460**	**187**	**125**
	Total	**8382**	**4075**	**2358**

Table 2: Specifics of the proprietary dataset.

Appendix C. Species-wise lifts with respect to clusCLS

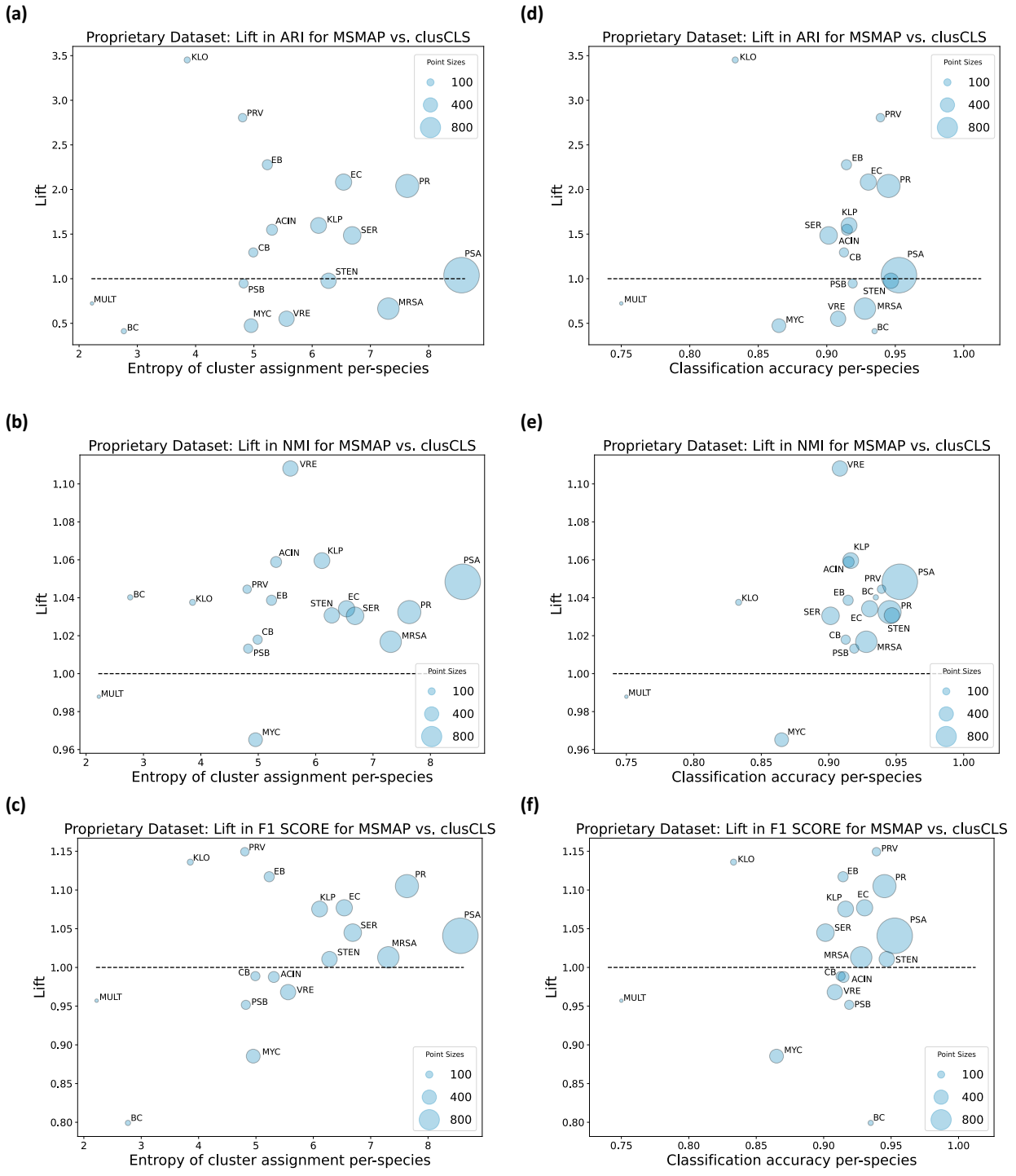

Figure 7: Lifts in ARI, NMI, and cluster F1 score between MSMAP and clusCLS for different species in the proprietary dataset, sorted by the entropy of ground truth outbreak clusters or the species classification accuracy. The dot size reflects the number of MALDI-TOF samples in that species. Species label descriptions are listed in Appendix B.