Leveraging MALDI-TOF Mass Spectroscopy for Pathogen Detection via Semi-Supervised Learning

Christian Belardi*
Stottler Henke Associates, Inc.
1650 S. Amphlett Blvd., Suite 300
San Mateo, CA 94402.
cbelardi@shai.com

Pusong Li*
Stottler Henke Associates, Inc.
1650 S. Amphlett Blvd., Suite 300
San Mateo, CA 94402.
alanli@shai.com

Abstract

In response to the growing global threat of bioterrorism, we explore the potential for machine learning to address the need for biosurveillance technology, specifically, technology capable of assessing the pathogenic potential of novel bacteria, for the Defense Advanced Research Projects Agency (DARPA). This paper investigates Matrix Assisted Laser Desorption/Ionization - Time Of Flight (MALDI-TOF) Mass Spectroscopy (MS) as a technique for exposing bacterial phenotype that can be used as an indicator of pathogenicity. Data constrained, we show that our neural architecture based on a variational autoencoder can learn a dimensionality reduction conducive to pathogen detection. Our encoder presses a 10,000 dimensional spectra down to a 10 dimensional embedding, which is much more easily classified using out of the box machine learning algorithms. We use both an unsupervised generative task and supervised discriminative task to train our encoder, which enables us to utilize both labeled and unlabeled examples during training. The ultimate result produces embeddings that out perform embeddings learned via the discriminative or generative task alone, showing the utility of a semi-supervised approach.

Introduction

The rapidly developing biotechnology industry presents a major opportunity for advancements that can greatly improve the way we live. From automated agricultural management for increased crop yields, to health tracking via biosensors, the potential benefits of biotechnology are vast. Increased interest in the industry and related fields has facilitated the flow of relevant information, enabling innovations beyond what was previously considered possible. However, these new technologies expose new vulnerabilities, which enable new threats to emerge. Proliferation of biotechnology continues to drive the necessity for biosurveillance/biodefense technology, which is of critical importance to national security. The problems that must be addressed in the development of biosurveillance technology have proven to be formidable, many of which are active areas of study within microbiology and infectious disease research. Machine learning has found success in modeling phenomena not yet well understood, particularly in the areas of computer vision and natural language processing. Thus a data driven approach to solving these complex biosurveillance challenges may prove beneficial.

Matrix Assisted Laser Desorption/Ionization - Time Of Flight (MALDI-TOF) Mass Spectroscopy (MS) has become an increasingly popular technique used across disciplines to characterize samples. For a given sample, MALDI-TOF MS can capture information regarding the lipids, peptides, and proteins present, revealing various molecular patterns (Dunham et al. 2016). Variation in the way different organisms survive, particularly pathogenic vs. non-pathogenic organisms, suggests the potential for these molecular patterns to vary as well. In regards to biosurveillance, MALDI-TOF MS is a strong candidate technique for monitoring microorganisms because it is a fast, high throughput, and cost effective technique that does not require trained laboratory personnel (Singhal et al. 2015). Though modern MALDI-TOF MS technology requires a significantly sized sample to achieve a quality measurement, emerging single cell proteomics technology suggests the potential for single cell MALDI-TOF MS replacements (Marx 2019).

The difficulty and importance of the biosurveillance/biodefense problem highlights it's significance and emphasizes the necessity that it is addressed as soon as possible. Advances in machine learning technology can help to bridge the gap between our understanding and the real world. We hope to take a step towards this goal in our research.

Related Work

Identification of Microorganisms

There has been a significant amount of work studying the use of MALDI-TOF MS for identification of microorganisms, particularly bacteria, in place of the standard 16S rRNA sequencing. Sauer *et al.* demonstrated accurate identification of phytopathogenic bacteria by applying hierarchical clustering and similarity scoring algorithms to MALDI-TOF MS data (Sauer et al. 2008). In another study of 132

*Authors contributed equally to this work

bacterial isolates, Hsieh *et al.* show human pathogens can be consistently identified by their MALDI-TOF MS spectra using machine learning. Furthermore, they found that their approach was capable of identifying all isolates in a mixed sample of six different species (Hsieh et al. 2008).

Leveraging MALDI-TOF based microorganism identification technology, the Planetary Protection Group at NASA's Jet Propulsion Laboratory (JPL) is developing a MALDI-TOF MS database to archive extremophiles found during the spacecraft assembly process. NASA has found MALDI-TOF MS to be an accurate, and high throughput technique for real-time identification of bacterial isolates. In the future, NASA plans to use this extremely cost-effective technique to identify potential contaminates and dangerous pathogens (Seuylemezian et al. 2018)

The demonstrated success of MALDI-TOF MS for microorganism identification has led to calls for wide spread adoption of this technology. Patel discusses the implications of this new technology in clinical microbiology. Modern genomics based identification tests require significant time to yield results, whereas MALDI-TOF MS can provide identifications much faster. Faster identifications will speed up the diagnosis process, which is critical in such a time-sensitive domain (Patel 2013).

Pathogen Detection

Alam *et al.* demostrate that MALDI-TOF MS is useful for detecting protein toxins such as staphylococcal enterotoxin B, botulinum neurotoxins, Clostridium perfringens epsilon toxin, shiga toxin, etc. In their work, they go on to define a simple method for identification of protein toxins by MALDI-TOF/TOF, which can theoretically be extended to all protein toxins (Alam, Kumar, and Kamboj 2012). Beyond Alam *et al.*, we are unaware of extensive work evaluating MALDI-TOF MS for human pathogen detection, particularly for the purpose of assessing unknown bacterial isolates.

Embeddings and Dimensionality Reduction

In many domains, extremely high dimensional data can have its most important features represented in a fraction of the size, making manual interpretation far easier and reducing burden on downstream algorithms (machine learning or otherwise) applied to the processed data. In a phenomena known as the *curse of dimensionality*, the quantity of data required to estimate a function to a particular degree of accuracy generally grows exponentially with respect to the input dimensionality (Bellman 1961). While neural networks appear less strongly susceptible to this "curse" than older machine learning methods, reducing dataset dimensionality still confers strong benefits in reducing quantity of training data required.

Representing data in a lower-dimensional form is known as *manifold learning*, as the basis of the approach is an assumption that data can be represented as points on a lower-dimensional manifold in the higher-dimensional space without losing important information. Examples of this range from the classical Principal Component Analysis (PCA) (Pearson 1901) to modern techniques based on using the last layer of a neural network trained to label data as a featurization (Sharif Razavian et al. 2014). Autoencoders and deep autoencoders have recently shown great promise in manifold learning (Wang et al. 2014), (Rifai et al. 2011), and while variational autoencoders (Kingma and Welling 2013) are often used for generative applications, part of this generation involves creating a meaningful latent space, which is a lower dimensional representation of data.

Problem Statement

Let $\mathbf{X}$ denote the set of all relevant MALDI-TOF data in general, residing in d-dimensional space $\mathbb{S}$. Each element $x_i \in \mathbf{X}$, has an associated binary Bio-Safety Level (BSL) label b_i (we are only considering BSL 1 and "more than 1" in this work, that is, we would like to determine if bacteria are pathogenic), which we collect into vector $\mathbf{b}$. We would like to create a function F that maps from $\mathbb{S}$ to a d'-dimensional space $\mathbb{S}'$, where $d' \ll d$. This mapping should be such that

$$\forall x_1, x_2 \in X,\ b_1 \neq b_2 \implies F(x_1) \neq F(x_2)$$

that is, the projection does not remove any information required to determine our target BSL label. Ideally, we would like the mapping to satisfy the stronger constraint

$$\forall x_1, x_2, x_3 \in X,\ (b_1 = b_2 \wedge b_1 \neq b_3) \implies \|F(x_1) - F(x_2)\| < \|F(x_1) - F(x_3)\|$$

That is, groups with the same BSL label should separate and not overlap in Euclidean space. This increases human interpretability, and allows simpler learning algorithms that do not need extensive training (such as k-nearest neighbors) to generate good results on the resulting embeddings.

To approximate this ideal F as $\hat{F}$, we use a labeled subset of the data X denoted $\{\hat{X}, \hat{b}\}$, and an unlabeled subset of the data $\hat{Y}$ in a semi-supervised setting.

Methodology

In order to create an $\hat{F}$, we use a neural-net architecture shown in Figure 1, as neural nets are known to be strong function approximators (Hornik 1991). The proposed architecture combines the idea of a variational autoencoder (Kingma and Welling 2013) with the insight that the last layer of a classifier can provide excellent features, even if the target domain is slightly different from the domain the classifier is trained for (Sharif Razavian et al. 2014). This bears some similarity to certain other work on semi-supervised learning for classification (Kingma et al. 2014), however, our goal is extracting high-quality embeddings with some label guidance instead of using unlabeled data to improve classifier accuracy.

Theoretical basis

In order to learn embeddings in a semi-supervised fashion, we utilize a variational autoencoder conditioned on the main objective, namely the BSL label. Variational autoencoders map from distribution to distribution, learning a conditional distribution $P(z \mid x)$ that maps a data distribution X to an embedding distribution Z. We define $G(z \mid x)$ as our

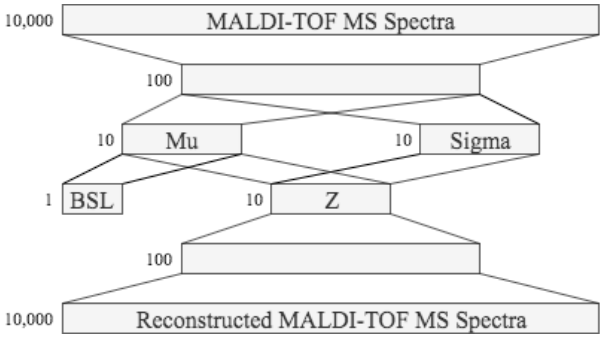

Figure 1: Image of neural architecture with layer sizes. In training, MALDI-TOF MS spectra are fed into the topmost layer, reconstruction loss is calculated from input spectra and output reconstructed spectra, variational loss from the μ and σ, and auxiliary loss from the BSL auxiliary output and the BSL label associated with the input spectra.

approximation of this conditional distribution. Then, using the KL divergence metric — which represents the "closeness" of the distributions by quantifying the expected log-likelihood ratio of data points between two distributions over the data space — as a minimization target to make $G(z \mid x)$ approximate $P(z \mid x)$ as closely as possible, the minimization objective given a data point x is:

$$\begin{aligned}\text{Obj}(x) &= \mathbb{KL}\left[P(z \mid x) \parallel G(z \mid x)\right] \\ &= \sum_{z \in Z} G(z \mid x) \log \frac{G(z \mid x)}{P(z \mid x)} \\ &= \mathop{\mathbb{E}}_{z \sim G(z|x)} \left[\log \frac{G(z \mid x)}{P(z \mid x)}\right] \\ &= \mathop{\mathbb{E}}_{z \sim G(z|x)} \left[\log G(z \mid x) - \log \frac{P(x \mid z) P(z)}{P(x)}\right]\end{aligned}$$

Which can be simplified to

$$\mathop{\mathbb{E}}_{z \sim G(z|x)} \left[\log G(z \mid x) - \log P(x \mid z) - \log P(z)\right]$$

by moving $\log P(x)$ out and dropping it from the objective, as it does not depend on z or the mapping G. This can be then written in terms of another KL divergence:

$$\begin{aligned}\text{Obj}(x) &= -\mathop{\mathbb{E}}_{z \sim G(z|x)} \left[\log P(x \mid z)\right] + \mathop{\mathbb{E}}_{z \sim G(z|x)} \left[\log \tfrac{G(z|x)}{P(z)}\right] \\ &= -\mathop{\mathbb{E}}_{z \sim G(z|x)} \left[\log P(x \mid z)\right] + \sum_{z \in Z} G(z \mid x) \log \tfrac{G(z|x)}{P(z)} \\ &= -\mathop{\mathbb{E}}_{z \sim G(z|x)} \left[\log P(x \mid z)\right] + \mathbb{KL}\left[G(z \mid x) \parallel P(z)\right]\end{aligned}$$

Here, the first term is the *reconstruction loss*, which encourages maximization of the likelihood the original data point is returned from reconstruction. The second term is the *variational regularizer*, which encourages the embeddings to be distributed according to *P(z)*. This stops the network from assigning one specific z for each x, and has the effect of encouraging similar x to have similar embeddings.

Given this formulation of the theoretical loss, we can lay out the neural network architecture and derive our loss in practice. The *encoder* network $E(x) \rightarrow (\mu, \sigma)$ represents $G(z \mid x)$, and maps an input data point x to a distribution of z by outputting a mean μ_i and sigma σ_i for each element z_i of z. In practice, setting $P(z)$ to be a multivariate standard normal serves as a good regularizer to encourage similar x to map to similar z. As such, our variational regularizer term can be written as:

$$\begin{aligned}\mathbb{KL}\left[G(z \mid x) \parallel P(z)\right] &= \sum_i \mathbb{KL}\left[G(z_i \mid x) \parallel P(z_i)\right] \\ &= \sum_i \mathbb{KL}\left[\mathcal{N}(\mu_i, \sigma_i) \parallel \mathcal{N}(0, 1)\right] \\ &= \frac{1}{2} \sum_i [-\log(\sigma_i^2) + \sigma_i^2 + \mu_i^2 - 1]\end{aligned}$$

To realize the reconstruction loss term, we must define a *decoder* network $D(z) \rightarrow x$. Given this network, log-likelihood over the distribution in latent space given by the encoder — that is, $\log\left(P(x \mid z)\right)$ — is $\mathcal{L}\left(D(z), x\right)$, where $\mathcal{L}$ is some formulation of log-likelihood of $D(z)$ given ground truth x. For normalized and standardized continuous-valued data (i.e. assumed to be standard Gaussian), this can be approximated by negative mean squared error. Our reconstruction loss term is then

$$\mathop{\mathbb{E}}_{z \sim \mathcal{N}(E(x))} \left[\frac{1}{n} \sum_{i=1 \ldots n} \left(D(z)_i - x_i\right)^2\right]$$

Where n is the dimensionality of x. In practice, this expectation is realized through Bayesian methods, with the loss evaluated on a z sampled from the normal distribution given by $\mathcal{N}(E(x))$.

Thus far, we have provided a quick theoretical basis of the variational autoencoder method in general, which our architecture extends. The variational autoencoder operates from a source *distribution* to a target embedding *distribution*, however, we would like deterministic embeddings. As such, we use the mean of the output embedding distribution, that is, $\hat{F}(x) := \mu$, where μ is the vector of means given by encoder network $E(x)$. From here, we motivate the auxiliary loss network $A(\hat{F}(x)) \rightarrow \hat{b}$ by considering our desired objectives. If $b_1 \neq b_2 \implies F(x_1) \neq F(x_2)$ is not satisfied, then $A(\hat{F}(x_1)) = A(\hat{F}(x_2))$, which would incur a loss, as the predicted values $\hat{b_1} = \hat{b_2}$, while in reality $b_1 \neq b_2$. Additionally, as we propose a direct connection from $\hat{F}(x)$ to the output $\hat{b_x}$ (i.e. a zero-layer neural network), $\hat{b_1}$ is a linear combination of $\hat{F}(x)$, meaning distance in $\hat{F}(x)$ space directly translates to distance in $\hat{b}$ space, fulfilling the second objective.

For an alternate interpretation of this, instead of examining auxiliary loss network $A(\hat{F}(x)) \rightarrow \hat{b}$, we look instead at the zero-layer network A by itself, taking in inputs μ. The direct connection from μ to A means that in this neural network wiring, $A(\mu) = M_A * \mu + c$, where M_A is the weights vector and c is the bias constant (A has scalar output). To fulfill our stated goals, we fit a maximum-margin hyperplane (linear SVM) in order to encourage maximum separation between the two BSL classes. With this motivation, we chose

the hinge loss as our auxiliary loss term:

$$\max\left(0, 1 - b_x * A\left(\hat{F}(x)\right)\right)$$

Where $b_x \in \{1, -1\}$ for the two BSL classes.

Our final training loss is the sum of the reconstruction loss, variational regularizer, and our additional auxiliary loss. For forward-pass encoding of input spectra x, we use the μ output of encoder network E, which we have designated as $\hat{F}(x)$.

Dataset

We compiled a dataset of 6264 spectra – spanning over 200 different species of microorganisms – from version 3 of the Robert Koch Institute's (RKI) MALDI-TOF MS database of highly pathogenic microorganisms (Lasch, Stämmler, and Schneider 2018). BSL was chosen as the target variable for our supervised machine learning task because it is the most accessible pathogenicity scoring metric available to us. In order to attain labeled spectra, we hand labeled a subset of the RKI database using the DSMZ catalogue of microorganisms to associate BSL with strain number. Strains identified by DSM number, or in some cases ATCC number, were labeled with their respective BSL. Ultimately 176 strains were labeled: 97 BSL-1, 79 BSL-2, 0 BSL-3, and 0 BSL-4. The resulting dataset contains 1359 labeled and 4905 unlabeled spectra.

Preprocessing

We used RKI's Microbe-MS MATLAB software to apply preprocessing to the data using the default functions and parameters. The following is a list of the functions and parameters applied via Microbe-MS. Full explanations of the operations performed in the functions listed below can be found on the Microbe-MS wiki.

1. Smoothing (21 smooth points)
2. Baseline Correction (100 intervals)
3. Normalization
4. Cut (2000-12000 m/z)

After running the data through Microbe-MS, we performed two additional preprocessing steps to facilitate learning in downstream algorithms. Each spectra is a collection of intensity (au), mass-to-charge (m/z) ratio coordinate pairs. Ideally, we could compare intensity vectors where each index into the vector corresponds to a specific m/z value. This is not possible in this format because spectra may be sampled at different m/z values. In order to achieve a uniformly sampled vector for each spectra, we interpolated the spectra from the au, m/z coordinate pairs and then resampled uniformly. Finally we normalized each vector to a unit vector using the L1 norm.

Experiments

In order to assess the validity of our results, we divide our dataset into an unlabeled training, and labeled training, validation, and test sets partitioned by microorganism strain. Setting the 4905 unlabeled spectra aside as our unlabeled training set, we setup our analysis by partitioning the labeled data into training, validation, and test sets, following the 70-15-15 splitting convention. The resulting training, validation, and test sets are as follows.

Sets	Strains	Spectra
Train	123	976
Validation	26	226
Test	27	157

We assert that the set of strains in each of the unlabeled training, and labeled training, validation, and test sets are disjoint. That is, no strain that appears in one set will appear in any of the others.

We evaluate how well an algorithm has embedded the original spectra by training a simple discriminator model on the embeddings and then evaluating the model's performance on the test set. We choose the K-Nearest Neighbors (KNN) algorithm for our discriminator because it is a simple and effective classifier when distances reflect semantically meaningful information. The KNN algorithm is ideal for our purposes because our embedding objective is essentially to construct a semantically meaningful low dimensional space. Furthermore, KNN accuracy on the training and validation sets can be viewed as a metric of how well the embeddings cluster, while accuracy on the test set will indicate how well the embeddings generalize.

We benchmark performance of embeddings generated by a series of neural network based approaches as well as PCA, against the performance of KNN on the original spectra (i.e. no embeddings). For network based approaches the validation set is used to measure model convergence, which will determine at what epoch the model should be evaluated on the test set. All network based approaches feature an identical encoder stack, which is composed of two fully connected layers, sizes 100 and 10, that embed the 10000 dimensional input vector. Furthermore, all networks that optimize for reconstruction feature an identical decoder stack. In the case of the discriminator only network, we follow best practices for binary classification problems using Binary Cross Entropy (BCE) to train the encoder stack. As discussed in the methodology section, our architecture uses both reconstruction and hinge loss to train the encoder stack. See Figures 1 and 2 for diagrams of the three network architectures evaluated.

We find that while all methods besides the unguided VAE do extremely well in training accuracy, all peaking at more than 95% training accuracy, most of the methods tested

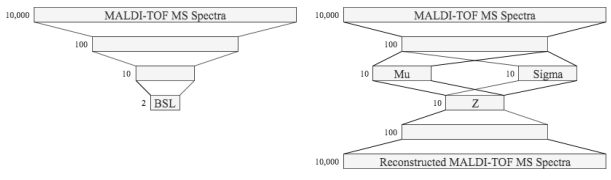

Figure 2: Diagrams of the discriminator architecture (left) and VAE architecture (right), which we benchmarked our auxiliary VAE architecture against.

Embedding Scheme	Epochs	Acc	Sensitivity	Specificity	True Pos	True Neg	False Pos	False Neg
Auxiliary VAE	25	**.917**	**.904**	**.929**	.420	.497	.038	.045
VAE (no auxiliary loss)	41	.548	.740	.381	.344	.204	.331	.121
Neural-net discriminator	3	.713	.644	.774	.299	.414	.121	.166
PCA	-	.777	.890	.679	.414	.363	.172	.051
No embeddings	-	.822	.781	.857	.363	.459	.076	.102

Table 1: Test accuracies for classification by the simple KNN classifier with specified embeddings at best validation epoch. BSL 2 or higher (pathogenic) is taken as the True case and BSL 1 (non-pathogenic) is taken as the False case. The embeddings generated by our auxiliary VAE beat all baselines in accuracy, sensitivity, and specificity

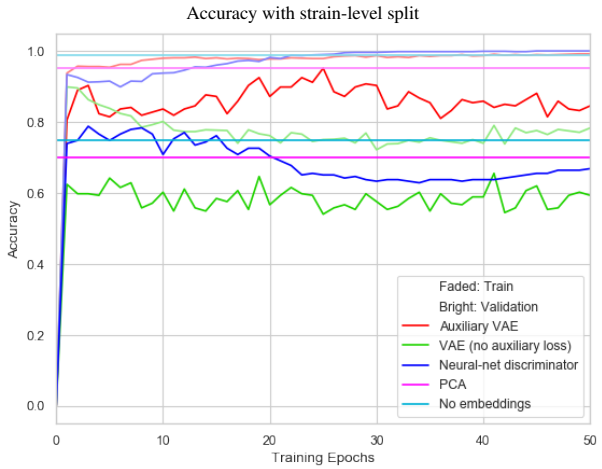

Figure 3: Training and validation classification accuracy for the Auxiliary VAE and various baseline methods vs epochs, on a validation set consisting of strains not present in the training set. The embeddings generated by each method are used to classify data using K-nearest neighbors on the training set with k=5. PCA and using raw data are methods that do not have epochs and are represented as horizontal lines.

perform significantly worse on test and validation. This is likely because while KNN works well for classifying spectra within strains even without embeddings, generalizing to new strains is difficult and requires dimensionality reduction that creates good clustering of the two biosafetly level classes. Additionally, the neural-net discriminator with no VAE component begins to heavily overfit, reducing its validation accuracy in later epochs. The VAE with auxiliary loss performs consistently well in validation. A graph of training epochs vs validation accuracy is included in Figure 3.

The three neural architectures (two baselines and our architecture) are then stopped at the epoch at which validation accuracy was highest, then run on the test set for a final evaluation. The results of this, along with the results of running the two one-shot baselines, are shown in Table 1.

Identifying Significant Peaks

In the interest of helping to better understand the MALDI-TOF MS signatures relevant to pathogenicity, we perform post-hoc analysis of the embedded space learned by our auxiliary VAE model in the strain partitioned experiment. We interpret specific spectra-to-embedding mappings with layer-wise relevance propagation. Given a spectra, layer-wise relevance propagation examines network weights, biases, and activations, as well as the corresponding embedding to propagate relevance backwards through the network (Bach et al. 2015). The resulting relevance vector represents the magnitude of positive or negative support each m/z value provides to the embedded spectra. See Figure 5 for a chart of MALDI-TOF spectra and their corresponding relevance vectors calculated using our auxiliary VAE architecture.

Species Experiment

The problem of determining BSL (i.e. pathogenic potential) becomes significantly harder when the dataset is partitioned at a species level, such that the set of species in each of the training, validation, and test sets are disjoint.

Sets	Species	Spectra
Train	93	926
Validation	20	223
Test	23	210

We also benchmarked our algorithms on this substantially harder problem in an effort to get an approximation of how our algorithms would generalize to novel/unknown microorganism. In this setting, accuracy overall is much lower, and often very close to random guessing. Accuracy is also highly dependant on the species present in the evaluation set, as seen in the large difference between the test and validation set accuracies for PCA and no embedding KNN models, which on an ideal dataset should have the same (or very similar) performance in these two scenarios, as these formulations do not use the validation set to tune hyperparameters. This variance is likely due to the relatively small number of species included in the validation and test sets, due to lack of data. Training and validation accuracy plotted over epochs is shown in Figure 4, and final test accuracy is shown in Table 2

Discussion

We see in the experiments section that our architecture out performs the discriminator and VAE models, showing that the semi-supervised approach beats both approaches it is based on. An interesting benefit we see from the semi-supervised approach is that the test error does not drop off significantly as does the discriminator model. We believe this is likely because our architecture is trying to optimize

Embedding Scheme	Epochs	Acc	Sensitivity	Specificity	True Pos	True Neg	False Pos	False Neg
Auxiliary VAE	2	**.681**	**.988**	.476	.395	.286	.314	.005
VAE (no auxiliary loss)	36	.500	.750	.333	.300	.200	.400	.100
Neural-net discriminator	1	.667	.917	**.500**	.367	.300	.300	.033
PCA	0	.590	.821	.437	.329	.262	.338	.071
No embeddings	0	.633	**.988**	.397	.395	.238	.362	.005

Table 2: Test accuracies for classification by the simple KNN classifier with specified embeddings at best validation epoch. BSL 2 or higher (pathogenic) is taken as the True case and BSL 1 (non-pathogenic) is taken as the False case. While the Auxiliary VAE has the highest overall accuracy, it is beaten or tied in specificity and sensitivity separately.

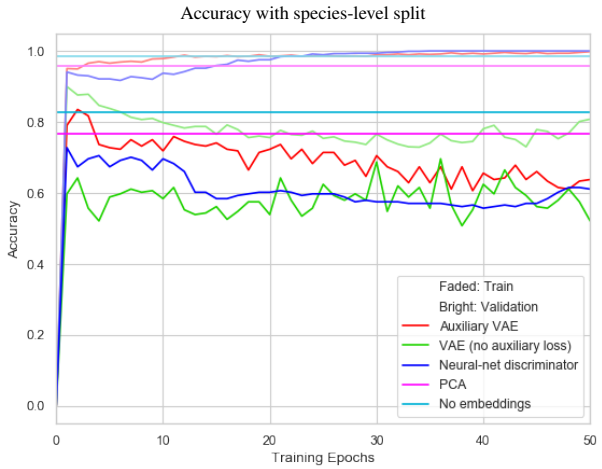

Figure 4: Training and validation classification accuracy for the Auxiliary VAE and various baseline methods vs epochs, on a validation set that has entirely species not present in the training set. The embeddings generated by each method are used to classify data using K-nearest neighbors on the training set with k=5. PCA and using raw data are methods that do not have epochs and are represented as horizontal lines.

for two objectives at once, which results in the two objectives regularizing each other. Another nice quality of our architecture is that a simple discriminative model trained on our network's embeddings has a very high sensitivity to pathogens. We observe this in both the strain and species experiments, see Tables 1 and 2. This is a critical metric for evaluating whether a biosurveillance system is fieldable, since it is much worse to call a pathogen non-pathogenic than it is to make the opposite mistake. Additionally, the layer-wise relevance propagation provided interesting insight into how our model processes the spectra. It was unexpected to find that the largest peaks were not always the most relevant to a spectra's embedding.

Future Work

We see multiple directions for future research and improvement building off of the work presented in this paper. First, we are particularly interested to know if the significant peaks identified via post-hoc analysis are indicative of biological phenomena associated with pathogenicity, such as protein toxins. Second, a necessary extension to our work here is reformulating the problem as a regression instead of a binary classification problem, in order to not only say whether or not an organism is pathogenic but also to say how pathogenic we think it might be. Third, the test sets used in this study were limited due to lack of data, resulting in some artifacts in validation/test accuracy — such as PCA and no embeddings baselines doing unusually well on the harder problem of pathogen detection on new species. Acquiring more data and evaluating models on these larger sets would decrease this variance and provide stronger evidence of the value of this approach. Finally, the largest barrier to a deployable pathogen detection system is likely the challenge of building a model that can generalize to divergent inputs. Of the estimated trillions of microbes on the planet we are familiar with a fraction of a percentage, thus it is critical to develop tools that will remain robust to the outliers. We believe a good first step at addressing this challenge is to improve the model to better handle data split at a species level, where the held-out species' spectra are pseudo-novel test points. While the current architecture does beat baselines in accuracy in this setting, 68.1% accuracy is not high enough for a deployable system.

Figure 5: Salient peak identification with layer-wise relevance propagation. The charts above show the normalized spectra (orange) and relevance of each m/z (purple) for various pathogenic and non-pathogenic strains of bacteria.

Acknowledgments

This material is based upon work supported by DARPA under Contract No. 140D6319C0030. The views, opinions, and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. This document is approved for public release, distribution unlimited.

References

Alam, S. I.; Kumar, B.; and Kamboj, D. V. 2012. Multiplex detection of protein toxins using MALDI-TOF-TOF tandem mass spectrometry: application in unambiguous toxin detection from bioaerosol. *Analytical chemistry* 84(23):10500–10507.

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10(7):e0130140.

Bellman, R. 1961. *Adaptive Control Processes: A Guided Tour*. Princeton University Press.

Dunham, S. J.; Ellis, J. F.; Li, B.; and Sweedler, J. V. 2016. Mass spectrometry imaging of complex microbial communities. *Accounts of chemical research* 50(1):96–104.

Hornik, K. 1991. Approximation capabilities of multilayer feedforward networks. *Neural networks* 4(2):251–257.

Hsieh, S.-Y.; Tseng, C.-L.; Lee, Y.-S.; Kuo, A.-J.; Sun, C.-F.; Lin, Y.-H.; and Chen, J.-K. 2008. Highly efficient classification and identification of human pathogenic bacteria by MALDI-TOF MS. *Molecular & cellular proteomics* 7(2):448–456.

Kingma, D. P., and Welling, M. 2013. Auto-encoding variational bayes. *CoRR* abs/1312.6114.

Kingma, D. P.; Mohamed, S.; Rezende, D. J.; and Welling, M. 2014. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, 3581–3589.

Lasch, P.; Stämmler, M.; and Schneider, A. 2018. Version 3 (20181130) of the MALDI-TOF mass spectrometry database for identification and classification of highly pathogenic microorganisms from the Robert Koch-Institute (RKI) [data set]. Zenodo. https://doi.org/10.5281/zenodo.1880975.

Marx, V. 2019. A dream of single-cell proteomics. *Nature methods* 1–4.

Patel, R. 2013. Matrix-assisted laser desorption ionization–time of flight mass spectrometry in clinical microbiology. *Clinical infectious diseases* 57(4):564–572.

Pearson, K. 1901. LIII. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2(11):559–572.

Rifai, S.; Vincent, P.; Muller, X.; Glorot, X.; and Bengio, Y. 2011. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 833–840. Omnipress.

Sauer, S.; Freiwald, A.; Maier, T.; Kube, M.; Reinhardt, R.; Kostrzewa, M.; and Geider, K. 2008. Classification and identification of bacteria by mass spectrometry and computational analysis. *PloS one* 3(7):e2843.

Seuylemezian, A.; Aronson, H. S.; Tan, J.; Lin, M.; Schubert, W.; and Vaishampayan, P. 2018. Development of a custom MALDI-TOF MS database for species-level identification of bacterial isolates collected from spacecraft and associated surfaces. *Frontiers in microbiology* 9:780.

Sharif Razavian, A.; Azizpour, H.; Sullivan, J.; and Carlsson, S. 2014. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 806–813.

Singhal, N.; Kumar, M.; Kanaujia, P. K.; and Virdi, J. S. 2015. MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis. *Frontiers in microbiology* 6:791.

Wang, W.; Huang, Y.; Wang, Y.; and Wang, L. 2014. Generalized autoencoder: A neural network framework for dimensionality reduction. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 490–497.