Extended-Spectrum Beta-Lactamase and Carbapenemase-Producing prediction in *Klebsiella pneumoniae* based on MALDI-TOF mass spectra

A. Guerrero-López, C. Sevilla-Salcedo, A. Candela, M. Hernández-García, E. Cercenado, P. M. Olmos, R. Cantón, P. Muñoz, V. Gómez-Verdejo, R. del Campo, and B. Rodríguez-Sánchez

***Abstract*— Matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) Mass Spectrometry (MS) is a reference method for microbial identification. Currently, machine learning techniques are used to predict Antibiotic Resistance (AR) based on MALDI-TOF data. However, current solutions need costly preprocessing steps, their reproducibility is difficult due to hyperparameter cross-validation, they do not provide interpretable results, and they do not take into account the epidemiological difference inherent to data coming from different laboratories. In this paper, we validate a multi-view heterogeneous Bayesian model (SSHIBA) for AR mechanism prediction based on MALDI-TOF MS. This novel approach allows exploiting local epidemiology differences between data sources, gets rid of preprocessing steps, is easily reproducible because hyperparameters are optimized by Bayesian inference, and provides interpretable results. To validate this model and its advantages, we present two domains of *Klebsiella pneumoniae* isolates: 282 samples of Hospital General Universitario Gregorio Marañón (GM) domain and 120 samples for Hospital Universitario Ramón y Cajal (RyC) domain that discriminates between Wild Type (WT), Extended-Spectrum Beta-Lactamases (ESBL)-producers and ESBL + Carbapenemases (ESBL+CP)-producers. Experimental results prove that SSHIBA outperforms state-of-the-art (SOTA) algorithms by exploiting the multi-view approach that allows it to distinguish between data domains, avoiding local epidemiological problems. Moreover, it shows that there is no need to preprocess MALDI-TOF data. Its implementation in microbiological laboratories could improve the detection of multi-drug resistant isolates, optimizing the therapeutic decision and reducing the time to obtain results of the resistance mechanism. The proposed model implementation, specifically adapted to AR prediction, and data collections are publicly available on GitHub at: github.com/alexjorguer/RMPrediction**

***Index Terms*— Semi-supervised, missing data, heterogeneous, Bayesian, ESBL, CP, MALDI-TOF, Klebsiella pneumoniae, antibiotic resistance prediction**

Submitted on XX/XX/XXXX. This work was supported by Spanish MINECO (Agencia Estatal de Investigación) [TEC2017-92552-EXP to P.O., RTI2018-099655-B-100 to P. O., TEC2017-83838-R to A. G., C. S. and V. G., PID2020-115363RB-I00 to A. G., C. S. and V. G.]; and Comunidad de Madrid [IND2017/TIC-7618, IND2018/TIC-9649, IND2020/TIC-17372, Y2018/TCS-4705 to P. O.]; and the BBVA Foundation under the Domain Alignment and Data Wrangling with Deep Generative Models (Deep-DARWiN) project to P. O.; and the European Union (European Regional Development Fund and the European Research Council) through the European Union's Horizon 2020 Research and Innovation Program [714161 to P. O.]; and Intramural Program of the Gregorio Marañón Health Research Institute to A. G.; and Health Research Fund (Instituto de Salud Carlos III. Plan Nacional de I+D+I 2013-2016) of the Carlos III Health Institute (ISCIII, Madrid, Spain) [PI15/01073, PI18/00997 to A. C. and B. R.] partially financed by the European Regional Development Fund (FEDER) 'A way of making Europe'; and Health Research Fund Miguel Servet contract [CPII19/00002 to B. R.]

A. Guerrero-López and P. M. Olmos are with the Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Leganés, 28911, Spain and with the Gregorio Marañón Health Research Institute, Hospital General Universitario Gregorio Marañón, Madrid, 28009, Spain (e-mail: alexjorguer@tsc.uc3m.es, pamartin@ing.uc3m.es).

C. Sevilla-Salcedo and V. Gómez-Verdejo are with the Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Leganés, 28911, Spain (e-mail: casevill@pa.uc3m.es, vanessag@ing.uc3m.es)

A. Candela, E. Cercenado, P. Muñoz, and B. Rodríguez-Sánchez are with the Gregorio Marañón Health Research Institute, Hospital General Universitario Gregorio Marañón, Madrid, 28009, Spain (email: acandelagon@gmail.com, emilia.cercenado@salud.madrid.org, pamunoz@iisgm.com, mbelen.rodriguez@iisgm.com)

M. Hernández-García, R. Cantón, and R. del Campo are with the Department of Microbiology, Hospital Ramón y Cajal and Instituto Ramón y Cajal de Investigación Sanitaria, Madrid, 28034, Spain and with the CIBER en Enfermedades Infecciosas, Madrid, 28034, Spain (e-mail: martahernandez1986@gmail.com, rafael.canton@salud.madrid.org, rosacampo@yahoo.com)

I. Introduction

Multidrug-resistant *Klebsiella pneumoniae* is considered a global public health threat by major international health organizations due to its rapid spread, high morbidity and mortality, and the economic burden associated with its treatment and control [1]–[3]. Resistance to carbapenems is a major challenge, as recognized by the World Health Organization (WHO) [4], where some carbapenemases have been shown to hydrolyze almost all beta-lactam antibiotics. Concretely, *K. pneumoniae* has shown a great capability to acquire antibiotic resistant mechanisms, mainly beta-lactamases and carbapenemases, over the last two decades. The presence of *K. pneumoniae* isolates hosting these resistance mechanisms complicates the treatment options and patients' outcome [5]. Thus, besides the routinely Antimicrobial Susceptiblity Testing (AST), rapid diagnostic methods such as Matrix-assisted laser desorption

ionization time-of-flight (MALDI-TOF) Mass Spectrometry (MS), beyond identification, should be implemented in clinical microbiology laboratories for early detection of multidrug-resistant isolates.

MALDI-TOF MS is designed for microbial identification, but also allows the detection of ESBL and Carbapenemases (CP) due to the different molecular weight of the antibiotic after its hydrolysis by resistant bacteria [6]. This approach is faster than conventional AST (30-60 min vs. 18-24 h), but requires highly trained personnel and its use is limited to clinical laboratories. However, Machine Learning (ML) approaches can automatically analyze and predict AR based on the MALDI-TOF MS protein profiles, as suggested in [7]. The most relevant limitation for this methodology is the high complexity of the MALDI-TOF spectrum, which is also influenced by the particularities of each lineage and its accessory genome. In this sense, two isolates carrying the same CP gene could differ in their MS due to their particular genetic background. Therefore, useful ML models have to be able to jointly learn from their MALDI-TOF MS and their epidemiology to model these particularities.

In the literature, ML has been widely used for microbial species identification [8]–[13]. However, there is a lack of quality publications on prediction of AR using ML. A first preprocessing step is commonly performed by using the R package *MALDIquant* (MQ) (MQ) [14]. After this first step, supervised ML methods such as Support Vector Machine (SVM) or Random Forest (RF) are applied for the detection of different antibiotic resistances in *S. aureus* [15]–[17]. One of the most commonly used tools is the ClinProTools software [18] which contains plug-and-play classification models such as Genetic Algorithm (GA), Supervised Neural Network (SNN), SVM or Quick Classifier (QC) [19], although they are used as a black box. Other previous work dealing specifically with AR prediction in *K. pneumoniae* [20] points out that RF is the best model for predicting the production of CP, although it is worth mentioning that the data set was small (less than 100 samples). Although supervised SOTA models are a powerful classification tool, they are not designed to deal with high dimensionality data such as the MALDI-TOF MS. Consequently, it is necessary to reduce the dimensionality of these data; in this regard, some authors [21], [22] proposed the use of a GA as a dimenionality reduction technique. Then, they used a SVM as a classifier to perform susceptibility prediction of *S. aureus*. Other approach [23], proposed using a RF as a peak selector to subsequently predict AR using simpler methods such as Logistic Regressor (LR) or Linear Discriminat Analysis (LDA). Other authors [24], used unsupervised learning to identify relevant features and then applied Binary Discriminant Analysis (BinDA) [25] and SVM as classifiers. Nowadays, Bayesian models are starting to be applied in this field, as they get rid of cross-validation problems and can provide a predictive distribution with a measure of confidence.

A recent study [26] proposed a new SOTA approach where they first reduce the dimensionality of the MALDI-TOF to only 200 peaks per sample by implementing topological peak filtering. Then, they proposed a specifically tailored non-linear kernel which exploits the correlation between the peaks intensity and their relative position in m/z dimension. Finally, they perform the prediction by a Bayesian probabilistic method, such as a Gaussian Process (GP).

In this paper, we propose to predict the AR using MALDI-TOF by means of a new Bayesian multivariate model called Kernelized Sparse Semi-Supervised Interbattery Bayesian Analysis (KSSHIBA) [27], [28] which is able to outperform SOTA algorithms in the prediction of CP and ESBL susceptibility while:

- **Avoids preprocessing steps**: the model is able to efficiently handle the raw MALDI-TOF data with only Total Ion Current (TIC) normalization, so it does not need to use external preprocessing packages such as MQ.
- **Eliminates hyperparameter cross-validation**: since the hyperparameters can be automatically optimized by maximizing the variational lower bound.
- **Achieves dimensionality reduction**: the model is able to use kernelized data representation (reducing the data dimension to the number of data points) and, besides, it obtains a low dimensional data representation by projecting all data views to a common latent space.
- **Provides interpretability**: since it computes a weight matrix associated with each view, capable of explaining how the different views correlate with each other.

II. Methods

A. Isolates selection and processing

We include two different data domains. The first data domain has 282 consecutive clinical *K. pneumoniae* isolates collected between 2014 and 2019 and later isolated at Hospital General Universitario Gregorio Marañón (GM). Therefore, this data domain is called from now on the GM domain. The second data domain has 120 isolates which were characterized in surveillance programs (STEP and SUPERIOR) [29], [30] sourcing from 8 Spanish and 11 Portuguese hospitals. The AST determination of these 120 isolates has been performed in the Hospital Universitario Ramón y Cajal (RyC). Therefore, this data domain is called from now on the RyC domain.

The AST determination has been performed separately in their origin center by the automated broth microdilution method Microscan® System (Beckman-Coulter, CA, USA), using EUCAST (2021) common criteria. The presence of ESBL/CP genetic resistant mechanisms has been corroborated by molecular testing. Each isolated is labelled as Wild Type WT), ESBL-producers or ESBL+CP-producers as shown in Table I.

TABLE I
DATASET DETAILED BY DOMAIN AND LABEL TYPES

Dataset	Label	Samples
GM	WT	85
	ESBL	6
	ESBL+CP	191
RyC	WT	9
	ESBL	58
	ESBL+CP	53

Isolates have been kept frozen at -80ºC in skimmed milk and, after thawing, they have been cultured overnight at 37ºC

in Columbia Blood agar (bioMérieux, Lyon, France) during 3 subcultures for metabolic activation. The MS analysis has been centralized and performed by the same operator using a MBT Smart MALDI Biotyper mass spectrometer (Bruker Daltonics, Bremen), in 6 separated replicates (2 positions on 3 consecutive days). The protein extraction has been performed adding 1µl 100% formic acid and then drying at room temperature. Next, 1µl of HCCA matrix solution (Bruker Daltonics) have been added to each spot. The MALDI-TOF spectra has been acquired in positive linear mode in the range of 2,000 to 20,000 Da, using default settings [31], although only data between 2,000-12,000 m/z [32], [33] has been used.

The Ethics Committees of both GM and RyC (codes MICRO.HGUGM.2020-002, and 087-16, respectively) have approved this study. The study has been performed on microbiological samples, not human products, and patient informed consent has not been required.

B. Proposed model

In this context, we propose the use of SSHIBA [27], a Bayesian semi-supervised model that assumes that there is a latent representation capable of generating the different views of the heterogeneous data. We consider a multi-view problem composed of 3 views: one view is the MALDI-TOF MS data kernelized, another view is the AR we want to predict, and another is the domain label representing the hospital where the data belong. This can be seen in the graphical model presented in Fig. 1[1].

The MALDI-TOF data, i.e. the first view, is kernelized to reduce its dimension. Our observations are denoted as $\mathbf{K} \in \mathbb{R}^{N\times N}$, where N is the number of samples in each experiment, since it is a kernel matrix computed between all the MALDI-TOF MS samples. Then, each row represents a kernelized observation, denoted as$\mathbf{k}_{\mathrm{n},:}$:

$$\mathbf{k}_{\mathrm{n},:} = [\mathrm{kf}(\mathrm{MD}_1, \mathrm{MD}_1), ..., \mathrm{kf}(\mathrm{MD}_1, \mathrm{MD}_N)] \tag{1}$$

where $\mathrm{kf}(\mathrm{MD}_a, \mathrm{MD}_b)$ is a kernel function between MD_a and MD_b, which are an arbitrary pair of MALDI-TOF mass spectra.

Using the kernelized extension of the model called KSSHIBA [28], we consider that there exists some low-dimensional latent variable vector $\mathbf{z}_{\mathrm{n},:} \in \mathbb{R}^{1\times K_c}$ which is linearly combined with a set of dual variables $\mathbf{A} \in \mathbb{R}^{N\times K_c}$, where K_c is the dimension of the low-dimensional latent space, and a zero-mean Gaussian noise $\tau^{(\mathrm{M})}$ to generate each row of the kernelized observations $\mathbf{k}_{\mathrm{n},:}$, as:

$$\mathbf{k}_{\mathrm{n},:} \;=\; \mathbf{z}_{\mathrm{n},:}\,\mathbf{A}^{\mathrm{T}} + \tau^{(\mathrm{M})} \tag{2}$$

where the prior over the latent space is given by $\mathbf{z}_{\mathrm{n},:} \sim \mathcal{N}(0, I_{K_c})$ and the prior over the dual variables is given by $\mathbf{A} \sim \mathcal{N}(0, (\alpha_{\mathrm{k}})^{-1} I_{K_c})$. The random variable $\alpha_{\mathrm{k}} \sim \Gamma(a, b)$ follows an Automatic Relevance Determination (ARD) prior [34] over the columns of $\mathbf{A}$ to automatically select the columns of $\mathbf{z}_{\mathrm{n},:}$ that are indeed relevant to explain the current data view.

[1]Icons images were provided by smart.servier.com

For the AR observations, denoted as T, we propose to use a one-hot encoding for the WT, ESBL and ESBL+CP tags. Likewise, for the data domain, denoted as D, we consider binary encoding where a 0 value means that the data come from the GM domain and a value of 1 means that the data come from the RyC domain.

To accommodate for these two binary observations, we first consider that there exists two real latent variables $\mathbf{X}^{(\mathrm{m})}$, $m \in \{\mathrm{T}, D\}$, that are generated by the common low-dimensional latent variables $\mathbf{z}_{\mathrm{n},:} \in \mathbb{R}^{N\times K_c}$, which as before, are linearly combined with a projection matrix $\mathbf{W}^{(\mathrm{m})} \in \mathbb{R}^{D\times K_c}$ (where D is the observation dimension) and a zero-mean Gaussian noise $\tau^{(\mathrm{m})}$, as follows:

$$\mathbf{x}^{(\mathrm{m})}_{\mathrm{n},:} = \mathbf{z}_{\mathrm{n},:}\,\mathbf{W}^{(\mathrm{m})^{\mathrm{T}}} + \tau^{(\mathrm{m})} \;\text{ for } m \in \{\mathrm{T}, D\} \tag{3}$$

where $\mathbf{W}^{(\mathrm{m})}$'s prior is identical to $\mathbf{A}$'s to automatically select which columns of $\mathbf{z}_{\mathrm{n},:}$ are needed to explain these two views. Then, we are able to generate $\mathbf{T}^{(\mathrm{m})}$ by conditioning to this new latent representation $\mathbf{X}^{(\mathrm{m})}$ using and independent Bernoulli probability model, for AR view, as:

$$p\left(\mathrm{t}_{\mathrm{n},:} \,|\, \mathbf{x}^{(\mathrm{m})}_{\mathrm{n},:}\right) = \prod_{d=1}^{3} p\left(\mathrm{t}_{\mathrm{n,d}} \,|\, \mathrm{x}^{(\mathrm{T})}_{\mathrm{n,d}}\right) \tag{4}$$

where:

$$p\left(\mathrm{t}_{\mathrm{n,d}} \,|\, \mathrm{x}^{(\mathrm{T})}_{\mathrm{n,d}}\right) = e^{\mathrm{x}^{(\mathrm{T})}_{\mathrm{n,d}}\,\mathrm{t}_{\mathrm{n,d}}}\sigma\left(-\mathrm{x}^{(\mathrm{T})}_{\mathrm{n,d}}\right) \tag{5}$$

and for domain view, as it is binary:

$$p\left(\mathrm{d}_{\mathrm{n},1} \,|\, \mathbf{x}^{(\mathrm{m})}_{\mathrm{n},1}\right) = e^{\mathrm{x}^{(\mathrm{D})}_{\mathrm{n},1}\,\mathrm{d}_{\mathrm{n},1}}\sigma\left(-\mathrm{x}^{(\mathrm{D})}_{\mathrm{n},1}\right) \tag{6}$$

The model is trained by evaluating the posterior distribution of all the random variables posteriors given the observed data. These posteriors are approximated through mean-field variational inference [35] maximizing the evidence lower bound (ELBO). For more details, see [27], [28]. Furthermore, the Bayesian nature of the model allows it to work in a semi-supervised fashion, using all available information to determine the approximate distribution of the variables. In turn, the model can marginalize out any type of missing values in the data, as well as predict the test samples for AR by sampling its variational distribution.

C. Model training and validation

We study two different scenarios: (1) training and testing in each domain separately (intra-domain analysis), and (2) training and testing in both domains simultaneously (inter-domain analysis). For each analysis, we divide each domain into 5 random training-test folds. Due to the label imbalance seen in Table 1, we correct it in each training fold by oversampling the minority class, which ultimately result in stratified folds with a consistent class ratio.

For inter-domain analysis, we merge the two randomized 5-folds of training previously used in the first analysis in two different ways: (1) directly combining both domains, i.e., train with only two views (kernelized MALDI-TOF view and AR view) and (2) combining both domains by adding a third view indicating to which domain each data belongs. In other words,

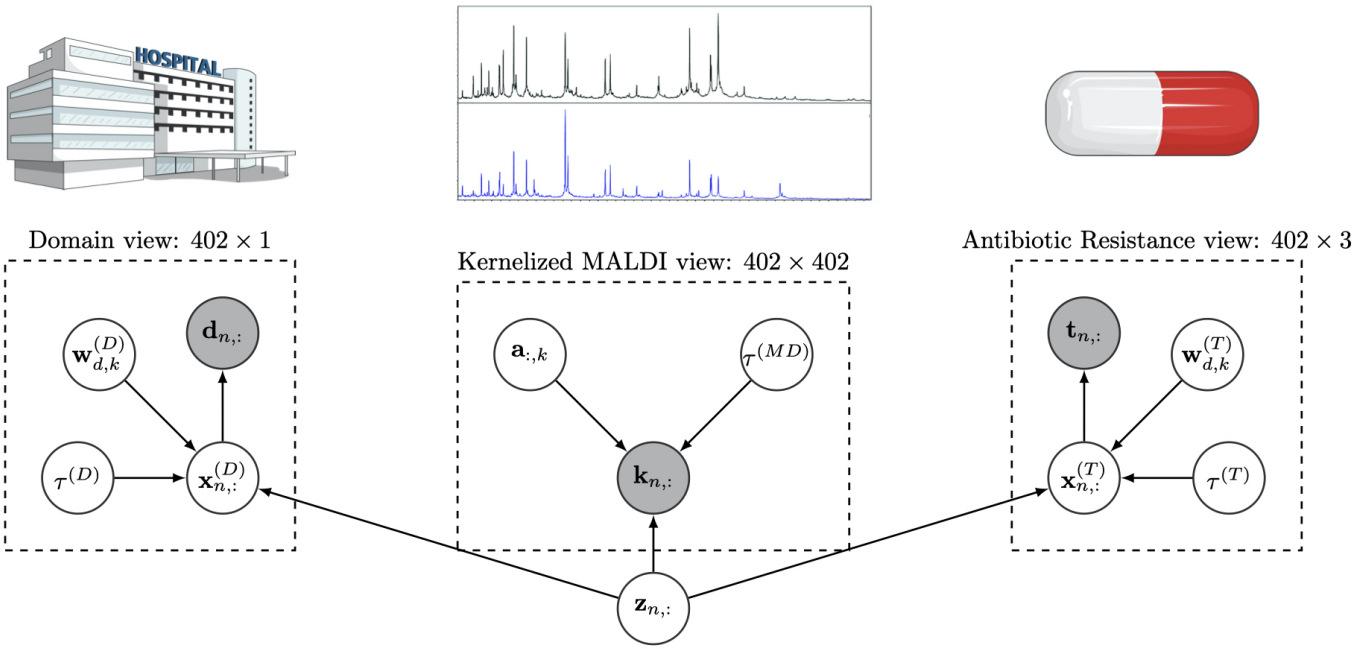

Fig. 1. Probabilistic graphical model for the evaluated data set: view $\mathbf{D}$ corresponds to the label of the domain they come from (GM or RyC), view $\mathbf{M}$ corresponds to the kernelized MALDI-TOF MS data, and view $\mathbf{T}$ corresponds to the AR (WT, ESBL or ESBL+CP). The white circles represent random variables that the model learns, while the gray circles represent the observations.

in the first case we do not use the D observations, meanwhile in the second case we are using the D observations. In this way, we analyze the importance of knowing the origin of the data.

Finally, we measure the performance in terms of Area Under the ROC Curve (AUC) of AR prediction over the test folds.

D. Methods under study

We compare KSSHIBA with a SVM and a GP since all three models can work with kernel formulations. As a kernel function, we first test a nonlinear approach, such as Radial Basis Function (RBF) that is given by:

$$kf(\mathrm{MD}_a, \mathrm{MD}_b) = \exp\left(-\frac{||\mathrm{MD}_a - \mathrm{MD}_b'||^2}{2\sigma^2}\right) \quad (7)$$

where σ is the variance hyperparameter. Then, we also test a linear kernel which follows:

$$kf(\mathrm{MD}_a, \mathrm{MD}_b) = \mathrm{MD}_a^T \mathrm{MD}_b \quad (8)$$

Finally, we work with a SOTA kernel function called Peak Information KErnel (PIKE) [26], which exploits the nonlinear correlations between the MALDI-TOF peaks as follows:

$$kf(\mathrm{MD}_a, \mathrm{MD}_b) = \frac{1}{2\sqrt{2\pi t}} \sum_{i,j}^{Peaks} \lambda_i \lambda_j' \exp\left(-\frac{(p_i - p_j')^2}{8t}\right) \quad (9)$$

where t is a smoothing parameter that has to be cross-validated, $\lambda_{i,j}$ correspond to the intensity value of each pair of peaks, and $p_{i,j}$ is their m/z position in the spectra. Let's remind that each MALDI-TOF consists of 12,000 different peaks. Due to the computational cost to evaluate Eq. 9 in that amount of peaks, the spectra are preprocessed beforehand by topological peak selection keeping only 200 peaks per sample, as the authors of [26] indicate in their paper.

Since we are solving a multi-class classification problem, we train the SVMs and GPs in a one-vs-all scheme. Besides, we also compare ourselves against a multitask RF.

Regarding cross-validation, we use an inner 5-fold over the training folds to validate all hyperparameters. We cross-validate the C value (0.01, 0.1, 1, 1, 10) for the SVM and both the number of estimators (50, 100, 150) and the maximum number of features (auto, log2) for the RF. For both KSSHIBA and GP, the hyperparameters are optimized by maximizing the ELBO and the marginal log likelihood of the data, respectively. In case of using PIKE, we also cross-validate t smoothing value.

Finally, when we use the MQ package for preprocessing the MS we denoted it using the prefix *MQ-*, e.g. MQ-KSSHIBA-RBF.

III. Experiments

In this section, we present the results obtained using the proposed model and the different SOTA algorithms. First, we study the classification performance in the intra-domain scenario. Then, we analyze the performance in the inter-domain scenario to evaluate the advantages of working with multi-view data sources. Finally, we study the latent space projection learned by KSSHIBA to understand the correlation between the source domain and the labels.

A. Intra-domain scenario

Table II summarizes the results obtained by training and testing independent models for each domain (GM and RyC).

TABLE II
RESULTS OF ALL MODELS IN THE INTRA-DOMAIN SCENARIO IN TERMS OF MEAN AUC AND STANDARD DEVIATION W.R.T. THE 5 RANDOM SPLITS. THE BEST RESULT FOR EACH CASE IS SHOWN IN BOLD.

Dataset	Label	KSSHIBA RBF	KSSHIBA LINEAR	KSSHIBA PIKE	GP LINEAR	MQ-GP PIKE [26]	SVM RBF	RF
GM	WT	0.61±0.14	0.70±0.15	0.71±0.16	0.70±0.18	**0.75±0.11**	0.67±0.12	0.70±0.17
	ESBL	**0.57±0.28**	0.46±0.19	0.56±0.32	0.54±0.18	0.35±0.14	0.40±0.29	0.39±0.21
	ESBL+CP	**0.85±0.14**	0.77±0.16	0.78±0.09	0.80±0.20	0.79±0.07	0.82±0.19	0.80±0.19
RyC	WT	0.47±0.35	0.49±0.22	**0.64±0.19**	0.48±0.28	0.56±0.20	0.45±0.15	0.57±0.26
	ESBL	0.70±0.10	0.59±0.08	0.43±0.09	0.58±0.14	0.43±0.11	**0.72±0.14**	0.69±0.10
	ESBL+CP	0.67±0.12	0.66±0.05	0.43±0.09	0.62±0.06	0.55±0.05	0.71±0.17	**0.71±0.07**

For the GM domain, as shown in Table II, KSSHIBA outperforms the baselines in terms of AUC in both ESBL and ESBL+CP prediction. Nonlinear kernels provide the best results in the three tasks. Specifically, the RBF kernel is the best choice for both ESBL and ESBL+CP while the PIKE kernel outperforms all other functions in WT prediction.

In contrast, the RyC domain turns out to be the most complex to deal with, since no model works correctly on all three labels simultaneously due to the overall sample size. Despite this, we observe the same results as for GM: nonlinear techniques such as PIKE, RBF or RF performed better than the linear ones.

B. Inter-domain scenario

Table III shows the results obtained by different models when trained jointly on both GM and RyC domains. The name of every model is constructed by four terms, indicating *Preprocess-Model-Kernel-Views*. For example, KSSHIBA-LINEAR-DOMAIN means that we use raw MALDI-TOF data, a KSSHIBA model with a linear kernel function, and we add the D labels. In contrast, MQ-GP PIKE means that we use MQ package to preprocess MQ data, and a GP model with a PIKE kernel function without using the D observations.

By combining both domains, the linear version of KSSHIBA DOMAIN outperforms all SOTA models for both WT and ESBL+CP predictions. Besides, KSSHIBA DOMAIN also outperforms all models in the previous experiment which were targeted to each particular domain. Since KSSHIBA works with heterogeneous data, such as the domain label, it is able to exploit the information inherent in both data sets.

Regarding the kernel function, when using the RBF kernel the results worsen significantly making it clear that both distributions are far apart and, therefore, a linear kernel can better explain these differences. This is due to the fact that, setting a common γ parameter in the RBF kernel that adequately explains the similarity of the data in both domains is not feasible. Thus, a simpler linear kernel gets better results.

If we compare the versions of KSSHIBA with domain view and without it, e.g. KSSHIBA LINEAR DOMAIN versus KSSHIBA LINEAR, the improvement in the first case indicates that the model is able to get rid of the possible local epidemiological bias induced by the domain and properly merges both datasets.

In addition, KSSHIBA shows a new advantage for WT and ESBL+CP prediction: no external preprocessing with MQ is required when we merge data coming from different domains. Although MQ preprocessing presents better results in the GM domain, it performs poorly in the RyC domain. The MQ pipeline defines reference peaks based on most frequent peaks. We hypothesize that this pipeline leads the data to be biased to the bigger domain, being GM in our scenario. As we can see, using the raw data allows us to maintain an almost identical performance in the GM domain while we improve all predictions for the RyC.

The remaining models, apart from not outperforming KSSHIBA, refute the linear kernel hypothesis: these kernels are better for prediction of both WT and ESBL+CP. This indicates that in the first experiment, by only taking into account one domain, the model overfitted the samples in that domain. Whereas, by adding out-of-distribution data, the linear kernel is able to generalize better.

Regarding ESBL prediction, there is no model that achieves high AUC values without losing in all other predictions. Therefore, we would need more data from this label to be able to generalize correctly.

C. Latent space analysis

Since KSSHIBA LINEAR DOMAIN exhibits the best performance, we analyze the latent space projection learned to understand the importance of the domain label.

KSSHIBA automatically determines the number of latent features needed to explain at the same time all three views by performing dimensionality reduction.

Fig. 2 represents the average weight of each latent factor $\mathbf{W}^{(\mathrm{m})}$ for $m \in \{\mathrm{MD}, \mathrm{T}, D\}$, i.e. the average across every column. We recover $\mathbf{W}^{(\mathrm{MD})}$ by moving from the dual space $\mathbf{A}$ variables to primal space as follows:

$$\mathbf{W}^{(\mathrm{MD})} = \mathrm{MD}^T \mathbf{A} \tag{10}$$

where MD are the raw MALDI-TOF observations. Due to the sparsity induced by the ARD prior explained above, $\mathbf{W}^{(\mathrm{m})}$ automatically selects the k features of $\mathbf{z}_{\mathrm{n},:}$ that are only relevant for each m data view.

In this case, KSSHIBA decided that only 76 latent features are needed, as shown in Fig. 2. It is noteworthy that, from these 76 variables, only 13 latent factors are used to predict the label AR. From these 13, all of them share information with the MALDI-TOF view but only 3 of them simultaneously correlate all available information. Finally, note that 51 private latent features are necessary for the MALDI-TOF view, which corresponds to an unsupervised projection of the data that

TABLE III

RESULTS OF ALL MODELS IN THE INTER-DOMAIN SCENARIO IN TERMS OF MEAN AUC AND STANDARD DEVIATION W.R.T THE 5 RANDOM SPLITS. THE BEST RESULT FOR EVERY CASE IS SHOWN IN BOLD.

Dataset	Label	KSSHIBA LINEAR DOMAIN	KSSHIBA PIKE DOMAIN	KSSHIBA RBF DOMAIN	MQ-KSSHIBA LINEAR DOMAIN	KSSHIBA LINEAR	GP LINEAR	SVM LINEAR	MQ-GP PIKE [26]	SVM RBF
GM	WT	**0.77±0.11**	0.64±0.17	0.59±0.16	**0.78±0.06**	0.72±0.14	0.76±0.10	0.62±0.13	0.74±0.14	0.65±0.13
	ESBL	0.46±0.19	0.44±0.37	0.32±0.23	0.51±0.25	0.39±0.21	0.43±0.20	0.39±0.21	0.40±0.12	0.40±0.19
	ESBL+CP	**0.88±0.08**	0.73±0.12	0.81±0.12	**0.90±0.05**	0.86±0.08	0.86±0.08	0.85±0.08	0.77±0.10	0.85±0.08
RyC	WT	**0.70±0.16**	0.51±0.14	0.63±0.07	0.48±0.28	0.66±0.16	0.68±0.17	0.59±0.20	0.62±0.22	0.57±0.26
	ESBL	0.55±0.09	0.39±0.14	0.66± 0.05	0.53±0.07	0.49±0.09	0.60±0.10	**0.69±0.12**	0.49±0.09	0.69±0.12
	ESBL+CP	**0.68±0.10**	0.51±0.14	0.59±0.11	0.63±0.10	0.64±0.06	0.64±0.04	0.66±0.14	0.59±0.06	0.68±0.14

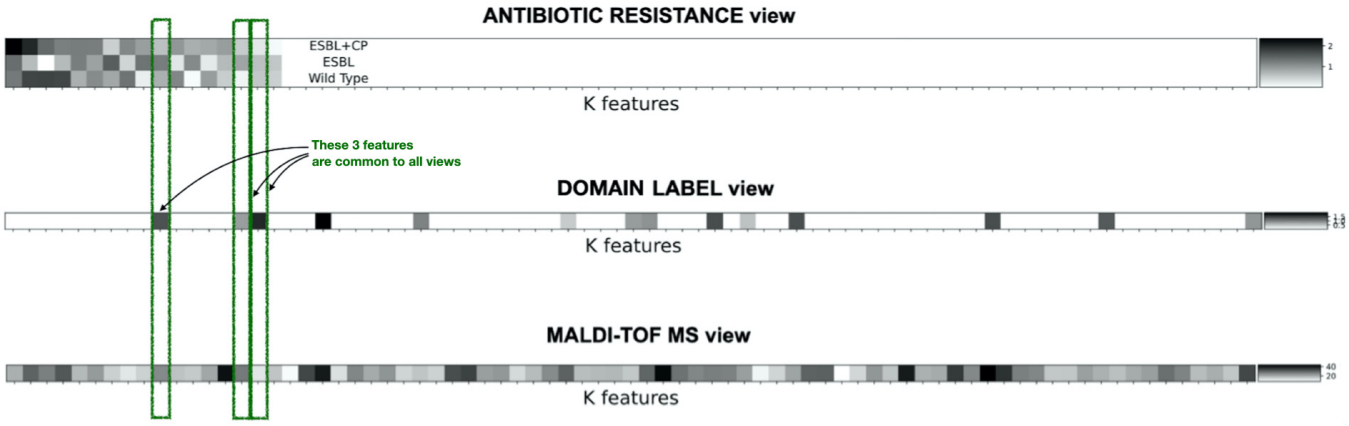

Fig. 2. Each row represents the mean of each $\mathbf{w}_{\mathbf{d},:}^{(\mathbf{m})}$ d-row having then $\mathbf{76}$ values, one per each k latent feature. Each subplot represents one $\mathbf{W}^{(\mathbf{m})}$ matrix per view. The most important features (the highest weight value) are represented in black, and the least important features (the lowest weight value) are represented in white. Finally, the features were ordered by their relevance to the prediction task

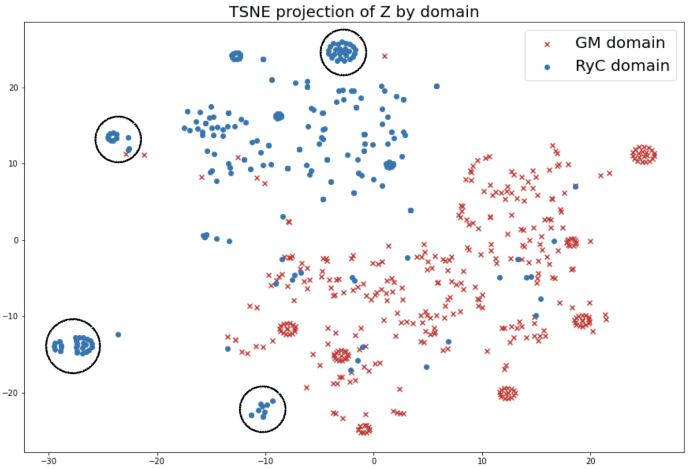

Fig. 3. The 14 latent variables of $\mathbf{Z}$ that are relevant for the domain view has been projected to a 2-dimensional using a t-SNE. Every cross stand for a $\mathbf{z}_{\mathbf{n},:}$ whose observation comes from GM domain, while every dot stand for a $\mathbf{z}_{\mathbf{n},:}$ whose observation comes from RyC domain.

exclusively models the behavior of the MALDI-TOF view, as a Principal Component Analysis could do.

Fig. 2 also shows that there is a correlation between the origin domain of each strain and its AR, as they have 3 shared latent features. In addition, the domain label is also used to explain the projection of the MS, proving that the MALDI-TOF distributions differs based on epidemiology.

Regarding the low dimension projection, in Fig. 3 we use a T-distributed Stochastic Neighbor Embedding (t-SNE) to project the 14 latent features that $\mathbf{w}_{d,k}^{(2)}$ indicates are relevant to explain the domain view to a 2-dimensional space. It can be seen that both domains can be expressed by two different distributions and that a simple linear classifier can separate both distributions. The GM domain presents a more compact space as they are all coming from the same hospital. In contrast, as RyC domain is a collection of 19 different hospitals, we can see that it presents a sparser distribution with different clusters of data points. In fact, the four black circles in Fig. 3 correspond to four different hospitals of the RyC domain collection. This means that our model clusters the data not only by domain but also by hospital without knowing that information.

KSSHIBA is able to is able to reduce the dimensionality from an initial MALDI-TOF of 10,000 features to only 76 latent features. This low-dimensional latent space representation justifies the better results of KSSHIBA over all SOTA approaches in the inter-domain scenario, as it is able to detect the differences between the MALDI-TOF depending on their domain learning to generalize better the prediction.

IV. CONCLUSION

In this work, two different bacterial assemblages were presented: one from the same hospital domain without any particular inclusion criteria (GM); and another that groups strains from 18 geographically dispersed hospitals, selected by their phenotypic and genotypic resistance to beta-lactams

(RyC). The latter domain was characterized by whole genome sequencing and includes both frequent and rare clonal lineages. Although the same AST methodology was used in both settings, a different person performed it, thus, we cannot rule out possible discrepancies linked to each facility/person. On the contrary, all the MALDI-TOF spectra were performed by a single worker.

We fitted and validated a Bayesian semi-supervised model to clinical isolates of *K. pneumoniae* for prediction of ESBL and CP production based on both the MALDI-TOF spectra and their origin.

As we could see, the intra-domain experiment (Table II) yielded a first hypothesis where the nonlinear kernels performed better for each separate domain. In contrast, the second experiment disproved this hypothesis when dealing with both domains simultaneously. Table III demonstrates that one cannot take lightly how to deal with different distributions. Therefore, the KSSHIBA LINEAR DOMAIN version shows that exploiting a multi-view heterogeneous model is better when dealing with inter-domain data. The domain information improved the learning process by making KSSHIBA able to properly model the different data distributions by getting rid of the bias introduced by the data itself. This is because, when dealing with different distributions, we can overfit to one of the distributions if it presents unbalanced data. Moreover, this last experiment make it clear that linear kernels area better choice when dealing with out-of-distribution data as they can generalize better.

Therefore, the application of multi-view Bayesian models, such as KSSHIBA, to this type of problem represents a step forward in the SOTA AR prediction. We outpeform, in terms of AUC, to all previously used models in this field, at the same time that: (i) we get rid of parameter cross-validation facilitating the reproducibility of the study; (ii) we deal with the high dimensionality of the MALDI-TOF by kernels; (iii) we provide dimensionality reduction by projecting all views to a common low-dimensional latent space; (iv) we get rid of external preprocessing being now able to work with raw data; (v) we provide interpretability by means of a weight matrix $\mathbf{W}^{(\mathrm{m})}$ associated to each view; and finally, (vii) we provide a multi-view approach where we exploit epidemiological information resolving the overfitting to one distribution. Our contribution is, therefore, a step forward towards the goal of reducing ineffective antibiotic prescribing by being able to predict possible resistance mechanisms in *K. pneumoniae*. Its implementation in microbiological laboratories could improve the detection of multidrug-resistant isolates, optimizing the therapeutic decision and reducing the time to obtain resistance mechanism results.

As a future work, *K. pneumoniae* isolates identified by MALDI-TOF could be automatically classified by KSSHIBA. The results obtained (WT, ESBL-producer or ESBL+CP-producer) would be validated during a period of time using *K. pneumoniae* isolates from different geographic origin. Finally, it could be implemented as a rapid AST method in laboratories equipped with MALDI-TOF.

ACKNOWLEDGMENT

We acknowledge the SUPERIOR Study Group which provided the RyC domain data and includes the following members: Antonio Oliver and Xavier Mulet (Hospital Universitario Son Espases, Palma, Spain); Emilia Cercenado (Hospital General Universitario Gregorio Marañón, Madrid, Spain); Germán Bou and M. Carmen Fernández (Hospital Universitario A Coruña, A Coruña, Spain); Álvaro Pascual and Mercedes Delgado-Valverde (Hospital Universitario Virgen Macarena, Sevilla, Spain); Concepción Gimeno and Nuria Tormo (Consorcio Hospital General Universitario de Valencia, Valencia, Spain); Jorge Calvo, Jesús Rodríguez-Lozano and Ana Ávila Alonso (Hospital Universitario Marqués de Valdecilla, Santander, Spain); Jordi Vila, Francesc Marco and Cristina Pitart (Hospital Clínic, Barcelona, Spain); and María García del Castillo, Sergio García-Fernández, Marta Hernández-García, Marta Tato and Rafael Cantón (Hospital Universitario Ramón y Cajal, Madrid, Spain). This study was sponsored by MSD Spain. The STEP Study Group includes the following members: José Melo-Cristino (Serviço de Microbiologia Centro Hospitalar Lisboa Norte, Lisboa, Portugal); Margarida F. Pinto, Cristina Marcelo, Helena Peres, Isabel Lourenço, Isabel Peres, João Marques, Odete Chantre and Teresa Pina (Laboratório de Microbiologia, Serviço de Patologia Clínica, Centro Hospitalar Universitário Lisboa Central, Lisboa, Portugal); Elsa Gonçalves and Cristina Toscano (Laboratório de Microbiologia Clínica Centro Hospitalar de Lisboa Ocidental, Lisboa, Portugal); Valquíria Alves (Serviço de Microbiologia, Unidade Local de Saúde de Matosinhos, Matosinhos, Portugal); Manuela Ribeiro, Eliana Costa and Ana Raquel Vieira (Serviço Patologia Clínica, Centro Hospitalar Universitário São João, Porto, Portugal); Sónia Ferreira, Raquel Diaz and Elmano Ramalheira (Serviço Patologia Clínica, Hospital Infante Dom Pedro, Aveiro, Portugal); Sandra Schäfer, Luísa Tancredo and Luísa Sancho (Serviço de Patologia Clínica, Fernando Fonseca, Amadora, Portugal); Ana Rodrigues and José Diogo (Serviço de Microbiologia, Hospital Garcia de Orta, Almada, Portugal); Rui Ferreira (Serviço de Patologia Clínica—Microbiologia—CHUA—Unidade de Portimão, Portugal); Helena Ramos, Tânia Silva and Daniela Silva (Serviço de Microbiologia, Centro Hospitalar Universitário do Porto, Porto, Portugal); Catarina Chaves, Carolina Queiroz and Altair Nabiev (Serviço de Microbiologia, Centro Hospitalar Universitário de Coimbra, Coimbra, Portugal); Leonor Pássaro, Laura Paixao, João Romano and Carolina Moura (MSD Portugal, Paço de Arcos, Portugal).

REFERENCES

[1] CDC, "Cdc biggest threats: Carbapenem-resistant Enterobacteriaceae (CRE)," Centers for Disease Control and Prevention, p. 2, 2017.

[2] E. Tacconelli, "GLOBAL PRIORITY LIST OF ANTIBIOTIC-RESISTANT BACTERIA TO GUIDE RESEARCH, DISCOVERY, AND DEVELOPMENT OF NEW ANTIBIOTICS," World Health Organization, p. 7, 2017.

[3] R. Edward, "Carbapenem-resistant Enterobacteriaceae - Second update," European Centers for Disease Control and Prevention, p. 17, 2019.

[4] E. Tacconelli, E. Carrara, A. Savoldi, S. Harbarth, M. Mendelson, D. L. Monnet, C. Pulcini, G. Kahlmeter, J. Kluytmans, Y. Carmeli, M. Ouellette, K. Outterson, J. Patel, M. Cavaleri, E. M. Cox, C. R. Houchens, M. L. Grayson, P. Hansen, N. Singh, U. Theuretzbacher,

N. Magrini, A. O. Aboderin, S. S. Al-Abri, N. Awang Jalil, N. Benzonana, S. Bhattacharya, A. J. Brink, F. R. Burkert, O. Cars, G. Cornaglia, O. J. Dyar, A. W. Friedrich, A. C. Gales, S. Gandra, C. G. Giske, D. A. Goff, H. Goossens, T. Gottlieb, M. Guzman Blanco, W. Hryniewicz, D. Kattula, T. Jinks, S. S. Kanj, L. Kerr, M.-P. Kieny, Y. S. Kim, R. S. Kozlov, J. Labarca, R. Laxminarayan, K. Leder, L. Leibovici, G. Levy-Hara, J. Littman, S. Malhotra-Kumar, V. Manchanda, L. Moja, B. Ndoye, A. Pan, D. L. Paterson, M. Paul, H. Qiu, P. Ramon-Pardo, J. Rodríguez-Baño, M. Sanguinetti, S. Sengupta, M. Sharland, M. Si-Mehand, L. L. Silver, W. Song, M. Steinbakk, J. Thomsen, G. E. Thwaites, J. W. van der Meer, N. Van Kinh, S. Vega, M. V. Villegas, A. Wechsler-Fördös, H. F. L. Wertheim, E. Wesangula, N. Woodford, F. O. Yilmaz, and A. Zorzet, "Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis," The Lancet Infectious Diseases, vol. 18, no. 3, pp. 318–327, Mar. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1473309917307533

[5] G. Daikos and A. Markogiannakis, "Carbapenemase-producing klebsiella pneumoniae:(when) might we still consider treating with carbapenems?" Clinical Microbiology and Infection, vol. 17, no. 8, pp. 1135–1141, 2011.

[6] M. Oviaño and G. Bou, "Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry for the Rapid Detection of Antimicrobial Resistance Mechanisms and Beyond," Clinical Microbiology Reviews, vol. 32, no. 1, pp. e00 037–18, Nov. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6302359/

[7] C. Lange, S. Schubert, J. Jung, M. Kostrzewa, and K. Sparbier, "Quantitative Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry for Rapid Resistance Detection," Journal of Clinical Microbiology, vol. 52, no. 12, pp. 4155–4162, Dec. 2014, publisher: American Society for Microbiology. [Online]. Available: https://journals.asm.org/doi/10.1128/JCM.01872-14

[8] G. A. Satten, S. Datta, H. Moura, A. R. Woolfitt, M. d. G. Carvalho, G. M. Carlone, B. K. De, A. Pavlopoulos, and J. R. Barr, "Standardization and denoising algorithms for mass spectra to classify whole-organism bacterial specimens," Bioinformatics, vol. 20, no. 17, pp. 3128–3136, 2004.

[9] L. Deng, Y. Zhong, M. Wang, X. Zheng, and J. Zhang, "Scale-adaptive deep model for bacterial raman spectra identification," IEEE Journal of Biomedical and Health Informatics, 2021.

[10] T. Zhang, J. Ding, X. Rao, J. Yu, M. Chu, W. Ren, L. Wang, and W. Xue, "Analysis of methicillin-resistant staphylococcus aureus major clonal lineages by matrix-assisted laser desorption ionization–time of flight mass spectrometry (maldi–tof ms)," Journal of microbiological methods, vol. 117, pp. 122–127, 2015.

[11] N. Esener, M. J. Green, R. D. Emes, B. Jowett, P. L. Davies, A. J. Bradley, and T. Dottorini, "Discrimination of contagious and environmental strains of streptococcus uberis in dairy herds by means of mass spectrometry and machine-learning," Scientific reports, vol. 8, no. 1, pp. 1–12, 2018.

[12] H.-Y. Wang, W.-C. Li, K.-Y. Huang, C.-R. Chung, J.-T. Horng, J.-F. Hsu, J.-J. Lu, and T.-Y. Lee, "Rapid classification of group b streptococcus serotypes based on matrix-assisted laser desorption ionization-time of flight mass spectrometry and machine learning techniques," BMC bioinformatics, vol. 20, no. 19, pp. 1–17, 2019.

[13] C.-R. Chung, H.-Y. Wang, F. Lien, Y.-J. Tseng, C.-H. Chen, T.-Y. Lee, T.-P. Liu, J.-T. Horng, and J.-J. Lu, "Incorporating statistical test and machine intelligence into strain typing of staphylococcus haemolyticus based on matrix-assisted laser desorption ionization-time of flight mass spectrometry," Frontiers in microbiology, vol. 10, p. 2120, 2019.

[14] S. Gibb and K. Strimmer, "Maldiquant: a versatile r package for the analysis of mass spectrometry data," Bioinformatics, vol. 28, no. 17, pp. 2270–2271, 2012.

[15] K. Sogawa, M. Watanabe, T. IshigE, S. Segawa, A. Miyabe, S. Murata, T. Saito, A. Sanda, K. Furuhata, and F. Nomura, "Rapid discrimination between methicillin-sensitive and methicillin-resistant staphylococcus aureus using maldi-tof mass spectrometry," Biocontrol science, vol. 22, no. 3, pp. 163–169, 2017.

[16] C. A. Mather, B. J. Werth, S. Sivagnanam, D. J. SenGupta, and S. M. Butler-Wu, "Rapid detection of vancomycin-intermediate staphylococcus aureus by matrix-assisted laser desorption ionization–time of flight mass spectrometry," Journal of clinical microbiology, vol. 54, no. 4, pp. 883–890, 2016.

[17] K. Asakura, T. Azechi, H. Sasano, H. Matsui, H. Hanaki, M. Miyazaki, T. Takata, M. Sekine, T. Takaku, T. Ochiai et al., "Rapid and easy detection of low-level resistance to vancomycin in methicillin-resistant staphylococcus aureus by matrix-assisted laser desorption ionization time-of-flight mass spectrometry," PloS one, vol. 13, no. 3, p. e0194212, 2018.

[18] R. Ketterlinus, S.-Y. Hsieh, S.-H. Teng, H. Lee, and W. Pusch, "Fishing for biomarkers: analyzing mass spectrometry data with the new clinprotools™ software," Biotechniques, vol. 38, no. S6, pp. S37–S40, 2005.

[19] P.-L. Ho, C.-Y. Yau, L.-Y. Ho, J. H. Chen, E. L. Lai, S. W. Lo, W. Cindy, and K.-H. Chow, "Rapid detection of cfia metallo-β-lactamase-producing bacteroides fragilis by the combination of maldi-tof ms and carbanp," Journal of Clinical Pathology, vol. 70, no. 10, pp. 868–873, 2017.

[20] T.-S. Huang, S. S.-J. Lee, C.-C. Lee, and F.-C. Chang, "Detection of carbapenem-resistant Klebsiella pneumoniae on the basis of matrix-assisted laser desorption ionization time-of-flight mass spectrometry by using supervised machine learning approach," PLoS ONE, vol. 15, no. 2, p. e0228459, Feb. 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7004327/

[21] J. H. K. Chen, K. K. K. She, O.-Y. Wong, J. L. L. Teng, W.-C. Yam, S. K. P. Lau, P. C. Y. Woo, V. C. C. Cheng, and K.-Y. Yuen, "Use of MALDI Biotyper plus ClinProTools mass spectra analysis for correct identification of Streptococcus pneumoniae and Streptococcus mitis/oralis," Journal of Clinical Pathology, vol. 68, no. 8, pp. 652–656, Aug. 2015, publisher: BMJ Publishing Group Section: Original article. [Online]. Available: https://jcp.bmj.com/content/68/8/652

[22] J. Bai, Z. Fan, L. Zhang, X. Xu, and Z. Zhang, "Classification of methicillin-resistant and methicillin-susceptible staphylococcus aureus using an improved genetic algorithm for feature selection based on mass spectra," in Proceedings of the 9th International Conference on Bioinformatics and Biomedical Technology, 2017, pp. 57–63.

[23] M. Delavy, L. Cerutti, A. Croxatto, G. Prod'hom, D. Sanglard, G. Greub, and A. T. Coste, "Machine learning approach for candida albicans fluconazole resistance detection using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry," Frontiers in microbiology, vol. 10, p. 3000, 2020.

[24] W. Tang, N. Ranganathan, V. Shahrezaei, and G. Larrouy-Maumus, "Maldi-tof mass spectrometry on intact bacteria combined with a refined analysis framework allows accurate classification of mssa and mrsa," PloS one, vol. 14, no. 6, p. e0218951, 2019.

[25] S. Gibb and K. Strimmer, "Differential protein expression and peak selection in mass spectrometry data by binary discriminant analysis," Bioinformatics, vol. 31, no. 19, pp. 3156–3162, 2015.

[26] C. Weis, M. Horn, B. Rieck, A. Cuénod, A. Egli, and K. Borgwardt, "Topological and kernel-based microbial phenotype prediction from maldi-tof mass spectra," Bioinformatics, vol. 36, no. Supplement_1, pp. i30–i38, 2020.

[27] C. Sevilla-Salcedo, A. Guerrero-López, P. M. Olmos, and V. Gómez-Verdejo, "Bayesian Sparse Factor Analysis with Kernelized Observations," arXiv:2006.00968 [cs, stat], Jan. 2021, arXiv: 2006.00968. [Online]. Available: http://arxiv.org/abs/2006.00968

[28] C. Sevilla-Salcedo, V. G'omez-Verdejo, and P. Olmos, "Sparse Semi-supervised Heterogeneous Interbattery Bayesian Analysis," Pattern Recognit., 2021.

[29] M. Hernández-García, S. García-Fernández, M. García-Castillo, G. Bou, E. Cercenado, M. Delgado-Valverde, X. Mulet, C. Pitart, J. Rodríguez-Lozano, N. Tormo, D. López-Mendoza, J. Díaz-Regañón, and R. Cantón, "WGS characterization of MDR Enterobacterales with different ceftolozane/tazobactam susceptibility profiles during the SUPERIOR surveillance study in Spain," JAC-Antimicrobial Resistance, vol. 2, no. 4, p. dlaa084, Oct. 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8210196/

[30] M. Hernández-García, S. García-Fernández, M. García-Castillo, J. Melo-Cristino, M. F. Pinto, E. Gonçalves, V. Alves, E. Costa, E. Ramalheira, L. Sancho, J. Diogo, R. Ferreira, T. Silva, C. Chaves, L. Pássaro, L. Paixão, J. Romano, and R. Cantón, "Confronting Ceftolozane-Tazobactam Susceptibility in Multidrug-Resistant Enterobacterales Isolates and Whole-Genome Sequencing Results (STEP Study)," International Journal of Antimicrobial Agents, vol. 57, no. 2, p. 106259, Feb. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924857920304799

[31] B. Rodríguez-Sánchez, M. Marín, C. Sánchez-Carrillo, E. Cercenado, A. Ruiz, M. Rodríguez-Créixems, and E. Bouza, "Improvement of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry identification of difficult-to-identify bacteria and its impact in the workflow of a clinical microbiology laboratory," Diagnostic Microbiology and Infectious Disease, vol. 79, no. 1, pp. 1–6, May 2014.

[32] C. Rodrigues, V. Passet, A. Rakotondrasoa, and S. Brisse, "Identification of Klebsiella pneumoniae, Klebsiella quasipneumoniae, Klebsiella variicola and Related Phylogroups by MALDI-TOF Mass Spectrometry," Frontiers in Microbiology, vol. 9, p. 3000, 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fmicb.2018.03000

[33] M. E. Zvezdanova, M. J. Arroyo, G. Méndez, A. Candela, L. Mancera, J. G. Rodríguez, J. L. Serra, R. Jiménez, I. Lozano, C. Castro, C. López, P. Muñoz, J. Guinea, P. Escribano, B. Rodríguez-Sánchez, W. Sánchez-Yebra, J. Sánchez-Gómez, I. Lozano, E. Marfil, M. Muñoz de la Rosa, R. T. García, F. Cobo, C. Castro, C. López, A. Rezusta, T. Peláez, C. Castelló-Abietar, I. Costales, J. L. Serra, R. Jiménez, C. L. Echeverría, C. L. Pérez, G. Megías-Lobón, B. Lorenzo, F. Sánchez-Reus, J. Ayats, M. T. Martín, I. Vidal, V. Sánchez-Hellín, E. Ibáñez, J. Pemán, M. Fajardo, C. Pazos, M. Rodríguez-Mayo, A. Pérez-Ayala, E. Gómez, J. Guinea, P. Escribano, J. Serrano, E. Reigadas, B. Rodríguez, E. Zvezdanova, J. Díaz-García, A. Gómez-Núñez, J. G. Leiva, M. Machado, P. Muñoz, I. Sánchez-Romero, J. García-Rodríguez, J. Luis del Pozo, M. R. Vallejo, C. Ruiz de Alegría-Puig, L. López-Soria, J. M. Marimón, D. Vicente, M. Fernández-Torres, and S. Hernáez-Crespo, "Detection of azole resistance in Aspergillus fumigatus complex isolates using MALDI-TOF mass spectrometry," Clinical Microbiology and Infection, Jun. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1198743X21003220

[34] R. M. Neal, Bayesian learning for neural networks. Springer Science & Business Media, 2012, vol. 118.

[35] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," Journal of the American statistical Association, vol. 112, no. 518, pp. 859–877, 2017.