Clinical Microbiology with Multi-view deep probabilistic models

Alejandro Jorge Guerrero López

A dissertation submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in
Multimedia and Communications

Dept. de Teoría de la Señal y Comunicaciones
Universidad Carlos III de Madrid, Spain

Directora:
Vanessa Gómez Verdejo

Director:
Pablo Martínez Olmos

March 2023

ACKNOWLEDGEMENTS

QUIÉN iba a decir que yo escribiría una tesis. Seguro que mi familia no, por eso son los primeros a los que quiero agradecer esto. Primero a mis padres, que sin ellos no estaría aquí en todos los sentidos de la palabra. Quién os iba a decir a vosotros que no solo iba a terminar la ESO, sino que acabaría siendo Doctor en algo. A mi madre, por sacar siempre una sonrisa hasta en los momentos más difíciles y hacer que, para mí, la vida fuera siempre más sencilla. Por mantener la familia unida, por quererme siempre. A mi padre, por ser la persona más trabajadora que conozco y siempre encontrar una forma de llegar a final de mes contra todo pronóstico. Por quererme como nadie. Cuántas veces dijiste que cómo iba a hacer bachillerato si yo tenía que ir a la obra contigo y yo te contesté que quería estudiar para no tener que trabajar nunca. Lo he conseguido, papá. A mis hermanos que, si no fuera porque vosotros cuidáis de ellos, yo no hubiera tenido esta oportunidad. I, finalment, a la meva segona família, Francis, Rafel i Maria Antònia, que durant tots aquests anys m'han acollit com a un més dels seus.

A Sergi, que desde los 3 años siempre ha estado escuchándome, aunque no sepa qué demonios hago en Madrid ni en qué me estoy doctorando. Gracias por recordarme siempre quién soy y de dónde vengo.

A Mari y Ramiro, compañeros de máster, doctorado y, sobre todo, amigos. Ramiro, gracias por ser el guía en este camino. El único que tenía clarísimo que quería ser docente e investigador. Gracias a tus consejos y charlas en el Yacón supe que no estaba loco. Mari, gracias por tu tiempo, espacio y cariño. Por compartir tardes, fiestas, paseos, picnics, clases, cubículo, jefes, asignaturas, proyectos, en definitiva, por ser mi pack en esta travesía.

A los compañeros de comida que son penitentes de este andar, gracias por vuestro tiempo, quejas y felicidad a partes iguales: Emese, Josemi, David, Dani, Peis, Jose Carlos, Jesús y Lorena.

A Pablo y Vanessa que, tras recibir tropecientos mails pasivo-agresivos de un alumno pidiendo por llegar a final de mes, dijeron "vamos a darle una oportunidad a este chico". Por ser unos directores de tesis ejemplares. A Pablo, por estar siempre allí cuando nos quedábamos atascados en cualquiera de los proyectos que llevamos. Porque, aunque yo estuviera perdido y a veces me preguntara a mí mismo "para qué me habré metido yo en esto" aparecía él, con una risa que se escucha y se contagia en todo el departamento, diciéndome que esto es facilísimo, que en dos días lo tengo. Y resultaba ser verdad. A Vanessa que en momentos de crisis existencial te envía un mensaje "baja 5 min" y, después de 1h y media de charlar, por fin

llegamos al tema que queríamos tratar y lo soluciona todo. Por revisarlo todo con lupa que, aunque proteste (porque me encanta protestar), agradezco que quieras dedicar tu tiempo en mí. Gracias a los dos por enseñarme que se puede ser buen profesor e investigador si pasas gusto por lo que haces.

A Belén y David, por acogerme en el hospital, por enseñarme desde cero lo que es la biología sin ser yo nada de eso. Por tener paciencia cuando no entiendo vuestras conversaciones médicas. Gracias por las reuniones-desayuno para ponerme al día de todos los movimientos del hospital. Gracias por vuestra ilusión en la IA aplicada que me enseñó que lo que hago es útil e importante.

Y por supuesto, a la persona que me lleva acompañando desde que tengo 18 años y quiso compartirlo todo conmigo. A la persona que me soportó en mis peores y mejores momentos. La que no solo no le bastó con tenerme de compañero de clase, sino que me quiso de compañero de biblioteca, de noches de estudio, de findes programando, de citas de vino y queso, de piso, de tiempo, de viajes, de sueños, en resumen, gracias por quererme de compañero de vida. Que si estoy aquí es gracias a ti. Que esta tesis es tan tuya como mía. Que te quiero.

PUBLISHED AND PRESENTED CONTENTS

The following list of works is a brief bibliography of journal articles and preprints that have arisen from this thesis. In each contribution, its presence is indicated as either complete or partial in this thesis. Finally, I declare that the material from this source included in the thesis is not marked by typographical means or references.

JOURNALS.

1. Candela, A., Guerrero-López, A., Mateos, M., Gómez-Asenjo, A., Arroyo, M.J., Hernandez-García, M., del Campo, R., Cercenado, E., Cuénod, A., Méndez, G. and Mancera, L., 2023. Automatic discrimination of species within the Enterobacter cloacae complex using MALDI-TOF Mass Spectrometry and supervised algorithms. Accepted for publication in *Journal of Clinical Microbiology*; [pdf] - Partly included in Chapter 1. Its presence is indicated as a footnote in Chapter 1.

2. Sevilla-Salcedo, C., Guerrero-López, A., Olmos, P. M., & Gómez-Verdejo, V. (2022). Bayesian sparse factor analysis with kernelized observations. *Neurocomputing, 490, 66-78.* [pdf] – Partly included in Chapter 3. Its presence is indicated in the introduction of Chapter 3.

3. Guerrero-López, A., Sevilla-Salcedo, C., Gómez-Verdejo, V., & Olmos, P. M. (2022). Multimodal hierarchical Variational AutoEncoders with Factor Analysis latent space. *arXiv preprint arXiv:2207.09185 (first review round in Information Sciences).* [pdf] – Completely included in Chapter 4. Its presence is indicated in the introduction of Chapter 4.

4. Guerrero-López, A., Candela, A., Sevilla-Salcedo, C., Hernández-García, M., Martinez-Olmos, P., Canton, R., Muñoz, P., Gómez-Verdejo, V., del Campo, R., & Rodriguez-Sanchez, B. (2023). Automatic antibiotic resistance prediction in Klebsiella pneumoniae based on MALDI-TOF mass spectra. *Engineering Applications of Artificial Intelligence, 118, 105644.* [pdf] – Completely included in Chapter 5. Its presence is indicated in the introduction of Chapter 5.

CONGRESSES.

1. Blázquez-Sánchez, M., Guerrero-López, A., Candela, A., Jiménez-Rosillo, L., Rodríguez-Temporal, D., Arroyo, M.J., Martín, A., Jiménez-Navarro, L., Méndez, G., Mancera, L., Muñoz, P., Alcalá, L., Rodríguez-Sánchez, B. (2023). Desarrollo de modelos de clasificación automática basados en espectrometría de masas MALDI-TOF y Machine Learning para el tipado rápido de Clostridioides difficile. Oral comunication. *XXVI Congreso de la Sociedad Española de Enfermedades Infecciosas y Microbiología Clínica* – Completely included in Chapter 6. Its presence is indicated in the introduction of Chapter 6.

ABSTRACT

CLINICAL microbiology is one of the critical topics of this century. Identification and discrimination of microorganisms is considered a global public health threat by the main international health organisations, such as World Health Organisation (WHO) or the European Centre for Disease Prevention and Control (ECDC). Rapid spread, high morbidity and mortality, as well as the economic burden associated with their treatment and control are the main causes of their impact. Discrimination of microorganisms is crucial for clinical applications, for instance, *Clostridium difficile* (*C. diff*) increases the mortality and morbidity of healthcare-related infections. Furthermore, in the past two decades, other bacteria, including *Klebsiella pneumoniae* (*K. pneumonia)*, have demonstrated a significant propensity to acquire antibiotic resistance mechanisms. Consequently, the use of an ineffective antibiotic may result in mortality. Machine Learning (ML) has the potential to be applied in the clinical microbiology field to automatise current methodologies and provide more efficient guided personalised treatments.

However, microbiological data are challenging to exploit owing to the presence of a heterogeneous mix of data types, such as real-valued high-dimensional data, categorical indicators, multilabel epidemiological data, binary targets, or even time-series data representations. This problem, which in the field of ML is known as multi-view or multi-modal representation learning, has been studied in other application fields such as mental health monitoring or haematology. Multi-view learning combines different modalities or views representing the same data to extract richer insights and improve understanding. Each modality or view corresponds to a distinct encoding mechanism for the data, and this dissertation specifically addresses the issue of heterogeneity across multiple views.

In the probabilistic ML field, the exploitation of multi-view learning is also known as Bayesian Factor Analysis (FA). Current solutions face limitations when handling high-dimensional data and non-linear associations. Recent research proposes deep probabilistic methods to learn hierarchical representations of the data, which can capture intricate non-linear relationships between features. However, some Deep Learning (DL) techniques rely on complicated representations, which can hinder the interpretation of the outcomes. In addition, some inference methods used in DL approaches can be computationally burdensome, which can hinder their practical application in real-world situations. Therefore, there is a demand for more interpretable, explainable, and computationally efficient techniques for high-dimensional data. By combining multiple views representing the same information,

such as genomic, proteomic, and epidemiologic data, multi-modal representation learning could provide a better understanding of the microbial world. Hence, in this dissertation, the development of two deep probabilistic models, that can handle current limitations in state-of-the-art of clinical microbiology, are proposed. Moreover, both models are also tested in two real scenarios regarding antibiotic resistance prediction in *K. pneumoniae* and automatic ribotyping of *C. diff* in collaboration with the Instituto de Investigación Sanitaria Gregorio Marañón (IISGM) and the Instituto Ramón y Cajal de Investigación Sanitaria (IRyCIS).

The first presented algorithm is the Kernelised Sparse Semi-supervised Heterogeneous Interbattery Bayesian Analysis (SSHIBA). This algorithm uses a kernelised formulation to handle non-linear data relationships while providing compact representations through the automatic selection of relevant vectors. Additionally, it uses an Automatic Relevance Determination (ARD) over the kernel to determine the input feature relevance functionality. Then, it is tailored and applied to the microbiological laboratories of the IISGM and IRyCIS to predict antibiotic resistance in *K. pneumoniae*. To do so, specific kernels that handle Matrix-Assisted Laser Desorption Ionization (MALDI)-Time-Of-Flight (TOF) mass spectrometry of bacteria are used. Moreover, by exploiting the multi-modal learning between the spectra and epidemiological information, it outperforms other state-of-the-art algorithms. Presented results demonstrate the importance of heterogeneous models that can analyse epidemiological information and can automatically be adjusted for different data distributions. The implementation of this method in microbiological laboratories could significantly reduce the time required to obtain resistance results in 24-72 hours and, moreover, improve patient outcomes.

The second algorithm is a hierarchical Variational AutoEncoder (VAE) for heterogeneous data using an explainable FA latent space, called FA-VAE. The FA-VAE model is built on the foundation of the successful KSSHIBA approach for dealing with semi-supervised heterogeneous multi-view problems. This approach further expands the range of data domains it can handle. With the ability to work with a wide range of data types, including multilabel, continuous, binary, categorical, and even image data, the FA-VAE model offers a versatile and powerful solution for real-world data sets, depending on the VAE architecture. Additionally, this model is adapted and used in the microbiological laboratory of IISGM, resulting in an innovative technique for automatic ribotyping of *C. diff*, using MALDI-TOF data. To the best of our knowledge, this is the first demonstration of using any kind of ML for *C. diff* ribotyping. Experiments have been conducted on strains of Hospital General Universitario Gregorio Marañón (HGUGM) to evaluate the viability of the proposed approach. The results have demonstrated high accuracy rates where KSSHIBA even achieved perfect accuracy in the first data collection. These models have also been tested in a real-life outbreak scenario at the HGUGM, where successful classification of all outbreak samples has been achieved by FA-

VAE. The presented results have not only shown high accuracy in predicting each strain's ribotype but also revealed an explainable latent space. Furthermore, traditional ribotyping methods, which rely on PCR, required 7 days while FA-VAE has predicted equal results on the same day. This improvement has significantly reduced the time response by helping in the decision-making of isolating patients with hyper-virulent ribotypes of *C. diff* on the same day of infection. The promising results, obtained in a real outbreak, have provided a solid foundation for further advancements in the field. This study has been a crucial stepping stone towards realising the full potential of MALDI-TOF for bacterial ribotyping and advancing our ability to tackle bacterial outbreaks.

In conclusion, this doctoral thesis has significantly contributed to the field of Bayesian FA by addressing its drawbacks in handling various data types through the creation of novel models, namely KSSHIBA and FA-VAE. Additionally, a comprehensive analysis of the limitations of automating laboratory procedures in the microbiology field has been carried out. The shown effectiveness of the newly developed models has been demonstrated through their successful implementation in critical problems, such as predicting antibiotic resistance and automating ribotyping. As a result, KSSHIBA and FA-VAE, both in terms of their technical and practical contributions, signify noteworthy progress both in the clinical and the Bayesian statistics fields. This dissertation opens up possibilities for future advancements in automating microbiological laboratories.

RESUMEN

La microbiología clínica es uno de los temas críticos de este siglo. La identificación y discriminación de microorganismos se considera una amenaza mundial para la salud pública por parte de las principales organizaciones internacionales de salud, como la Organización Mundial de la Salud (OMS) o el Centro Europeo para la Prevención y Control de Enfermedades (ECDC). La rápida propagación, alta morbilidad y mortalidad, así como la carga económica asociada con su tratamiento y control, son las principales causas de su impacto. La discriminación de microorganismos es crucial para aplicaciones clínicas, como el caso de *Clostridium difficile* (*C. diff*), el cual aumenta la mortalidad y morbilidad de las infecciones relacionadas con la atención médica. Además, en las últimas dos décadas, otros tipos de bacterias, incluyendo *Klebsiella pneumoniae* (*K. pneumonia*), han demostrado una propensión significativa a adquirir mecanismos de resistencia a los antibióticos. En consecuencia, el uso de un antibiótico ineficaz puede resultar en un aumento de la mortalidad. El aprendizaje automático (ML) tiene el potencial de ser aplicado en el campo de la microbiología clínica para automatizar las metodologías actuales y proporcionar tratamientos personalizados más eficientes y guiados.

Sin embargo, los datos microbiológicos son difíciles de explotar debido a la presencia de una mezcla heterogénea de tipos de datos, tales como datos reales de alta dimensionalidad, indicadores categóricos, datos epidemiológicos multietiqueta, objetivos binarios o incluso series temporales. Este problema, conocido en el campo del aprendizaje automático (ML) como aprendizaje multimodal o multivista, ha sido estudiado en otras áreas de aplicación, como en el monitoreo de la salud mental o la hematología. El aprendizaje multivista combina diferentes modalidades o vistas que representan los mismos datos para extraer conocimientos más ricos y mejorar la comprensión. Cada vista corresponde a un mecanismo de codificación distinto para los datos, y esta tesis aborda particularmente el problema de la heterogeneidad multivista.

En el campo del aprendizaje automático probabilístico, la explotación del aprendizaje multivista también se conoce como Análisis de Factores (FA) Bayesianos. Las soluciones actuales enfrentan limitaciones al manejar datos de alta dimensionalidad y correlaciones no lineales. Investigaciones recientes proponen métodos probabilísticos profundos para aprender representaciones jerárquicas de los datos, que pueden capturar relaciones no lineales intrincadas entre características. Sin embargo, algunas técnicas de aprendizaje profundo (DL) se basan en representaciones complejas, dificultando así la interpretación de los resultados. Además,

algunos métodos de inferencia utilizados en DL pueden ser computacionalmente costosos, obstaculizando su aplicación práctica. Por lo tanto, existe una demanda de técnicas más interpretables, explicables y computacionalmente eficientes para datos de alta dimensionalidad. Al combinar múltiples vistas que representan la misma información, como datos genómicos, proteómicos y epidemiológicos, el aprendizaje multimodal podría proporcionar una mejor comprensión del mundo microbiano. Dicho lo cual, en esta tesis se proponen el desarrollo de dos modelos probabilísticos profundos que pueden manejar las limitaciones actuales en el estado del arte de la microbiología clínica. Además, ambos modelos también se someten a prueba en dos escenarios reales relacionados con la predicción de resistencia a los antibióticos en *K. pneumoniae* y el ribotipado automático de *C. diff* en colaboración con el IISGM y el IRyCIS.

El primer algoritmo presentado es Kernelised Sparse Semi-supervised Heterogeneous Interbattery Bayesian Analysis (SSHIBA). Este algoritmo utiliza una formulación kernelizada para manejar correlaciones no lineales proporcionando representaciones compactas a través de la selección automática de vectores relevantes. Además, utiliza un Automatic Relevance Determination (ARD) sobre el kernel para determinar la relevancia de las características de entrada. Luego, se adapta y aplica a los laboratorios microbiológicos del IISGM y IRyCIS para predecir la resistencia a antibióticos en *K. pneumoniae*. Para ello, se utilizan kernels específicos que manejan la espectrometría de masas Matrix-Assisted Laser Desorption Ionization (MALDI)-Time-Of-Flight (TOF) de bacterias. Además, al aprovechar el aprendizaje multimodal entre los espectros y la información epidemiológica, supera a otros algoritmos de última generación. Los resultados presentados demuestran la importancia de los modelos heterogéneos ya que pueden analizar la información epidemiológica y ajustarse automáticamente para diferentes distribuciones de datos. La implementación de este método en laboratorios microbiológicos podría reducir significativamente el tiempo requerido para obtener resultados de resistencia en 24-72 horas y, además, mejorar los resultados para los pacientes.

El segundo algoritmo es un modelo jerárquico de Variational AutoEncoder (VAE) para datos heterogéneos que utiliza un espacio latente con un FA explicativo, llamado FA-VAE. El modelo FA-VAE se construye sobre la base del enfoque de KSSHIBA para tratar problemas semi-supervisados multivista. Esta propuesta amplía aún más el rango de dominios que puede manejar incluyendo multietiqueta, continuos, binarios, categóricos e incluso imágenes. De esta forma, el modelo FA-VAE ofrece una solución versátil y potente para conjuntos de datos realistas, dependiendo de la arquitectura del VAE. Además, este modelo es adaptado y utilizado en el laboratorio microbiológico del IISGM, lo que resulta en una técnica innovadora para el ribotipado automático de *C. diff* utilizando datos MALDI-TOF. Hasta donde sabemos, esta es la primera demostración del uso de cualquier tipo de ML para el ribotipado de *C. diff*. Se han realizado experimentos en cepas

del Hospital General Universitario Gregorio Marañón (HGUGM) para evaluar la viabilidad de la técnica propuesta. Los resultados han demostrado altas tasas de precisión donde KSSHIBA incluso logró una clasificación perfecta en la primera colección de datos. Estos modelos también se han probado en un brote real en el HGUGM, donde FA-VAE logró clasificar con éxito todas las muestras del mismo. Los resultados presentados no solo han demostrado una alta precisión en la predicción del ribotipo de cada cepa, sino que también han revelado un espacio latente explicativo. Además, los métodos tradicionales de ribotipado, que dependen de PCR, requieren 7 días para obtener resultados mientras que FA-VAE ha predicho resultados correctos el mismo día del brote. Esta mejora ha reducido significativamente el tiempo de respuesta ayudando así en la toma de decisiones para aislar a los pacientes con ribotipos hipervirulentos de *C. diff* el mismo día de la infección. Los resultados prometedores, obtenidos en un brote real, han sentado las bases para nuevos avances en el campo. Este estudio ha sido un paso crucial hacia el despliegue del pleno potencial de MALDI-TOF para el ribotipado bacteriana avanzado así nuestra capacidad para abordar brotes bacterianos.

En conclusión, esta tesis doctoral ha contribuido significativamente al campo del FA Bayesiano al abordar sus limitaciones en el manejo de tipos de datos heterogéneos a través de la creación de modelos noveles, concretamente, KSSHIBA y FA-VAE. Además, se ha llevado a cabo un análisis exhaustivo de las limitaciones de la automatización de procedimientos de laboratorio en el campo de la microbiología. La efectividad de los nuevos modelos, en este campo, se ha demostrado a través de su implementación exitosa en problemas críticos, como la predicción de resistencia a los antibióticos y la automatización del ribotipado. Como resultado, KSSHIBA y FA-VAE, tanto en términos de sus contribuciones técnicas como prácticas, representan un progreso notable tanto en los campos clínicos como en la estadística Bayesiana. Esta disertación abre posibilidades para futuros avances en la automatización de laboratorios microbiológicos.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Machine Learning (ML) and Deep Learning (DL) are widely used to solve real-world problems in a variety of applications. For example, DL models can be used to perform weather forecasting using temporal models such as Long Short-Term Memory (LSTM) networks [12]. In finance, ML approaches have been proposed for dealing with high-dimensional data, where the number of features often exceeds the number of available samples. For instance, a Random Forest (RF) regressor is used to predict monetary policies and macroeconomic risks in Chinese financial data [13]. In recent years, several major research organisations have made significant contributions to the DL community. DeepMind's AlphaFold [14] is a DL model that accurately predicts 3D protein structures, while researchers at OpenAI have developed transformer-based DL models such as Codex [15], which is based on the Generative Pretrained Transformer (GPT) model [16] and trained on all open source code on GitHub, providing a tool to help developers write code more easily. Another well-known GPT-based model is ChatGPT, which has been trained using reinforcement learning from human feedback and has become a chatbot that can easily handle a wide range of prompts, including literature generation [17], medical exam questions [18], and treatment options [19].

Both ML and DL, are scientific disciplines focused on analysing data by building models that can understand their distribution and predict their future behaviour. DL involves the use of artificial neural networks, which are algorithms designed to recognise patterns and relationships in data. DL algorithms consist of multiple layers of artificial neurones, or nodes, that process and transform input data through a series of computations. Each layer processes the input data in a different way, and the output of one layer becomes the input for the next layer. This hierarchical structure allows DL algorithms to learn and extract features from data in a more powerful manner and flexible than traditional ML algorithms.

Real-world data are often heterogeneous and contain a mix of data types, such as real-valued, categorical, multilabel, binary, or time-series data, which can make it challenging to exploit. These types of data can be found in various domains, such as finance [20], weather [21], and health [22]. In the 1990s, it was proposed that more information could be extracted to better understand the world by combining multiple modalities, or views, representing the same information [23]. Thus, in humans, using multiple sensors makes it easier to understand the world, such

as combining the ability to read lips with the ability to hear voices to improve communication [24]. In the ML field, this technique is also known as multi-view or multi-modal representation learning [25]. The terms *modality* or *view* refer to a particular encoding mechanism of the same information. For example, to describe a car, we could have an image of the car (a 2D real-valued matrix), technical specifications such as the engine power (a real-valued scalar), colour (a categorical value), price (a positive real-valued scalar), and year (an ordinal value). Each of these is a different view of the same data point, hence a multi-modal heterogeneous approach to describe the same object. This dissertation is focused on this concept of multi-modal heterogeneity.

In the medical field, multi-modal learning is a key component to correctly generalising knowledge. For instance, in mental health monitoring, learning from multiple input sources has helped stress detection; as in [26], the authors combine data from video cameras, accelerometers and pressure-sensitive touchscreens to detect stress in patients; whereas other authors [27] propose mixing data from electrodermal activity, photoplethysmograph, microphone and accelerometers to detect it. Other works [28] combine written information from ecological momentary assessment and electronic health records to predict suicidal ideation. In addition, in neuroimaging, multi-modal learning has also helped the state-of-the-art to improve learning. In particular, for brain tumour segmentation, learning from Magnetic Resonance Images (MRI) and combining them with first-order statistics, shape features, and texture features improves survival prediction [29]. Other studies, such as [30], fuse different types of MRI, such as T1, Fractional Anisotropy, and Positron Emission Tomography (PET) images to predict standard dose PET (S-PET) which allows for reduced radiation risk for patients. Furthermore, multi-modal information fusion has also helped classify patients according to disease risk. In [31] they propose to combine a large number of different sources of information, such as Computed Tomography scans, clinical laboratory measurements, genetic data, metabolome data, magnetic resonances, and microbiomes of the same patient. In conclusion, to effectively address the problems today, specifically in the clinical field, models capable of correctly combining heterogeneous and multiple data types are needed.

1.1 Clinical microbiology

Clinical microbiology is one of the urgent topics of the present century [32, 33, 34, 35]. The identification and discrimination of microorganisms are considered global public health threats by the main international health organisations, such as World Health Organisation (WHO) [36] or the European Centre for Disease Prevention and Control (ECDC) [37], due to their rapid spread, high morbidity and mortality, as well as the economic burden associated with their treatment and control [38].

Discrimination of microorganisms is crucial for clinical applications, e.g., the *Cryptosporidium parvum* protozoan parasite contaminates water and produces diarrhoea in animals and humans; therefore, rapid detection of these types of pathogens is critical [39]. In the food industry, spoilage microorganisms such as *Bacillus subtilis* or *Escherichia coli* are a great concern [40], and their rapid detection and discrimination are mandatory. Another bacterium, such as *Clostridium difficile* contributes to the mortality and morbidity of healthcare-related infections in the United States [41, 42]. Furthermore, in the past two decades, another bacterium such as *K. pneumoniae* has shown a great ability to acquire antibiotic-resistant mechanisms, mainly beta-lactamases and carbapenemases [43, 38, 44] which implies that the administration of an inefficient antibiotic could lead to the death of the patient.

1.1.1 The clinical microbiology workflow

Regarding the *K. pneumoniae* example commented in the above section, Fig. 1.1 shows the workflow in a real hospital situation when a patient presents a possible infection. Given a patient with potential pneumonia, a doctor determines to collect samples from them, which can be, such as urine or blood. The samples are then analysed by the hospital's technical consultant in the laboratory. The common pipeline to analyse this sample is to incubate them for 12-24h to grow bacteria colonies. After this time, technicians compute mass spectrometry using the Matrix-Assisted Laser Desorption Ionization Time-Of-Flight (MALDI-TOF) technique, which is detailed in Section 1.1.2 of this dissertation. Currently, MALDI-TOF automatically discriminates microorganisms by consulting private databases provided by Vitek MS [1] (bioMérieux, France) or MALDI Biotyper[2] (Bruker Daltonics, Germany), depending on the MALDI-TOF MS machine the laboratory has. These commercial platforms present reference spectra for the most common microorganisms in the clinic and update their database periodically. This MALDI-TOF directly identifies species, genera, and family, thus revealing that a *K. pneumoniae* is causing the infection.

However, as mentioned above, *K. pneumoniae* usually acquires antibiotic resistance mechanisms [43, 38, 44], and the antibiotic administered by the doctor when the patient came to the hospital may not be effective. Therefore, it is necessary to perform an Antimicrobial Susceptibility Testing (AST) to know which antimicrobial treatment actually works, which requires about 24-72 additional hours. Once AST is performed, a personalised decision is made for the current patient. Unfortunately, from the first time the patient is hospitalised, 96 hours, that is, 4 days, have passed

[1] https://www.biomerieux.es/diagnostico-clinico/productos/vitekr-ms

[2] https://www.bruker.com/en/products-and-solutions/microbiology-and-diagnostics/microbial-identification.html

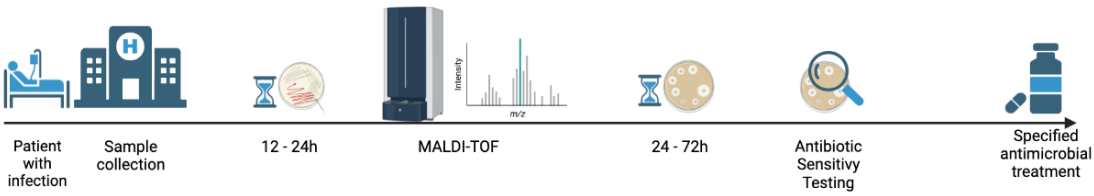


Figure 1.1: Infection identification and treatment determination workflow in a real hospital.

until the exact antimicrobial treatment is applied. Hence, the main limitation of the current method is the time consumption that requires to administrate a proper antibiotic.

The current MALDI-TOF MS solutions fall short in discriminating subspecies, contagious or virulent isolates, or antibiotic resistance, limiting the accuracy of diagnoses and the development of appropriate treatment plans. To overcome this challenge, researchers are turning to ML techniques to analyse the information contained in the spectra at a deeper level. While some studies claim to automatically perform the AST [45, 11, 46], their limited dataset sizes and lack of multi-modal data hinder their generalisation power, highlighting a gap in the current literature. Addressing this gap requires the development of multi-modal models capable of leveraging heterogeneous data to improve accuracy and generalisation power. The implementation of such models in microbiological laboratories has the potential to enhance the detection of multi-drug resistant isolates, optimise therapeutic decisions, and significantly reduce the time needed to obtain results on resistance mechanisms compared to current manual methods. The widespread adoption of these models could reduce costs by avoiding the administration of inefficient antibiotics and, most importantly, save lives by enabling personalised and timely treatment.

1.1.2 Microorganism identification through mass spectrometry

Mass Spectrometry (MS) is a technology that has been widely used in laboratories for the analysis of bio-molecules since the 1970s [47, 48] to nowadays [49, 50, 51, 52]. This technique ionises the molecules to a gas state with a positive charge and then differentiates them by their mass-to-charge ratio (m/z) using a detector. Therefore, an MS has three elements, as seen in Figure 1.2: an ioniser, a mass analyser and a detector, all contained in a void atmosphere to work in absolute values [53].

The ioniser electrically charges molecules that are generated by the excess or loss of electrons [54]. The most common ionisation uses a desorption process. This means that each sample is transformed into gas ions, which are intended for

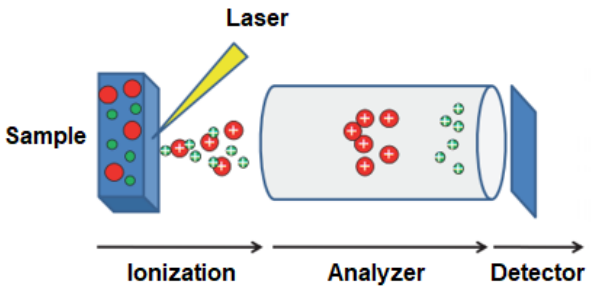


Figure 1.2: Mass Spectrometry scheme. Source: [1]

non-volatile and thermally unstable samples with molecular weights greater than 105 Dalton (Da). This kind of ionisation is called Matrix-Assisted Laser Desorption Ionization (MALDI).

When MALDI technique is used, the sample is placed in a conductive material and then $1\mu l$ matrix solution is added to each spot. The matrix is a substance capable of absorbing energy when irradiated with a laser of varying wavelengths. Nitrogen lasers, which emit at a wavelength of 337 nm, are generally used. The interaction between the matrix molecules and the laser photons triggers gas phase sublimation of the matrix, which is immediately followed by the ionisation of the sample [55, 56]. Specifically, the most suitable matrix to identify microorganisms and proteins is α-cyano-4-hydroxy-transcinnamic acid [57, 58]. Later, the ions are accelerated by an electrical field directed to the mass analyser and detector.

The mass analyser has two objectives: first, to discriminate the ions with respect to their mass/charge (m/z) ratio and then to direct them to the detector. In the case of MALDI, the most common mass analyser used is the Time-Of-Flight (TOF) analyser. In the TOF mass analyser (see Figure 1.3), the ions go through a void flight tube due to the acceleration provided by the electrical field previously mentioned. Regarding the m/z ratio, since the MALDI charges are positive and their kinetic energy is constant, the time of flight depends only on the mass of the peptides present. Hence, lighter particles reach the detector earlier [56, 53, 59] which introduces a temporal component to the m/z axis, as it records the time of detection.

Finally, the ions impact the detector, as seen in Figure 1.3. From this detector, a mass spectrogram is created called MALDI-TOF Mass Spectrometry (MS), as shown in Figure 1.4. The Y axis shows the number of ions detected in an arbitrary unity called *intensity*. The X axis presents their mass/charge ratio (m/z). MALDI-TOF MS usually are composed of $20,000$ m/z elements with a specific intensity for each one, thus a high-dimensional data vector. Then, each ion appears as

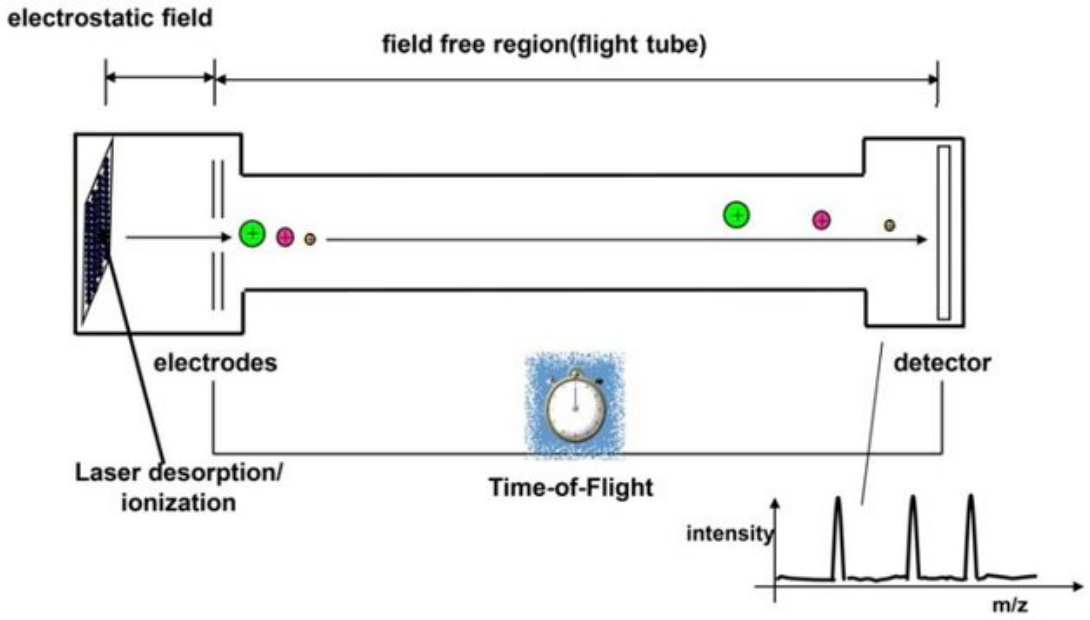


Figure 1.3: TOF mass analyser. Source: [1]

several peaks grouped around the same m/z position representing the statistical distribution of the different isotopes of the ion [53, 60].

Clinicians benefit from the advantages offered by this technique, which include high specificity in molecular weight determination, versatility in detecting different compounds like proteins, peptides, toxins, or nucleic acids, and flexibility to analyse various sample types, including volatile, non-volatile, polar, apolar, solids, liquids, and gases.

1.1.3 Machine Learning in clinical microbiology

In the field of microbiology, various techniques are used to identify and study bacteria. One such method is the use of MALDI-TOF for the determination of bacterial species by comparing spectra against proprietary databases provided by companies such as MALDI Biotyper (MBT), ASTA MicroIDys, or Vitek MS. However, this approach is limited to species-level identification and does not facilitate subspecies determination. Additionally, conventional methods for determining antibiotic resistance mechanisms in bacteria can take up to 96 hours. Furthermore, traditional approaches for automatic ribotyping of hypervirulent strains, Polymerase Chain Reaction (PCR), can take more than 7 days to yield results. Hence, in the current century, ML has been applied to solve such problems.

MALDI-TOF MS is commonly known to be called *fat data* meaning that in real-world problems the datasets are usually smaller than 500 samples and each MALDI-

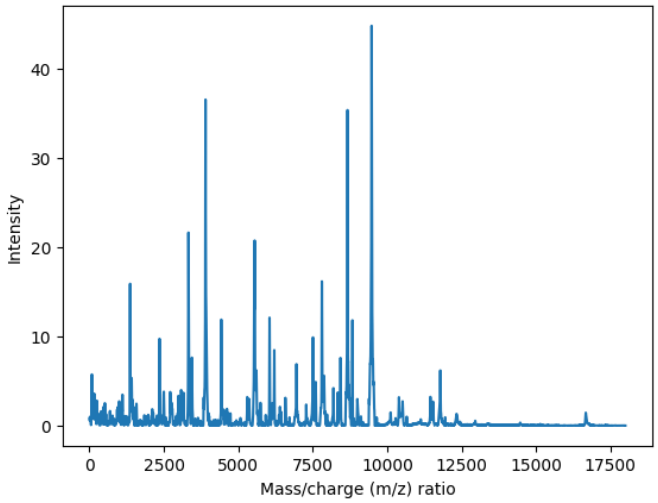


Figure 1.4: An example of a MALDI-TOF MS representing a *Clostridium difficile* isolate. In case of *C. difficile* only the first $18K$ m/z position are relevant for its determination.

TOF MS has $20,000$ features, i.e., $D >> N$. In [61], they identified different clonal lineages of methicillin-resistant *S. aureus* by using ClinProTools [62] black-box private software. The same software was used in [63] to discriminate between contagious and environmental strains of *Streptococcus uberis*. However, prior studies preferred to use open-source models. For example, in [64] the discrimination between *B. anthracis, E. coli, S. pneumoniae 18C-A, and S.pyogenes* based on their MALDI-TOF was perfectly performed by a RF. Other authors, such as [65], used both an RF and a Support Vector Machine (SVM) to classify different serotypes of Group B *Streptococcus* (GBS). Other works, such as [66], also used an RF, a SVM, and a multi-Logistic Regressor (LR) to perform strain typing of *S. haemolyticus*. Other approaches, intended for high-dimensional data, such as [67], proposed using sparse SVMs to classify the intestinal bacterial composition. Posterior research, such as [68], used an RF to identify subspecies of *Mycobacterium abscessus* based on their MALDI-TOFs. However, other authors have proposed using different spectroscopy methods to identify pathogenic bacteria. In [69] the authors used Raman spectroscopy images combined with a Convolutional Neural Network (CNN) for the rapid identification of pathogenic bacteria at the single-cell level. Following the same idea, in [70] they also used Raman spectroscopy, combined with LSTMs and CNNs models, to identify *Urechis unicinctus*. Other authors [71] have identified bacteria phenotypes also using Raman spectroscopy and a Transformer model.

Our first collaboration[3] with the group led by Belén Rodríguez started with

[3]This work is accepted for publication in Journal of Clinical Microbiology from the American

the work presented in Candela-Guerrero et al. [72]. During this international project, the viability of applying ML techniques to MALDI-TOF MS data for the detection of bacterial subspecies within the *Enterobacter cloacae complex* (EEC) was studied. The EEC is a group of closely related Gram-negative bacteria that belong to the genus Enterobacter. They are commonly found in the environment and human gut [73] and can cause a variety of infections, particularly in people with compromised immune systems [74, 75]. They can be difficult to distinguish from other members of the genus Enterobacter [76, 77] and can be antibiotic-resistant [78]. A total of 357 isolates were employed and different models were attempted. The models were first trained using 90 bacterial isolates from the Ramón y Cajal University Hospital (UHRyC), and were subsequently tested on a national scale using 126 different UHRC bacterial isolates, as well as on an international scale using 141 samples from the University Hospital of Basel (UHB), Switzerland. As a result, poor performance was observed with linear methods, such as Partial Least Squares-Discriminant Analysis (PLS-DA); however, non-linear approaches, such as SVM with RBF kernel or RF, yielded superior results by correctly classifying 122 out of 126 (96.8%) isolates from UHRC and 136 out of 141 (96.4%) isolates from UHB. These results demonstrate the effectiveness of MALDI-TOF MS in identifying different species within a large cluster of similar bacteria. Furthermore, these promising results have served as the foundation for the ongoing collaboration between both departments, TSC at UC3M and Microbiology at IISGM.

During this collaboration, two main projects have been developed. Firstly, probabilistic ML methods have been applied for the prediction of antibiotic resistance in *K. pneumoniae*. Additionally, automatic ribotyping of *Clostridium difficile* has been performed by applying probabilistic deep learning methods. In the following sections, the current state-of-the-art for both bacteria is reviewed, with a particular focus on the number of samples used, the software/technique proposed, and their reproducibility.

Machine Learning applied to *K. pneumoniae* antibiotic resistance determination

It has been established in the literature that one of the major challenges in the field of microbiology is the emergence of multidrug-resistant bacteria [38, 36, 37]. In particular, multidrug-resistant *K. pneumoniae* is considered a global public health threat by major international health organisations due to its rapid spread, high morbidity and mortality, as well as the economic burden associated with its treatment and control [38, 36, 37].

K. pneumoniae was first described by Carl Friedlander in 1882 in the lungs of people who had died from pneumonia. *K. pneumoniae* is a Gram-negative bacteria

Society for Microbiology (ASM).

and belongs to the Enterobacteriaceae family. It is important in the day-to-day work of microbiology due to its resistance to multiple drugs and, nowadays, it is resistant to most available antibiotics. This bacterium causes different infections such as respiratory and urinary tract, bloodstream, or surgical site [79, 80]. In addition, it is a common nosocomial infection, those that occur when a patient is hospitalised and infected inside the hospital. These infections are generally treated with β-lactam antibiotics [81, 82] such as Amoxicillin-Clavulanate. But there exist different isolates of *K. pneumoniae* which produce β-lactamase enzymes, as Extended-Spectrum Beta-Lactamases (ESBL)s [83] which make them resistant to antibiotics, restricting the possible treatments. Therefore, *K. pneumoniae* started to be treated with carbapenem antibiotics, such as imipenem and meropenem. However, several isolates of *K. pneumoniae* started to develop resistance mechanisms of Carbapenemases (CP). This resistance to carbapenems is a major challenge, as recognised by the World Health Organization (WHO) [44], since some carbapenemases can hydrolyse almost all *beta*-lactam antibiotics, making this bacteria resistant to both *beta*-lactams and carbapenems. Nowadays, in clinical laboratories, as seen in Fig. 1.1, 96 hours are needed to perform AST, which is routine.

MALDI-TOF is designed for microbial identification, but prior studies have found that detection of the resistance mechanism ESBL and CP can be inferred from these data due to different molecular weights after hydrolysis by resistant bacteria [84]. As suggested in [85], ML approaches can automatically analyse and predict Antibiotic Resistance (AR) based on the MALDI-TOF MS protein profiles. Therefore, different studies have analysed the use of ML approaches over MALDI-TOF data to reduce the time needed to detect these resistances from 96 to 24 hours.

The existing literature focuses mainly on classical ML models such as RF, SVM, or Genetic Algorithm (GA). Additionally, research conducted in recent years tends to involve small datasets and results that are not reproducible, owing to the use of private datasets and proprietary software. In fact, a limited number of studies have been carried out using accessible databases and open-source code.

ClinProTools®3.0 (Bruker, Daltonics, Bremen, Germany) is one of the widely private software used in the current literature. It is developed by the same company that manufactures MALDI-TOF machines. This tool provides several traditional ML models such as GA, Supervised Neural Network (SNN), SVM, or Quick Classifier (QC). Researchers at Beijing Tongren Hospital [86] proposed the use of ClinProTools to classify 143 isolates of multidrug-resistant bacteria, including CP and non-CP strains of *K. pneumoniae*. However, only 22 samples were indeed *K. pneumoniae*. Furthermore, another private software, flexAnalysis, was used to manually select 9 peaks for the detection of meropenem resistance. Finally, a GA model claimed a perfect classification of the 22 isolates. As previously noted, *K. pneumoniae* can also exhibit resistance to β-lactam antibiotics, such as amoxicillin, called ESBL-producer

bacteria. In their study, Li et al. [87] sought to differentiate between ESBL- and non-ESBL-producing bacteria by examining a combination of *E. coli* and *K. pneumoniae* isolates. Specifically, with respect to *K. pneumoniae*, the authors studied 25 isolates again using GA, SNN, and QC from ClinProoTools. In [88], the samples were automatically identified by the MALDI biotyper (Bruker). The authors then used ClinProTools to perform a PCA and a dendrogram to group 25 carbapenem-resistant *K. pneumoniae* isolates, resulting in two distinct groups. In doing so, they identified two separate nosocomial infections that are epidemiologically distinct from each other. ClinProTools was also used by [89] where they differentiate between normal *K. pneumoniae*, which typically affects patients in intensive care units [90], and hypervirulent, which can infect healthy patients. The authors reported that the use of SVM on 43 isolates successfully differentiated healthy and hypervirulent strains. Researchers from the Azienda Ospedaliero-Universitaria Pisana [91] apply ClinProTools to detect resistance to 139 *K. pneumoniae* isolates. The process involves the common steps of baseline removal and smoothing of the MALDI-TOF MS data, normalisation through their own Total Ion Current (TIC), and finally, using a GA in conjunction with a K-Nearest Neighbors (KNN), both available through ClinProTools, to differentiate resistance among isolates. Recently, in [92], researchers from Wuhan University applied a GA to detect CP over 175 isolates. The preprocessing pipeline they used consisted of various steps, such as baseline subtraction, spectra averaging, smoothing, and recalibration. Therefore, they divided the isolates into two groups and performed a binary classification between CP and non-CP, where non-CP includes other types of antibiotic-resistant bacteria that were not specifically focused on carbapenem resistance. In summary, the use of black-box private software is a proliferation in the field of microbiology. Lately, a Spanish company has developed a new private software equivalent to ClinProtools called CLOVER Bioanalytical Software ®(Clover Biosoft). This software incorporates more advanced ML models such as RF, SVM, Hierachical Clustering Analysis (HCA), Partial Least Square Discriminant Analysis (PLS-DA), KNN or Light Gradient Boosting Machine (LightGBM). In [93], researchers from Spain used this software to conduct an analysis of 162 isolates of *K. pneumoniae* in order to discriminate between CP-resistant and non-CP-resistant bacteria. They used both PLS-DA and RF to classify these isolates, eventually achieving a perfectly accurate score between CP and non-CP.

Rather than relying on proprietary, black-box software, other researchers have used open-source programming languages, such as Python or R, to automatically differentiate between isolates. This approach has the advantage of increasing the reproducibility of the experiments, as the underlying code is accessible and can be executed. An example of this approach is the study by Kaohsiung Veterans General Hospital in Taiwan [94], which used open source Spanish software, *Mass-Up*, programmed in R [95] to perform an analysis of 95 *K. pneumoniae* isolates to

discriminate between carbapenem-resistant and sensitive strains. From Mass-Up, they used a SVM, a KNN, an RF and an LR classifier. Other studies, such as in [96], conducted a longitudinal study to detect resistance to ciprofloxacin in 15782 isolates of *K. pneumoniae* collected between June 2013 and February 2018. In this study, they used an XGBOOST and an SVM model to predict resistance to ciprofloxacin. Moreover, a preprocessed version of the MALDI-TOF MS was made available consisting of 133 peaks selected using the peak-noise ratio. However, no code implementation was provided for reproducibility. Other open-source approaches are based on Python implementation, such as the work presented by Wang et al. [68] from Anhui Medical University Hospital in China, where they discriminated between resistant and sensitive carbapenem in 171 isolates of *K. pneumoniae* by using SVM and RF. The authors claim that the data is available upon request. However, in this dissertation, we attempted to contact them and did not receive a response. In contrast, researchers at ETH Zurich have created a large real-world clinical data set of MALDI-TOF, consisting of 5554 *K. pneumoniae* samples, which is fully shared with the community [97]. From this vast dataset, they performed different ML and DL techniques to detect resistance to ceftriaxone in *K. pneumoniae* using LightGBM and MultiLayer Perceptron (MLP) models.

Table 1.1: Literature review w.r.t. *K. pneumoniae* antibiotic resistance detection. Regarding data availability, *Yes** indicates that the data is available upon request.

Work	# of samples	Software used	Open source	Probabilistic	Data availability	Reproducible
Wang et al. 2013 [86]	22	ClinProTools	No	No	No	No
Li et al. 2014 [87]	25	ClinProTools	No	No	No	No
Angeletti et al. 2015 [88]	25	ClinProTools	No	No	No	No
Huang et al. 2015 [89]	43	ClinProTools	No	No	No	No
Giordano et al. 2018 [91]	139	ClinProTools	No	No	No	No
Gato et al. 2021 [93]	162	Clover BioSoft	No	No	Yes	No
Huang et al. 2020 [94]	95	Mass-Up	Yes	No	Yes*	No
Weis et al. 2020 [11]	1769	Python	Yes	Yes	Yes	Yes
Huang et al. 2022 [92]	175	ClinProTools	No	No	Yes*	No
Wang et al. 2022 [68]	171	Python	Yes	No	Yes*	Yes
Weis et al. 2022 [97]	5554	Python	Yes	No	Yes	Yes
Weis et al. 2022 [46]	1769	Python	Yes	Yes	Yes	Yes
Wang et al. 2022 [96]	15782	R	Yes	No	Yes	Yes
Total			6/13	2/13	8/13	5/13

In summary, a recurrent characteristic across previous studies on the detection of antibiotic resistance in *K. pneumoniae* is the use of deterministic models. The existing literature does not incorporate probabilistic approaches in this area of research. However, an exception to this trend is the work of the research team from ETH Zurich, who has applied Gaussian Process (GP) in the prediction of antibiotic resistance in *K. pneumoniae* using a large real-world clinical dataset obtained from MALDI-TOF [97]. Their first study presented was in 2021 [11] where they created a new kernel function specifically tailored for MALDI-TOF MS data. This new

kernel is called Peak Information KErnel (PIKE) and for each pair of MALDI-TOF called $\mathrm{MD}_a, \mathrm{MD}_b$:

$$\mathrm{PIKE}(\mathrm{MD}_a, \mathrm{MD}_b) = \frac{1}{2\sqrt{2\pi t}} \sum_{i,j}^{\#Peaks} \lambda_i^a \lambda_j^b \exp\left(- \frac{(p_i^a - p_j^b)^2}{8t} \right). \tag{1.1}$$

By taking a closer look at this function, we can first notice that λ_i^a and λ_j^b denote the intensity of a pair of peaks, and their multiplication is a measure of the interaction between them. Secondly, as p_i^a and p_j^b denote the position m/z of this pair of peaks, their interaction is multiplied by the square difference between their m/z positions. This means that two peaks that are a greater distance from each other will be multiplied by 0, and hence their influence is irrelevant. However, two peaks that are close will be more relevant. This difference is smoothed by a parameter t that the authors fix to 5 in their article. In summary, PIKE computes a kernel by comparing the MALDI-TOF MS by pairs where the exponential Euclidean distance helps to compare similar peaks, as they have similar p_i and p'_j values. However, computing this PIKE scales with the number of peaks that each MD has; hence the spectra must preprocessed topological peak selection. This peak selection is a simple peak detection method based on the persistence concept from computational topology which automatically results in a peak detection because local maxima exhibit high persistence values by construction in MALDI-TOF data. Following the indications of the authors of [11], only 200 peaks are kept per sample. Once the PIKE is constructed, they proposed using a GP classifier to detect 3 different antibiotic resistances in 1769 *K. pneumoniae* samples from the DRIAMS [97] dataset. Because they used a probabilistic approach, they could analyse class probabilities and decision boundaries with uncertainty measurement of antimicrobial resistant prediction. Following this probabilistic approach, they improved their own results in a consecutive study [46] in which the authors tackled the high standard deviation values in the prediction in different train and test splits. To do so, they applied hierarchical agglomerative clustering to infer the underlying phylogenetic structure between microbial samples. Thus, they again calculated the PIKE and used a GP to predict the same bacteria mentioned in the previous paper, now achieving better overall performance.

In conclusion, previous studies have demonstrated reproducibility on the determination of antibiotic resistance in *K. pneumoniae* is limited. Table 1.1 summarises the literature review performed in previous paragraphs. As shown, of the 13 reviewed articles, only a small fraction, 5, are fully reproducible, and only 8 make their data available, with 3 of them being accessible upon request. Efforts to obtain data from the corresponding authors were met with limited success, with only one, Huang et al. [94], providing it. This lack of reproducibility presents a significant obstacle for researchers verifying and replicating previous findings.

Furthermore, deterministic approaches have been extensively explored, with 11 out of 13 reviewed papers utilising them. Only two studies proposed a probabilistic perspective. Moreover, a gap in the literature exists regarding the proposal of multi-modal DL models to solve the problem. Other authors have also acknowledged this gap and have suggested that the combination of multi-modal data can aid in the identification and discrimination of bacteria. For example, Kalkan et. al. [98] statistically demonstrated the ability to analyse the bacterial population of the Black Sea by combining information from MALDI-TOF and rRNA sequencing. Additionally, Tressler et al. [99] obtained metabolomic and lipidomic measurements of human breast cancer by manually combining Nuclear Magnetic Resonance (NMR) spectroscopy and MALDI-TOF imaging. These studies highlight the potential of multi-modal data to discriminate and identify bacteria, but to date, neither ML nor DL methods have been applied.

Given the promising results obtained by the probabilistic approaches of Table 1.1 and the current gap in applying multi-modal learning, this thesis proposes an extension of Factor Analysis (FA) to determine antibiotic resistance mechanisms automatically. Therefore, in Chapter 5, we tailor and apply the kernelised model developed in Chapter 3, to the prediction of the resistance mechanism in *K. pneumoniae*.

Machine Learning applied to *C. diff* determination

In the 2000s, there was a significant increase in the incidence of *Clostridium difficile* (*C. diff* or *C. difficile*) in the United States, resulting in a 400% increase and excess costs of $4.8 billion [100]. Similarly, in Europe, between 2011 and 2012, *C. diff* was responsible for 48% of all gastrointestinal infections [101], with a high prevalence among patients who have recently been treated with antibiotics [102]. These infections, particularly severe cases, can lead to life-threatening complications such as *sepsis* (a systemic inflammatory response to infection) [103] or *colitis* (inflammation of the colon) [104].

In the field of microbiology, *C. difficile* is a widely recognised and serious pathogen, first reported in the literature by Hall and O'Toole in 1935 [105]. It is a Gram-positive, spore-forming bacterium belonging to the *Peptostreptococcaceae* family and is known to be one of the main causes of antibiotic-associated diarrhoea. *C. diff* has a clonal population with various Sequence Types (ST); the largest study conducted in the UK identified 69 different ST [106]. In particular, the ribotype 027 (RT027) is often referred to as hypervirulent [107, 108] due to its association with severe colitis and higher mortality rates [109, 110, 111]. This is because it has a deletion in the tcdC gene, which regulates the production of toxins responsible for causing intestinal damage, namely Toxin B and Toxin Binary.

Currently, laboratories rely on the use of a real-time PCR assay, specifically

GeneXpert *C. difficile*, to detect the presence of toxins B, binary and deletion in the tcdC gene in *C. difficile* strains [112]. However, the emergence of new ribotypes, such as RT181, which share characteristics with hypervirulent RT027, presents challenges for accurate detection using PCR methods. Studies have yet to confirm whether RT181 is indeed hypervirulent, despite being similar to RT027 [113, 114]. This highlights a limitation in current methodologies, as current PCR methods cannot differentiate between RT181 and RT027. However, as Cuénod et al. [115] noted in their literature review, MALDI-TOF MS can be used for ribotyping of *C. difficile*. This was previously demonstrated by Reil et al. in 2011 [116] through the analysis of 355 *C. diff* samples using MALDI-TOF MS, where biomarkers were found for the manual identification of RT001, RT027 and RT078/126 in the mass range between $3K - 13K$ Da. Rizzardi et al. in 2015 [117] also noted that extended MALDI-TOF MS, focusing on the mass range between $30K - 50K$ Da, was able to identify biomarkers for various ribotypes, including RT010, RT011, RT012, RT015, RT017, and RT020, among others.

Existing literature suggests that ML techniques can identify patterns in the microbiota of *C. diff* RT027 strains [118], or to predict the outcomes of patient infections caused by *C. diff* RT027 strains [119, 120]. However, the application of ML for the automatic ribotyping of *C. diff* has not yet been proven. In a study by Calderaro et al. [121], MALDI-TOF MS was used on a set of 29 *C. diff* samples to classify them as epidemic or non-epidemic, but no conclusive results were obtained regarding automatic ribotyping. Furthermore, the authors did not provide any accessible code or data.

Up to this day, no studies have applied ML methods to perform automatic ribotyping of *C. diff*. In light of this gap in the literature, in Chapter 6, a preliminary study is conducted in which the models presented in Chapter 3 and Chapter 4 are tailored and applied for the automatic ribotyping of *C. diff* between RT027, RT181, and other strains.

1.2 Motivation

In recent years, multi-modal learning has been claimed to be an improvement in different fields of study. In practise, in bacteria discrimination, classical learning methods based only on MALDI-TOF are not enough to make a precise determination of antibiotic resistance, as in [11] where using a GP Weis et al. achieve AUPRC values of 0.55 and 0.56 in determining antibiotic resistance to ciprofloxacin or piperacillin/tazobactam.

Therefore, we propose that multi-modal learning methods must be developed and used in hospital infection treatment workflows, following Fig. 1.5, to enhance bacterial identification and discrimination. To do so, these methods must be

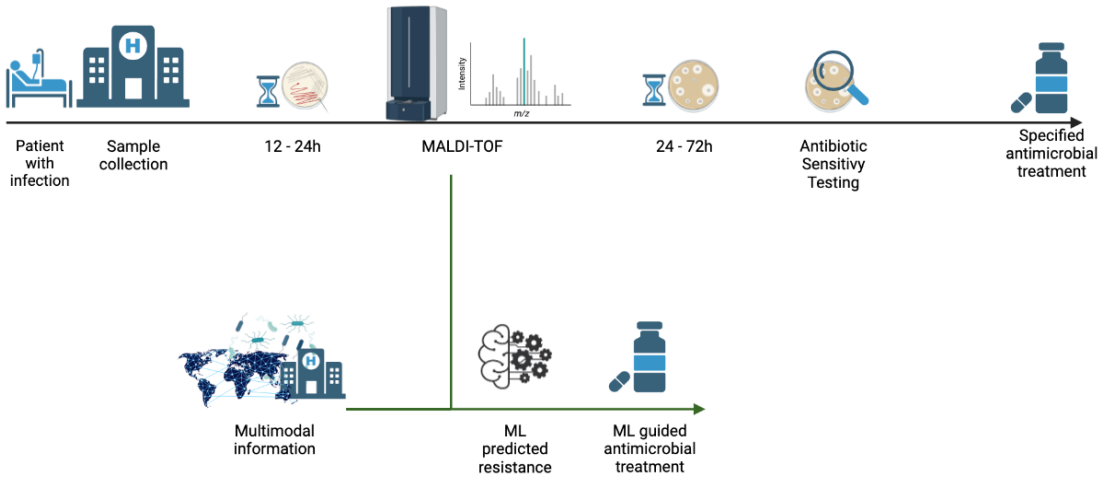


Figure 1.5: Infection treatment workflow incorporating ML and multi-modal information

developed specifically tailored to cover the needs of bacteria and medical data, which are:

- Dealing with **heterogeneous and multi-modal data**, such as real, categorical, binary, or images. For example, combining real information, such as the MALDI-TOF; with categorical information, such as epidemiological information about the hospital where the samples are coming from; binary information regarding the resistance to different antibiotics; or even images of the agar-agar plates.

- Dealing with **high-dimensional fat data**, which means that the number of features exceeds the number of samples. Particularly, in MALDI-TOF, where each spectrum contains $20K$ features and the samples are usually $N < 500$.

The available multi-modal learning solutions rely on complex hierarchical deep latent representations to project data correlation among views. This complexity imposes a lack of modularity to add new data views when available and difficulty missing view handling. Moreover, current complex models hinder inference training and obscure the interpretability of the model. Hence, a model intended for medical data needs to be modular to easily adapt to new data types changing its architecture while keeping a simple relation that can be interpreted.

In this dissertation, we collaborate with the IISGM in Area 4, which focuses on clinical microbiology, infectious diseases, and Human Immunodeficiency Virus. Specifically, we collaborate with Belén Rodríguez Sánchez, PhD, from the Infectious Disease group. As such, each technical contribution developed in this thesis has

been applied to microbiology problems, particularly in the identification and discrimination of different types of bacteria.

1.3 Objectives

For this reason, in this thesis, we aim to:

- Develop **interpretable** deep-generative models to analyse the behaviour of the data.
- Implement **modular** formulations that can adapt to different medical data and problems.
- Provide efficient and accurate **models** that can be used in real microbiology laboratories.
- Provide open-source, available and reproducible models to the community.

To achieve these objectives, we extend an existing FA model designed for heterogeneous data, the Sparse Semi-supervised Heterogeneous Interbattery Bayesian Analysis (SSHIBA) model, to the needs of microbiological data using two different approaches.

First, we start with a Bayesian formulation of FA and tailor it to a kernel formulation. This allows us to propose a new model that can handle high-dimensional data while exploiting the correlation between multiple views for multi-modal learning. By using kernel methods, the model can operate in dual space and incorporate non-linearities specifically tailored to microbiological data, such as specific kernels [11]. Therefore, we use an FA model whose input is the kernelised view of the data, rather than the raw data. These dual space variables are linearly combined using an FA approach with the available multi-view information. Additionally, we implement automatic relevance vector selection to further reduce the input space. Finally, the multi-modal learning nature of the model enables us to exploit all the available information about a given problem.

Secondly, we extend the above model by integrating VAE generative models. This adds modularity and flexibility to the global framework. VAEs are powerful generative models known for their versatility; that is, their encoder-decoder structure can be easily modified to handle new types of data without affecting any other part of the Bayesian FA formulation. This means that we are adding a new modular and flexible component that can handle large amounts of different data, such as temporal data or image data, simply by changing the encoder-decoder architecture.

In collaboration with the IISGM, we make a significant contribution to the field by demonstrating the feasibility of our proposed models on real-world bacteria discrimination problems. Our first model, Kernelised SSHIBA (KSSHIBA), incorporates high-dimensional MALDI-TOF MS data ($\mathbb{R}^{N\times 20K}$ where $N < 500$), epidemiological data (hospital source of each bacteria), and categorical information on antibiotic resistance to predict antibiotic resistance in *Klebsiella pneumoniae*. Our second model, Factor Analysis Variational AutoEncoder (FAVAE), uses specially designed VAE encoder-decoder architectures for MALDI-TOF data to automatically ribotype *Clostridium difficile* bacteria in a preliminary study, which is then tested on a real outbreak that occurred on January 24th at the Hospital General Universitario Gregorio Marañón (HGUGM).

1.4 Organisation

This dissertation is divided into seven chapters, which are reviewed in the following paragraphs. All published articles are included in the introductory lines of their respective chapter.

Chapter 2. Background. In this chapter, the necessary technical background required for a comprehensive understanding of the dissertation is provided. The first part reviews FA models while the second part details VAEs and hierarchical VAEs in detail, pointing out the current limitations of both models.

Chapter 3. Kernelised SSHIBA. Once the background have been established, our first technical contribution is presented. In this chapter, we review the kernelised version of SSHIBA presented in [122]. First, we present the technical advances of KSSHIBA in terms of dealing with kernelised views, automatic Bayesian relevance vector selection, and automatic feature-relevance determination. Then, we provide experimental results that demonstrate all the functionalities in various high-dimensional datasets. The final section provides a summary of the key findings and concluding remarks.

Chapter 4. Factor Analysis Variational AutoEncoder. Our second technical contribution is presented in this chapter. First, we introduce the theoretical foundations and mathematical formulation of our Factor Analysis Variational AutoEncoder (FA-VAE) [123]. Then, the following sections present experimental results to demonstrate the efficacy of the approach, including conditioning a pretrained VAE to a specific label, domain adaptation between distinct datasets and styles, and using the approach as a transfer learning tool between generative models. The last part offers a synopsis of the main discoveries and ultimate comments.

Chapter 5. Automatic antibiotic resistance prediction using KSSHIBA. In this chapter, we demonstrate the practical application of the first technical contribution of this dissertation, KSSHIBA. Specifically, we adapt the model presented in Chapter 3 to a real-world scenario involving the prediction of antibiotic resistance mechanisms in *K. pneumoniae*. To address this challenge, we leverage two different epidemiological datasets, one from the HGUGM and the other collected from 20 different hospitals in Spain and Portugal. We then describe how we tailor KSSHIBA to this type of data by integrating multi-modal information such as MALDI-TOF MS, epidemiological data, and antibiotic resistance mechanisms. Through experimental results, we demonstrate that KSSHIBA outperforms existing state-of-the-art methods by exploiting multi-modal learning.

Chapter 6. Automatic ribotyping based on probabilistic techniques. This chapter presents a practical application of the two technical contributions, KSSHIBA and FA-VAE, introduced in Chapters 3 and 4, respectively. We demonstrate how both models can be tailored to solve the automatic ribotyping problem of *Clostridium difficile*. By leveraging probabilistic techniques, we propose an approach to automatically ribotype three different classes: hyper-virulent RT027, 027-like RT181, and other ribotypes. Our method is tested on a dataset consisting of 275 samples collected at HGUGM where FA-VAE outperforms other methods. We also apply our approach to a real outbreak that occurred on January 24th and show that FA-VAE can instantly perform ribotyping on the first day of the outbreak, whereas traditional PCR techniques require seven days to complete.

Chapter 7. Discussions and Conclusions. We conclude summarising the several technical and application contributions to the field of microbiology presented in this thesis. Specifically, we have introduced two novel techniques, KSSHIBA and FA-VAE, and demonstrated their effectiveness in solving practical problems encountered in hospitals. These contributions represent a first step towards the use of multi-modal learning for bacterial identification and discrimination. Furthermore, we have identified several potential avenues for future research to further advance this field. Overall, this dissertation represents an important contribution to the microbiology field and provides a foundation for further exploration and innovation.

CHAPTER 2

BACKGROUND

The following chapter aims to provide a thorough technical background that is essential for comprehending this dissertation. The chapter is divided into two main sections. Firstly, a detailed review of FA models is presented. Secondly, VAE and hierarchical VAEs are studied. Lastly, the current limitations of the state-of-the-art models applied to the clinical microbiology field are highlighted.

2.1 Bayesian Linear Regression

Traditionally, most supervised ML methods consist in developing algorithms able to learn a function, $f_{\mathbf{w}}$, which maps the input variable $\mathbf{x}$ into an output or target variable y, that is,

$$y = f_{\mathbf{w}}(\mathbf{x}) \tag{2.1}$$

where $\mathbf{w}$ represents the parameters, or, weights, needed to translate the input information into the desired output. To optimise these parameters, a cost or loss function L is defined and its averaged value over the training data is minimised w.r.t. $\mathbf{w}$. This process is known as the training phase. A significant number of conventional ML algorithms adopt this approach. A well-known example is the Linear Regression (LR) model. The goal of LR is to find the optimal linear relationship between the input variables $\mathbf{x}$ and the output variable y. This linear relationship is represented by an equation of the form:

$$y = w_0 + w_1 x_1 + ... + w_d x_d = \mathbf{w}^T \mathbf{x} + w_0 \,, \tag{2.2}$$

where y is the output variable, $\mathbf{x} = [x_1, \ldots, x_d]^T$ are the input variables, and $w_0, \ldots, w_d$ are the regression coefficients of shape $D \times 1$. For model training, the LR method uses the mean least square cost function, that is, LR finds the values of $\mathbf{w}$ coefficients that minimise the sum of the squared differences between the predicted and actual values of the outputs over the training data.

Once the model weights $\mathbf{w}$, are learnt, LR can compute the output for new data $\mathbf{x}^*$ as

$$f^*_{\mathbf{w}} = \mathbf{w}^T \mathbf{x}^* + w_0 \,. \tag{2.3}$$

So, for a given input, a deterministic output is computed. However, neither the uncertainty about the value of the calculated $\mathbf{w}$ nor how it affects the predicted output is taken into account.

To address this issue, the literature proposed Probabilistic ML (PML). PML is a subfield of ML that deals with the development of models that can make predictions and decisions based on uncertain or probabilistic information. In PML, a model is trained not only to produce a single output, but also to generate a probability distribution over possible outputs. In traditional ML, models are typically trained to minimise risk, which is defined as the expected value of a loss function. On the contrary, PML models are trained to maximise the likelihood of the parameters, which is the probability of the data, $\mathbf{x}$, given the parameters, $\mathbf{w}$ of the model. In PML, some usually used models are GPs or Bayesian models. In particular, in Bayesian models, the model parameters $\mathbf{w}$, are treated as random variables (rv), and prior distributions are specified for these variables to represent prior knowledge or beliefs about their values. These prior distributions are used to update the posterior distribution when new data is observed, following Bayes' theorem:

$$p(\mathbf{w}|\mathbf{x}) = p(\mathbf{x}|\mathbf{w})\frac{p(\mathbf{w})}{p(\mathbf{x})} \propto p(\mathbf{x}|\mathbf{w})p(\mathbf{w})\,, \tag{2.4}$$

where $p(\mathbf{x})$ is often considered a normalisation constant. The posterior distributions represent the updated beliefs about the model parameters given the observed data.

Following the previous example, the Bayesian approach can be applied to the classical LR model. Bayesian LR extends Eq. (2.2) defining the model as

$$y = \mathbf{w}^T\mathbf{x} + \epsilon\,, \tag{2.5}$$

where $\mathbf{w}$ are the same coefficients defined in classic LR and ϵ is zero-mean noise with diagonal covariance matrix $\sigma^2 I$. Thus, the likelihood function for y follows

$$p(y|\mathbf{x}, \mathbf{w}) \sim \mathcal{N}(y|\mathbf{x}\mathbf{w}, \sigma^2 I)\,, \tag{2.6}$$

and then a prior probability distribution is assumed to $\mathbf{w}$ taking a Normal Gaussian distribution:

$$p(\mathbf{w}) \sim \mathcal{N}(\mathbf{w}|0, \lambda^{-1} I)\,. \tag{2.7}$$

With this Bayesian approach, we can compute the posterior distribution of the parameters following:

$$p(\mathbf{w}|y, \mathbf{x}) = p(y|\mathbf{x}, \mathbf{w})\frac{p(\mathbf{w})}{p(y|\mathbf{x})}\,, \tag{2.8}$$

where $p(\mathbf{w}|y, \mathbf{x})$ is the posterior distribution of the parameters given the data and $p(y|\mathbf{x})$ is the marginal likelihood of the data. Hence, the posterior distribution is Gaussian given by

$$p(\mathbf{w}|y, \mathbf{x}) \sim \mathcal{N}(\mathbf{w}|\mu, \Lambda^{-1})\,, \tag{2.9}$$

where $\mu = \Lambda(\lambda^{-1}\mathbf{w} + \sigma^2\mathbf{x}^T y)$ and $\Lambda = \lambda^{-1} + \sigma^2\mathbf{x}^T X$ being the mean and precision matrices, respectively.

Finally, we can compute the posterior distribution of the parameters and make predictions for new data points by obtaining the predictive distribution analytically. This strategy allows us to quantify the uncertainty in the model parameters and predictions and to incorporate prior knowledge or beliefs about the parameters if available.

This thesis seeks to advance the understanding of Bayesian models through a rigorous analysis of FA and VAE models. The applications proposed in this dissertation will focus on high-dimensional and heterogeneous multi-view problems and will evaluate the efficacy of FA models in addressing diverse data types. Firstly, we will examine the classical methods of FA, including Probabilistic Principal Component Analysis (PPCA) and its Bayesian counterpart Bayesian PPCA [124]. We will then delve into Bayesian InterBattery Factor Analysis (BIBFA) [125], an extension of BPPCA that addresses multi-view data handling. Finally, we will analyse an extension of BIBFA that incorporates sparsity, Automatic Relevance Determination (ARD) and handling of missing data, called SSHIBA [2]. These analyses will provide a comprehensive understanding of FA models and their applications in data analysis.

2.2 FA

Factor Analysis (FA) aims to identify the latent variables that explain the covariance structure of a set of observed variables. For this purpose, FA is based on the assumption that observed variables, $\mathbf{X} \in \mathbb{R}^{N\times D}$, are linear combinations of a smaller set of hidden or latent factors, $\mathbf{G} \in \mathbb{R}^{N\times K}$, where $K << D$, plus some added noise, $\epsilon \in \mathbb{R}^{N\times D}$. That is,

$$\mathbf{X} = \mathbf{G}\,\mathbf{W}^T + \epsilon\,. \tag{2.10}$$

where $\mathbf{W} \in \mathbb{R}^{D\times K}$ contains the coefficients that describe the relationship between the observed variables $\mathbf{X}$ and the latent variables $\mathbf{G}$. The underlying factors are not directly observable, but their effects are present in the observed variables. The objective of FA is to find the hidden factors $\mathbf{G}$ and the projection matrix $\mathbf{W}$.

FA is a versatile technique with a wide range of applications, such as data compression, visualisation, dimensionality reduction, feature extraction, and the derivation of interpretable features from high-dimensional data. Probabilistic FA has been developed to provide a probabilistic interpretation of results, allowing not only the characterisation of the data, but also an estimation of their uncertainty.

For example, PPCA is a probabilistic approach to the classical PCA method. In a classical PCA, the goal is to find a linear transformation of the data that

maximises the variance of the projected data. The PPCA model assumes that the data is generated by a linear process with Gaussian noise. Specifically, the model assumes that each data point $\mathbf{x}_{\mathrm{n},:} = [x_{n,1}, \ldots, x_{n,d}]$ is generated by

$$\mathbf{x}_{\mathrm{n},:} = \mathbf{g}_{\mathrm{n},:}\mathbf{W}^T + \epsilon\,, \tag{2.11}$$

where $\mathbf{x}_{\mathrm{n},:}$ is a D-dimensional data point, $\mathbf{W}$ is a $D \times K$ deterministic matrix of loading vectors and $\mathbf{g}_{\mathrm{n},:} = [g_{n,1}, \ldots, g_{n,k}]$ is a K-dimensional latent variable associated with $\mathbf{x}_{\mathrm{n},:}$ and ϵ is a D-dimensional Gaussian noise term with zero mean and covariance matrix $\tau^{-1}I_D$. The latent variable $\mathbf{g}_{\mathrm{n},:}$ is assumed to have a Gaussian prior distribution with zero mean and identity covariance matrix,

$$\mathbf{g}_{\mathrm{n},:} \sim \mathcal{N}(0, I_K)\,, \tag{2.12}$$

where K is a hyperparameter that is $K << D$ to perform dimensionality reduction.

Given the data, the goal of PPCA is to infer the posterior distribution of the latent variables $\mathbf{g}_{\mathrm{n},:}$ and the parameters of the model ($\mathbf{W}$ and Σ). This can be done using the Expectation Maximisation (EM) algorithm or Variational Inference (VI).

In the following subsections, we shall delve into the various Bayesian formulations of FA. The techniques discussed will be Bayesian PPCA, BIBFA, and finally Sparse Semi-supervised Heterogeneous Interbattery Bayesian Analysis (SSHIBA), which this dissertation extends and adapts to microbiological data.

2.2.1 Bayesian Probabilistic Principal Component Analysis

Bayesian Probabilistic Principal Component Analysis (BPPCA) [124] is a Bayesian extension of PPCA. In PPCA [126], the authors assumed that, given the data point $\mathbf{x}_{\mathrm{n},:}$, an independent latent variable $\mathbf{g}_{\mathrm{n},:} \in \mathbb{R}^{1\times K}$ exists and its prior is distributed as an isotropic Gaussian:

$$\mathbf{g}_{\mathrm{n},:} \sim \mathcal{N}(0, I_K)\,, \tag{2.13}$$

being, as explained above, $K << D$. As such, the data point $\mathbf{x}_{\mathrm{n},:}$ can be generated as a linear combination of the latent random variable $\mathbf{g}_{\mathrm{n},:}$ and the projection random variable $\mathbf{W}$, specifically given by

$$\mathbf{x}_{\mathrm{n},:} \,|\, \mathbf{g}_{\mathrm{n},:} \sim \mathcal{N}\big(\mathbf{g}_{\mathrm{n},:}\,\mathbf{W}^{\mathrm{T}}, \tau^{-1}I_D\big)\,. \tag{2.14}$$

The BPPCA [124] extension incorporates prior probability distributions on the model parameters $\mathbf{W}$ and another on the noise parameter ϵ with precision τ, transforming both into rv, as seen in Fig.2.1. Regarding the projection matrix the projection matrix $\mathbf{W}$, a Gaussian prior distribution is assumed following $\mathbf{W} \sim \mathcal{N}(0, I_K)$, and over the noise, a Gamma prior distribution as $\tau \sim \Gamma(a^\tau, b^\tau)$.

In traditional PPCA [126], τ is a parameter learnt by optimising marginal likelihood and inference has an exact closed form solution. However, in BPPCA,

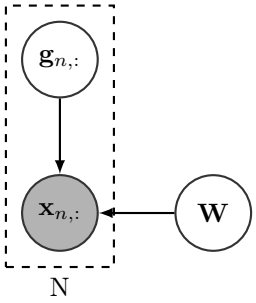


Figure 2.1: Graphical model of BPPCA. Observations are represented by grey circles, whereas random variables are represented by white circles.

τ is assumed to have a Gamma prior distribution, being so a random variable. Hence, approximate inference is used. Let us denote as $\mathbf{\Theta}$ the set of all previously defined rvs by BPPCA, that is, $\{\mathbf{G}, \mathbf{W}, \tau\} \subset \mathbf{\Theta}$. BPPCA proposes to infer the posterior distribution of $\mathbf{\Theta}$, namely $p(\mathbf{\Theta}|\,\mathbf{X})$. Nevertheless, the direct calculation of this posterior distribution is intractable due to the calculation of $p(\mathbf{X})$. As an alternative, VI [127] is used to approximate $p(\mathbf{\Theta}|\,\mathbf{X})$ by utilising a simpler and more tractable distribution $q(\mathbf{\Theta})$.

Variational Inference

In VI, the proposal is to minimise the Kullback-Leibler (KL) divergence between $q(\mathbf{\Theta})$ and $p(\mathbf{\Theta}|\,\mathbf{X})$. The KL measures the divergence between two probability distributions, i.e, their dissimilarity, meaning that when $\mathrm{KL}(q(\mathbf{\Theta}||p(\mathbf{\Theta}|\,\mathbf{X})) = 0$, the approximate posterior $q(\mathbf{\Theta})$ is equal to the true posterior $p(\mathbf{\Theta}|\,\mathbf{X})$. The KL, defined as,

$$KL(q||p) = \int q(\mathbf{\Theta}) \log\left(\frac{q(\mathbf{\Theta})}{p(\mathbf{\Theta}|\,\mathbf{X})}\right) d\mathbf{\Theta}\,, \tag{2.15}$$

quantifies the dissimilarity between q and p. By developing this formula, we can:

$$\begin{aligned}
KL(q||p) &= \int q(\boldsymbol{\Theta}) \log\left(\frac{q(\boldsymbol{\Theta})}{p(\boldsymbol{\Theta}|\,\mathbf{X})}\right) d\boldsymbol{\Theta} \\
&= \int q(\boldsymbol{\Theta}) \overbrace{\left(\log q(\boldsymbol{\Theta}) - \log p(\boldsymbol{\Theta}|\,\mathbf{X})\right)}^{\text{Prop. of logarithms}} d\boldsymbol{\Theta} \\
&= \int q(\boldsymbol{\Theta}) \log q(\boldsymbol{\Theta}) d\boldsymbol{\Theta} - \int q(\boldsymbol{\Theta}) \log p(\boldsymbol{\Theta}|\,\mathbf{X}) d\boldsymbol{\Theta} \\
&= \overbrace{\mathbb{E}_q[\log q(\boldsymbol{\Theta})] - \mathbb{E}_q[\log p(\boldsymbol{\Theta}|\,\mathbf{X})]}^{\text{Def. of expectation}} \\
&= \mathbb{E}_q[\log q(\boldsymbol{\Theta})] - \overbrace{\mathbb{E}_q\left[\log \frac{p(\mathbf{X}, \boldsymbol{\Theta})}{p(\mathbf{X})}\right]}^{\text{Bayes rule}} \\
&= \mathbb{E}_q[\log q(\boldsymbol{\Theta})] - \overbrace{\mathbb{E}_q[\log p(\mathbf{X}, \boldsymbol{\Theta}) - \log p(\mathbf{X})]}^{\text{Prop. of logarithms}} \\
&= \overbrace{\mathbb{E}_q[\log q(\boldsymbol{\Theta}) - \log p(\mathbf{X}, \boldsymbol{\Theta})] + \mathbb{E}_q[\log p(\mathbf{X})]}^{\text{Prop. linearity expectation}} \\
&= \mathbb{E}_q[\log q(\boldsymbol{\Theta}) - \log p(\mathbf{X}, \boldsymbol{\Theta})] + \underbrace{\log p(\mathbf{X})}_{intractable},
\end{aligned} \tag{2.16}$$

following VI, group the intractable terms and the tractable terms as follows:

$$\underbrace{\mathbb{E}_q[\log p(\mathbf{X}, \Theta) - \log q(\boldsymbol{\Theta})]}_{tractable} = \underbrace{\log p(\mathbf{X}) - KL\left(q(\boldsymbol{\Theta})||p(\boldsymbol{\Theta}|\,\mathbf{X})\right)}_{\text{intractable}} . \tag{2.17}$$

The key concept of VI is that the maximisation of the left part of Eq. (2.17) is equivalent to the maximisation of the right part. By definition, the KL divergence is a non-negative function, i.e., $KL \geq 0$. Therefore, the maximisation of the tractable part of Eq. (2.17) causes the KL to be 0, resulting in the left part of the formula becoming a lower bound of the log evidence $\log p(\mathbf{X})$, referred to as the Evidence LowerBOund (ELBO), as represented in the following equation:

$$L(q) = -KL\left(q(\boldsymbol{\Theta})||p(\boldsymbol{\Theta}|\,\mathbf{X})\right) + \log p(\mathbf{X}) \leq \log p(\mathbf{X}) \tag{2.18}$$

where $L(q)$ is defined by completely tractable terms:

$$L(q) = \mathbb{E}_q[\log p(\mathbf{X}, \Theta) - \log q(\boldsymbol{\Theta})] . \tag{2.19}$$

Consequently, the maximisation of $L(q)$ results in making the approximation $q(\boldsymbol{\Theta})$ closer to the true posterior $p(\boldsymbol{\Theta}|\,\mathbf{X})$, and even, at its maximum point, being $q(\boldsymbol{\Theta}) = p(\boldsymbol{\Theta}|\,\mathbf{X})$.

One method to calculate this ELBO is, for example, using the mean-field approximation [127]. This technique considers that the approximated distribution $q(\mathbf{\Theta})$ can be fully factorised as

$$q(\mathbf{\Theta}) = \prod_i q(\theta_i) = \prod_i q_i \,, \tag{2.20}$$

where θ_i is the i-th rv of the model and its approximate distribution is denoted as q_i. Thus, we can apply this trick to maximise $L(q)$ w.r.t each q_j factor as

$$\begin{aligned}
L(q_j) &= \int q(\mathbf{\Theta}) \left[\log p(X, \mathbf{\Theta}) - \log q(\mathbf{\Theta})\right] d\Theta \\
&= \overbrace{\int \prod_i q_i \left[\log p(X, \mathbf{\Theta}) - \sum_i \log q_i\right] d\mathbf{\Theta}}^{\text{Mean-field trick}} \\
&= \overbrace{\int q_j \left(\prod_{i\neq j} q_i\right) \log p(X, \mathbf{\Theta}) d\mathbf{\Theta} - \int q_j \left(\prod_{i\neq j} q_i\right) \log q_j d\mathbf{\Theta}}^{\text{Separate the j-th r.v.}} \\
&- \int q_j \left(\prod_{i\neq j} q_i\right) \left(\sum_{i\neq j} \log q_i\right) d\mathbf{\Theta} \\
&= \int q_j \left[\int \prod_{i\neq j} q_i \log p(X, \mathbf{\Theta}) d\theta_i\right] d\theta_j - \int q_j \log q_j d\theta_j + \text{cte} \\
&= \int q_j \log f_j d\theta_j - \int q_j \log q_j d\theta_j + \text{cte} \\
&= -\text{KL}\left(q_j || f_j\right)
\end{aligned} \tag{2.21}$$

where f_j is defined by

$$\log(f_j) = \mathbb{E}_{-q_j}[\log p(\mathbf{X}, \mathbf{\Theta})] + \text{cte} \tag{2.22}$$

where $-q_j$ represents the expectation of all variables except a specific θ_j. Therefore, the KL divergence minimisation between q_j and f_j, maximises the ELBO $L(q_j)$. Thus, we can compute that the optimal value for $\log q_j$ that minimises this KL is given by

$$\log q_j^* = \mathbb{E}_{-q_j}\big[\log p(X, \mathbf{\Theta})\big] + \text{constant.} \tag{2.23}$$

In other words, we are going to optimise q_j by calculating the expectation over all the remaining variables in the model as the mean-field follows [128, 129].

Following this theory, in BPPCA the posterior distribution can be approximated by mean-field VI. Applying it, the posterior $p(\mathbf{\Theta}|\,\mathbf{X})$ is approximated as

$$p(\mathbf{\Theta}|\,\mathbf{X}) \approx q(\mathbf{\Theta}) = q(\mathbf{W})q(\tau)\prod_{n} q(\mathbf{g}_{\mathrm{n},:})\,. \tag{2.24}$$

Now, to obtain each q_j, we apply Eq. (2.23) iteratively to each rv of Eq. (2.24). To do so, one of the rvs is fixed (called q_j) and the expectations of the rest are calculated. Following this approach, the BPPCA update rules for each q_j are calculated and shown in Table 2.1 where $\langle\rangle$ is the expectation operator.

Table 2.1: The expressions of the updated distributions for all BPPCA's rvs obtained by mean-field.

Variable	q^* **distribution**	**Parameters**
$\mathbf{g}_{\mathrm{n},:}$	$\mathcal{N}\big(\mathbf{g}_{\mathrm{n},:}\,\vert\langle\mathbf{g}_{\mathrm{n},:}\rangle, \Sigma_{\mathbf{G}}\big)$	$\langle\mathbf{g}_{\mathrm{n},:}\rangle = \langle\tau\rangle\mathbf{X}\langle\mathbf{W}\rangle\Sigma_{\mathbf{G}}$ $\Sigma_{\mathbf{G}}^{-1} = \mathbf{I}_{\mathrm{K_c}} + \langle\tau\rangle\langle\mathbf{W}^{\mathrm{T}}\,\mathbf{W}\rangle$
$\mathbf{W}^{(\mathrm{m})}$	$\prod_d \mathcal{N}\big(\mathbf{W}\,\vert\langle\mathbf{W}\rangle, \Sigma_{\mathbf{W}}\big)$	$\langle\mathbf{W}\rangle = \langle\tau\rangle\mathbf{X}^T\langle\mathbf{G}\rangle\Sigma_{\mathbf{W}}$ $\Sigma_{\mathbf{W}}^{-1} = I_K + \langle\tau\rangle\langle\mathbf{G}^T\,\mathbf{G}\rangle$
τ	$\Gamma\big(\tau\vert a_\tau, b_\tau\big)$	$a_\tau = \frac{DN}{2} + a^\tau$ $b_\tau = b^\tau + \frac{1}{2}\Big(\sum_n\sum_d \mathbf{x}_{n,d}^2 - \mathrm{Tr}\left(\langle\mathbf{W}\rangle\langle\mathbf{G}^T\rangle\mathbf{X}\right)$ $+\frac{1}{2}\,\mathrm{Tr}\left(\langle\mathbf{W}^{\mathrm{T}}\,\mathbf{W}\rangle\langle\mathbf{G}^{\mathrm{T}}\,\mathbf{G}\rangle\right)\Big)$

2.2.2 Bayesian Inter-Battery FA

In the present work, we have shown the importance of multimodal learning in real-world situations. Building upon BPPCA and extending PPCA, Klami et al. [125] proposed a multimodal version of this technique, named BIBFA. In BIBFA, the latent dimension of $\mathbf{g}_{\mathrm{n},:} \in \mathbb{R}^{1\times K}$, K, is automatically learnt.

In BIBFA, the BPPCA generative model is extended to include multiple input views. Each view is denoted as the m-th view, with a total of M different modalities that can be learnt. The technique proposes learning of a common latent projection space for all M modalities by identifying the inter- and intra-view data correlation. The global shared latent space, represented by $\mathbf{G}$, is integrated with a set of unique projection matrices $\mathbf{W}^{(\mathrm{m})}$, where m ranges from 1 to M, to produce the respective

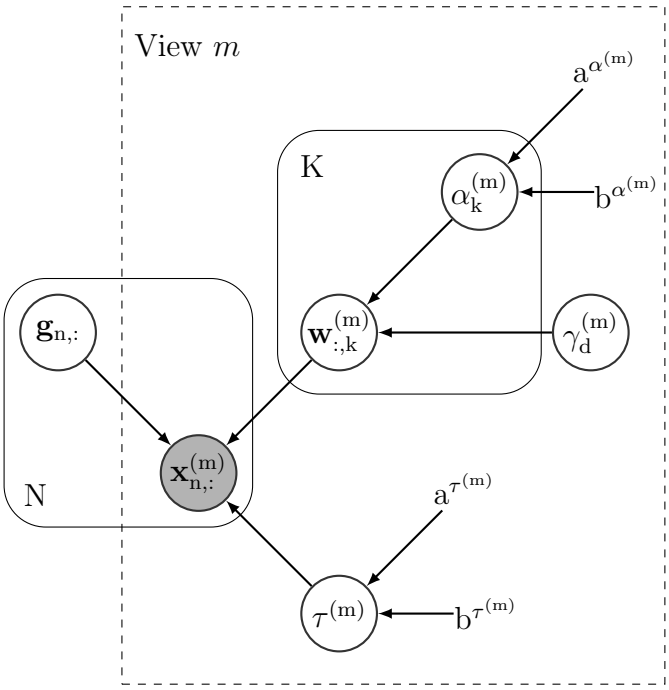


Figure 2.2: Graphical model of BIBFA. Observations are represented by dark circles, whereas random variables are represented by white ones. The rest denotes hyper-parameters

views of the data, denoted as $\mathbf{X}^{(\mathrm{m})}$. Thus, BIBFA defines the joint probability distribution as follows:

$$\mathbf{g}_{\mathrm{n},:} \sim \mathcal{N}(0, I_K)\,, \tag{2.25}$$

$$\mathbf{w}_{:,\mathrm{k}}^{(\mathrm{m})} \sim \mathcal{N}(0, \alpha_{\mathrm{k}}^{(\mathrm{m})^{-1}} I_K)\,, \tag{2.26}$$

$$\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})} \,|\, \mathbf{g}_{\mathrm{n},:} \sim \mathcal{N}(\mathbf{g}_{\mathrm{n},:}\, \mathbf{W}^{(\mathrm{m})^{\mathrm{T}}}, \tau^{(\mathrm{m})^{-1}} I_K)\,, \tag{2.27}$$

$$\alpha_{\mathrm{k}}^{(\mathrm{m})} \sim \Gamma(a^{\boldsymbol{\alpha}^{(\mathrm{m})}}, b^{\boldsymbol{\alpha}^{(\mathrm{m})}})\,, \tag{2.28}$$

$$\tau^{(\mathrm{m})} \sim \Gamma(a^{\tau^{(\mathrm{m})}}, b^{\tau^{(\mathrm{m})}})\,, \tag{2.29}$$

where differences from BPPCA can be observed, as illustrated in the graphical model presented in Fig. 2.2.

In BIBFA, specific parameters $\mathbf{W}^{(\mathrm{m})}$ are used for each view, which is a natural assumption. As the data will differ between views, it follows that $\mathbf{X}^{(\mathrm{m})}$ must also be different for each m modality. Given that each $\mathbf{X}^{(\mathrm{m})}$ corresponds to a different modality, the global $\mathbf{G}$ must be combined differently to generate each modality.

Additionally, a new random variable $\alpha_{\mathrm{k}}^{(\mathrm{m})}$ is introduced, which serves as an ARD prior that induces column sparsity in the columns of the projection matrix $\mathbf{W}^{(\mathrm{m})}$

[130]. To do so, each k-column of $\mathbf{W}^{(\mathrm{m})}$, $\mathbf{w}_{:,\mathrm{k}}^{(\mathrm{m})}$, follows a Gaussian distribution with zero mean and precision $\alpha_{\mathrm{k}}^{(\mathrm{m})}$. For this $\alpha_{\mathrm{k}}^{(\mathrm{m})}$ we assume a Gamma prior distribution allowing the precision to achieve high values. When this occurs, i.e., $\alpha_{\mathrm{k}}^{(\mathrm{m})}$ value increase significantly, all elements of $\mathbf{w}_{:,\mathrm{k}}^{(\mathrm{m})}$ tends to 0. This means that the k-th component of $\mathbf{G}$ does not affect to the generation of $\mathbf{X}^{(\mathrm{m})}$, thus obtaining an automatic latent feature selection. As the $\alpha_{\mathrm{k}}^{(\mathrm{m})}$ values are different for each m modality, this helps to induce projections of different modalities by imposing four possible scenarios:

1. For a specific k column, if $\mathbf{w}_{:,\mathrm{k}}^{(\mathrm{m})} \neq 0$ for all m modalities, it means that this k latent feature is **shared** among all views, i.e., it represents common information between different modalities.

2. For a specific k column, if $\mathbf{w}_{:,\mathrm{k}}^{(\mathrm{m})} = 0$ for all m modalities, it means that this k latent feature is not utilised by **any** views and can be pruned, reducing the dimension of the latent space.

3. For a specific k column, if $\mathbf{w}_{:,k}^{(m=j)} \neq 0, \mathbf{w}_{:,k}^{(m\neq j)} = 0$, it means that this k latent feature is **private** to modality $m = j$.

4. And, for a specific k column, if $\mathbf{w}_{:,k}^{(M_1)} \neq 0, \mathbf{w}_{:,k}^{(M_2)} = 0$, where M_1, M_2 are two different subsets of modalities, it means that this latent feature k is shared by the subset of modalities $[M_1, M_2]$.

The generative model proposed by BIBFA aims to enhance the learning of information from multimodal data while providing interpretability. BIBFA can automatically prune the latent space dimension by identifying latent features that are not used by any view in the reconstruction process.

Variational inference

As was proposed in BPPCA, mean-field is again used to calculate the approximate posterior distribution of all r.v. $\mathbf{\Theta}$. Therefore, starting from Eq. (2.24) it can be extended to cover multiple views and also the new ARD prior as follows:

$$p(\mathbf{\Theta}|\, \mathbf{X}^{(\mathrm{m})}) \approx q(\mathbf{\Theta}) = \prod_{m=1}^{M} \left(q(\mathbf{W}^{(\mathrm{m})}) q(\tau^{(\mathrm{m})}) \prod_{k}^{K} q(\alpha_{\mathrm{k}}^{(\mathrm{m})}) \right) \prod_{n=1}^{N} q(\mathbf{g}_{\mathrm{n},:}) \,, \tag{2.30}$$

where the main differences are that now BIBFA proposes two different products, one per modality that contains its private rv $\tau^{(\mathrm{m})}, \mathbf{W}^{(\mathrm{m})}, \alpha_{\mathrm{k}}^{(\mathrm{m})}$, and another by samples that only refer to the global shared latent space.

Moreover, in BIBFA, they proposed a prediction formulation where, given a new test sample, denoted $\mathbf{x}_{*,:}$, and a subset of observed views M_o, a predictive

distribution for an unobserved view u can be obtained. To do so, the posterior of the unobserved latent variable $\mathbf{g}_{*,:}$ is firstly calculated as:

$$\begin{aligned} p\big(\mathbf{g}_{*,:}\,|\mathbf{x}_{*,:}^{(M_o)}, D_{\text{train}}\big) &= \int p\big(\mathbf{g}_{*,:}\,|\mathbf{x}_{*,:}^{(M_o)}, \boldsymbol{\Theta}\big) p(\boldsymbol{\Theta}|D_{\text{train}}) d\boldsymbol{\Theta} \\ &\approx \overbrace{p\big(\mathbf{g}_{*,:}\,|\mathbf{x}_{*,:}^{(M_o)}, \langle\boldsymbol{\Theta}\rangle\big)}^{\text{Point estimate}} \\ &= \mathcal{N}(\mathbf{g}_{*,:}\,|\langle\mathbf{g}_{*,:}\rangle, \Sigma_{\mathbf{g}_{*,:}})\,, \end{aligned} \tag{2.31}$$

where, using point estimation Monte-Carlo approximation, we assume that the mean value of the $\boldsymbol{\Theta}$ is rich enough to compute the prediction being now a Gaussian distribution. Finally, by applying the Bayes rule, we can define the posterior $p\big(\mathbf{g}_{*,:}\,|\mathbf{x}_{*,:}^{M_o}, \boldsymbol{\Theta}\big)$ as a Gaussian distribution

$$\langle\mathbf{g}_{*,:}\rangle = \sum_{m}^{M_o} \big(\langle\tau^{(\mathrm{m})}\rangle \mathbf{x}_{*,:}^{m} \langle\mathbf{W}^{(\mathrm{m})}\rangle\big) \Sigma_{\mathbf{g}_{*,:}}\,, \tag{2.32}$$

$$\Sigma_{\mathbf{g}_{*,:}}^{-1} = I_K + \sum_{m}^{M_o} \Big(\langle\tau^{(\mathrm{m})}\rangle \langle\mathbf{W}^{(\mathrm{m})^{\mathrm{T}}}\,\mathbf{W}^{(\mathrm{m})}\rangle\Big)\,, \tag{2.33}$$

being $\langle\mathbf{g}_{*,:}\rangle$ the mean and $\Sigma_{\mathbf{g}_{*,:}}$ the covariance.

Now, using this posterior, the predictive posterior distribution of the unobserved modalities can be computed as follows:

$$p\big(\mathbf{x}_{*,:}^{(u)}|\mathbf{x}_{*,:}^{(M_o)}, \boldsymbol{\Theta}\big) = \int p\big(\mathbf{x}_{*,:}^{(u)}|\,\mathbf{g}_{*,:}, \boldsymbol{\Theta}\big) p\big(\mathbf{g}_{*,:}\,|\mathbf{x}_{*,:}^{(M_o)}, \Theta\big) d\Theta\,. \tag{2.34}$$

As both are Gaussians, any prediction of an unobserved modality can be sampled from another Gaussian distribution following:

$$\langle\mathbf{x}_{*,:}^{(u)}\rangle = \langle\mathbf{g}_{*,:}\rangle \langle\mathbf{W}^{(u)^T}\rangle\,, \tag{2.35}$$

$$\Sigma_{\mathbf{x}_{*,:}^{(u)}}^{-1} = \langle\tau^{(u)^{-1}}\rangle I_D + \langle\mathbf{W}^{(u)}\rangle \Sigma_{\mathbf{g}_{*,:}} \langle\mathbf{W}^{(u)^T}\rangle\,. \tag{2.36}$$

The BIBFA model is presented as a method for learning from multimodal data and making predictions of unobserved parts, with the ability to prune latent factors automatically. However, it is acknowledged that this model has limitations in handling heterogeneous data types and dealing with missing data in semi-supervised environments which are common problems in real case scenarios.

2.2.3 Sparse Semi-supervised Heterogeneous Interbattery Bayesian Analysis

In [2], an extension of BIBFA is proposed, called Sparse Semi-supervised Heterogeneous Interbattery Bayesian Analysis (SSHIBA). This work extends the limitations

of BIBFA to handle different heterogeneous data types, such as multilabel and binary, and allows it to operate in semi-supervised scenarios with missing data.

The SSHIBA [2] model presents a solution to heterogeneous multi-view problems with samples represented in M different modalities. Each m modality, or view, can be a multilabel, binary, real, or categorical object. The general model framework, shown in Fig. 2.3a, considers that the sample n-th of the m-th view, $\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})} \in \mathbb{R}^{1\times D}$, can be projected into a latent space of lower dimension, $\mathbf{g}_{\mathrm{n},:} \in \mathbb{R}^{1\times K}$ where K is the number of latent factors of this common space. For this low-dimensional latent space, the following prior is assumed:

$$\mathbf{g}_{\mathrm{n},:} \sim \mathcal{N}(0, I_K)\,. \tag{2.37}$$

As seen in Fig. 2.3b, the data point n corresponding to the view m, $\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})}$, can be generated by linearly combining $\mathbf{g}_{\mathrm{n},:}$ with a projection matrix $\mathbf{W}^{(\mathrm{m})} \in \mathbb{R}^{D\times K}$, that is,

$$\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})} = \mathbf{g}_{\mathrm{n},:}\,\mathbf{W}^{(\mathrm{m})^{\mathrm{T}}} + \boldsymbol{b}^{(\mathrm{m})} + \epsilon^{(m)}\,, \tag{2.38}$$

where $\epsilon^{(m)}$ is Gaussian noise with zero mean and whose noise power, $\tau^{(\mathrm{m})}$ follows a Gamma distribution:

$$\tau^{(\mathrm{m})} \sim \Gamma\left(a^{\tau^{(\mathrm{m})}}, b^{\tau^{(\mathrm{m})}}\right). \tag{2.39}$$

Therefore, we assume that $\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})}$ given $\mathbf{g}_{\mathrm{n},:}$ follows:

$$\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})} \,|\, \mathbf{g}_{\mathrm{n},:} \sim \mathcal{N}\left(\mathbf{g}_{\mathrm{n},:}\,\mathbf{W}^{(\mathrm{m})^{\mathrm{T}}} + \boldsymbol{b}^{(\mathrm{m})}, \tau^{(\mathrm{m})^{-1}} I_{D_m}\right). \tag{2.40}$$

SSHIBA also includes a double ARD prior [124], column- and row-wise, on each view's projection $\mathbf{W}^{(\mathrm{m})}$ matrix:

$$\mathrm{w}_{\mathrm{d,k}}^{(\mathrm{m})} \sim \mathcal{N}\left(0, (\gamma_{\mathrm{d}}^{(\mathrm{m})}\,\alpha_{\mathrm{k}}^{(\mathrm{m})})^{-1}\right), \tag{2.41}$$

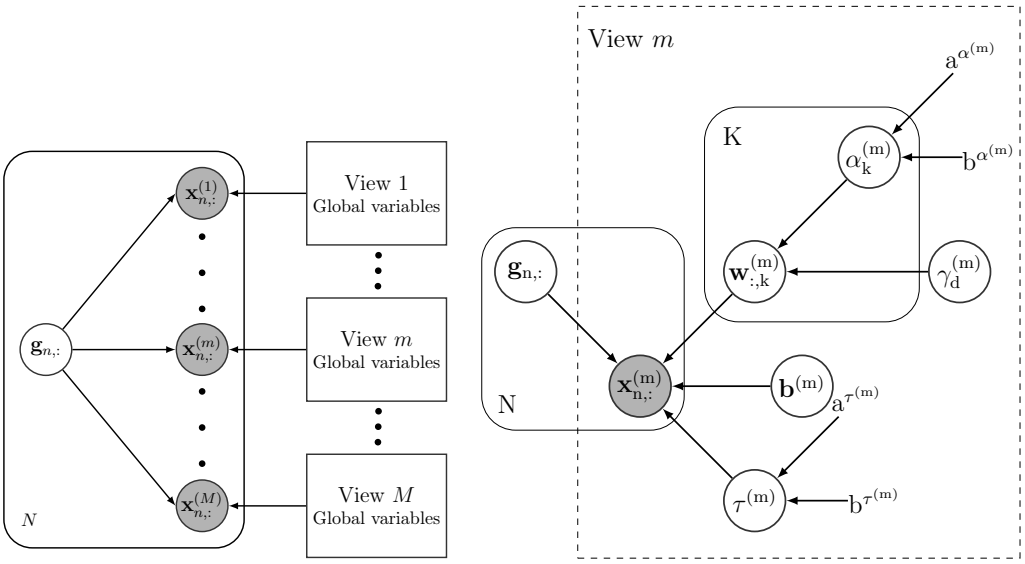
$$\boldsymbol{b}^{(\mathrm{m})} \sim \mathcal{N}(0, I_{D_m})\,, \tag{2.42}$$

$$\alpha_{\mathrm{k}}^{(\mathrm{m})} \sim \Gamma\left(a^{\boldsymbol{\alpha}^{(\mathrm{m})}}, b^{\boldsymbol{\alpha}^{(\mathrm{m})}}\right), \tag{2.43}$$

$$\gamma_{\mathrm{d}}^{(\mathrm{m})} \sim \Gamma\left(a^{\boldsymbol{\gamma}^{(\mathrm{m})}}, b^{\boldsymbol{\gamma}^{(\mathrm{m})}}\right). \tag{2.44}$$

On the one hand, $\alpha_{\mathrm{k}}^{(\mathrm{m})}$ eliminates the latent factors cross-validation process. It implies that each $\mathbf{W}^{(\mathrm{m})}$ effectively selects which part of the global latent space $\mathbf{G}$ is specific to each m view (intraview) or shared among views (interview). On the other hand, $\gamma_{\mathrm{d}}^{(\mathrm{m})}$ creates sparsity over the D features, thus proposing a feature selection.

Given the observed data, the model is trained by evaluating the posterior distribution of all rv. However, these posteriors cannot be calculated precisely, and



(a) Model multi-view. (b) Particular m view.

Figure 2.3: Basic structure of SSHIBA. Observations are indicated by dark spheres, rv by white ones and hyper-parameters by non-circles. While Figure 2.3a depicts the latent space projection $\mathbf{g}_{\mathrm{n},:}$ encompassing all m input perspectives, Figure 2.3b exemplifies how an actual m-view is modeled. Source: [2]

thus are estimated through the mean-field VI [127] to approximate the posterior distribution with a completely factorised variational family q as:

$$p(\mathbf{\Theta}|\mathbf{X}^{(m)}) \approx \prod_{m=1}^{M}\left(q(\mathbf{W}^{(\mathrm{m})})q(\tau^{(\mathrm{m})})\prod_{k=1}^{K}q(\alpha_{\mathrm{k}}^{(\mathrm{m})})\prod_{d=1}^{D_m}q(\gamma_{\mathrm{d}}^{(\mathrm{m})})\right)\prod_{n=1}^{N}q(\mathbf{g}_{\mathrm{n},:}), \tag{2.45}$$

where $\mathbf{\Theta}$ is a vector of all rv in the model, shown in Fig. 2.3b. Therefore, combining the mean-field approximation of the posterior distribution with the ELBO creates a workable coordinate optimisation procedure, in which each rv is computed while keeping the rest constant, as detailed below:

$$q^*(\theta_i) \propto \mathbb{E}_{\mathbf{\Theta}_{-i}}[\log p(\mathbf{\Theta}_{-i}, \theta_{\mathbf{i}1,:}, ..., \theta_{\mathbf{i}N,:})]\,, \tag{2.46}$$

with $\mathbf{\Theta}_{-i}$ represents the set of all possible rv excluding θ_i. The rules for updating every rv can be seen in Table 2.2.

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 marginalise any missing values in the data and predict test samples for any view by sampling from its variational distribution.

Table 2.2: Distribution for all Θ obtained by mean-field approximation. Source: [2]

Variable	q^* distribution	Parameters
$\mathbf{g}_{\mathrm{n},:}$	$\mathcal{N}\big(\mathbf{g}_{\mathrm{n},:}\,\big\vert\mu_{\mathbf{g}_{\mathrm{n},:}}, \Sigma_{\mathbf{G}}\big)$	$\mu_{\mathbf{g}_{\mathrm{n},:}} = \sum_{\mathrm{m}=1}^{\mathrm{M}} \Big(\langle \tau^{(\mathrm{m})} \rangle \mathbf{X}^{(m)} \langle \mathbf{W}^{(\mathrm{m})} \rangle \Sigma_{\mathbf{G}} \Big)$ $\Sigma_{\mathbf{G}}^{-1} = I_K + \sum_{\mathrm{m}=1}^{\mathrm{M}} \Big(\langle \tau^{(\mathrm{m})} \rangle \langle \mathbf{W}^{(\mathrm{m})^{\mathrm{T}}} \mathbf{W}^{(\mathrm{m})} \rangle \Big)$
$\mathbf{W}^{(\mathrm{m})}$	$\prod_{\mathrm{d}=1}^{\mathrm{D_m}} \mathcal{N}\big(\mathbf{w}_{\mathrm{d},:}^{(\mathrm{m})}\,\big\vert\mu_{\mathbf{w}_{\mathrm{d},:}^{(\mathrm{m})}}, \Sigma_{\mathbf{W}^{(\mathrm{m})}}\big)$	$\mu_{\mathbf{w}_{\mathrm{d},:}^{(\mathrm{m})}} = \langle \tau^{(\mathrm{m})} \rangle \mathbf{X}^{(m)^T} \langle \mathbf{G} \rangle \Sigma_{\mathbf{W}^{(\mathrm{m})}}$ $\Sigma_{\mathbf{W}^{(\mathrm{m})}}^{-1} = \mathrm{diag}(\langle \boldsymbol{\alpha}^{(\mathrm{m})} \rangle) \langle \gamma_{\mathrm{d}}^{(\mathrm{m})} \rangle + \langle \tau^{(\mathrm{m})} \rangle \langle \mathbf{G}^{\mathrm{T}} \mathbf{G} \rangle$
$\boldsymbol{\alpha}^{(\mathrm{m})}$	$\prod_{\mathrm{k}=1}^{\mathrm{K_c}} \Gamma\big(\alpha_{\mathrm{k}}^{(\mathrm{m})}\,\big\vert a_{\alpha_{\mathrm{k}}^{(\mathrm{m})}}, b_{\alpha_{\mathrm{k}}^{(\mathrm{m})}}\big)$	$a_{\alpha_{\mathrm{k}}^{(\mathrm{m})}} = \frac{D_m}{2} + a^{\boldsymbol{\alpha}^{(\mathrm{m})}}$ $b_{\alpha_{\mathrm{k}}^{(\mathrm{m})}} = b^{\boldsymbol{\alpha}^{(\mathrm{m})}} + \frac{1}{2} \sum_{\mathrm{d}=1}^{\mathrm{D_m}} \langle \gamma_{\mathrm{d}}^{(\mathrm{m})} \rangle \langle \mathrm{w}_{\mathrm{d,k}}^{(\mathrm{m})}\, \mathrm{w}_{\mathrm{d,k}}^{(\mathrm{m})} \rangle$
$\tau^{(\mathrm{m})}$	$\Gamma\big(\tau^{(\mathrm{m})}\,\big\vert a_{\tau^{(\mathrm{m})}}, b_{\tau^{(\mathrm{m})}}\big)$	$a_{\tau^{(\mathrm{m})}} = \frac{D_m N}{2} + a^{\tau^{(\mathrm{m})}}$ $b_{\tau^{(\mathrm{m})}} = b^{\tau^{(\mathrm{m})}} + \frac{1}{2} \Big(\sum_{\mathrm{n}=1}^{\mathrm{N}} \sum_{\mathrm{d}=1}^{\mathrm{D_m}} \mathrm{x}_{\mathrm{n,d}}^{(\mathrm{m})^2}$ $-2\,\mathrm{Tr}\,\big(\langle \mathbf{W}^{(\mathrm{m})} \rangle \langle \mathbf{G}^{\mathrm{T}} \rangle \mathbf{X}^{(m)}\big)$ $+\,\mathrm{Tr}\,\big(\langle \mathbf{W}^{(\mathrm{m})^{\mathrm{T}}} \mathbf{W}^{(\mathrm{m})} \rangle \langle \mathbf{G}^{\mathrm{T}} \mathbf{G} \rangle\big)\Big)$

Table 2.3 presents an overview of the features and limitations of current FA models. These models generally utilise linear decomposition methods, with some incorporating additional features. For example, the BIBFA model handles multi-view data, while the SSHIBA model incorporates both heterogeneity and semi-supervised learning. Despite these improvements, there are limitations to the versatility of these models in certain contexts, specifically in microbiology scenarios. In particular, these models may not be suitable for data that exhibit non-linear relationships and may struggle with high-dimensional data where the number of variables, D, exceeds the number of samples, N. This poses a computational bottleneck for SSHIBA, as it requires the inversion of the matrix $\Sigma_{\mathbf{W}^{(\mathrm{m})}}^{-1} \in \mathbb{R}_{+}^{D \times D}$. To address this limitation, this thesis presents various extensions to the SSHIBA model to better suit it for microbiological data. In Chapter 3, a kernel-based extension, KSSHIBA [122], is proposed to handle high-dimensional data, such as that obtained from MALDI-TOF MS, while exploiting non-linear relationships in the data. Then, in Chapter 4, a VAE-based extension, FA-VAE [123], is proposed to exploit non-linear relationships and further expand the types of heterogeneous data to handle.

Table 2.3: Features and limitations of current FA models.

	BPPCA	BIBFA	SSHIBA
Bayesian FA	✓	✓	✓
Multiview	✗	✓	✓
Heterogeneous	✗	✗	✓
Semi-supervised	✗	✗	✓
High-dimesional data	✗	✗	✗
Non-linear data	✗	✗	✗

2.3 Probabilistic Deep Learning

Probabilistic Deep Learning (DL) is a subfield of DL that deals with the integration of probabilistic methods with deep Neural Network (NN). In probabilistic DL, models are trained to make probabilistic predictions that provide a measure of the uncertainty associated with the model's output. This uncertainty can be used to estimate the confidence in the model predictions, which can be particularly useful in decision-making applications where the cost of a wrong prediction can be high.

One of the key examples is Deep Gaussian Processes (DGP). A DGP is a hierarchical Bayesian model that extends the capabilities of traditional GP algorithms [131] by incorporating multiple layers of latent variables. The hierarchical structure of DGP allows for the modelling of complex, non-linear relationships in the data. DGP models have been used in various tasks such as regression [132] or active learning [133].

Finally, VAEs are a prominent class of generative models in the field of probabilistic DL. These models are trained to reconstruct input data using a probabilistic approach, and consist of two main components: an encoder network that maps input data to a set of latent variables and a decoder network that maps the latent variables back to the original data space. VAEs have gained recognition for their flexibility in handling a wide range of data types, including both continuous and discrete variables, owing to their encoder-decoder architecture. The architecture of VAEs can be easily adapted to new types of data, such as weather prediction by using a MLP as shown in [134] or detect traffic issues from real videos by using a CNN as shown in [135]. Furthermore, VAEs have also been applied to time series analysis in anomaly detection problems by using a LSTM architecture as shown in [136].

The VAE model is going to be the focal point of a portion of the models analysed in this dissertation. We will conduct an in-depth analysis of VAE models to assess their adaptability in handling diverse types of data. Subsequently, the following sections will explore the nowadays techniques to handle multi-view information in

a hierarchical manner. This analysis will provide a comprehensive understanding of VAE models and their capabilities in data analysis.

2.3.1 Variational AutoEncoders

In order to gain a proper understanding of a VAE, it is important to first consider the basics of the non-probabilistic AutoEncoder (AE). A non-probabilistic AE consists of an encoder and a decoder. The goal of the AE is, using an encoder, to create a smaller representation of the input information that still contains critical information about it so that the original input can be reconstructed by a decoder. An intuition for how this process works can be found in Fig. 2.4, which illustrates how an AE encodes a face into latent attributes that characterise the image, allowing a decoder to reconstruct the same image.

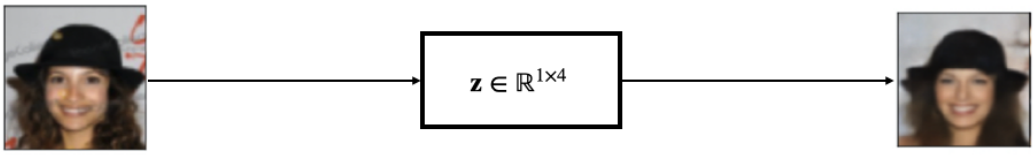


Figure 2.4: AutoEncoder example

A VAE is a probabilistic approach to describing an observation in a latent space. Rather than describing an observation as a fixed set of attributes, a VAE describes a probability distribution for each latent attribute. This can be visualised in Fig. 2.5, where, instead of a deterministic value for each attribute, we have a Gaussian probability distribution that describes each attribute. Given a face, it is encoded into a latent Gaussian probability distribution that characterises each attribute. The decoder thus samples that distribution and generates an infinite number of images that are similar to the original one.

Thus, a VAE that assumes that there exists a hidden latent variable $\mathbf{z} \in \mathbb{R}^{1\times K}$ that is capable of generating observations $\mathbf{x} \in \mathbb{R}^{1\times D}$ through a non-linear model. It is worth noting that since it is compressed, $K << D$, meaning that the latent dimension is always lower than the original, as its purpose is to condense the information. Following the intuition of Figs. 2.4 and 2.5, the observation $\mathbf{x} \in \mathbb{R}^{1\times64\times64}$ is an image of a face with 64×64 pixels and compressed into a latent variable $\mathbf{z} \in \mathbb{R}^{1\times4}$ where we have four different latent variables that can be interpreted to explain specific attributes of the image such as gender or smiling. In other words, D represents the original observation dimension (the number of pixels in the image), and K is a hyperparameter that denotes the dimension of the

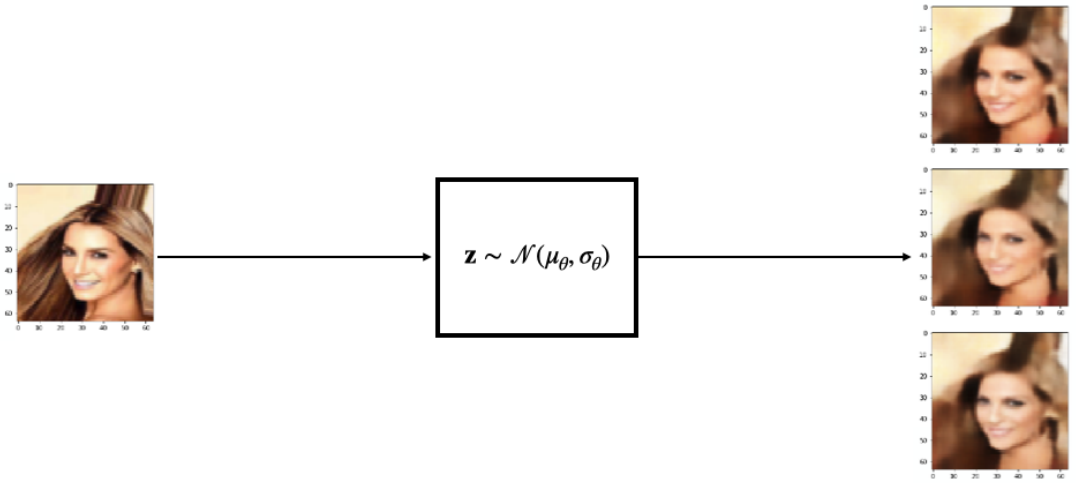


Figure 2.5: Variational AutoEncoder example. Given an example image, various images are generated by sampling from a multivariate Gaussian distribution in the latent space.

hidden latent features. These latent variables are inferred from the observations as follows:

$$\overbrace{p_\phi(\mathbf{z}|\mathbf{x})}^{encoder} = \frac{\overbrace{p_\theta(\mathbf{x}|\mathbf{z})}^{decoder} p(\mathbf{z})}{p(\mathbf{x})}\,, \tag{2.47}$$

where, $p_\theta(\mathbf{x}|\mathbf{z})$ represents the decoder, which is a parametric NN with parameters θ whose architecture is a hyperparameter; $p(\mathbf{z})$ represents the prior assumed over the latent variables, where in the classical VAE it is typically assumed to be a normal distribution $\mathcal{N}(0, 1)$ according to [137]; and $p(\mathbf{x})$ is the marginal likelihood of the data. However, computing this marginal likelihood is intractable and, thus, it is not possible to use Eq. (2.47) to construct the encoder. Instead, we use VI to approximate it with a tractable parameterised family of distributions defined as $q_\eta(\mathbf{z}|\mathbf{x})$. To do this, we minimise the KL divergence between these two distributions,

$$\min_\eta \mathrm{KL}(q_\eta(\mathbf{z}|\mathbf{x})||p_\phi(\mathbf{z}|\mathbf{x}))\,, \tag{2.48}$$

which is equivalent to maximising the following ELBO [137]:

$$\mathcal{L}_{\theta,\eta} = \underbrace{\mathbb{E}_{q_\eta(\mathbf{z}|\mathbf{x})}[\log\big(p_\theta(\mathbf{x}|\mathbf{z})\big)]}_{\mathbf{I}} - \underbrace{\mathrm{KL}\big(q_\eta(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\big)}_{\mathbf{II}}\,, \tag{2.49}$$

where now $q_\eta(\mathbf{z}|\mathbf{x})$ is the encoder which is a parametric NN with parameters η whose architecture is a hyperparameter; the likelihood term **I** drives the reconstruction of the observations given by the encoder; and **II** ensures that the learnt

encoder distribution $q_\eta(\mathbf{z}|\mathbf{x})$ is restricted to the prior distribution $p(\mathbf{z})$ working as a regularisation term.

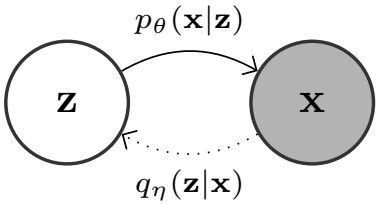


Figure 2.6: VAE structure, where $q_\eta(\mathbf{z}|\mathbf{x})$ is the encoder network and $p_\theta(\mathbf{z}|\mathbf{x})$ is the decoder network.

Since Kingma and Welling introduced the original VAE in 2013 [137], numerous variants and improvements have been developed. In this thesis, we focus on β-VAE [138, 139], a modification of the original VAE that introduces a hyperparameter β to the regularisation term in the ELBO (term **II** in Eq. (2.49)). This hyperparameter, which is typically set to $\beta>1$, leads to more disentangled latent representations of the latent space $\mathbf{z}$.

It is worth noting that, traditionally, both vanilla VAE and β-VAE are not designed for heterogeneous and mixed data types. They are mainly applied to a single type of data. In recent years, various researchers have conducted a deep analysis that extends the applicability of VAEs to heterogeneous problems. The next subsection reviews the background of different heterogeneous VAEs.

2.3.2 Heterogeneous Variational AutoEncoders

In the literature, different approaches have addressed the problem of heterogeneous data. With respect to this problem, VAEs have been widely explored and extended to deal with mixed data types. The key problem when dealing with heterogeneous data types in VAEs is the approach that the author takes to mix their latent representation. As explained above, given input data $\mathbf{X}$, the vanilla VAE encodes a latent representation $\mathbf{Z}$. However, given multiple input data modalities, where $\mathbf{X}^{(m)}$ is a specific modality of M possibles, the key question is how to mix different encodings without losing information. Previous research has shown that there are two current approaches nowadays: (i) differentiating each modality in both input and latent space using specific likelihoods for each data type, and (ii) only differentiating them in the input space. The following subsections detail both schemes.

Specific likelihoods per data type

Regarding the first approach, the most known heterogeneous VAE is the Heterogeneous Incomplete VAE (HIVAE) [3]. HIVAE proposes to use different likelihoods for each data type, that is, feeding the encoder with the different data types and then using a specific likelihood for real ($\mathbb{R}$), positive ($\mathbb{R}^{++}$), count ($\mathcal{N}$), categorical, and ordinal data. Thus, an ELBO is maximised for all observations where each data type is fused by combining the different likelihoods. Its graphical generative model is shown in Fig. 2.7 where γ_{nd} determines the specific parameters of each likelihood type and $\mathbf{s}_n$ determines which component of the Gaussian mixture generates each latent space $\mathbf{z}_n$. Then, they induce the hierarchy by means of $\mathbf{y}_n$, which is an intermediate representation of all data types.

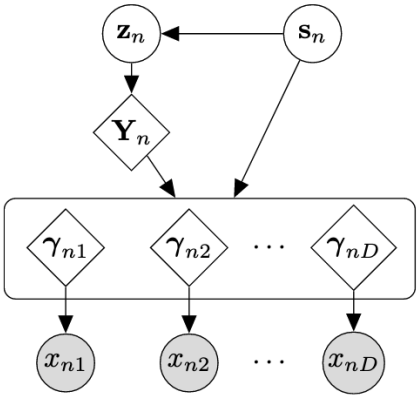


Figure 2.7: HIVAE graphical model. Source: [3].

However, the idea of combining information directly in the latent space is not novel, as it has been previously studied in the Joint Multimodal Variational Autoencoder (JMVAE) proposed by Suzuki et al. in [140]. The graphical model of the JMVAE is depicted in Figure 2.8. In this work, the authors proposed the use of different encoders for each type of data and specific likelihoods, such as Gaussian for real data and Bernoulli for binary data.

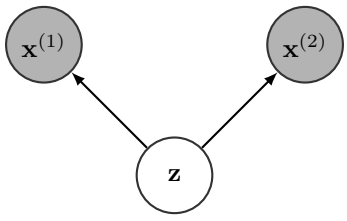


Figure 2.8: JMVAE graphical generative model.

Other authors, such as Barrejón et al. [141], extended HIVAE [3] to work with sequential data driven by LSTMs. Moreover, they also provided a robust model that can handle missing data in heterogeneous time series by adding an amortised VI, i.e., introducing a NN that estimates the missing values. As in HIVAE, they proposed having a specific likelihood for each data type.

In summary, previous studies have typically maximised an ELBO through a weighted sum of partial ELBOs or losses, one per data type. However, as pointed out by Javaloy et al. in [142], this approach can lead to negative transfer. To address this issue, Javaloy et al. proposed the use of a multitask learning approach that homogenises the gradients between tasks before the weighted sum, thus ensuring that no task is overlooked in the optimisation process.

Common likelihoods for all data types

Other authors point out that it is not necessary to differentiate each data type in the latent space. They claim that combining different likelihoods is difficult due to different magnitudes and training speeds. Instead, they proposed to handle each data type by different and specific encoder-decoder architectures and then fuse the latent spaces directly.

Building upon the idea of JMVAE [140], which utilises different likelihoods for each data type, Vedantam et al. [143], proposed an extension. Here the authors model each likelihood as Gaussian and mix them by a Product of Experts (PoE) [144], following an intuition from FA [145]. Additionally, to make the posterior more well-behaved and close to spherical, they introduce a prior $p(\mathbf{z})$ to the PoE as a universal expert.

A similar approach was used in Variational Selective AE (VSAE) [4]. The authors factorised the latent space as a product of Gaussians, decoupling the heterogeneity in the input space but integrating it into the latent space. As seen in Fig. 2.9, the authors proposed a specific encoder for each data type, indicated by red, blue, and green. Additionally, a separate encoder is used to handle missing information, and a gate is utilised to decide whether to use the normal encoder or the missing encoder. Afterwards, the model samples from each posterior distribution q and concatenates all latent spaces into a common one. Finally, specific decoders handle each data type.

The idea of creating a specific private latent space for each type of data and then mixing them was previously proposed by other authors, for instance, in the Multimodal Variational Autoencoder (MVAE) proposed by Wu et al. in [5]. The key difference is that they fused all private information using a PoE [144], as depicted in Fig. 2.10. This is an efficient way of training the joint probability of all private spaces, creating a common shared space.

Other authors proposed to follow the intuition of FA and separate private and

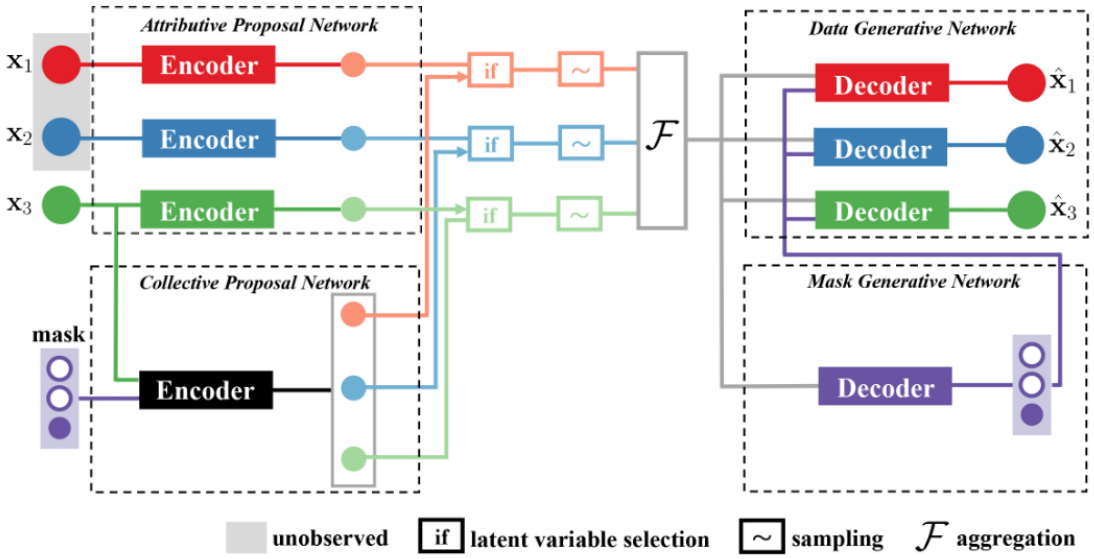


Figure 2.9: VSAE model overview. Source: [4].

common information shared between views. In [6] the authors determine that the different data modalities share common information but also have private information that defines why they are different modalities. For example, a flower can be described by both an image and text, but each provides specific information. Therefore, they proposed a private latent variable for each type of data and a shared one between all modalities, as shown in Fig. 2.11. They follow the same approach as in the MVAE [5], defining specific encoders for every data type, and the shared latent space is constructed by PoE of all private representations. While the MVAE decoder only uses the shared information, the Disentangled Multimodal

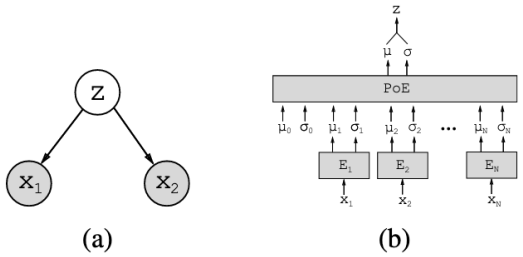


Figure 2.10: MVAE model. (a) represents the generative model, equal to JMVAE but (b) denotes the difference in construction, as seen a PoE is used to mix the different private latent representations. Source: [5].

VAE (DMVAE) decoder uses both the shared and private representations as input, as they believe that combining both improves disentangled representations.

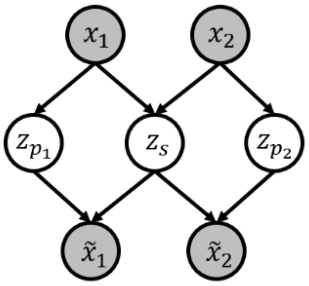


Figure 2.11: DMVAE graphical model. Source: [6].

An extension of the DMVAE is the Multi-VAE [7] where also a private latent space for each view is used, but the shared latent space is now a categorical latent variable following a Gumbel softmax reparametrization trick [146], as seen in Fig. 2.12. Thus, they simply concatenate both latent spaces and proposed different decoders per view.

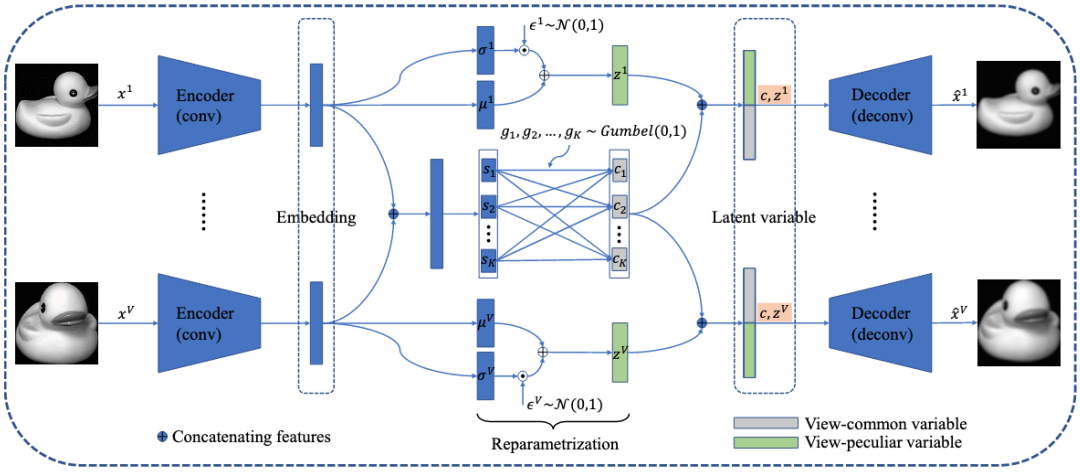


Figure 2.12: Multi-VAE architecture proposal. Source: [7].

Yuge Shi et al. generalised MVAE [5] into MMVAE [147]. Instead of following the product of experts as seen in DMVAE [6] or MVAE [5], they outperformed the cross-generation of images between different data modalities by proposing a change from PoE to a Mixture of Experts (MoE) claiming that PoE leads to a miscalibration of experts. They also proposed using Importance Weighted AutoEncoder (IWAE) [148] instead of vanilla VAE [137] for each specific data type, as it computes a

tighter ELBO by weighting importance and provides a higher entropy.

Other approaches simplify the architecture; for example in AMVAE [8] the authors substitute the PoE of DMVAE [6] or the MoE of MMVAE [147] by a fully connected layer that yields promising results, as seen in Fig. 2.13.

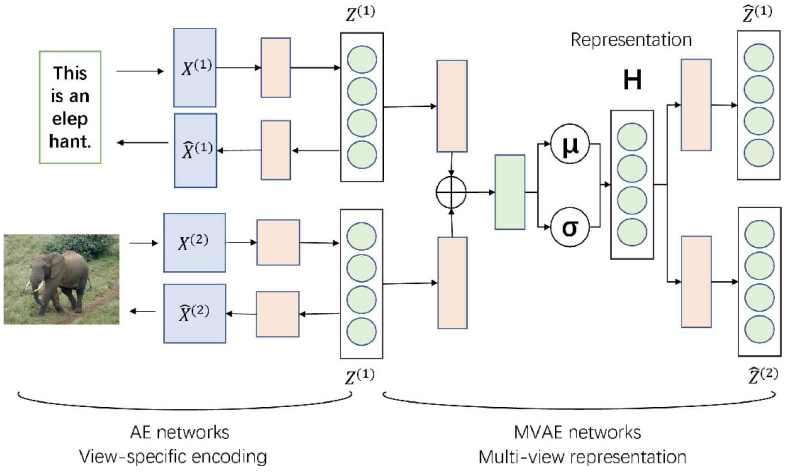


Figure 2.13: AMVAE architecture proposal. Source: [8].

However, other researchers argue that simplicity is insufficient to capture the complexity inherent in the data. The VAEM [9] is an example of these complex models. This technique utilises a hierarchical structure to handle common information between data types and private information that characterises each type. The authors proposed a training process divided into two stages. In the first phase, an individual marginal VAE is trained for every variable, adapting to the nature of each data type, as represented by the inner square of Figure 2.14. In the second stage, a top-level VAE is trained whose inputs are all the hidden representations created by each marginal VAE, represented by h_n in Figure 2.14. The VAEM thus consists of a hierarchical structure of VAEs, where the top-level VAE uses a VampPrior [149] to handle the hierarchical dependency as a mixture of Gaussians.

Peis et al. 2022 [10] improved the VAEM by substituting the second phase with an arbitrary hierarchy of latent variables. While VAEM (Fig. 2.14) uses a one-level hierarchy called h_n, HHVAEM (Fig. 2.15) generalises it to any arbitrary level (L) of hierarchy. This new hierarchical structure better captures the dependencies between each dimension by balanced Gaussian likelihoods. Additionally, they improved posterior sampling by means of Hamiltonian Monte Carlo. Finally, they can handle missing data represented as y_v in Fig. 2.15.

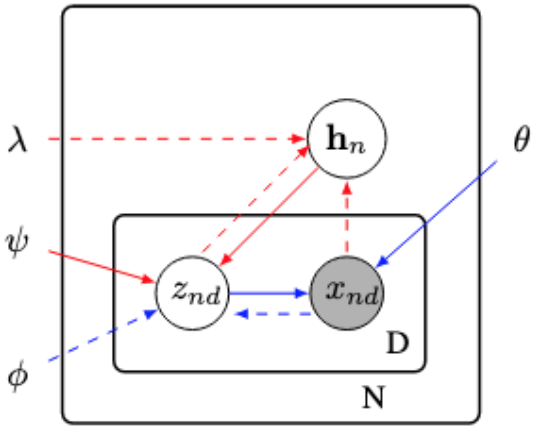


Figure 2.14: VAEM graphical model. Source: [9].

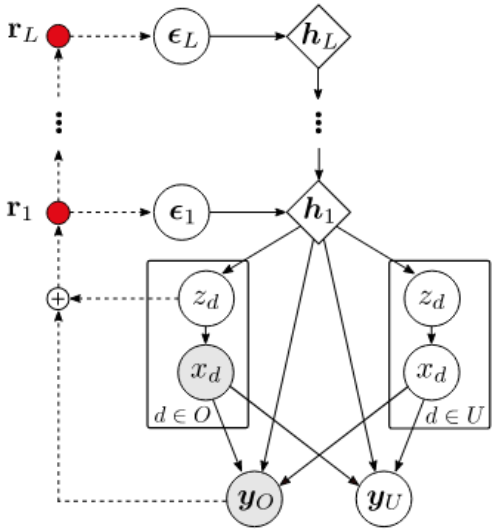


Figure 2.15: HHVAEM graphical model. Source: [10].

As seen in this literature review, there are different approaches to handling multi-view data. However, some of them rely on a complex hierarchical deep-latent representation that projects the data correlation among views. Other works proposed an inference approach that hinders training and obscures the interpretability of the model. Finally, others consist of a simpler solution that works on well-defined benchmark datasets, but its implementation in real-world problems may not work.

In this study, a novel paradigm is put forth that incorporates and extends existing literature proposals. Drawing from the hierarchy in the latent space presented in HIVAE [3] and its extensions, and the intuition of utilising specific encoder-decoder architectures for each data view exemplified in MVAE [5] and subsequent models, our approach relies on FA for its ability to provide a modular and interpretable latent space. By combining the versatility of VAEs in handling various data types with the interpretability of FA, we proposed the FA Variational Autoencoder (FA-VAE) model in Chapter 4.

CHAPTER 3

KERNELISED SSHIBA

In this chapter, we introduce a new technical contribution in the field of Bayesian statistics. Our approach involves the extension of SSHIBA to a kernelised version, which enables the handling of non-linear high-dimensional data. Unlike other Bayesian kernel methods, KSSHIBA not only inherits the advantages of SSHIBA, such as multi-view modularity, the capability to handle missing data and heterogeneous data sources in an explainable way, but also adds the exploitation of non-linear relationships in the data. The proposed method addresses the growing challenge of handling heterogeneous multi-view non-linear data in various disciplines, particularly in medicine, by introducing a more advanced and adaptable approach. Through rigorous experimentation and analysis, we demonstrate the superiority of KSSHIBA over current models in terms of compression efficiency, prediction accuracy, and interpretability of the results. This novel approach has significant potential to foster new insights and drive progress in various fields such as neuroscience [150] and microbiology [45].

This model, KSSHIBA, is published at Neurocomputing journal from Elsevier [122]. KSSHIBA was developed in collaboration with Dr. Carlos Sevilla-Salcedo from Aalto University, Finland. While Dr. Sevilla-Salcedo applied the extension to high-dimensional neuroscience data, we focus on its use in exploiting non-linear relationships between MALDI-TOF MS data.

Thus, in this chapter, we provide a comprehensive overview of the theoretical model of KSSHIBA and in Chapter 5 we present its application to bacteria resistance prediction to antibiotics. Building upon the background in Bayesian statistics and FA introduced in Chapter 2, we thoroughly describe the method, including its model formulation, inference algorithm, and learning algorithm. A comprehensive evaluation of KSSHIBA on various real-world datasets is also presented, along with a detailed analysis of the results. In accordance with the open science philosophy upheld in this thesis, the implementation of the KSSHIBA model and all associated experimentation detailed in this chapter are readily accessible through a public repository on GitHub, under the link [1].

The chapter is organised as follows: Section 3.1 presents the kernelised formulation of SSHIBA, as well as the proposed formulations for RVs and feature selection. Then, in Section 3.2 we demonstrate how this new formulation performs

[1] https://github.com/sevisal/K-SSHIBA

in experimentation. Ultimately, in Section 3.3 the principal findings are highlighted.

3.1 Bayesian sparse factor analysis with kernelised observations

In the field of multi-view analysis, a common challenge is to find a shared latent representation, both inter- and intra-view, for N data samples represented in M different modalities, denoted as $\{\mathbf{X}^{(\mathrm{m})}\}_{m=1}^{M}$. The objective is to compress the information contained in $\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})} \in \mathbb{R}^{D_m}$, which represents a data point n of the view m, into a low-dimensional space of size $K << (D_1, \ldots, D_M)$, while considering the correlations among the data. This latent representation, referred to as $\mathbf{G}$, contains the shared information between all data modalities. $\mathbf{G}$ is obtained through a combination of its kernel representation, $\mathbf{k}_{\mathrm{n},:}^{(\mathrm{m})}$, and dual variables, $\mathbf{A}^{(\mathrm{m})}$ in traditional MultiVariate Analysis (MVA) techniques, such as Canonical Correlation Analysis (CCA). Our proposed approach reforms this idea from a probabilistic perspective.

In this work, we build upon the SSHIBA algorithm [2] and propose the use of latent variables $\mathbf{g}_{\mathrm{n},:}$ that are merged with dual variables $\mathbf{A}^{(\mathrm{m})}$ to generate the kernel vector $\mathbf{k}_{\mathrm{n},:}^{(\mathrm{m})}$. The relationship between $\mathbf{g}_{\mathrm{n},:}$, $\mathbf{A}^{(\mathrm{m})}$, and $\mathbf{k}_{\mathrm{n},:}^{(\mathrm{m})}$ is expressed by:

$$\mathbf{k}_{\mathrm{n},:}^{(\mathrm{m})} \;=\; \mathbf{g}_{\mathrm{n},:}\,\mathbf{A}^{(\mathrm{m})^{\mathrm{T}}} + \tau^{(\mathrm{m})}\,, \tag{3.1}$$

which implies that $\mathbf{k}_{\mathrm{n},:}^{(\mathrm{m})}$ is equal to the dot product of $\mathbf{g}_{\mathrm{n},:}$ and the transpose of $\mathbf{A}^{(\mathrm{m})}$, augmented by Gaussian random noise with a zero mean and power determined by a Gamma distribution, specified by $\tau^{(\mathrm{m})}$. The mapping function $\phi(\cdot)$ and its corresponding kernel function $K(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^{\top}\phi(\mathbf{x}')$ are used to calculate $\mathbf{k}_{\mathrm{n},:}^{(\mathrm{m})}$, which represents the kernel between $\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})}$ and all training data points.

The matrix $\mathbf{A}^{(\mathrm{m})}$ is used as a dual projection matrix in the algorithm, and its structure is defined using the structured ARD prior used in both BIBFA and SSHIBA. This prior promotes sparsity in the columns of the full matrix, by assigning a Gaussian distribution to $\mathbf{a}_{:,\mathrm{k}}^{(\mathrm{m})} \sim \mathcal{N}\left(0, \left(\alpha_{\mathrm{k}}^{(\mathrm{m})}\right)^{-1}\mathbf{I}_{\mathrm{K_c}}\right)$, and a Gamma distribution to $\alpha_k^{(m)} \sim \Gamma\left(a^{\alpha^{(m)}}, b^{\alpha^{(m)}}\right)$. The fusion of these distributions leads to the creation of sparse latent factors, enabling the selection of the most pertinent ones [151]. Consequently, each $\mathbf{A}^{(\mathrm{m})}$ determines the intra-view information of each $\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})}$ contained in the shared space $\mathbf{g}_{\mathrm{n},:}$. By employing this approach, we can capture inter-view information in $\mathbf{g}_{\mathrm{n},:}$ while selecting the intra-view information through $\mathbf{A}^{(\mathrm{m})}$.

The graphical representation of KSSHIBA is shown in Fig. 3.1. There are two types of data views considered in KSSHIBA, linear and kernelised. For linear

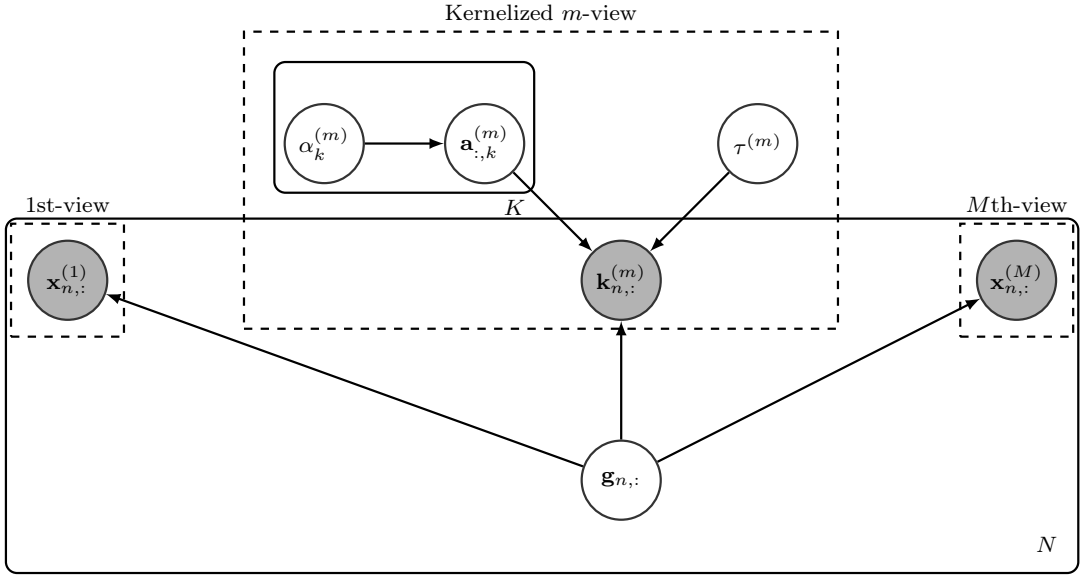
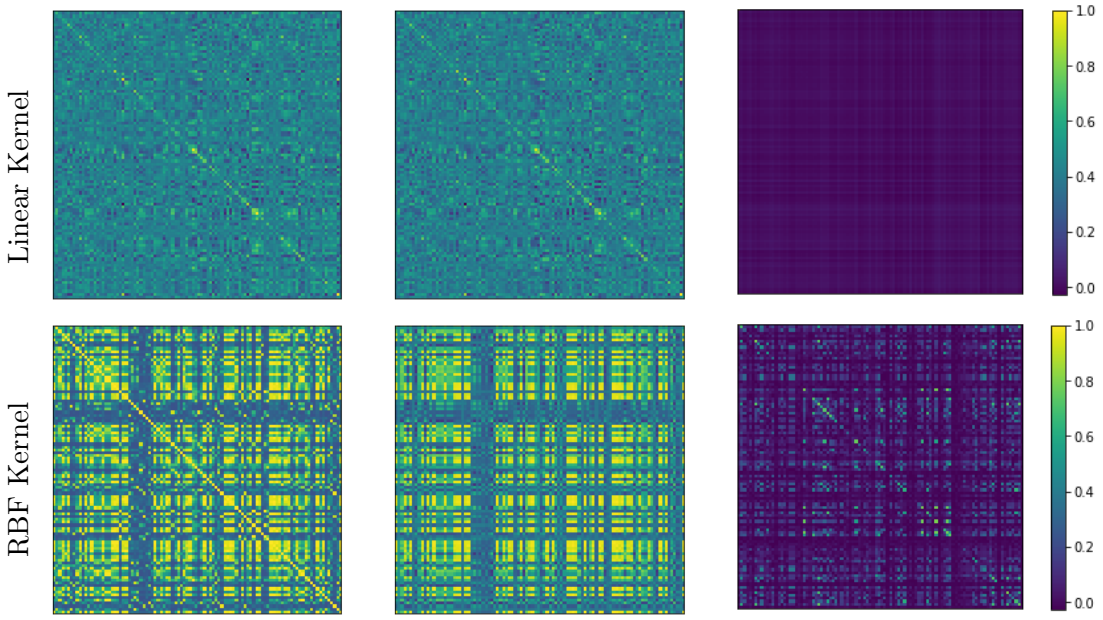


Figure 3.1: Graphic model of KSSHIBA.

views, the standard SSHIBA generative model can be used, where the relationship between the data and the latent variables is represented as $\mathbf{x}_{\mathrm{n,:}}^{(\mathrm{m})} = \mathbf{g}_{\mathrm{n,:}}\,\mathbf{W}^{(\mathrm{m})^{\mathrm{T}}} + \tau^{(\mathrm{m})}$. The weight matrix $\mathbf{W}^{(\mathrm{m})}$ follows the structured ARD prior discussed previously. On the other hand, for the kernelised views, the relationship between the data and the latent variables is modelled through a kernel function as described in Eq. (3.1). This type of representation is used when non-linear relationships within and between the views need to be considered or when the dimensionality of the view is large relative to the number of data points and working in the dual space is preferred to reduce the number of parameters. When both linear and kernelised views are present, the latent projection $\mathbf{g}_{\mathrm{n,:}}$ is learned to accurately reconstruct both types of data representations.

It should be noted that simply drawing a sample from the model in Eq. (3.1) does not guarantee that the resulting kernel matrix will be positive semidefinite. The kernel matrix is treated as a type of observation, known as a *kernelised observation*, and the parameters of the model are selected based on the goal of minimising reconstruction error. As shown in Fig. 3.2, a graphical illustration is provided that demonstrates how the kernelised observation and its reconstruction through Eq. (3.1) using the mean of the posterior distribution of $\mathbf{g}_{\mathrm{n,:}}$ are related. The illustration shows that the kernelised observations are effectively reconstructed. While more suitable models could be employed to customise the observation model, such as specifying the noise distribution as an inverse Wishart or modelling its covariance as the product of two low-rank matrices, they were found to result in

a less flexible overall model and therefore were not considered in this study. The focus here is on the model presented in Eq. (3.1), and it is acknowledged that further investigations in this field could lead to even more promising outcomes.



(a) Real matrix (b) Approximated matrix (c) Reconstruction error

Figure 3.2: KSSHIBA's generative properties demonstrated through an example of complete kernel matrix reconstruction.

The posterior distribution of a kernelised m-view can be evaluated as:

$$p(\Theta|\mathbf{k}_{1,:},\dots,\mathbf{k}_{N,:}) = \frac{\prod_{n=1}^{N} p(\mathbf{k}_{n,:}|\Theta)p(\Theta)}{p(\mathbf{k}_{1,:},\dots,\mathbf{k}_{N,:})}, \tag{3.2}$$

$$p(\mathbf{k}_{1,:},\dots,\mathbf{k}_{N,:}) = \int p(\Theta,\mathbf{k}_{1,:},\dots,\mathbf{k}_{N,:})d\Theta, \tag{3.3}$$

where Θ refers to all random variables (rv). We employ a mean-field VI technique [127] and optimise the ELBO of Eq. (3.3) as:

$$\log p(\mathbf{k}_{1,:},\dots,\mathbf{k}_{N,:}) \geq \int q(\Theta)\log\left(\frac{\prod_{n=1}^{N} p(\mathbf{k}_{n,:}|\Theta)p(\Theta)}{q(\Theta)}\right)d\Theta \tag{3.4}$$

with a completely factorised variational family, permitting us to estimate the

posterior outlined in Eq (3.2) as

$$p(\Theta|\mathbf{K}^{(m)}) \approx \prod_{m=1}^{M}\left(q\left(\mathbf{A}^{(\mathrm{m})}\right)q\left(\tau^{(\mathrm{m})}\right)\prod_{\mathrm{k}=1}^{\mathrm{K_c}}q\left(\alpha_{\mathrm{k}}^{(\mathrm{m})}\right)\right)\prod_{n=1}^{N}q(\mathbf{g}_{\mathrm{n},:})\,, \tag{3.5}$$

where the stack version of the kernel for each sample, $\mathbf{k}_{n,:}^{(m)}$, of the m-view, with dimension $N \times D_m$, is represented by $\mathbf{K}^{(m)}$.

The optimisation algorithm for the approximate inference approach in the KSSHIBA framework utilises the mean-field posterior structure and the ELBO expressed in Eq. (3.4). This algorithm is computationally efficient, as it resembles a coordinate-ascent method and optimises each factor in the approximate posterior distribution in Eq. (3.5) applying:

$$q^*(\theta_i) \propto \mathbb{E}\Theta{-}i\left[\log p(\Theta, \mathbf{k}1, :, \ldots, \mathbf{k}N, :)\right],, \tag{3.6}$$

where θ_i refers to a specific parameter, and Θ_{-i} represents all other parameters in the set Θ.

A reformulation of the input matrix in terms of a kernel matrix results in a similar structure to SSHIBA, allowing for the preservation of previous model capabilities. The mean-field factor update rules for KSSHIBA are displayed in Table 3.1. The expected value is denoted by <>. Further explanations of the calculations involved can be found in the supplementary material of the KSSHIBA paper [122].

Whereas the GP computational cost is equivalent to $O(N^3)$, KSSHIBA delivers a more systematic and efficient optimisation process of the order of $O(N^2K + K^3)$. This significant improvement in computational efficiency enables the exploration of even more complex and non-linear relationships in data, making KSSHIBA a highly flexible and powerful method for machine learning and data analysis.

Instead of relying on heuristics [152, 153] or a two-step optimisation processes [154], KSSHIBA employs a mean-field update rule to determine kernel parameters. This approach offers several benefits, including the ability to perform semi-supervised learning and feature selection, as well as incorporating the benefits of the SSHIBA formulation.

3.1.1 Automatic relevance vector determination

The proposed model features a double ARD prior for its dual variables $\mathbf{A}^{(\mathrm{m})}$:

$$\mathrm{a}_{\mathrm{n,k}}^{(\mathrm{m})} \sim \mathcal{N}\left(0, \left(\gamma_{\mathrm{n}}^{(\mathrm{m})}\,\alpha_{\mathrm{k}}^{(\mathrm{m})}\right)^{-1}\right). \tag{3.7}$$

Table 3.1: Table of mean-field approximated q distribution rules for variables in the KSSHIBA model.

Variable	q^* distribution	Parameters
$\mathbf{g}_{\mathrm{n},:}$	$\mathcal{N}\left(\mathbf{g}_{\mathrm{n},:}\,\middle\vert\,\mu_{\mathbf{g}_{\mathrm{n},:}}, \Sigma_{\mathbf{G}}\right)$	$\mu_{\mathbf{g}_{\mathrm{n},:}} = \sum_{\mathrm{m}=1}^{\mathrm{M}} \left(\langle\tau^{(\mathrm{m})}\rangle\, \mathbf{K}^{(\mathrm{m})} \langle\mathbf{A}^{(\mathrm{m})}\rangle \Sigma_{\mathbf{G}}\right)$ $\Sigma_{\mathbf{G}}^{-1} = \left(\mathbf{I}_{\mathrm{K_c}} + \sum_{\mathrm{m}=1}^{\mathrm{M}} \left(\langle\tau^{(\mathrm{m})}\rangle \langle\mathbf{A}^{(\mathrm{m})\mathrm{T}}\, \mathbf{A}^{(\mathrm{m})}\rangle\right)\right)$
$\mathbf{A}^{(\mathrm{m})}$	$\prod_{\mathrm{n}=1}^{\mathrm{N}} \left(\mathcal{N}\left(\mathbf{a}_{\mathrm{n},:}^{(\mathrm{m})}\,\middle\vert\,\mu_{\mathbf{a}_{\mathrm{n},:}^{(\mathrm{m})}}, \Sigma_{\mathbf{A}^{(\mathrm{m})}}\right)\right)$	$\mu_{\mathbf{a}_{\mathrm{n},:}^{(\mathrm{m})}} = \langle\tau^{(\mathrm{m})}\rangle\, \mathbf{K}^{(\mathrm{m})\mathrm{T}} \langle\mathbf{G}\rangle \Sigma_{\mathbf{A}^{(\mathrm{m})}}$ $\Sigma_{\mathbf{A}^{(\mathrm{m})}}^{-1} = \left(\mathrm{diag}(\langle\boldsymbol{\alpha}^{(\mathrm{m})}\rangle) + \langle\tau^{(\mathrm{m})}\rangle \langle\mathbf{G}^{\mathrm{T}}\, \mathbf{G}\rangle\right)$
$\alpha_{\mathrm{k}}^{(\mathrm{m})}$	$\Gamma\left(\alpha_{\mathrm{k}}^{(\mathrm{m})}\,\middle\vert\, a_{\alpha_{\mathrm{k}}^{(\mathrm{m})}}, b_{\alpha_{\mathrm{k}}^{(\mathrm{m})}}\right)$	$a_{\alpha_{\mathrm{k}}^{(\mathrm{m})}} = \frac{N}{2} + a^{\boldsymbol{\alpha}^{(\mathrm{m})}}$ $b_{\alpha_{\mathrm{k}}^{(\mathrm{m})}} = b^{\boldsymbol{\alpha}^{(\mathrm{m})}} + \frac{1}{2}\langle\mathbf{A}^{(\mathrm{m})\mathrm{T}}\, \mathbf{A}^{(\mathrm{m})}\rangle_{k,k}$
$\tau^{(\mathrm{m})}$	$\Gamma\left(\tau^{(\mathrm{m})}\,\middle\vert\, a_{\tau^{(\mathrm{m})}}, b_{\tau^{(\mathrm{m})}}\right)$	$a_{\tau^{(\mathrm{m})}} = \frac{N^2}{2} + a^{\tau^{(\mathrm{m})}}$ $b_{\tau^{(\mathrm{m})}} = b^{\tau^{(\mathrm{m})}} + \frac{1}{2}\left(\sum_{\mathrm{n}=1}^{\mathrm{N}} \sum_{\tilde{\mathrm{n}}=1}^{\tilde{\mathrm{N}}} \mathrm{k}_{\mathrm{n,n}}^{(\mathrm{m})2}\right.$ $-2\,\mathrm{Tr}\left\{\langle\mathbf{A}^{(\mathrm{m})}\rangle \langle\mathbf{G}^{\mathrm{T}}\rangle\, \mathbf{K}^{(\mathrm{m})}\right\}$ $\left.+\,\mathrm{Tr}\left\{\langle\mathbf{A}^{(\mathrm{m})\mathrm{T}}\, \mathbf{A}^{(\mathrm{m})}\rangle \langle\mathbf{G}^{\mathrm{T}}\, \mathbf{G}\rangle\right\}\right)$

allowing for both row-wise by $\alpha_{\mathrm{k}}^{(\mathrm{m})}$ and column-wise sparsity induced by $\gamma_{\mathrm{n}}^{(\mathrm{m})}$ as

$$\gamma_{\mathrm{n}}^{(\mathrm{m})} \;\sim\; \Gamma\left(a^{\gamma^{(\mathrm{m})}}, b^{\gamma^{(\mathrm{m})}}\right). \tag{3.8}$$

This leads to a more condensed representation of the data, which in turn reduces the effective number of latent factors and automatically determines the quantity of factors that have an impact on the final model. The mean-field factor update rules for this double ARD case are presented in Table 3.2. Additional information regarding the calculation of these expressions can be located in the supplementary section of the KSSHIBA study [122].

3.1.2 Automatic feature selection

However, the previous double ARD method only selects relevance vectors, not features as intended in SSHIBA. To achieve automatic feature selection, we can modify the ARD kernel by multiplying each feature of the original observations with a variable $\lambda_{\mathrm{d}}^{(\mathrm{m})}$ in the kernel definition. This allows for the determination of

Table 3.2: Updated q distribution for automatic RV selection.

Variable	q^* distribution	Parameters
$\mathbf{A}^{(\mathrm{m})}$	$\prod_{\mathrm{n}=1}^{\mathrm{N}} \mathcal{N}\left(\mathbf{a}_{\mathrm{n},:}^{(\mathrm{m})} \middle\vert \mu_{\mathbf{a}_{\mathrm{n},:}^{(\mathrm{m})}}, \Sigma_{\mathbf{a}_{\mathrm{n},:}^{(\mathrm{m})}}\right)$	$\mu_{\mathbf{A}^{(\mathrm{m})}} = \langle \tau^{(\mathrm{m})} \rangle \mathbf{X}^{(m)^T} \langle \mathbf{G} \rangle \Sigma_{\mathbf{A}^{(\mathrm{m})}}$ $\Sigma^{-1}_{\mathbf{a}_{\mathrm{n},:}^{(\mathrm{m})}} = \mathrm{diag}(\langle \boldsymbol{\alpha}^{(\mathrm{m})} \rangle) \langle \gamma_{\mathrm{n}}^{(\mathrm{m})} \rangle + \langle \tau^{(\mathrm{m})} \rangle \langle \mathbf{G}^{\mathrm{T}}\, \mathbf{G} \rangle$
$\boldsymbol{\gamma}^{(\mathrm{m})}$	$\prod_{\mathrm{n}=1}^{\mathrm{N}} \Gamma\left(\gamma_{\mathrm{n}}^{(\mathrm{m})} \middle\vert a_{\gamma_{\mathrm{n}}^{(\mathrm{m})}}, b_{\gamma_{\mathrm{n}}^{(\mathrm{m})}}\right)$	$a_{\gamma_{\mathrm{n}}^{(\mathrm{m})}} = \frac{K}{2} + a^{\gamma^{(\mathrm{m})}}$ $b_{\gamma_{\mathrm{n}}^{(\mathrm{m})}} = b^{\gamma^{(\mathrm{m})}} + \frac{1}{2} \sum_{\mathrm{k}=1}^{\mathrm{K_c}} \langle \alpha_{\mathrm{k}}^{(\mathrm{m})} \rangle \langle \mathrm{a}_{\mathrm{n,k}}^{(\mathrm{m})}\, \mathrm{a}_{\mathrm{n,k}}^{(\mathrm{m})} \rangle$

feature relevance through the ARD structure. For instance, a Radial Basis Function (RBF) kernel can be expressed as

$$\mathrm{k}_{\mathrm{n,n}}^{(\mathrm{m})} \;=\; \exp\left(-\sum_{\mathrm{d}=1}^{\mathrm{D_m}} \left(\mathrm{x}_{\mathrm{n,d}}^{(\mathrm{m})} - \mathrm{x}_{\mathrm{n,d}}^{(\mathrm{m})} \right)^2 \lambda_{\mathrm{d}}^{(\mathrm{m})} \right), \tag{3.9}$$

thus optimisation of $\boldsymbol{\lambda}^{(\mathrm{m})}$ is performed by maximising the lower bound of the mean-field approach. If the m-th view is kernelised, the only term in the lower bound that is affected by the ARD is $\mathbb{E}_q\big[\ln\big(p\big(\mathbf{K}^{(\mathrm{m})} \,|\Theta\big)\big)\big]$. Hence, the optimised ELBO regarding the kernel is now[2]

$$\mathrm{ELBO}_K \;=\; -\frac{\langle \tau^{(\mathrm{m})} \rangle}{2} \sum_{\mathrm{n}=1}^{\mathrm{N}} \sum_{\mathrm{u}=1}^{\mathrm{N}} \left(\mathrm{k}_{\mathrm{n,u}}^{(\mathrm{m})^2} - 2\, \mathrm{k}_{\mathrm{n,u}}^{(\mathrm{m})} \langle \mathbf{a}_{\mathrm{u},:}^{(\mathrm{m})} \rangle \langle \mathbf{g}_{\mathrm{n},:}^{\mathrm{T}} \rangle + \langle \mathbf{a}_{\mathrm{u},:}^{(\mathrm{m})^{\mathrm{T}}}, \mathbf{a}_{\mathrm{u},:}^{(\mathrm{m})} \rangle \langle \mathbf{g}_{\mathrm{n},:}^{\mathrm{T}}, \mathbf{g}_{\mathrm{n},:} \rangle \right), \tag{3.10}$$

The mean-field updates over the variational bound and the direct maximisation of the lower bound are performed alternately using gradient ascent methods such as Pytorch and Adam. By setting a threshold for $\boldsymbol{\lambda}^{(\mathrm{m})}$, the model can automatically select the most relevant features during training.

3.2 Results

In this section, we carry out a comprehensive evaluation of the performance of the KSSHIBA model on benchmark datasets. The objective of this evaluation is to study the behaviour of the model with respect to various datasets and problems. The aim is to thoroughly examine how the extensions proposed in this chapter perform in different scenarios and to assess if the goals of their proposals are achieved. In Chapter 5, the KSSHIBA model is applied to a real-world scenario

[2]The full derivation can be found in [122]

involving the prediction of antibiotic resistance based on MALDI-TOF MS and heterogeneous data. Additionally, in Chapter 6, the performance of the KSSHIBA is compared to that of the FA-VAE model, which is introduced in Chapter 4, in the context of automatic ribotyping of bacteria based on MALDI-TOF MS data.

3.2.1 Experimental setup

The objective of this section is to assess the effectiveness of the KSSHIBA framework by means of a comparison with other leading algorithms that share analogous capabilities with respect to multi-dimensional regression and categorical classification. To perform the regression task, we use eight datasets from the Mulan repository [155, 156, 157], which are described in detail in Table 3.3. We compare the KSSHIBA with KPCA, KCCA, MRD, Support Vector Regression (SVR) [158], MLP [159], and Heterogeneous Incomplete Variational AutoEncoder (HI-VAE) [3]. For the classification task, we use four image datasets to study the feature selection method of KSSHIBA and validate its interpretability. Finally, we use an heterogeneous dataset of arrhythmia to compare the latent factors learned by KSSHIBA with other methods such as SVM or CCA+SVM.

In our evaluation of the multi-regression performance, we compared the KSSHIBA framework to KPCA and KCCA using a radial basis function (RBF) kernel. To determine the best kernel hyperparameter value for each of these models, we considered 20 values of γ on a log scale, ranging from $\frac{10^{-10}}{C}$ to $\frac{10^{0.5}}{C}$, where C is the number of tasks. For the KPCA model, predictions were made by combining it with linear regression (KPCA+LR), while KCCA was used on its own and in conjunction with linear regression (KCCA+LR) to provide a comprehensive comparison. The latent factors number was fixed to either the maximum possible value of C for KCCA or to explain 95% of the variance in KPCA. We also evaluated the performance of SVR, and like KCCA and KPCA, we used an RBF kernel. Additionally, we evaluated the regularisation parameter λ, exploring 11 values on a logarithmic scale from $[10^{-4}$ to $10^{4}]$.

To evaluate the performance of KSSHIBA in relation to other methods, we compared it to MRD [160] in both multiregression and latent space interpretability tasks. In this comparison, the Matlab library found in reference [160] was utilised, and the quantity of latent variables was established to be double the quantity of tasks $(2 * C)$. The models were trained using an RBF kernel with automatic relevance determination, and the training epochs were constrained to 100 because of the time-consuming nature of the training process for MRD.

Additionally, we compared KSSHIBA to an heterogeneous VAE approach known as HI-VAE, which, like KSSHIBA, is capable of modelling heterogeneous data, dealing with missing values, and finding low-dimensional data representations. We chose to use the same kernel as we used for other baselines and followed the

layer configuration suggested by the authors, which consisted of three layers with dimensions 50-50-50.

Finally, we also compared our model to an MLP neural network in two different scenarios. First, to force a dimension bottleneck of C in the hidden layers, similar to KCCA, we tested two approaches: (i) a dense layer with C neurons and (ii) two dense layers with C neurons and D_m neurons, respectively. Second, a scenario without a bottleneck was considered, with three configurations being validated: (i) 100 neurons unique dense layer, (ii) 100-50 dense layers, and (iii) 10-50-100 bottleneckled dense layers.

In regards to KSSHIBA, we utilised its semi-supervised capability for output prediction by incorporating the test samples without their targets during the training phase. The labels of these samples were then predicted using the mean of the posterior distribution. The inference process of KSSHIBA used a convergence criteria based on the evolution of the lower bound to determine the number of iterations. The algorithm was terminated when the mean value of the previous lower bound values was greater than the lower bound in the last iteration minus 10^{-4}, or when it reached 10,000 iterations. Ten random initialisation were performed for each KSSHIBA model, and the one with the best lower bound was chosen. The ARD priors included in the projection matrix $\mathbf{W}^{(\mathrm{m})}$ were used to automatically prune latent factors in KSSHIBA.

We conducted a nested 10-folds Cross-Validation (CV) to determine the hyper-parameters of each model. The outer CV split the dataset into training and test sets, while the inner CV split the training set into a validation set and a second training set. This approach enabled us to assess the overall performance of the framework and validate the model parameters.

To evaluate the performance of the various method variations and optimise the hyperparameters of the method through CV, we relied on the coefficient of determination (R2) for multiregression and accuracy in case of classification.

3.2.2 Performance evaluation of KSSHIBA for multi-dimensional regression

The purpose of this section is to evaluate the effectiveness of KSSHIBA in semi-supervised multidimensional regression and compare it to some of the current leading methods. For this comparison, we have utilised eight multidimensional regression datasets from the Mulan repository [155, 156, 157], the key characteristics of which are outlined in Table 3.3.

The results of the proposed model in two scenarios, one where the number of latent factors was learned using the ARD prior and another where it was set to the number of tasks (C), are presented in Table 3.4. These results are compared to those of other methods, including KCCA, KPCA, MRD, SVR, MLP, and HI-VAE.

In particular, we can see that KSSHIBA consistently outperforms most reference methods in every database, pointing out the performance advantages obtained in *edm* and *oes97*. Additionally, it is notable that this performance improvement is accomplished with an effective dimensional reduction, since KSSHIBA, applying a feature extraction, is able to outperform both an SVR and an MLP that use all the original features. When looking at the models that find latent representations (KPCA+LR, KCCA+LR, and HI-VAE), none of them provides consistently good

Table 3.3: Characteristics of the multitask databases used in this work.

Database	Samples	Features	Tasks
at7pd	296	411	6
at1pd	337	411	6
edm	154	16	2
enb	768	8	2
jura	359	15	3
oes10	403	298	16
oes97	334	263	16
wq	1,060	16	14

Table 3.4: R2 scores expressed as mean ± standard deviation (white) and latent factor (light gray) are depicted in each sub-row, respectively, for the KSSHIBA and the various methods under examination on the multitask databases. The data has been normalised and, if a kernel is used, it has been centred.

	KSSHIBA	KSSHIBA $K = C$	MRD	KPCA + LR	KCCA	KCCA + LR	SVR RBF	MLP $K = C$	MLP
at1pd	0.77 ± 0.09	$\mathbf{0.78 \pm 0.09}$	0.67 ± 0.07	0.67 ± 0.12	0.45 ± 0.05	0.75 ± 0.11	0.01 ± 0.05	0.75 ± 0.09	0.77 ± 0.12
	53 ± 8	6	12	22 ± 10	6	6		6	
at7pd	0.48 ± 0.26	0.52 ± 0.13	0.48 ± 0.12	0.39 ± 0.19	0.24 ± 0.05	$\mathbf{0.57 \pm 0.16}$	0.01 ± 0.03	0.29 ± 0.33	0.35 ± 0.69
	53 ± 11	6	12	21 ± 1	6	6		6	
oes97	0.63 ± 0.16	$\mathbf{0.69 \pm 0.10}$	0.34 ± 0.07	0.45 ± 0.20	0.30 ± 0.08	0.36 ± 0.09	0.39 ± 0.10	0.57 ± 0.22	0.58 ± 0.21
	108 ± 11	16	32	12 ± 7	16	16		16	
oes10	0.79 ± 0.08	$\mathbf{0.80 \pm 0.07}$	0.38 ± 0.07	0.59 ± 0.15	0.35 ± 0.17	0.43 ± 0.12	0.47 ± 0.12	0.77 ± 0.07	0.76 ± 0.08
	104 ± 22	16	32	14 ± 7	16	16		16	
edm	0.37 ± 0.19	0.21 ± 0.09	-0.17 ± 0.45	$\mathbf{0.38 \pm 0.19}$	0.26 ± 0.18	0.18 ± 0.26	0.35 ± 0.19	0.14 ± 0.17	0.26 ± 0.21
	17 ± 2	2	4	16 ± 5	2	2		2	
jura	$\mathbf{0.61 \pm 0.10}$	0.30 ± 0.10	0.57 ± 0.06	0.38 ± 0.11	0.11 ± 0.08	0.18 ± 0.15	0.60 ± 0.05	0.32 ± 0.12	$\mathbf{0.61 \pm 0.06}$
	64 ± 7	3	6	23 ± 1	3	3		3	
wq	0.12 ± 0.01	0.12 ± 0.01	-0.35 ± 0.08	0.09 ± 0.02	-0.01 ± 0.01	-0.01 ± 0.01	0.08 ± 0.02	0.10 ± 0.02	$\mathbf{0.13 \pm 0.03}$
	48 ± 3	14	28	29 ± 0.98	14	14		14	
enb	$\mathbf{0.99 \pm 0.01}$	0.86 ± 0.02	0.91 ± 0.01	0.86 ± 0.01	0.96 ± 0.01	0.98 ± 0.01	$\mathbf{0.99 \pm 0.01}$	0.89 ± 0.01	0.99 ± 0.08
	118 ± 4	2	4	13 ± 1	2	2		2	

results over all databases, while KSSHIBA provides outstanding results due to its ease of adaptation to different scenarios. At the same time, the results obtained by KSSHIBA with $K = C$ reveal that standard KSSHIBA is too conservative in the number of extracted features and we could force a more restrictive pruning without degrading the final performance (note that KSSHIBA with $K = C$ only deteriorates in the problems with only 2 or 3 (*edm, jura* and *enb*) since in these cases the number of latents is extremely reduced.

3.2.3 Evaluation of the solution in terms of RVs

The purpose of the KSSHIBA approach is to demonstrate its ability to construct compact solutions by selecting a subset of training points, known as RVs. This was tested using the same databases and experimental setup as in the previous section, but with comparison to the KPCA + LR and KCCA + LR models using the Nyström subsampling technique. The optimal percentage of RVs was determined by cross-validation by exploring values from 1 to 100 percent of the total number of training data.

The results shown in Table 3.5 indicate that the inclusion of automatic RV selection in KSSHIBA maintained or improved performance for most of the databases, while reducing the complexity of the model. The reduction in the number of RVs also resulted in a further reduction in the number of latent factors. In comparison to KPCA + LR and KCCA + LR, KSSHIBA showed a lower percentage of RVs needed to describe the kernel, as it learns the relevance of each element and eliminates it accordingly, while KPCA and KCCA obtain the compact solution through a random selection of RVs.

For the purpose of comparison, we also included the results of the MRD approach when the number of inducing points was varied. However, we observed that the behaviour of MRD with respect to the number of inducing points is unstable and highly dependent on the database. This is because the position of the inducing points is determined using a regular grid, which does not allow for optimisation, resulting in significant fluctuations in performance.

In contrast, the results of KPCA combined with LR and KCCA combined with LR may vary and require modifications to the quantity of RVs to attain accurate outcomes, whereas KSSHIBA demonstrates a comparatively stable coefficient of determination (R2) value. This stability can be attributed to KSSHIBA's capability to determine the significance of each RV and factor this information into all model inferences when updating its parameters.

Finally, Fig. 3.3 provides a analysis of the effect of the number of RVs or inducing points (MRDs) on the evaluation of the solution in terms of RVs. The results for MRD are not included for the *wq* database as the model iterations have not been completed at the time this material is being written due to the high

computational time required by the library.

3.2.4 Analysis of the feature relevance

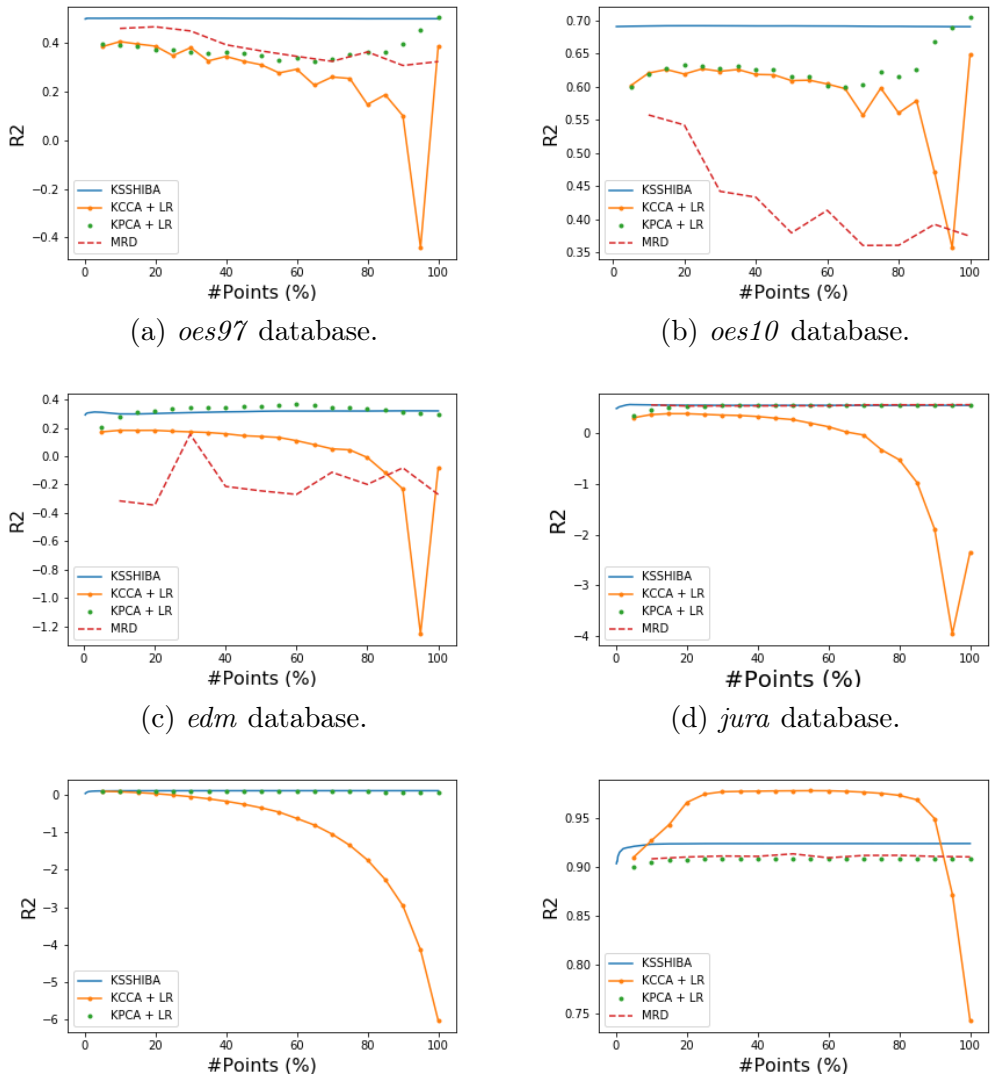
In order to evaluate the feature relevance of the KSSHIBA approach (as outlined in Section 3.1.2), we conduct an experiment using image classification datasets. The input view in this experiment consists of images, while the output view is the corresponding category label. This setup provides a visual representation of the relevance of each pixel as a feature.

For this experiment, we use three image datasets: *warpAR10P* (60×40 pixels), *Yale* (32×32 pixels), and *Olivetti* (32×32 pixels). These datasets can be found in the Feature Selection Repository[3] and the aligned version of the Labelled Faces in the Wild (LFW) dataset [161] obtained by [162]. The characteristics of these datasets are listed in Table 3.6.

[3] http://featureselection.asu.edu/datasets.php

Table 3.5: Outcome of the automatic RV selection experiment on the multitask database. The first subcolumn illustrates the average and deviation of the R2 score in white, with the light grey indicating the actual quantity of latent variables (K). The relevant vectors are represented as a percentage of the total samples.

	Sparse KSSHIBA		KPCA + LR		KCCA + LR	
	R2 - K	$\%RVs$	R2 - K	$\%RVs$	R2 - K	$\%RVs$
at1pd	0.77 ± 0.09 41 ± 11	18.4 ± 24.1	0.78 ± 0.09 87 ± 35	69.7 ± 32.9	$\mathbf{0.80 \pm 0.09}$ 6	84.8 ± 27.5
at7pd	0.55 ± 0.15 70 ± 27	18.5 ± 26.3	0.56 ± 0.18 90 ± 37	79.7 ± 31.7	$\mathbf{0.60 \pm 0.12}$ 6	73.9 ± 34.1
oes97	$\mathbf{0.58 \pm 0.15}$ 61 ± 7	38.6 ± 24.5	0.52 ± 0.24 124 ± 34	81.7 ± 27.8	0.42 ± 0.30 16	23.9 ± 27.8
oes10	$\mathbf{0.77 \pm 0.11}$ 74 ± 6	44.4 ± 38.4	0.71 ± 0.12 132 ± 53	71.9 ± 11.6	0.66 ± 0.10 16	57.8 ± 35.2
edm	$\mathbf{0.42 \pm 0.21}$ 13 ± 4	53.8 ± 28.5	0.41 ± 0.26 29 ± 14	52.5 ± 30.5	0.20 ± 0.14 2	22.7 ± 13.6
jura	$\mathbf{0.58 \pm 0.14}$ 30 ± 4	48.7 ± 38.4	0.57 ± 0.10 59 ± 14	60.7 ± 28.9	0.36 ± 0.09 3	18.9 ± 7.5
wq	$\mathbf{0.12 \pm 0.01}$ 21 ± 2	58.1 ± 33.2	0.12 ± 0.02 96 ± 49	22.9 ± 15.9	0.10 ± 0.01 14	5.9 ± 3.1
enb	$\mathbf{0.99 \pm 0.01}$ 78 ± 8	19.5 ± 12.8	0.91 ± 0.01 28 ± 1	48.9 ± 32.9	0.97 ± 0.01 2	41.9 ± 12.2



(a) *oes97* database. (b) *oes10* database.

(c) *edm* database. (d) *jura* database.

(e) *wq* database. (f) *enb* database.

Figure 3.3: R2 results analysis for several % of RVs in KCCA+LR, KSSHIBA, KPCA+LR, and MRD. For the last, they are inducing points instead.

To prepare the LFW dataset for our analysis, we had to perform some processing steps. This involved cropping the images to remove any unnecessary information and reducing the size of the images to 60×40 pixels to lower the computational demands of the training process. To make the dataset more manageable, we only selected images of the 7 individuals with the largest number of images in the database.

The KSSHIBA approach used in this study involved incorporating an ARD kernel in the input images to conduct a feature relevance analysis. As a result, Fig. 3.4 displays the relevance masks learned by the model on each dataset, with lighter colours representing relevant pixels and darker colours indicating that a pixel is not relevant. By using these datasets, we can clearly see the face shape that has been learned by the model, which identifies the pixels that the model should focus on and pays less attention to the background. In some cases, the model can even clearly define facial features such as the nose, cheek, or chin, producing a face mask.

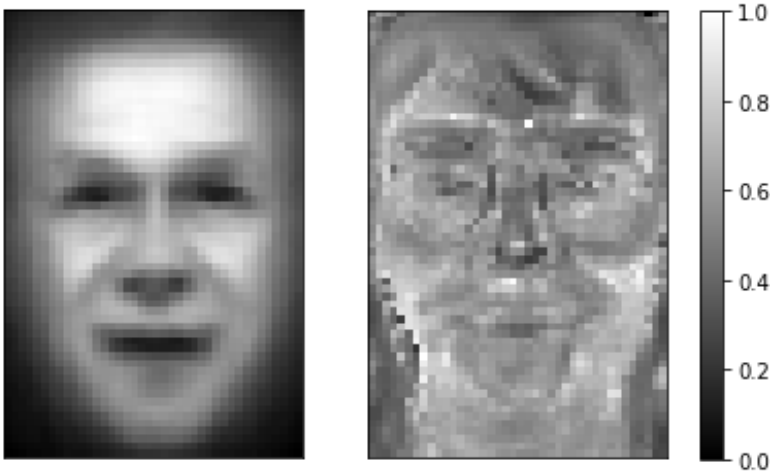
The masks produced by KSSHIBA reveal that the relevances learned for the LFW dataset are sharper and more detailed, likely due to the larger number of samples in this database. The masks for the warpAR10P and Olivetti datasets show that the model tends to focus on the area surrounding glasses, likely because a significant number of images in both databases feature people wearing glasses. Additionally, the model recognises that if a subject wears glasses in one image, they will likely wear them in others, making this feature relevant for subject classification in these datasets. The masks also suggest that the model tends to downplay the importance of mouth and eye features, while placing greater emphasis on hair, cheeks, and facial shape. This is not as pronounced in the warpAR10P dataset because many images feature cloth covering the face below the nose.

3.2.5 Analysis of the extracted latent factors

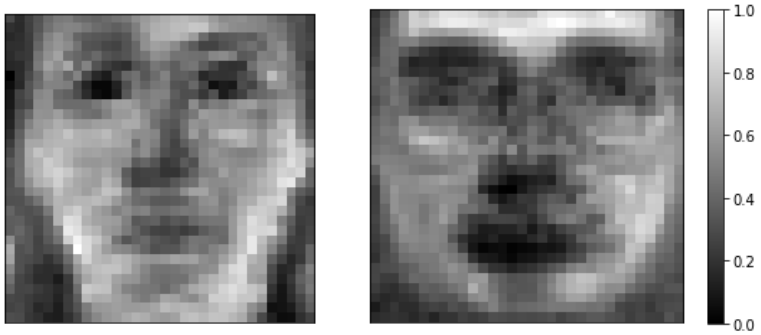
The aim of this section is to compare the interpretability of the extracted latent factors generated by the MRD approach based on shared GPLVMs with those generated by the proposed KSSHIBA model. To do this, we will use the *Oil* classification database [163] as the test case, which consists of 2,000 samples, 12 features, and 3 output classes. The models will be trained with 15 latent factors (the number of features plus the number of output classes), and with the addition of ARD latent factor selection. The KSSHIBA model uses an RBF kernel exclusively for the input view, while MRD uses it for both the input and output views. The

Table 3.6: Characteristics of the face databases used in this work.

Database	Samples	Features	Classes
LFW	1,277	2,400	7
warpAR10P	130	2,400	10
Yale	165	1,024	15
Olivetti	400	1,024	40



(a) *LFW* (b) *warpAR10P*

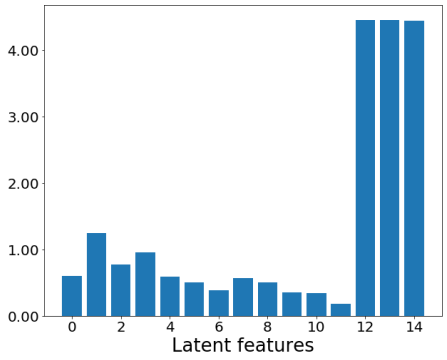


(c) *Yale* (d) *Olivetti*

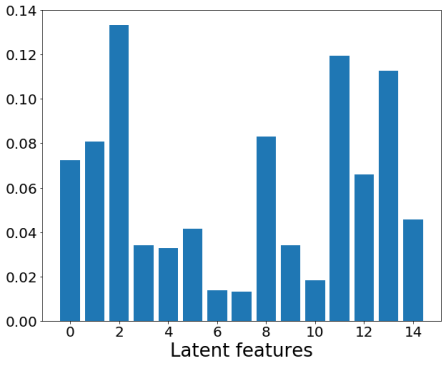
Figure 3.4: The feature selection extension of KSSHIBA has learnt feature masks for various face recognition problems, and these masks show the significance of each pixel. Lighter hues indicate higher relevance of the pixel, while darker shades signify lower relevance.

results showed that MRD achieved an accuracy of 99.0%, while KSSHIBA achieved an accuracy of 99.4%. It should be noted that the available MRD implementation in Matlab is not scalable for large datasets, leading to longer computational times.

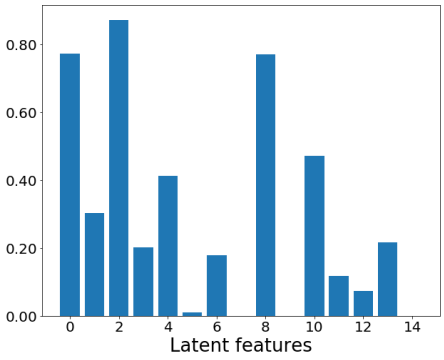
The relevance of each latent factor for MRD and KSSHIBA is shown in Fig. 3.5. With MRD, the number of latent factors to be used for input and output data is predetermined. Fig. 3.5a displays the relevance of all these common factors, where the first 12 are related to the input view and the last three to the output, as the model primarily focuses on the latter. In contrast, the KSSHIBA model displays separate weights for each perspective (as depicted in Fig.s 3.5b and 3.5c). This signifies that the latent factors might not be important and could be discarded



(a) MRD - common



(b) KSSHIBA - input



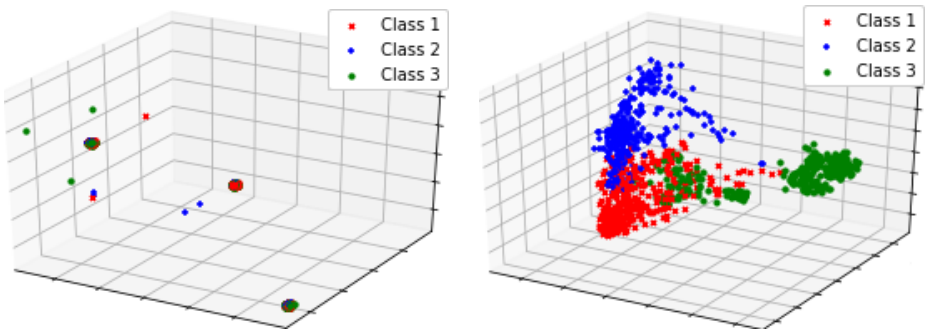
(c) KSSHIBA - output

Figure 3.5: Measure of relevance for each learnt latent factor on the *Oil* database. Fig. 3.5a shows the relevance of the commons for the MRD model (all latents have resulted to be shared by both views). Figs. 3.5b and 3.5c show, respectively, the relevance for the input view and the output view for KSSHIBA.

(latent 7), or they might only be relevant for one view (latents 5 and 14), or they could be significant for more than one view (highlighting latents 0, 2, and 8). The inclusion of an ARD prior to the $\mathbf{A}^{(\mathrm{m})}$ projection matrices results in a more understandable model, as it provides valuable information about the latent factors.

Finally, the interpretability of the results was analysed through the examination of the most relevant extracted latent features, as depicted in Fig. 3.6. The results indicate that the KSSHIBA algorithm was able to find a subspace in which the classification problem could be easily solved with the use of just three of the most relevant common latent factors. In contrast, the MRD model was found to project most of the samples into a single point, thus requiring a greater number of latent factors to discriminate between the different classes. It may be inferred from the results that incorporating non-linearities through kernel methods can enhance the processing of input data in both MRD and KSSHIBA models. However, non-

linearities do not appear to be as useful for predicting output categories. In the MRD model, non-linearities are incorporated into both input and output views, whereas in KSSHIBA, only the input data is processed using a kernelised method. Nonetheless, KSSHIBA is able to perform a more informative and discriminative analysis by computing categories linearly.



(a) Latent space of MRD (b) Latent space of KSSHIBA

Figure 3.6: Learnt projections for the Oil database. Each figure shows the projections over the three most relevant factors: latents 12, 13 and 14 for MRD and latents 0, 2 and 8 for KSSHIBA.

3.3 Conclusions

In this chapter, we have proposed starting from the SSHIBA approach, which was already able to work with semi-supervised heterogeneous multi-view problems. Then, we have extended its formulation to handle non-linear data relationships, to provide compact representations with an automatic selection of RVs, and to obtain the input feature relevance functionality by means of an ARD over the kernel.

The results prove the relevance of the proposed formulation, achieving not only competitive performance but also transforming the data into a reduced set of interpretable latent variables and a compact model consisting of a reduced subset of RVs. Furthermore, the feature relevance criteria are able to learn relevant masks that provide insight into the input space for the goal task.

In this study, the benchmarking results demonstrate the effectiveness of the KSSHIBA model in addressing non-linear data. As such, in Chapter 5, we further extend the application of KSSHIBA by customizing it to handle MALDI-TOF MS data within the domain of microbiology. By leveraging specific kernels in this field, we aim to demonstrate the versatility and robustness of the KSSHIBA model in addressing real-world data challenges.

While KSSHIBA's use of kernel functions enables efficient handling of non-linear

data, it may not be optimal for certain types of data. In light of this, we outline in the next chapter of this dissertation, Chapter 4, an extension to KSSHIBA and SSHIBA that incorporates Variational Autoencoders (VAEs) to provide a more comprehensive solution for a broader range of non-linear data, including image and temporal data that may require specialised kernels.

CHAPTER 4

FACTOR ANALYSIS VARIATIONAL AUTOENCODER

This chapter is motivated by the notable improvements achieved in the literature by generative models, such as VAEs, in representing high-dimensional complex data types such as images. In particular, we propose a novel Bayesian method, called Factor Analysis Variational AutoEncoder (FA-VAE), that combines the interpretability of Bayesian FA, the multi-view handling of MVA and the versatility of VAEs to address this problem. Our contribution is the integration of VAEs into the SSHIBA architecture to handle new input data types, making our method a powerful tool for tackling complex, high-dimensional and heterogeneous data problems. While the analysis of results in this chapter is focused on images to facilitate their interpretation, in Chapter 6 we analyse the versatility of this approach by extending its application to MALDI-TOF MS data.

The goal of the proposed method is to enhance the capabilities of existing techniques for handling diverse data types in real-world scenarios. Building upon the success of SSHIBA [2] and its kernelised version, KSSHIBA [122], which have proven effective for handling various data types including categorical, real, positive, binary, and kernelised data, we aim to push the boundaries of these approaches and tackle new and challenging data types such as images or time series, specifically using CNNs or RNNs. By doing so, we aim to develop a powerful and flexible tool that can adapt to any data structure.

For this purpose, we extend SSHIBA by introducing VAE versatility, while retaining the interpretability of the FA latent space. The proposed scheme is capable of assembling multiple VAEs, creating a model that can function with multiple data domains, depending on the VAE architecture. For instance, VAE with CNN encoder-decoder structure permits us to efficiently handle images. Specifically, we demonstrate that taking advantage of the multi-view SSHIBA architecture, we can condition a pretrained VAE to various labels or categories set in the other views. This way, using different VAEs per view, working in different domain, we perform domain adaptation between two distinct databases while conditioning them on external attributes surpassing SOTA procedures. Additionally, we illustrate how transfer learning between VAE models in different views can enhance the final model performance. Finally, we also show FA-VAE's flexibility, as a simple alteration in the encoder-decoder architecture enables it to function with new forms of data, such as time dependence using 1D-CNNs, as shown in Chapter 6.

The technical results presented in this chapter are under review at Information Sciences journal from Elsevier, whose preprint can be publicly accessed [123]. In accordance with the open science philosophy upheld in this thesis, the implementation of the FA-VAE model and all associated experiments detailed in this chapter are readily accessible through a public repository on GitHub, under the link [1].

The organisation of the chapter is as follows: In Section 4.1, we introduce the theoretical foundations and mathematical formulation of our proposed method. Subsequently, in Section 4.2, we conduct experiments to demonstrate the efficacy of our approach. Specifically, in Section 4.2.1, we illustrate the ability to condition a pretained VAE to a specific label. In Section 4.2.2, we showcase the use of our method for domain adaptation between distinct datasets and styles. And in Section 4.2.3, we propose using our approach as a transfer learning tool between generative models. Lastly, in Section 4.3, we summarise the key findings and provide concluding remarks.

4.1 Factor Analysis Variational AutoEncoder

Our contribution is the development of a novel hierarchical approach, known as Factor Analysis VAE (FA-VAE), that adapts the SSHIBA method to handle VAE networks to model new data types. To explain the FA-VAE method, we begin by summarising the generative model of the FA framework, in which each observation $\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})} \in \mathbb{R}^{1\times D_m}$ is associated with a global shared space $\mathbf{g}_{\mathrm{n},:} \in \mathbb{R}^{1\times K}$, where $K < \{D_1, D_2, \ldots, D_M\}$. The global shared space is assumed to follow a Gaussian distribution:

$$\mathbf{g}_{\mathrm{n},:} \sim \mathcal{N}(0, \mathbf{I}_K)\,. \tag{4.1}$$

These latent variables are combined with a set of unique projections matrices $\mathbf{w}_{:,\mathrm{k}}^{(\mathrm{m})} \in \mathbb{R}^{1\times K}$ assuming a Gaussian distribution

$$\mathbf{w}_{:,\mathrm{k}}^{(\mathrm{m})} \sim \mathcal{N}(0, {\alpha_{\mathrm{k}}^{(\mathrm{m})}}^{-1}\mathbf{I}_K)\,. \tag{4.2}$$

with an ARD $\alpha_{\mathrm{k}}^{(\mathrm{m})} \sim \Gamma(a^{\boldsymbol{\alpha}^{(\mathrm{m})}}, b^{\boldsymbol{\alpha}^{(\mathrm{m})}})$ prior over the columns. When the values of $\alpha_{\mathrm{k}}^{(\mathrm{m})}$ become high, it induces some k-columns to be set to 0, resulting in sparse representations. The combination of the shared global $\mathbf{g}_{\mathrm{n},:}$ latent variables with the projection matrices $\mathbf{W}^{(\mathrm{m})}$ generates the m-observations $\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})}$ as

$$\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})} \,|\, \mathbf{g}_{\mathrm{n},:} \sim \mathcal{N}(\mathbf{g}_{\mathrm{n},:}\,{\mathbf{W}^{(\mathrm{m})}}^{\mathrm{T}}, {\tau^{(\mathrm{m})}}^{-1}\mathbf{I}_K)\,. \tag{4.3}$$

where, due to the ARD induced by $\alpha_{\mathrm{k}}^{(\mathrm{m})}$, only the relevant k-columns of $\mathbf{g}_{\mathrm{n},:}$ for the current view are used to generate the current $\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})}$, disentangling private and shared information.

[1] https://github.com/aguerrerolopez/FA-VAE

Our approach involves the incorporation of latent information derived from multiple VAEs into the global $\mathbf{g}_{\mathrm{n},:}$ shared space. To do so, for a given m-view data point $\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})} \in \mathbb{R}^{1\times D_m}$ by means of a VAE-like structure, we obtain an embedded representation, $\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})} \in \mathbb{R}^{1\times D'_m}$, where $D'_m << D_m$. These embedded latent variables, $\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})}$ for $m = 1, ..., M$, contribute to the construction of the global latent variable by projecting their shared information into $\mathbf{g}_{\mathrm{n},:}$, where $K < \{D'_1, D'_2, \ldots, D'_M\}$. Given the nature flexibility of VAE-like structure, each individual pair of encoder-decoder for each m-th view can be chosen to handle different data types. This flexibility allows the FA-VAE model to effectively handle a wide range of heterogeneous observations, beyond those already handled by SSHIBA, such as images modelled using a CNN or sequential data modelled using an RNN.

To obtain the embedded latent representation $\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})}$ of the m-th view, we employ a VAE that encodes the observations using the variational distribution $\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})} \,|\, \mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})} \sim q_\eta(\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})})$, where

$$q_\eta(\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})}) \sim \mathcal{N}\left(\mu_\eta^{\mathbf{z}^{(\mathrm{m})}}(\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})}), \Sigma_\eta^{\mathbf{z}^{(\mathrm{m})}}(\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})})\right), \tag{4.4}$$

where $\mu_\eta^{\mathbf{z}^{(\mathrm{m})}}(\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})}), \Sigma_\eta^{\mathbf{z}^{(\mathrm{m})}}(\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})})$ are the output of an independent parametric encoder for each m-th view. Consequently, the FA framework interprets $\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})}$ as a pseudo-observation, whereby the generative FA model assumes that

$$\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})} \sim \mathcal{N}\left(\mathbf{g}_{\mathrm{n},:}\,\mathbf{W}^{(\mathrm{m})}, {\tau^{(\mathrm{m})}}^{-1}\,\mathbf{I}_{D'_m}\right), \tag{4.5}$$

and, thus, the observations $\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})}$ are generated from $\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})}$ using the VAE conditional distribution $\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})} \,|\, \mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})} \sim p_\theta(\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})} \,|\, \mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})})$ as

$$\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})} \sim \mathcal{N}\left(\mu_\theta^{\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})}}(\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})}), \sigma^{-1}\mathbf{I}_K\right) \tag{4.6}$$

where $\mu_\theta^{\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})}}(\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})})$ is the output of an independent parametric decoder for the m-th view and its precision has been fixed to σ.

The fusion of the VAE latent space and the FA framework is a non-trivial task. In a VAE, the maximisation of the ELBO ensures that the variational distribution $q_\eta(\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})} \,|\, \mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})})$ is close to the prior distribution $p(\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})})$ by minimising their KL divergence, as demonstrated in Eq. (2.49). However, in a standard VAE, the prior distribution is normally distributed, $p(\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})}) \sim \mathcal{N}\left(0, \mathbf{I}_{D'_m}\right)$,which imposes continuity and completeness on the latent space. In other words, any point in the latent space can be decoded to produce a meaningful output, and a series of samples can be decoded in a progressive manner from the training space. When dealing with a multi-modal problem, this regularisation term does not take into account that the embedded latent space $\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})}$ must share information with other views to learn them

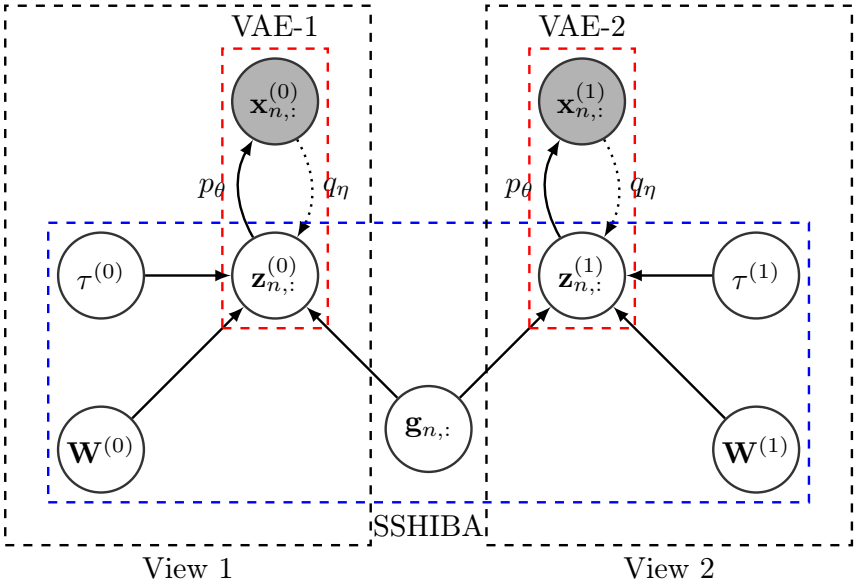


Figure 4.1: FA-VAE graphical model example with two VAEs. The blue dotted rectangle denotes the SSHIBA rv while the red rectangles indicate the two VAEs structures, one per view. Gray circles denote observations, and white circles represent rv.

by the shared global $\mathbf{g}_{\mathrm{n},:}$. If this issue is not addressed, the latent space $\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})}$ will only contain information for reconstructing the input, similar to an unsupervised vanilla VAE.

In order to create a $\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})}$ that shares information with the available multi-modal data, a possible approach is to impose a specific distribution as a regularisation term. Following this idea, we can use the distribution $\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})} \,|\, \mathbf{g}_{\mathrm{n},:}$ given by SSHIBA in Eq. (4.5) and employ it as a regularisation term of the VAE loss. This way, we propose using the following ELBO for each m-VAE:

$$\begin{aligned}\mathcal{L}^{(m)} = &\mathbb{E}_{q_\eta\left(\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})} \,|\, \mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})}\right)} \log\left(p_\theta(\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})} \,|\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})})\right) \\ &- \beta \mathrm{KL}\left(q_\eta(\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})} |\, \mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})}) || \mathcal{N}\left(\mathbf{g}_{\mathrm{n},:}\, \mathbf{W}^{(\mathrm{m})}, {\tau^{(\mathrm{m})}}^{-1}\right)\right),\end{aligned} \tag{4.7}$$

where the first term is the Gaussian Log-Likelihood (GLL) between the true observations $\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})}$ and their reconstruction sampled from $p_\theta(\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})} \,|\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})})$, and the second term minimises the KL divergence between the variational distribution $q_\eta(\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})} \,|\, \mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})})$ and the prior imposed in Eq. (4.5). Note that the ARD presented in the columns of $\mathbf{W}^{(\mathrm{m})}$ determines the relevant information that $\mathbf{g}_{\mathrm{n},:}$ contains w.r.t. our current view $\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})}$. The hyperparameter β serves as a balance between

two conflicting objectives: (i) the degree of fidelity in the reconstructed data and (ii) maximising the amount of shared information with the other views in the multi-modal dataset.

Regarding variational inference, let us define all random variables of the FA-VAE as three sets: $\theta_M \in \{\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})}, \mathbf{W}^{(\mathrm{m})}, \tau^{(\mathrm{m})}\}$ where m denote all possible views; $\theta_V \in \{\mathbf{Z}^{(v)}\}$ where v denote all VAE-like views; and $\theta_G \in \{\mathbf{g}_{\mathrm{n},:}\}$ which contains the global variables for all views. Hence, the set of random variables is $\Theta \in \{\theta_M, \theta_V, \theta_G\}$. SSHIBA updates the first set θ_M for all real, multi-label, binary or categorical data views by the mean-field variational inference explained in Chapter 2 in Table 2.2. The second set of variables, θ_V, are updated by the posterior of each v-VAE model, $q_\eta(\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})})$, defined in Eq. (4.4). Finally, the global latent variables θ_G are updated by mean-field approach but now differentiating between VAE and non-VAE views following

$$\mathbf{g}_{\mathrm{n},:} \sim \mathcal{N}\big(\mathbf{g}_{\mathrm{n},:} | \mu_{\mathbf{g}_{\mathrm{n},:}}, \Sigma_{\mathbf{Z}}\big)\,, \tag{4.8}$$

where

$$\mu_{\mathbf{g}_{\mathrm{n},:}} = \underbrace{\sum_{\mathrm{m}=1}^{\mathrm{M}} \left(\langle \tau^{(\mathrm{m})} \rangle \mathbf{X}^{(m)} \langle \mathbf{W}^{(\mathrm{m})} \rangle \Sigma_{\mathbf{Z}} \right)}_{\mathbf{I}} + \underbrace{\sum_{v=1}^{V} \left(\langle \tau^{(v)} \rangle \langle \mathbf{Z}^{(v)} \rangle \langle \mathbf{W}^{(v)} \rangle \Sigma_{\mathbf{Z}} \right)}_{\mathbf{II}}, \tag{4.9}$$

$$\Sigma_{\mathbf{Z}}^{-1} = \mathbf{I}_{\mathrm{K_c}} + \sum_{\mathrm{m}=1}^{\mathrm{M}} \left(\langle \tau^{(\mathrm{m})} \rangle \langle \mathbf{W}^{(\mathrm{m})^{\mathrm{T}}} \mathbf{W}^{(\mathrm{m})} \rangle \right). \tag{4.10}$$

Note that, in comparison to Table 2.2, now the mean of the global latent variable is calculated by two terms. The term **I** comprises all the m-views already tackled by SSHIBA, that is, real, multi-label or binary views. The term **II** contains all the v-views tackled now by each v-VAE architecture. However, regarding the calculation of $\Sigma_{\mathbf{Z}}^{-1}$ there is no distinction between views.

Following this procedure, FA-VAE trains according to the steps detailed in Algorithm 1.

4.2 Experiments

Throughout this section, we demonstrate the flexibility of FA-VAE in addressing relevant problems in deep probabilistic modelling. Firstly, in Section 4.2.1, we exploit the multi-view framework of FA-VAE to condition a pretained VAE to multilabel targets. Next, in Section 4.2.2, we apply it to a domain adaptation problem and compare it with the Multi-VAE model [7]. In addition, we disentangle and analyse the shared latent variables. Finally, in Section 4.2.3 we use the proposed FA-VAE framework to perform transfer learning between multiple VAEs, showing

Initialise $\mathbf{g}_{\mathrm{n},:}, \mathbf{W}^{(\mathrm{m})}, \mathbf{z}_{n,:}^{(V)}, \alpha_{\mathrm{k}}^{(\mathrm{m})}, \tau^{(\mathrm{m})}$
while *FA-VAE not converge* **do**
  Update $q(\mathbf{g}_{\mathrm{n},:})$ following Eq. (4.8);
  for *each view* **do**
    Update $q(\mathbf{W}^{(\mathrm{m})})$ following 2nd row of Table 2.2;
    for *each epoch* **do**
      Maximise the v-VAE's ELBO in Eq. (4.7)) ;
    end
    Update $\langle \mathbf{z}_{n,:}^{(V)} \rangle$ sampling from $q_\eta(\mathbf{z}_{n,:}^{(V)})$ defined in Eq. (4.4) ;
    Update $q(\alpha_{\mathrm{k}}^{(\mathrm{m})})$ following 3th row of Table 2.2;
    Update $q(\tau^{(\mathrm{m})})$ following 4th row of Table 2.2;
  end
end

Algorithm 1: FA-VAE training algorithm

how transfer learning creates a more expressive and understandable latent space than other models such as β-VAE [139]. The code for reproducing the following experiments can be found in GitHub[2].

4.2.1 FA-VAE as a conditioned generative model

In this first experiment, we demonstrate the utility of the FA-VAE framework in adapting a pretrained unconditioned VAE, $p_\theta(\mathbf{x})$, to model a conditional distribution, $p_\theta(\mathbf{x}|\mathbf{a})$, using a given set of labelled data $\{\mathbf{x}_i, \mathbf{a}_i\}_{i=1}^N$, where $\mathbf{x}_i$ denotes an observation and $\mathbf{a}_i$ denotes an attribute. We conducted our analysis using the CelebA dataset [164], which consists of 30000 samples of celebrity faces characterised by different attributes. Specifically, we selected three attributes to consider in our analysis, namely wearing lipstick, gender, and smiling. We used RGB images of size $64 \times 64 \times 3$ and ensured that the three attributes were represented in a stratified manner across the dataset.

We propose to use the FA-VAE multi-view framework to model a two-view setup for a given problem shown in Fig. 4.2. The first view is responsible for modelling the attributes (A) as a multi-label vector with three attributes: wearing lipstick, gender, and smiling. We use the standard multilabel SSHIBA's configuration for this view, without any VAE. In this case, we generate the binary attributes denoted as $\mathbf{x}_{n,:}^{(A)}$ from the pseudo-observation $\mathbf{z}_{n,:}^{(A)}$ using a Bernoulli distribution.

[2] https://github.com/aguerrerolopez/FA-VAE

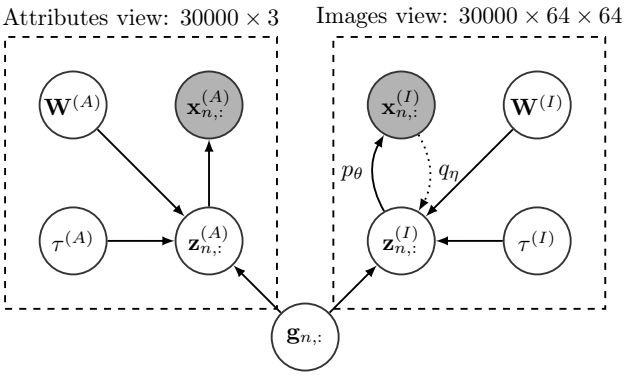


Figure 4.2: Conditioning a single VAE to a multilabel attribute vector using FA-VAE architecture where A denotes attributes view and I images views. Gray circles are observations, and white circles represent rv.

The mathematical expression for this distribution is given by

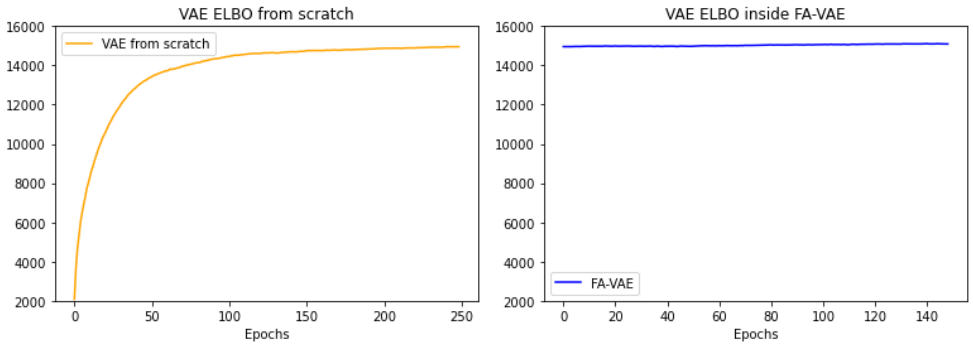
$$p\big(\mathbf{x}^{(A)}_{n,:}|\mathbf{z}^{(A)}_{n,:}\big) = \prod_{d=1}^{3} e^{z^{(A)}_{n,d} x^{(A)}_{n,d}} \sigma\Big(- z^{(A)}_{n,d}\Big), \tag{4.11}$$

where σ denotes the sigmoidal function.

In the second view (I), we consider observed RGB images denoted as $\mathbf{x}^{(I)}_{n,:} \in \mathbb{R}^{64\times64\times3}$. For the encoder and decoder, we adopt the network architecture proposed in the β-VAE paper [138]. Specifically, we utilize a CNN network with five convolutional layers, where the number of channels in each layer is $64, 128, 256, 512, 1024$, and the kernel size is 4, stride is 2, and padding is 1. Following the CNN layers, a fully connected layer is used to generate the parameters μ_η and Σ_η to infer the embedded latent variable $\mathbf{z}^{I}_{n,:} \in \mathbb{R}^{100}$. The decoder network follows the inverse structure of the encoder.

To train the FAVAE, we first train an unconditioned vanilla VAE on CelebA data until convergence, and then integrate it into FA-VAE's architecture to condition it on attributes. The convergence of these two steps is analysed in Figure 4.3. Specifically, Figure 4.3a shows the ELBO of the vanilla VAE trained from scratch, while Figure 4.3b shows the ELBO evolution during fine-tuning of FA-VAE. This indicates that the pre-trained VAE remains stable and does not lose reconstruction power, allowing for conditioning without interfering with its learning. Essentially, we demonstrate with these results that only 150 epochs are enough to condition an unsupervised model, such as a VAE, with FA-VAE.

The proposed model can be used to modify the attributes of a given image. Specifically, a CelebA image $\mathbf{x}^{(I)}_{n,:}$ is projected onto its embedded latent space $\mathbf{z}^{(I)}_{n,:}$ using $q_\eta(\mathbf{z}^{(I)}_{n,:}|\mathbf{x}^{(I)}_{n,:})$, and this latent space representation is fixed as the reference



(a) Vanilla VAE ELBO trained from scratch

(b) Pretained vanilla VAE ELBO inside FA-VAE's architecture

Figure 4.3: VAEs convergence by its own and inside FA-VAE's architecture. Fig. 4.3a shows the ELBO of a vanilla VAE trained on CelebA from scratch. In Fig. 4.3b we plug the vanilla VAE from Fig. 4.3a into FA-VAE's architecture.

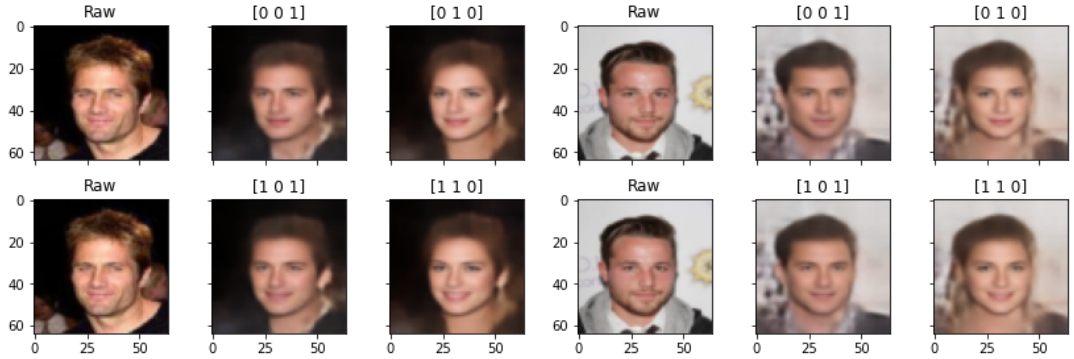
image. A set of attributes to be changed is selected to generate $\mathbf{z}_{n,:}^{(A)}$, and then the posterior predictive distribution of $\mathbf{g}_{\mathrm{n},:}$ (as shown in Table 2.2) is used, where:

$$\Sigma_{\mathbf{g}_{n,:}^*}^{-1} = \mathbf{I}_{\mathrm{K_c}} + \left(\langle\tau^{(\mathrm{A})}\rangle\langle\mathbf{W}^{(\mathrm{A})^{\mathrm{T}}}\mathbf{W}^{(\mathrm{A})}\rangle\right) + \left(\langle\tau^{(I)}\rangle\langle\mathbf{W}^{(I)^T}\mathbf{W}^{(I)}\rangle\right) , \tag{4.12}$$

$$\mu_{\mathbf{g}_{n,:}^*} = \left(\langle\tau^{(\mathrm{A})}\rangle\mathbf{z}_{n,:}^{(A)}\langle\mathbf{W}^{(\mathrm{A})}\rangle\Sigma_{\mathbf{g}_{n,:}^*} + \langle\tau^{(I)}\rangle\mathbf{z}_{n,:}^{(I)}\langle\mathbf{W}^{(I)}\rangle\Sigma_{\mathbf{g}_{n,:}^*}\right) . \tag{4.13}$$

We can utilise the proposed model to generate samples from the shared view $\mathbf{g}_{n,:}^*$ while keeping the private representations of the given attributes $\mathbf{z}_{n,:}^{(A)}$ fixed and $\mathbf{z}_{n,:}^{(I)}$ varied. Specifically, we sample $\mathbf{z}_{n,:}^{(I)}$ from a Gaussian distribution with mean $\mathbf{g}_{n,:}^*\mathbf{W}^{(I)}$ and variance $\tau^{(I)}$, and then generate new images using the generative distribution of FA-VAE $p_\theta(\mathbf{x}_{n,:}^{*(I)}|\mathbf{z}_{n,:}^{(I)})$. The resulting images are presented in Figs. 4.4a and 4.4b. These images demonstrate the ability of our model to alter attributes of the input image, such as gender (right column) or facial expression (smiling or neutral, top row vs. bottom row).

In addition, this model allows us to generate random conditioned facial images. The process for generating these images involves several steps. First, we create a random multilabel vector $\mathbf{x}_{n,:}^{(A)}$ using binary notation for attributes such as wearing lipstick, gender, and smiling. Next, we create the corresponding pseudo-observation $\mathbf{z}_{n,:}^{(A)}$ of this vector. Finally, we generate the posterior distribution of the global latent variable $\mathbf{g}_{\mathrm{n},:}$ given $\mathbf{z}_{n,:}^{(A)}$. This posterior distribution follows a Gaussian distribution



(a) Example 1 (b) Example 2

Figure 4.4: Different faces generated with FA-VAE by modifying their attributes. The left column of each subfigure represents the raw image. Each subfigure's centre and right columns represent the altered images by changing the different attributes indicated in the title, meaning [smile, lipstick, gender].

with parameters

$$\Sigma^{-1}_{\mathbf{g}^*_{n,:}} = I_{K_c} + \tau^{(\mathrm{A})}\, \mathbf{W}^{(\mathrm{A})^T}\, \mathbf{W}^{(\mathrm{A})}\,, \tag{4.14}$$

$$\mu_{\mathbf{g}^*_{n,:}} = \tau^{(\mathrm{A})}\, \mathbf{z}^{(A)}_{n,:}\, \mathbf{W}^{(\mathrm{A})}\, \Sigma_{\mathbf{g}^*_{n,:}}\,; \tag{4.15}$$

then we sample $\mathbf{z}^{(I)}_{n,:} \sim \mathcal{N}(\mathbf{g}^*_{n,:}\mathbf{W}^{(I)}, \tau^{(I)})$; and ultimately we use the generative distribution of FA-VAE $p_\theta(\mathbf{x}^{(I)}_{n,:}|\mathbf{z}^{(I)}_{n,:})$ to sample artificially generated conditioned images. Fig. 4.5 shows these images generated by 8 different configurations of the multilabel attributes $\mathbf{x}^{(A)}_{n,:}$.

4.2.2 Domain adaptation

In this study, we explore the modularity of FA-VAE and demonstrate its ability to combine multiple VAEs simultaneously. To illustrate this, we employ a three-view FA-VAE configuration, using the CelebA dataset [164] and the Google Cartoon Set [165] dataset. Specifically, we train a VAE with 10000 CelebA images in the first view and another VAE with 10000 Cartoon images in the second view. Additionally, we incorporate a binary label to model the hair color in the third view, using a categorical SSHIBA layer.

Our hypothesis is that the embedded variables $\mathbf{z}^{(m)}_{n,:}$ can capture domain information from face images (F), cartoon avatars (C), and hair color (H). The shared latent space variable, $\mathbf{g}_{\mathrm{n},:}$, serves as a bridge between domains, facilitating the adaptation of real-world faces to 2D cartoon avatars while accounting for hair color.

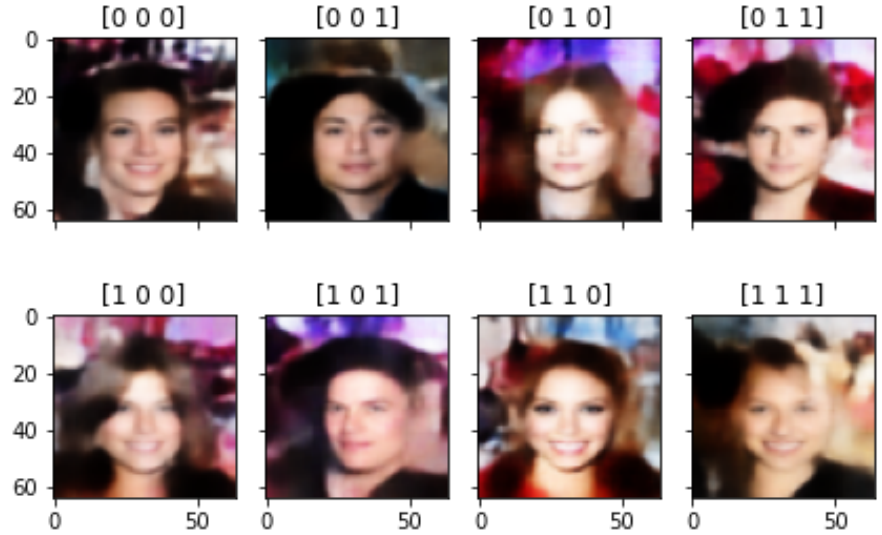


Figure 4.5: Fake faces generated by random $\mathbf{x}_{n,:}^{(A)}$ vectors. The title of each image indicates which attribute is activated: smiling, wearing lipstick, and gender. For example, [1 0 0] means a smile [1] without lipstick [0] on a woman's face [0], and [1 0 1] means a smile [1] without lipstick [0] on a male's face [1].

The architecture is depicted in Fig. 4.6. This approach has potential applications in various domains, including face recognition, character creation, and animation.

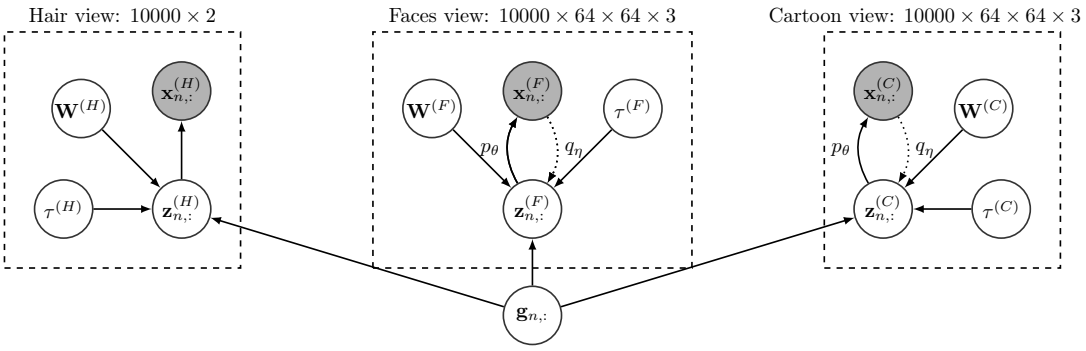


Figure 4.6: FA-VAE configuration to perform domain adaptation between two VAE-based views representing real-world faces (CelebA) and cartoon avatars (Cartoon) while conditioning to a third categorical view (hair). Grey circles denote observations, and white circles represent rv.

The CelebA faces view (F) comprises $10000 \times 64 \times 64 \times 3$ real celebrity face images, while the Cartoon view (C) consists of $10000 \times 64 \times 64 \times 3$ cartoon avatar

images. We use the β-VAE encoder-decoder configuration proposed in Section 4.2.1 by [139] with $\beta > 1$ for both views. The hair view (H) contains a binary label of 10000×2 indicating the hair color label: blond or brunette.

In this experiment we compare FA-VAE with the Multi-VAE [7] model, which utilises a discrete latent variable $\mathbf{c}_n$ to share the context of all views (see Fig. 4.7). In this experiment, we incorporate the hair label as an additional dimension in $\mathbf{x}_{n,:}^{(F)}$ and $\mathbf{x}_{n,:}^{(C)}$ for the Multi-VAE model.

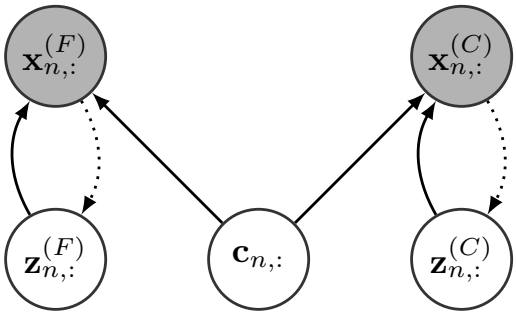


Figure 4.7: Multi-VAE configuration where two VAEs are conditioned with a categorical variable $\mathbf{c}_{n,:}$. CelebA images are represented by $\mathbf{x}_{n,:}^{(F)}$, Cartoon images are represented by $\mathbf{x}_{n,:}^{(C)}$, and $\mathbf{c}_{n,:}$ is the categorical variable which is shared by the two VAEs. Grey circles denote observations, and white circles represent rv.

Fig. 4.8 presents a comparison of the performance of FA-VAE and Multi-VAE models in the task of translating images from the CelebA domain to the Cartoon domain, based on hair colour information. The first row shows CelebA images, the second row displays the generated images by FA-VAE, and the third row shows the generated images by Multi-VAE. FA-VAE outperforms Multi-VAE in capturing inherent features such as sunglasses, in addition to hair colour. Multi-VAE fails to properly learn skin colour, as evidenced by Images 3 and 9, while FA-VAE produces high-quality 2D avatars without blur or artefacts.

The embedded latent variables $\mathbf{Z}^{(C)}$ and $\mathbf{Z}^{(F)}$ capture domain-specific information, while the global shared space $\mathbf{G}$ facilitates the transfer of information between domains, enabling domain adaptation to translate images from one domain to another. As an illustrative example, we begin by projecting two CelebA images into their embedded latent space, resulting in $\mathbf{z}_{1,:}^{(F)}$ and $\mathbf{z}_{2,:}^{(F)}$. We then project these into the shared latent space to obtain $\mathbf{g}_{1,:}$ and $\mathbf{g}_{2,:}$, which are used to sample the embedded latent space of the cartoon domain, resulting in $\mathbf{z}_{1,:}^{(C)}$ and $\mathbf{z}_{2,:}^{(C)}$. Finally, we interpolate the embedded representation of each pair of images in each domain using the convex combination:

$$\mathbf{z}_{\lambda,:}^{(v)} = \lambda \mathbf{z}_{1,:}^{(v)} + (1-\lambda)\mathbf{z}_{2,:}^{(v)}, \tag{4.16}$$

Figure 4.8: Domain adaptation from CelebA dataset to Cartoon dataset. The first row represents the original observations in CelebA dataset. In contrast, the second and third rows represent their translation to the Cartoon domain using **FA-VAE** (second row) and **Multi-VAE** (third row).

where v represents the domain (CelebA or Cartoon), and λ is a parameter that controls the degree of interpolation between the two images. Finally, using the embedded representation generated by $\mathbf{z}_{\lambda,:}^{(v)}$ we can use $p_\theta^v(\mathbf{x}_{\lambda,:}^{(v)}|\mathbf{z}_{\lambda,:}^{(v)})$ to generate a sample for each domain. The reconstructed sequences $\mathbf{x}_{\lambda,:}^{(v)}|\mathbf{z}_{\lambda,:}^{(v)}$ for different values of λ are shown in Fig. 4.9.

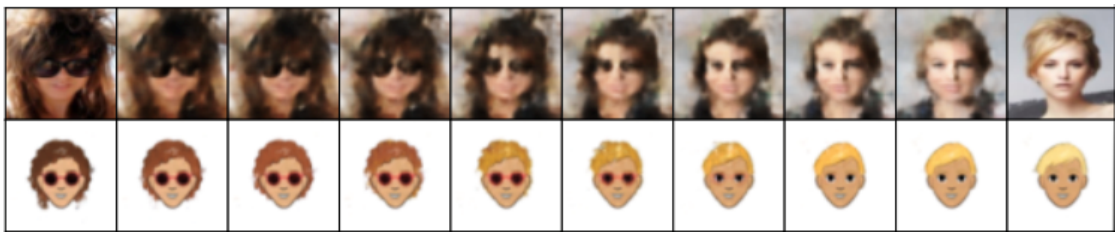


Figure 4.9: Examples of an image transformation and domain adaption application. The first and second rows show the evolution from $\mathbf{z}_{1,:}^{(F)}$ to $\mathbf{z}_{2,:}^{(F)}$ and, from $\mathbf{z}_{1,:}^{(C)}$ to $\mathbf{z}_{2,:}^{(C)}$, respectively.

Fig. 4.9 displays two rows of images showing a complete evolution from one image to the other through points sampled from the embedded space. The completeness of the embedded variables is demonstrated by the generation of meaningful images from any sampled point. Additionally, the gradual and smooth transitions between the images demonstrate the continuity of the embedded variables. In the cartoon domain, the images show a clear evolution in hairstyle, hair colour, and eyeglasses. Therefore, we can conclude that both embedded variables, $\mathbf{Z}^{(C)}$ and $\mathbf{Z}^{(F)}$, are informative and explainable, satisfying both completeness and continuity

criteria.

Similar to interpolating between points in an embedded latent space, we can interpolate directly from the global space by sampling from $\mathbf{g}_{n,:}$ and generating a sample that is interpretable by both domains. To illustrate this behaviour, we present two examples. We select two new CelebA observations $\mathbf{x}^{(F)}_{1,:}$ and $\mathbf{g}^{(F)}_{2,:}$ and project them to the global space $\mathbf{G}$ as $\mathbf{g}_{1,:}$ and $\mathbf{g}_{2,:}$. We then calculate $\mathbf{g}_{\lambda,:}$ using Eq. (4.16). Next, we project them back to their embedded spaces $\mathbf{z}^{(F)}_{\lambda,:}$ and $\mathbf{z}^{(C)}_{\lambda,:}$ by means of $\mathbf{W}^{(F)}$ and $\mathbf{W}^{(C)}$, respectively. Finally, each m-VAE decoder decodes $\mathbf{z}^{(F)}_{\lambda,:}$ and $\mathbf{z}^{(C)}_{\lambda,:}$ to create a pair of images. The reconstructed sequences $\tilde{\mathbf{x}}^{(m)}_{\lambda,:}|\mathbf{g}_{\lambda,:}$ are shown in Fig. 4.10.

Our analysis reveals a trade-off between the completeness and continuity of the generated content by each of the $\mathbf{g}_{\lambda,:}$. The CelebA faces domain exhibits better continuity resulting in a clear transition between the generated images. However, the Cartoon domain, characterised by a discrete set of features, shows better completeness with each $\mathbf{g}_{\lambda,:}$ generating a cartoon avatar without any distortion or artefact. We demonstrate that the global latent variable is both complete and continuous.

Figure 4.10: Example of an image transformation and domain adaptation application using a common global representation $\mathbf{g}_{\lambda,:}$. The first row shows images generated by the CelebA VAE while the second row shows images generated by the Cartoon VAE.

In this study, we utilise the FA-VAE model to analyse the interrelationships between multiple views of a dataset. To achieve this, we calculate the mean of each matrix $\mathbf{W}^{(\mathrm{m})}$ along its rows, resulting in a row vector of length K, where K is the number of latent factors in $\mathbf{g}_{n,:}$. This vector represents the relative importance of each latent factor in $\mathbf{g}_{n,:}$ for each view. The three resulting vectors are shown in Fig. 4.11, sorted by weight values in the hair view. The shared and unique latent features across views are then analysed based on their corresponding weights in the three views.

Our results show that the first 22 and the 25th latent features are shared between all three views, indicating that 23 latent features are required to explain the information shared across all three multimodal views. Additionally, latent feature 30 is shared exclusively between the CelebA and Cartoon datasets, but

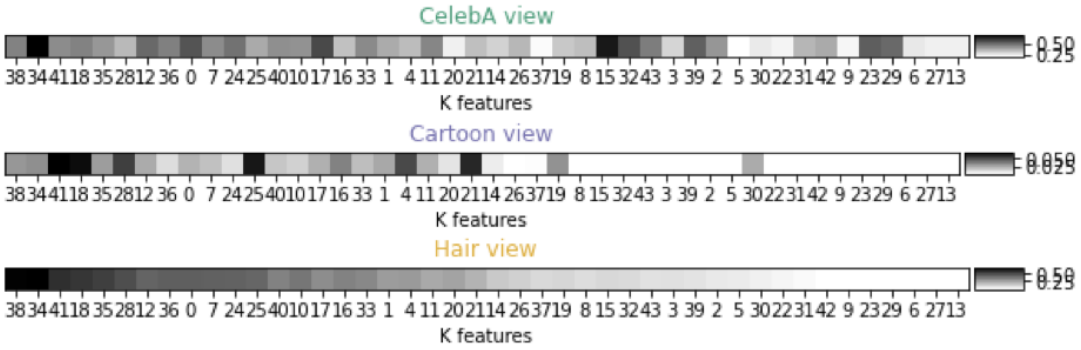


Figure 4.11: Mean over the rows of each $\mathbf{W}^{(\mathrm{m})}$ matrix. Each row is a vector representing the importance that each k latent feature of $\mathbf{g}_{\mathrm{n},:}$ has to reconstruct each view. The first row represents the Celeba view, the second row represents the Cartoon view, and the third row represents the Hair view.

it has no contribution to the hair view. Furthermore, we observe that five latent factors are exclusively used by the CelebA dataset, while almost all latent features are required to explain the complex CelebA dataset itself. Thus, the FA-VAE model allows for a visual analysis of the interrelationships between multiple views of a dataset, providing insights into the shared and unique information across the different views.

4.2.3 Transfer learning

The computational cost of training powerful VAEs is a significant challenge, as these deep networks require many epochs to train. To address this issue, we propose a novel approach to accelerate the training process using FA-VAE as a transfer learning tool between multiple VAEs operating on the same domain.

Consider the CelebA dataset as an illustrative example, with a two-view setup described by the graphical model shown in Figure 4.12. In the first view (V), we employ a pre-trained vanilla VAE, specifically the one trained in Section 4.2.1. As such, we only use the pre-trained encoder and decoder and do not train them again, resulting in static $\mathbf{z}_{n,:}^{(V)}$. For the second view (I), we begin with the architecture of the β-VAE discussed in Section 4.2.1 and add an additional final CNN layer to make it deeper. The final CNN layer has a channel size of 2048, kernel size of 4, stride of 2, and padding of 1 in the encoder, while the decoder exhibits an inverse structure.

Our hypothesis is that using the embedded space provided by a pre-trained vanilla VAE, $\mathbf{z}_{n,:}^{(V)}$, we can accelerate the training of a deeper VAE and potentially lead to a better global solution. In order to demonstrate this hypothesis, we

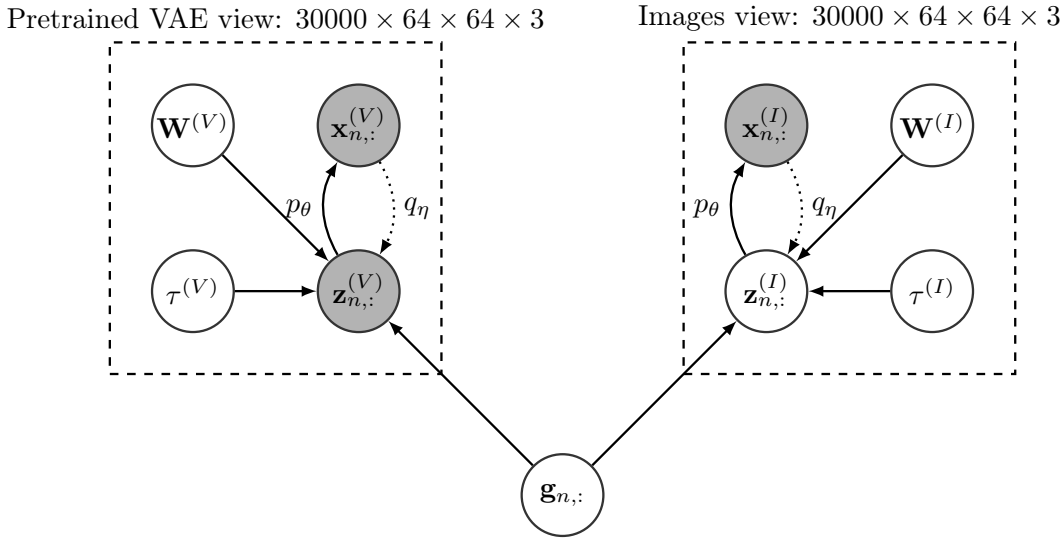
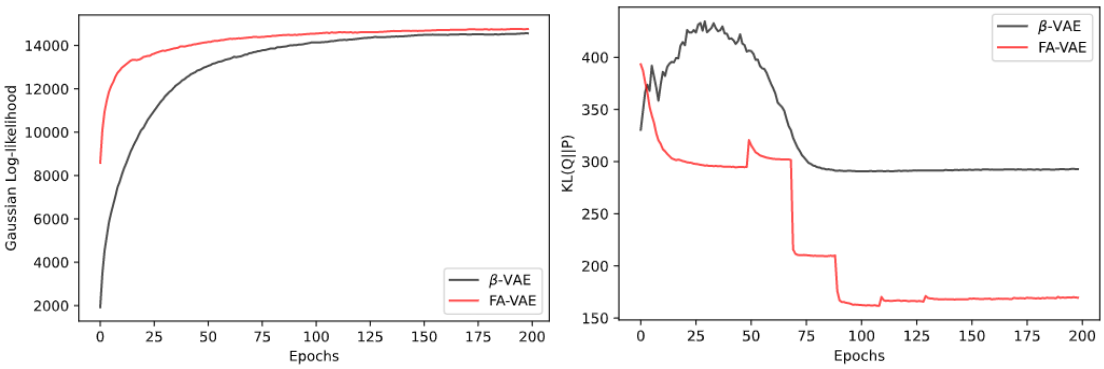


Figure 4.12: Transfer learning graphical model using FA-VAE. The V view represents information pre-learned by a vanilla VAE. As it is pretained, $\mathbf{z}^{(V)}_{n,:}$ is no longer a rv but an observation. The I view represents CelebA images using a β-VAE. Grey circles denote observations, and white circles represent rv.

compared two scenarios: our transfer learning approach using FA-VAE and a single β-VAE with the same structure as the second view of FA-VAE. We present the performance of both approaches in Figure 4.13. We observe that FA-VAE demonstrates accurate reconstruction capabilities in the initial epochs, indicating that the latent space of the vanilla VAE provides a good initialisation. Furthermore, FA-VAE achieves the same maximum GLL as β-VAE in approximately four times fewer epochs, demonstrating its faster speed. Additionally, FA-VAE achieves the highest absolute value in terms of GLL compared to β-VAE alone. The global term of KL for FA-VAE is lower than that for β-VAE, as shown in Figure 4.13b. The periodic spikes in the KL plot correspond to the times when the SSHIBA part of FA-VAE updates the VAE prior distribution, as seen in Algorithm 1. Meanwhile, for β-VAE, the KL starts increasing while FA-VAE directly decreases, which justifies the better behaviour since the beginning.

In Fig. 4.14, we present five randomly selected test images that were not used during the training phase. We apply three different models, namely β-VAE, FA-VAE, and Multi-VAE, to encode the images into their corresponding latent spaces, followed by reconstructing them back to their original domain. Since it might not be straightforward to discern the method that produces the best reconstruction, we provide the R2 score on the reconstruction of the 10,000 test images in Table 4.1. As can be seen from the table, FA-VAE outperforms the other models in terms of



(a) Reconstruction term measured by GLL.

(b) KL divergence term.

Figure 4.13: ELBO decomposition in the reconstruction term and the KL divergence term. The red line represents our approach, FA-VAE, while the black line represents β-VAE on its own.

the R2 score, indicating that transfer learning can enhance model performance.

Model	Samples	R2 score
Multi-VAE	10000	0.855 ± 0.154
β-VAE	10000	0.941 ± 0.032
FA-VAE	10000	$\mathbf{0.969 \pm 0.027}$

Table 4.1: Reconstruction performance measured in R2 score over 10,000 CelebA test samples.

FA-VAE offers the advantage of creating a more expressive and meaningful embedded latent representation of images compared to β-VAE. To demonstrate it, the 10 most relevant features are arbitrarily modified to analyse their visual impact on the reconstructed image. The 10 most relevant features were selected based on the absolute values in the weight matrix $\mathbf{W}^{(\mathrm{m})}$, for both embedded latent variables: $\mathbf{z} \in \mathbb{R}^{1x100}$ for β-VAE, and $\mathbf{z}^{(I)}_{n,:} \in \mathbb{R}^{1x100}$ for FA-VAE. These values were then randomly modified in the $[-20, 20]$ interval.

As shown in Fig. 4.15a, for β-VAE, each row represents one of the 10 most relevant features. The $\mathbf{z}$ latent space reveals that only three of the features have a visual interpretation. The blue rows show that facial hair can be controlled by increasing or decreasing its corresponding value. The green row also has an impact on the image contrast. However, the remaining features do not have a discernible visual impact on the images.

Figure 4.15b presents an evaluation of FA-VAE that demonstrates its superiority

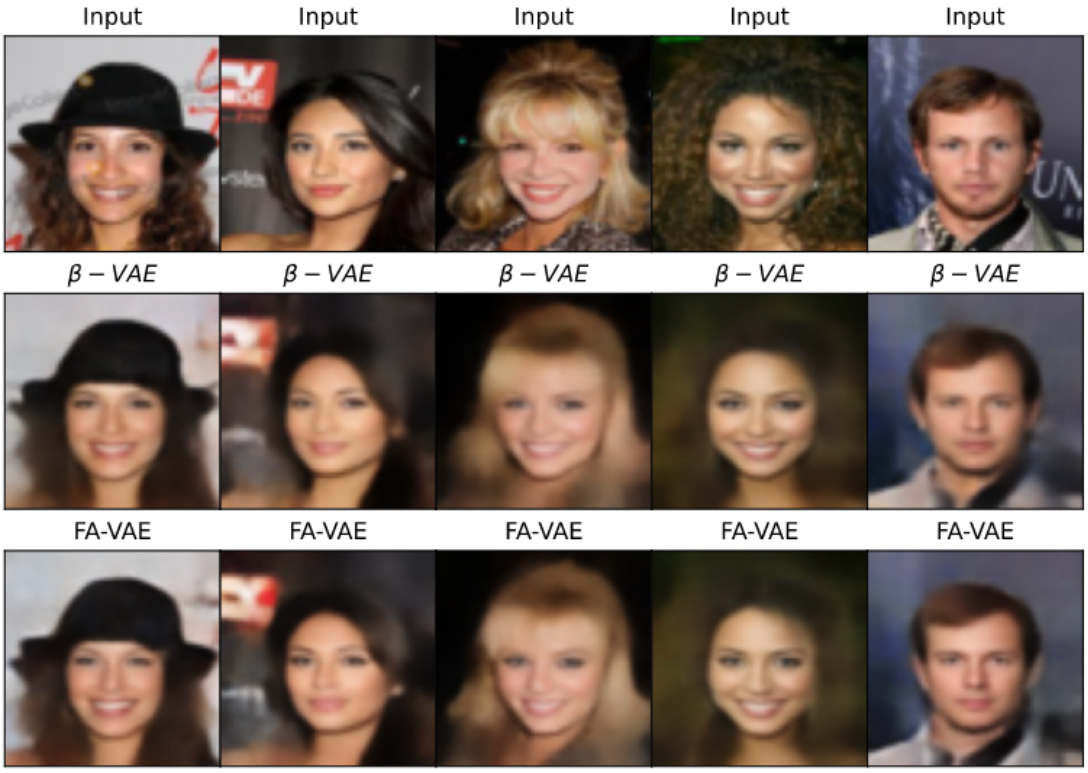


Figure 4.14: Images reconstructed by β-VAE and by FA-VAE

over β-VAE in terms of both interpretability and clarity of its embedded latent space. Among the 10 most relevant features identified, 9 are visually interpretable. For example, the golden-marked latent features control skin tone and facial rotation, the green rows determine gender and hairstyle, the blue row distinguishes between cold and warm background colours, and the grey-labelled latent features regulate smiling. Moreover, the study shows that FA-VAE can effectively capture facial information while filtering out noisy backgrounds in the latent space.

(a) β-VAE

(b) FA-VAE

Figure 4.15: Latent space evolution. Each row represents the 10 most relevant features based on absolute value. The red column represents the image generated by the model without any modification.

4.3 Conclusions

We have presented the first deep hierarchical VAE for mixed and heterogeneous data using an interpretable FA latent space in this work.

Having built on the foundation of the successful SSHIBA and KSSHIBA approaches in dealing with semi-supervised heterogeneous multi-view problems, we have developed the FA-VAE model to further expand the range of data domains it can handle. With the ability to work with a wide range of data types, including multilabel, continuous, binary, categorical, and even image data, depending on the VAE architecture, the FA-VAE model offers a versatile and powerful solution for real-world data sets.

Our results have shown that the FA-VAE model is capable of converting unsupervised problems into supervised ones, as demonstrated in Section 4.2.1, where a pretrained VAE has been accurately conditioned to different labels such as smiling, wearing lipstick, or new gender. Within just 150 epochs, the model has been able to readapt and condition itself to arbitrary labels. Additionally, we have demonstrated that the FA-VAE model can be used for domain adaptation between different image types, outperforming current SOTA approaches in translating real faces to emojis, as shown in Section 4.2.2. The FA-VAE model has also set a new standard in the field by being the first model capable of performing transfer learning between generative models, as demonstrated in Section 4.2.3. This has allowed for faster learning and improved performance, further solidifying its place as a robust solution for dealing with real-world data sets. To demonstrate the applicability of FA-VAE to real-world scenarios, in Chapter 6, we have applied it to automatically ribotyping *C. diff* isolates.

Overall, the FA-VAE model is a highly flexible and powerful supervised multi-VAE formulation for dealing with complex, high-dimensional data problems.

CHAPTER 5

AUTOMATIC ANTIBIOTIC RESISTANCE PREDICTION USING KSSHIBA

As part of the collaboration between the Department of Theory Signal and Communications from the UC3M and the Instituto de Investigación Sanitaria Gregorio Marañón, the technical contributions presented in this thesis have been tailored and adapted to actual microbiological contexts. Thus, in this chapter, we customise and implement the technical contributions explained in Chapter 3 to the prediction of antimicrobial resistant mechanisms in *K. pneumoniae*.

K. pneumoniae is widely recognised as a significant public health concern by major international health organisations, due to its rapid dissemination, substantial morbidity and mortality rates, and the financial impact associated with its treatment and control [38, 36, 37]. The existence of multidrug-resistant strains that harbour these resistance mechanisms complicates treatment options and the patient's prognosis [44, 43].

During the past decade, different outbreaks of *K. pneumoniae* were registered by Hospital General Universitario Gregorio Marañón (HGUGM) and various surveillance programmes (STEP and SUPERIOR) [166, 167] that were analysed in Hospital Universitario Ramón y Cajal (HURyC) creating two multidrug resistant data collections *K. pneumoniae*: one of 282 samples and another of 120 samples. Each bacteria strain is characterised by its MALDI-TOF MS, that is, a mass spectrometry containing 20k features for each sample.

As previously discussed in Chapter 2, current machine learning techniques have various limitations. As displayed in Table 1.1, certain studies [86, 87, 88, 89] concentrated on small datasets with less than 50 samples, suggesting solutions that lack the ability to generalise. Most of them used proprietary black-box tools such as ClinProTools [86, 87, 88, 89, 91, 93, 92]. However, only a few studies [11, 96, 68, 97] proposed open and reproducible research. Influenced by previous research [94, 11, 96, 68, 97] we proposed the use of open source code and publicly share the data used. Finally, drawing inspiration from [11, 46], we proposed the use of probabilistic modelling to conduct the prediction of *K. pneumoniae* resistance mechanisms.

Therefore, we apply and tailor KSSHIBA [122], our technical contribution presented in Chapter 3, to this microbiological scenario. As observed, the issue presented is a high-dimensional problem, suffering from the data curse, i.e., too

many features for a limited number of samples. Thus, we apply KSSHIBA directly, using well-established kernel functions such as linear kernel to reduce dimensionality; Radial Basis Function (RBF) kernel to to exploit non-linearity present in the data; or specifically tailored kernels for MALDI-TOF MS data such as PIKE [11] which we adapt to work within the KSSHIBA framework.

Using KSSHIBA, we propose to simplify the existing pipeline for determining antibiotic resistance mechanisms by leveraging epidemiological information using multimodal learning. In doing so, we aim to demonstrate two points: first, the applicability of KSSHIBA's model in real-world scenarios and its motivation; second, that KSSHIBA surpasses current state-of-the-art models, such as XGBoost [168], MLPs, RF, SVMs, or GPs. To do so, KSSHIBA conducts dimensionality reduction using kernel methods and automatically handles parameter selection using Bayesian optimisation. In particular, it explains each view by a linear product with a matrix weight that varies across views, enabling identification of which latent dimensions are explaining each of the views. We evaluated performance in two different bacterial domains: (1) using data from HGUGM and (2) grouping strains from 18 geographically dispersed hospitals, selected based on their phenotypic and genotypic resistance to beta-lactams and all analysed at HURyC. From the results, we conclude that a heterogeneous model with linear kernels must be used to predict susceptibility to AR.

The application results presented in this chapter are published at Engineering Applications of Artificial Intelligence from Elsevier [45]. In accordance with the open science philosophy upheld in this thesis, the implementation of the FA-VAE model and all associated experimentation detailed in this chapter are readily accessible through a public repository on GitHub, under the link [1]. The developed work was carried out in collaboration with the team led by Belén Rodríguez-Sánchez of Instituto de Investigación Sanitaria Gregorio Marañón (IISGM) and with the team led by Dra. Rosa del Campo of Instituto Ramón y Cajal de Investigación Sanitaria (IRyCIS).

The chapter is organised as follows: In Section 5.1, we provide the motivation for using KSSHIBA in the prediction of antibiotic resistance mechanisms. Then, in Section 5.2, we describe the *K. pneumoniae* collections, the technical adaptations made to KSSHIBA, and the models that were selected for comparison. In Section 5.3, we conduct several experiments to predict ESBL and CP resistance mechanisms. Finally, in Section 4.3, we draw conclusions and highlight the key findings of the study.

[1] https://github.com/aguerrerolopez/RMPrediction

5.1 Motivation to use KSSHIBA

Current SOTA solutions for the MALDI-TOF MS analysis tend to share the following common pipeline: (1) a time-consuming MALDI-TOF MS data preprocessing, e.g. using the *MALDIquant* (MQ) package; (2) a dimensionality reduction technique such as Principal Component Analysis (PCA); and (3) a classification/regression algorithm that has to be cross-validated to choose its hyperparameters. Moreover, this pipeline results in black-box private softwares that are not interpretable or lacks transparency in order to understand which variables or sources of information are performing the classification task, such as ClinProTools [86, 91, 92] or Clover BioSoft [93].

Our proposal simplifies this pipeline while provides explainable results, including epidemiological information, and outperforming the SOTA models in prediction AR. The use of KSSHIBA as a predictor for AR is motivated by several factors. Firstly, KSSHIBA can efficiently handle raw MALDI-TOF MS data, eliminating time-consuming pre-processing MQ, and avoiding external preprocessing. Secondly, KSSHIBA performs a double dimensionality reduction, obtaining a common low-dimensional latent representation of all input data sources through kernelised data representations that use matrices of the dimension number of samples (N) instead of the number of features (D). Thirdly, hyperparameter tuning is eliminated since the Bayesian nature of the model implies the ability to automatically optimize the model parameters by maximizing the variational lower bound. Fourthly, KSSHIBA provides interpretable solutions since it calculates a weight matrix for each view, capable of explaining how they correlate. Lastly, KSSHIBA tackles epidemiological differences by indicating the origin center of each spectrum, recognizing that two similar bacteria can be from different strains.

In this chapter, we propose to improve the efficiency of the current antibiotic prescription pipeline in hospitals by proposing a new pipeline as shown in Fig. 5.1. Our proposal involves the use of KSSHIBA to analyse multimodal information and reduce the time required for making antimicrobial treatment decisions by 24-72 hours. It is important to note that we do not suggest replacing the current methods of Antimicrobial Susceptibility Testing (AST), but rather to complement and accelerate the process. In our proposal, we suggest using ML-guided antimicrobial treatments during the waiting period for empiric results to improve the overall efficiency of the process.

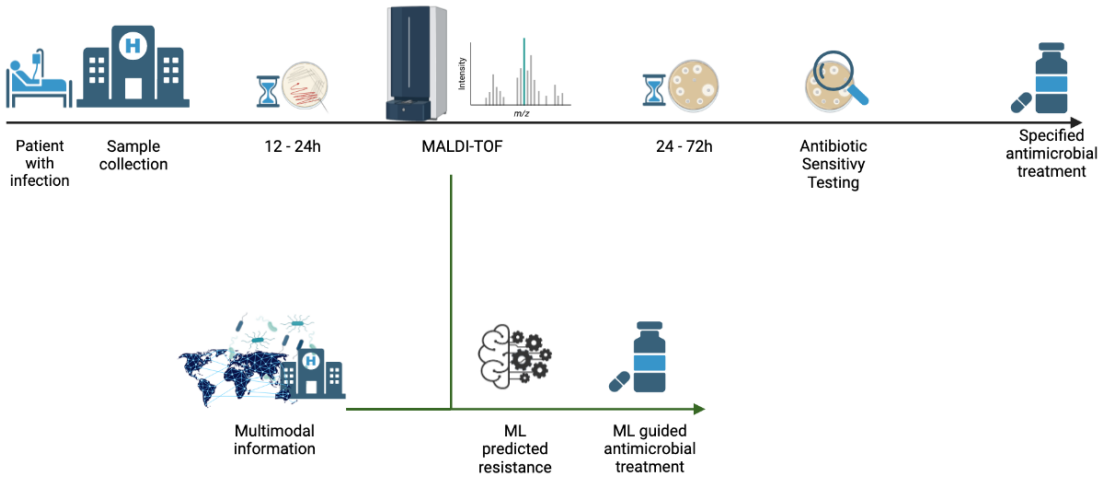


Figure 5.1: Infection treatment workflow enhanced by multimodal ML methods.

5.2 Materials and Methods

5.2.1 Isolates selection and processing

In this study, we have included two distinct data domains obtained from two different hospitals, namely Hospital General Universitario Gregorio Marañón (HGUGM) and Hospital Universitario Ramón y Cajal (HURyC). Due to differences in their origins and potential variations in analysis techniques, each domain has been treated as a separate entity to account for any potential epidemiological or technical differences.

The first data domain consists of 282 consecutive clinical *K. pneumoniae* isolates collected between 2014 and 2019, which were isolated and analysed at the HGUGM institution. This dataset is the most homogenised, as all isolates were collected and analysed in the same institution. Hereafter, we refer to this data domain as the HGUGM domain.

The second data domain comprises 120 isolates that were characterised in surveillance programs (STEP and SUPERIOR) [167, 166] obtained from 8 Spanish and 11 Portuguese hospitals. However, the MALDI-TOF MS spectra were all analysed at the HURyC institution. Therefore, this data domain is referred to as the HURyC domain.

Both domain datasets were kept frozen at $-80°C$ in skim milk and, after thawing, cultured overnight at $37°C$ on Columbia blood agar (bioMérieux, Lyon, France) for 3 subcultures for metabolic activation. The MS analysis has been centralised and performed by the same operator, at each institution, using an MBT Smart MALDI Biotyper mass spectrometer (Bruker Daltonics, Bremen), in 6 separate replicates (2 positions on 3 consecutive days). Protein extraction was

performed by adding 1µl 100% formic acid and then drying at room temperature. Next, 1µl of HCCA matrix solution (Bruker Daltonics) was added to each spot. The MALDI-TOF spectra have been acquired in a positive linear mode in the range of 2 kDa to 20 kDa, using default settings [169], although only data between 2,000-12,000 m/z [170, 171] has been used.

Both HGUGM and HURyC institutions utilised the same procedure to determine Antimicrobial Susceptibility Testing (AST) for their respective datasets, which involved the use of the automated broth microdilution method of the Microscan® System (Beckman-Coulter, CA, USA) in accordance with the common criteria EUCAST (2021). It is important to note that in this study, the term AST refers to the laboratory procedure used to empirically verify resistance mechanisms, while Antibiotic Resistance (AR) refers to the automatic prediction using machine learning techniques. The presence of Extended-Spectrum Beta-Lactamases (ESBL)/Carbapenemases (CP) genetic-resistant mechanisms was confirmed through molecular tests. Each isolate was labelled as Wild Type (WT), ESBL-producer, or ESBL+CP-producer, as shown in Table 5.1.

Table 5.1: 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

The Ethics Committees of both HGUGM and HURyC (codes MICRO.HGUGM.2020-002 and 087-16, respectively) have approved this study. The study was carried out on microbiological samples, not human products, and informed consent from the patient was not required.

5.2.2 Multi-view KSSHIBA for MALDI-TOF MS data

In this study, we propose adapting the KSSHIBA model [122] to effectively predict the susceptibility of CP and ESBL for each isolate using MALDI-TOF MS data. As explained in Chapter 3, KSSHIBA is a Bayesian multi-view semi-supervised model designed to address non-linearities in the data, such as those in MALDI-TOF MS. Furthermore, it can operate in a reduced dimensionality space ($N \times N$), where $N << D$, by leveraging kernel data representations and projecting all input views onto a shared, low-dimensional latent space. To exploit these functionalities in this problem, we explore two KSSHIBA architectures.

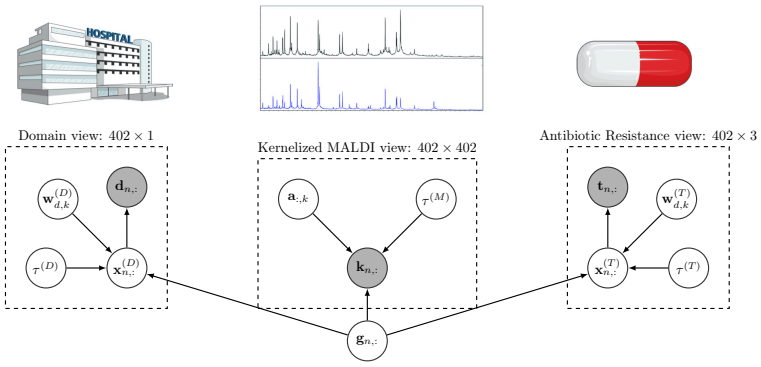


Figure 5.2: Probabilistic graphical model for the evaluated data set: view **D** corresponds to the label of the domain they come from (HGUGM or HURyC), view **M** corresponds to the kernelised MALDI-TOF MS data, and view **T** corresponds to the AR (WT, ESBL or ESBL+CP). The white circles represent random variables that the model learns, while the grey circles represent the observations.

For the first configuration, we propose using a multi-view framework composed of three data sources: the kernelised MALDI-TOF MS data, the AR to be predicted, and the domain label indicating the hospital where the data was analyzed. The graphical model representing this multi-view approach is depicted in Figure 5.2. By adopting this approach, we are able to effectively handle the unique characteristics of the MALDI-TOF MS data. The kernelised view is able to efficiently manage the non-linearities of the data. The domain view enables us to account for potential epidemiological differences between isolates, and the low-dimensional common space allows us to model the interaction between the domain, the data, and the AR in a transparent and explainable manner.

The MALDI-TOF data, i.e. the M view, is kernelised, $\mathbf{K} \in \mathbb{R}^{N \times N}$, where N is the number of samples in each experiment and $N << D$. Furthermore, choosing different kernel functions allows modelling the relations between peaks in different ways, such as using PIKE or RBF kernels. Then, each row of $\mathbf{K}$ represents a kernelised observation, denoted as $\mathbf{k}_{\mathrm{n},:}$:

$$\mathbf{k}_{\mathrm{n},:} = \left[\mathrm{kf}(M_n, M_1), ..., \mathrm{kf}(M_n, M_N)\right], \tag{5.1}$$

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

In this view, denotes as M, KSSHIBA considers that a common low-dimensional latent variable vector $\mathbf{g}_{\mathrm{n},:} \in \mathbb{R}^{1 \times K}$ is linearly combined with a set of dual variables

$\mathbf{A} \in \mathbb{R}^{N \times K}$, where K is the dimension of the low-dimensional latent space, and a zero mean Gaussian noise $\epsilon^{(M)}$ to generate each row of the kernelised observations $\mathbf{k}_{\mathrm{n},:}$, as:

$$\mathbf{k}_{\mathrm{n},:} \;=\; \mathbf{g}_{\mathrm{n},:}\,\mathbf{A}^{\mathrm{T}} + \epsilon^{(M)}\,, \tag{5.2}$$

where the prior over the latent space is given by $\mathbf{g}_{\mathrm{n},:} \sim \mathcal{N}(0, I_K)$; the prior over the k-th column of the dual variables $\mathbf{A}$, $\mathbf{a}_{:,k}^{(M)}$, is given by $\mathbf{a}_{:,k}^{(M)} \sim \mathcal{N}(0, (\alpha_{\mathrm{k}})^{-1} I_K)$; and the noise $\epsilon^{(M)} \sim \mathcal{N}\Big(0, \tau^{(\mathrm{M})^{-1}}\Big)$. Thus, the random variable $\alpha_{\mathrm{k}} \sim \Gamma(a, b)$ follows an ARD prior [130] over the columns of $\mathbf{A}$ to automatically select the columns of $\mathbf{g}_{\mathrm{n},:}$ (latent factors) that are indeed relevant to explain the current data view.

For the AR view, denoted as T, we propose to use a one-hot encoding for the Wild Type (WT), ESBL, and ESBL + Carbapenemases (ESBL+CP) tags. Similarly, for the data domain, denoted as D, we consider binary encoding where a value of 0 means that the data come from the HGUGM domain and a value of 1 means that the data come from the HURyC domain.

To accommodate for these two binary observations, we first consider that there exist two real latent variables $\mathbf{X}^{(\mathrm{m})}$, $m \in \{T, D\}$, which are generated by the common low-dimensional latent variable $\mathbf{g}_{\mathrm{n},:} \in \mathbb{R}^{1 \times K}$, and then are linearly combined with a projection matrix $\mathbf{W}^{(\mathrm{m})} \in \mathbb{R}^{D \times K}$ (where D_m is the observation dimension) and a Gaussian noise with zero mean $\epsilon^{(m)}$, as follows:

$$\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})} = \mathbf{g}_{\mathrm{n},:}\,\mathbf{W}^{(\mathrm{m})^{\mathrm{T}}} + \epsilon^{(m)} \text{ for } m \in \{T, D\}\,, \tag{5.3}$$

where $\mathbf{W}^{(\mathrm{m})}$'s prior is identical to $\mathbf{A}$'s to automatically select which columns of $\mathbf{g}_{\mathrm{n},:}$ 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 an independent Bernoulli probability model [129], for the AR view, as:

$$p\big(\mathrm{t}_{\mathrm{n},:} \,|\mathbf{x}_{n,:}^{(T)}\big) = \prod_{d=1}^{3} p\Big(\mathrm{t}_{\mathrm{n,d}} \,|\mathbf{x}_{n,d}^{(T)}\Big)\,, \tag{5.4}$$

where

$$p\Big(\mathrm{t}_{\mathrm{n,d}} \,|\mathbf{x}_{n,d}^{(T)}\Big) = e^{\mathbf{x}_{n,d}^{(T)}\,\mathrm{t}_{\mathrm{n,d}}}\sigma\Big(-\mathbf{x}_{n,d}^{(T)}\Big). \tag{5.5}$$

And for domain view, as it is binary:

$$p\Big(\mathrm{d}_{\mathrm{n},1} \,|\mathbf{x}_{n,1}^{(D)}\Big) = e^{\mathbf{x}_{n,1}^{(D)}\,\mathrm{d}_{\mathrm{n},1}}\sigma\Big(-\mathbf{x}_{n,1}^{(D)}\Big)\,. \tag{5.6}$$

The model is trained by evaluating the posterior distribution of all posteriors of random variables given the observed data. These posteriors are approximated through mean-field variational inference [127] maximising the Evidence Lower

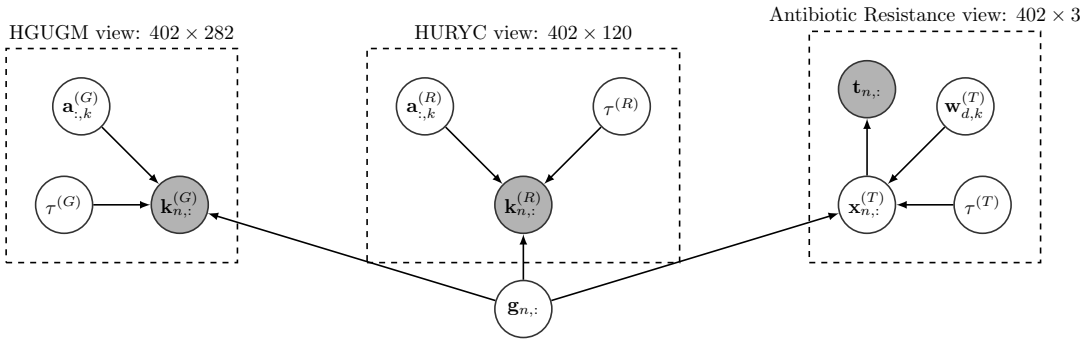


Figure 5.3: Probabilistic graphical model for multi-centre approach. The white circles represent random variables that the model learns, while the grey circles represent the observations.

BOund (ELBO). For more details, see [2, 122]. 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 marginalise any type of missing values in the data, as well as predict the test samples for AR by sampling its variational distribution. Table 5.2 shows the updated rules for the mean field factor for each random variable in the KSSHIBA model.

The utilisation of a single view for both data collection may prove too rigid, especially when using a parameterised kernel, such as an RBF one, which requires the same parameter value for both collections in the previous approach. As a result, we propose an alternative configuration, a novel multi-centre approach that employs a specific kernel for each centre. To facilitate this, we present a three-view scheme, as illustrated in Fig. 5.3.

The first view consists of MALDI-TOF MS data with Support Vectors (SV)s from the HGUGM, while the second view contains MALDI-TOF MS SVs from the HURyC. As in this problem the HGUGM contains 282 data points and the HURyC contains 120 data points, the construction of each kernel view is as follows: the HGUGM view compute a 402×282 kernel matrix, being the kernel between all the data and the HGUGMs SVs; however the cross-domain kernels (the kernel between HURyC data and HGUGM SVs) are set to zero, since the domain differences between these MALDI-TOF MS makes these cross-domain kernels be uninformative. This way, the last 120 rows of this kernel matris are zero. Similarly, the HURyC view comprises a 402×120 kernel matrix with the first 282 rows set to zero.

Finally, the third view comprises information on antibiotic resistance for all data points. By following this procedure, KSSHIBA can achieve domain adaptation

by translating MALDI-TOF MS data from one domain to another. Note that this configuration makes the latent variable $\mathbf{g}_{\mathrm{n},:}$ use the MALDI-TOF MS information related to each domain but the output space (the third view with the antibiotic resistance) is jointly learn by both domains.

5.2.3 Kernels for MALDI-TOF MS

As a kernel function, we first test a nonlinear approach, such as RBF [172] which is given by:

$$\mathrm{kf}(M_a, M_b) = \exp\left(- \frac{||M_a - M_b||^2}{2\sigma^2} \right), \tag{5.7}$$

where σ is the variance hyperparameter. Then, we also test a linear kernel [173] that follows:

$$\mathrm{kf}(M_a, M_b) = M_a^T M_b \,. \tag{5.8}$$

In both cases, M_a and M_b are a pair of MALDI-TOF spectra.

Finally, we work with a SOTA kernel function called PIKE [11], which exploits the nonlinear correlations between the MALDI-TOF peaks as follows:

$$\mathrm{kf}(M_a, M_b) = \frac{1}{2\sqrt{2\pi t}} \sum_{i,j}^{\#Peaks} \lambda_i \lambda'_j \exp\left(- \frac{(p_i - p'_j)^2}{8t} \right), \tag{5.9}$$

where t is a smoothing parameter that must be cross-validated, λ_i, λ_j correspond to the intensity values of each pair of peaks and p_i, p_j is their m/z position in the spectra. Recall that each MALDI-TOF consists of 12,000 different peaks. Due to the computational cost to evaluate Eq. (5.9) in that number of peaks, the spectra are preprocessed beforehand by topological peak selection [11]. This peak selection is a simple peak detection method based on the persistence concept from computational topology, which automatically results in peak detection because local maxima exhibit high persistence values by construction in MALDI-TOF data. Following the indications of the authors of [11], only 200 peaks are kept per sample.

Table 5.2: Updated rules, obtained by a mean field approximation, of q distribution for the different variables of the KSSHIBA model. The first row is common for all views. From row 2 to row 4 is only for the M view. Lastly, the three last rows are for views T and D.

Variable	$\boldsymbol{q^*}$ **distribution**	**Parameters**
$\mathbf{g}_{\mathrm{n},:}$	$\mathcal{N}(\mathbf{g}_{\mathrm{n},:}\,\vert\mu_{\mathbf{g}_{\mathrm{n},:}},\Sigma_{\mathbf{G}})$	$\mu_{\mathbf{g}_{\mathrm{n},:}}=\sum_{m\in\{M,T,D\}}\left(\langle\tau^{(\mathrm{m})}\rangle\,\mathbf{K}^{(\mathrm{m})}\langle\mathbf{A}^{(\mathrm{m})}\rangle\Sigma_{\mathbf{G}}\right)$ $\Sigma_{\mathbf{G}}^{-1}=\left(\mathbf{I}_{\mathrm{K_c}}+\sum_{m\in\{M,T,D\}}\left(\langle\tau^{(\mathrm{m})}\rangle\langle\mathbf{A}^{(\mathrm{m})^{\mathrm{T}}}\,\mathbf{A}^{(\mathrm{m})}\rangle\right)\right)$
$\mathbf{A}^{(\mathrm{m})}$	$\prod_{\mathrm{n}=1}^{\mathrm{N}}\left(\mathcal{N}\left(\mathbf{a}_{\mathrm{n},:}^{(\mathrm{m})}\,\vert\mu_{\mathbf{a}_{\mathrm{n},:}^{(\mathrm{m})}},\Sigma_{\mathbf{A}^{(\mathrm{m})}}\right)\right)$	$\mu_{\mathbf{a}_{\mathrm{n},:}^{(\mathrm{m})}}=\langle\tau^{(\mathrm{m})}\rangle\,\mathbf{K}^{(\mathrm{m})^{\mathrm{T}}}\langle\mathbf{G}\rangle\Sigma_{\mathbf{A}^{(\mathrm{m})}}$ $\Sigma_{\mathbf{A}^{(\mathrm{m})}}^{-1}=\left(\mathrm{diag}(\langle\boldsymbol{\alpha}^{(\mathrm{m})}\rangle)+\langle\tau^{(\mathrm{m})}\rangle\langle\mathbf{G}^{\mathrm{T}}\,\mathbf{G}\rangle\right)$
$\alpha_{\mathrm{k}}^{(\mathrm{m})}$	$\Gamma\left(\alpha_{\mathrm{k}}^{(\mathrm{m})}\,\vert a_{\alpha_{\mathrm{k}}^{(\mathrm{m})}},b_{\alpha_{\mathrm{k}}^{(\mathrm{m})}}\right)$	$a_{\alpha_{\mathrm{k}}^{(\mathrm{m})}}=\frac{N}{2}+a^{\boldsymbol{\alpha}^{(\mathrm{m})}}$ $b_{\alpha_{\mathrm{k}}^{(\mathrm{m})}}=b^{\boldsymbol{\alpha}^{(\mathrm{m})}}+\frac{1}{2}\langle\mathbf{A}^{(\mathrm{m})^{\mathrm{T}}}\,\mathbf{A}^{(\mathrm{m})}\rangle_{k,k}$
$\tau^{(\mathrm{m})}$	$\Gamma\left(\tau^{(\mathrm{m})}\,\vert a_{\tau^{(\mathrm{m})}},b_{\tau^{(\mathrm{m})}}\right)$	$a_{\tau^{(\mathrm{m})}}=\frac{N^2}{2}+a^{\tau^{(\mathrm{m})}}$ $b_{\tau^{(\mathrm{m})}}=b^{\tau^{(\mathrm{m})}}+\frac{1}{2}\Bigg(\sum_{\mathrm{n}=1}^{\mathrm{N}}\sum_{\tilde{\mathrm{n}}=1}^{\tilde{\mathrm{N}}}\mathrm{k}_{\mathrm{n,n}}^{(\mathrm{m})^2}$ $-2\,\mathrm{Tr}\left\{\langle\mathbf{A}^{(\mathrm{m})}\rangle\langle\mathbf{G}^{\mathrm{T}}\rangle\,\mathbf{K}^{(\mathrm{m})}\right\}$ $+\mathrm{Tr}\left\{\langle\mathbf{A}^{(\mathrm{m})^{\mathrm{T}}}\,\mathbf{A}^{(\mathrm{m})}\rangle\langle\mathbf{G}^{\mathrm{T}}\,\mathbf{G}\rangle\right\}\Bigg)$
$\mathbf{W}^{(\mathrm{m})}$	$\prod_{\mathrm{d}=1}^{\mathrm{D_m}}\mathcal{N}\left(\mathbf{w}_{\mathrm{d},:}^{(\mathrm{m})}\,\vert\mu_{\mathbf{w}_{\mathrm{d},:}^{(\mathrm{m})}},\Sigma_{\mathbf{W}^{(\mathrm{m})}}\right)$	$\mu_{\mathbf{w}_{\mathrm{d},:}^{(\mathrm{m})}}=\langle\tau^{(\mathrm{m})}\rangle\mathbf{X}^{(m)^T}\langle\mathbf{G}\rangle\Sigma_{\mathbf{W}^{(\mathrm{m})}}$ $\Sigma_{\mathbf{W}^{(\mathrm{m})}}^{-1}=\mathrm{diag}(\langle\boldsymbol{\alpha}^{(\mathrm{m})}\rangle)\langle\gamma_{\mathrm{d}}^{(\mathrm{m})}\rangle+\langle\tau^{(\mathrm{m})}\rangle\langle\mathbf{G}^{\mathrm{T}}\,\mathbf{G}\rangle$
$\alpha_{\mathrm{k}}^{(\mathrm{m})}$	$\prod_{\mathrm{k}=1}^{\mathrm{K_c}}\Gamma\left(\alpha_{\mathrm{k}}^{(\mathrm{m})}\,\vert a_{\alpha_{\mathrm{k}}^{(\mathrm{m})}},b_{\alpha_{\mathrm{k}}^{(\mathrm{m})}}\right)$	$a_{\alpha_{\mathrm{k}}^{(\mathrm{m})}}=\frac{D_m}{2}+a^{\boldsymbol{\alpha}^{(\mathrm{m})}}$ $b_{\alpha_{\mathrm{k}}^{(\mathrm{m})}}=b^{\boldsymbol{\alpha}^{(\mathrm{m})}}+\frac{1}{2}\sum_{\mathrm{d}=1}^{\mathrm{D_m}}\langle\gamma_{\mathrm{d}}^{(\mathrm{m})}\rangle\langle\mathrm{w}_{\mathrm{d,k}}^{(\mathrm{m})}\,\mathrm{w}_{\mathrm{d,k}}^{(\mathrm{m})}\rangle$
$\tau^{(\mathrm{m})}$	$\Gamma\left(\tau^{(\mathrm{m})}\,\vert a_{\tau^{(\mathrm{m})}},b_{\tau^{(\mathrm{m})}}\right)$	$a_{\tau^{(\mathrm{m})}}=\frac{D_m N}{2}+a^{\tau^{(\mathrm{m})}}$ $b_{\tau^{(\mathrm{m})}}=b^{\tau^{(\mathrm{m})}}+\frac{1}{2}\Bigg(\sum_{\mathrm{n}=1}^{\mathrm{N}}\sum_{\mathrm{d}=1}^{\mathrm{D_m}}\mathrm{x}_{\mathrm{n,d}}^{(\mathrm{m})^2}$ $-2\,\mathrm{Tr}\left(\langle\mathbf{W}^{(\mathrm{m})}\rangle\langle\mathbf{G}^{\mathrm{T}}\rangle\mathbf{X}^{(m)}\right)$ $+\mathrm{Tr}\left(\langle\mathbf{W}^{(\mathrm{m})^{\mathrm{T}}}\,\mathbf{W}^{(\mathrm{m})}\rangle\langle\mathbf{G}^{\mathrm{T}}\,\mathbf{G}\rangle\right)\Bigg)$

5.2.4 Model training and validation

In our study, we investigate two scenarios for analyzing our data: (1) training and testing on each domain separately, which is known as intra-domain analysis, and (2) training and testing on both domains together, which is known as inter-domain analysis.

For inter-domain analysis, we merge the datasets in two different ways: (1) by directly combining the two domains, using only the kernelised MALDI-TOF view and the AR view, and (2) by adding a third view that indicates the domain of each data point. In the first case, we do not use the D observations, while in the second case, we use them to study the importance of knowing the data origin. We divide each domain into five random training-test folds to evaluate performance. To correct for label imbalance, we oversample the minority class in each training fold to obtain consistent class ratios across the folds.

We compare kernelised Sparse Semi-Supervised Interbattery Bayesian Analysis (KSSHIBA) with an SVM and a GP since all three models can work with kernel formulations. Since we are solving a multiclass classification problem, we train the SVMs and GPs in a one-vs-all scheme. In addition, we also compare ourselves to multitasking RF.

Regarding cross-validation, we use an inner 5-fold over the training folds to validate all hyperparameters. To do so, we followed a 5-fold grid search cross-validation technique to cross-validate the parameter C in the set of values $\{0.01, 0.1, 1, 10\}$ for the SVM, and for the RF we adjust the number of estimators and the maximum number of features exploring the values $\{50, 100, 150\}$ and $\{auto, log2\}$, respectively. For both KSSHIBA and GP, the hyperparameters are optimised by maximising the ELBO and the marginal log-likelihood of the data, respectively. When using RBF kernel, we cross-validate γ parameter in a logspace with 13 steps in the range $\{-9, 3\}$. When using PIKE, we also cross-validate the t smoothing value in the range of $\{1, 5, 10\}$.

Finally, to demonstrate that KSSHIBA does not need external preprocessing, we use our model with and without MQ preprocessing. When we use MQ, we denote it using the prefix *MQ-*, e.g., MQ-KSSHIBA.

We measure performance in terms of the Area Under the ROC Curve (AUC) of the AR prediction on the test folds.

5.3 Results

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. Later, we analyse the performance in the inter-domain scenario to evaluate the advantages of working with multi-view data sources. Finally,

we study the projection of the latent space learnt by KSSHIBA to understand the correlation between the source domain and the labels.

Each model has a name that is made up of three terms, which refer to the preprocessing method, the type of model, and the kernel function used. For example, the name "KSSHIBA-RBF" means that the model uses raw MALDI-TOF data, the KSSHIBA model, and an RBF kernel function. On the other hand, the name "MQ-GP-PIKE" means that the model uses the MQ package to preprocess the MALDI-TOF data, a GP model, and the PIKE kernel function. In the intra-domain scenario, the multi-view nature of KSSHIBA is denoted by the term "Preprocessing-Model Kernel Domain". For instance, the model KSSHIBA-LINEAR-DOMAIN refers to the use of KSSHIBA model with a linear kernel function without any preprocessing, and with the addition of domain labels.

5.3.1 Intra-domain scenario

Tables 5.3 and 5.4 summarise the results obtained by training and testing independent models for each domain (HGUGM and HURyC).

Table 5.3: Results of nonlinear 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. The last row indicates the Weighted average AUC over all the data.

Dataset	Label	KSSHIBA RBF	KSSHIBA PIKE	MQ-GP PIKE [11]	SVM RBF	MQ-SVM RBF	MQ-RF	RF
GM	WT	0.61±0.14	0.71±0.16	**0.75±0.11**	0.67±0.12	0.71±0.18	0.74±0.15	0.70±0.17
	ESBL	**0.57±0.28**	0.56±0.32	0.35±0.14	0.40±0.29	0.53±0.21	0.45±0.21	0.39±0.21
	ESBL+CP	**0.85±0.14**	0.78±0.09	0.79±0.07	0.82±0.19	0.83±0.17	0.82±0.12	0.80±0.19
RyC	WT	0.47±0.35	**0.64±0.19**	0.56±0.20	0.45±0.15	0.45±0.21	0.52±0.09	0.57±0.26
	ESBL	0.70±0.10	0.43±0.09	0.43±0.11	**0.72±0.14**	0.52±0.12	0.58±0.13	0.69±0.10
	ESBL+CP	0.67±0.12	0.43±0.09	0.55±0.05	0.71±0.17	0.57±0.07	0.69±0.06	**0.71±0.07**
Weighted average AUC		**0.74 ± 0.13**	0.66 ± 0.18	0.68 ± 0.17	**0.74 ± 0.11**	0.71±0.16	0.73±0.12	0.71 ± 0.11

Table 5.4: Results of linear models in the intra-domain scenario in terms of mean AUC and standard deviation w.r.t. the 5 random splits. The last row indicates the weighted Weighted average AUC over all the data.

Dataset	Label	KSSHIBA LINEAR	GP LINEAR
GM	WT	0.70±0.15	0.70±0.18
	ESBL	0.46±0.19	0.54±0.18
	ESBL+CP	0.77±0.16	0.80±0.20
RyC	WT	0.49±0.22	0.48±0.28
	ESBL	0.59±0.08	0.58±0.14
	ESBL+CP	0.66±0.05	0.62±0.06
Weighted average AUC		0.70 ± 0.09	**0.71 ± 0.11**

According to the results presented in Table 5.3, when analysing the HGUGM domain, the KSSHIBA model outperforms the baselines in terms of AUC for both the prediction of ESBL and ESBL+CP. It is worth noting that non-linear kernels are more effective, as evidenced by the comparison between Table 5.3 and Table 5.4. Specifically, the RBF kernel is the most suitable option for predicting both ESBL and ESBL+CP, while the PIKE kernel is the best option for predicting WT.

In contrast, the HURyC domain presents significant challenges for modeling, as none of the models tested achieved satisfactory performance across all three labels. Nonetheless, it is noteworthy that the application of non-linear techniques, such as PIKE, RBF, and RF, yielded better results than their linear counterparts.

5.3.2 Inter-domain scenario

Table 5.5 and Table 5.6 show the results obtained by linear and nonlinear models, respectively, when trained jointly on both HGUGM and HURyC domains. When using the first approach, i.e., both MALDI-TOF MS data collections are completely combined in one view, we denote it as KSSHIBA-KERNEL-DOMAIN. In contrast, when using the second approach, i.e., each MALDI-TOF MS data collection is modelled by an independent kernel, we denote it as KSSHIBA-KERNEL-MULTI.

Table 5.5: Results of linear 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. The last row indicates the weighted Weighted average AUC over all the data.

Dataset	Label	KSSHIBA LINEAR MULTI	KSSHIBA LINEAR DOMAIN	MQ-KSSHIBA LINEAR DOMAIN	KSSHIBA LINEAR	GP LINEAR	SVM LINEAR
GM	WT	0.76±0.05	**0.77±0.11**	**0.78±0.06**	0.72±0.14	0.76±0.10	0.62±0.13
	ESBL	**0.57±0.24**	0.46±0.19	0.51±0.25	0.39±0.21	0.43±0.20	0.39±0.21
	ESBL+CP	0.82±0.03	**0.88±0.08**	**0.90±0.05**	0.86±0.08	0.86±0.08	0.85±0.08
RyC	WT	0.62±0.09	**0.70±0.16**	0.48±0.28	0.66±0.16	0.68±0.17	0.59±0.20
	ESBL	0.60±0.10	0.55±0.09	0.53±0.07	0.49±0.09	0.60±0.10	**0.69±0.12**
	ESBL+CP	0.55±0.13	**0.68±0.10**	0.63±0.10	0.64±0.06	0.64±0.04	0.66±0.14
Average performance		0.72±0.15	**0.77 ± 0.13**	**0.77 ± 0.18**	0.74 ± 0.17	0.76 ± 0.13	0.74 ± 0.13

Table 5.6: Results of nonlinear models in the inter-domain scenario in terms of mean AUC and standard deviation w.r.t. the 5 random splits. The MG-GP PIKE means reproducing the work done in [11]. The last row indicates the weighted Weighted average AUC over all the data.

Dataset	Label	KSSHIBA PIKE DOMAIN	KSSHIBA RBF DOMAIN	KSSHIBA RBF MULTI	MQ-GP PIKE [11]	SVM RBF
GM	WT	0.64±0.17	0.59±0.16	0.71±0.05	**0.74±0.14**	0.65±0.13
	ESBL	0.44±0.37	0.32±0.23	**0.53±0.28**	0.40±0.12	0.40±0.19
	ESBL+CP	0.73±0.12	0.81±0.12	0.80±0.05	0.77±0.10	**0.85±0.08**
RyC	WT	0.51±0.14	**0.63±0.07**	0.47±0.22	0.62±0.22	0.57±0.26
	ESBL	0.39±0.14	0.66± 0.05	**0.69±0.10**	0.49±0.09	0.69±0.12
	ESBL+CP	0.51±0.14	0.59±0.11	0.62±0.14	0.59±0.06	**0.68±0.14**
Average performance		0.62 ± 0.15	0.70 ± 0.13	0.73±0.11	0.69 ± 0.13	$\mathbf{0.75 \pm 0.12}$

Table 5.7: Results of SOTA non-kernel methods in terms of mean AUC and standard deviation w.r.t. the 5 random splits. The last row indicates the weighted Weighted average AUC over all the data.

Dataset	Label	LightGBM	MLP	XGBoost
GM	WT	0.58 ± 0.08	$\mathbf{0.80 \pm 0.08}$	0.76 ± 0.02
	ESBL	$\mathbf{0.68 \pm 0.09}$	0.50 ± 0.12	0.60 ± 0.17
	ESBL+CP	0.62 ± 0.04	$\mathbf{0.82 \pm 0.07}$	0.74 ± 0.08
RyC	WT	$\mathbf{0.75 \pm 0.21}$	0.52 ± 0.20	0.64 ± 0.33
	ESBL	0.41 ± 0.13	$\mathbf{0.57 \pm 0.09}$	0.52 ± 0.07
	ESBL+CP	0.60 ± 0.05	$\mathbf{0.68 \pm 0.06}$	0.64 ± 0.06
Average performance		0.58 ± 0.09	$\mathbf{0.74 \pm 0.12}$	0.69 ± 0.09

When both domains are combined into the same view, i.e., the first approach KSSHIBA-KERNEL-DOMAIN, the linear version with domain labels outperforms all SOTA models in predicting both WT and ESBL+CP, as shown in Table 5.5. Moreover, KSSHIBA-LINEAR-DOMAIN also outperforms all models used in the previous experiments that were targeted to each particular domain. With this configuration, the KSSHIBA model can handle heterogeneous data, such as the domain label, which allows it to exploit the information present in both data sets.

When comparing the Average Performance of the kernelised experiments, i.e., Table 5.5 and Table 5.6, the linear kernel outperformed all non-linear approaches, namely PIKE and RBF kernels. This may be induced because the distributions in the two domains are significantly different and a common parameter for the kernel that can capture these differences is not feasible. In fact, the multi-centre approach, KSSHIBA RBF MULTI, which uses different γ per view, namely the cross-validated ones in Section 5.3.1, outperformed KSSHIBA RBF DOMAIN where the same γ value is used as a unique kernel is presented.

When comparing KSSHIBA with and without a domain view, that is, KSSHIBA-KERNEL vs KSSHIBA-KERNEL-MULTI vs KSSHIBA-KERNEL-DOMAIN, the results pointed out that including domain information in the model improved its performance a both multi-view approaches outperform the KSSHIBA-KERNEL. We hypothesise that the domain information allowed the model to eliminate any possible local epidemiological bias and merge both datasets effectively.

Additionally, the KSSHIBA model had an advantage over other models in that it did not require external preprocessing with MQ. While MQ preprocessing performed well in the HGUGM domain, it performed poorly in the HURyC domain, likely due to biasing the data towards the larger domain. Using raw data instead of preprocessed data improved the performance of the model in the HURyC domain, while maintaining a similar performance in the HGUGM domain.

Regarding the multi-centre approach, the non-linear kernel showed a improvement in their performance as seen in Table 5.6. However, no global improvement has been accomplished in comparison to including the domain when using linear kernel, as seen in Table 5.5.

When comparing the performance of linear and non-linear kernel functions for the prediction of both WT and ESBL+CP, it was found that linear kernels were better suited for both tasks. This indicates that in the first experiment, where only one domain was considered, the model was over-fitted, and by incorporating out-of-distribution data, the linear kernel was able to generalise better.

After examining the relevant literature, we find that the current SOTA models for antibiotic resistance (AR) prediction include LightGBM, XGBoost, and MLP neural networks. In order to compare these models with our proposed KSSHIBA method, we conducted experiments as shown in Table 5.7. Our results indicate that both XGBoost and MLP show promising performance. However, our proposed KSSHIBA method achieves higher average performance scores than either of these models. Additionally, none of the models, including KSSHIBA, is able to achieve high AUC values for ESBL prediction across all tasks. We posit that this is due to a lack of sufficient data on ESBL, which is represented by only 64 samples in the dataset with a relatively low representation in the AR profile. Hence, additional data on ESBL is required to enable effective learning of its variability and generalisation to new data.

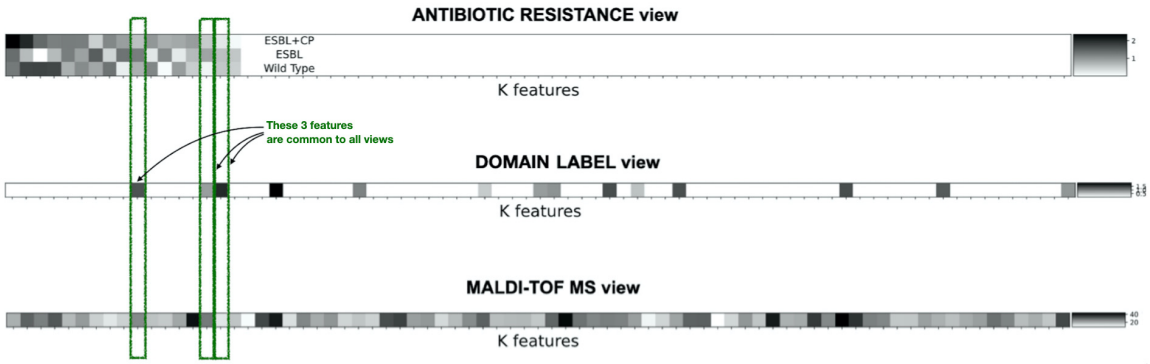


Figure 5.4: Latent space correlation between input views. Each row represents the mean of each $\mathbf{w}_{\mathrm{d},:}^{(\mathrm{m})}$ d-row having then 76 values, one per each k latent feature. Each subplot represents one $\mathbf{W}^{(\mathrm{m})}$ matrix per view. The most important features (the 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

5.3.3 Latent space analysis

Given that the KSSHIBA LINEAR DOMAIN model has shown the best performance among the considered models, we will now investigate the learned latent space projection to gain insights into the significance of the domain label.

Fig. 5.4 displays the average weight of each latent factor $\mathbf{W}^{(\mathrm{m})}$ for $m \in M, T, D$, which is calculated as the average across every column. The matrix $\mathbf{W}^{(M)}$ is obtained by moving from the dual space variables $\mathbf{A}$ to the primal space as shown in the following equation:

$$\mathbf{W}^{(M)} = M^T \mathbf{A}, , \tag{5.10}$$

where M corresponds to the raw MALDI-TOF observations. Because of the sparsity induced by the ARD prior, $\mathbf{W}^{(\mathrm{m})}$ automatically selects only the k relevant features of $\mathbf{g}_{\mathrm{n},:}$ for each m data view.

In this instance, the KSSHIBA model determined that a reduced set of 76 latent features was sufficient, as illustrated in Fig. 5.4. It is worth noting that of these 76 features, only 14 are relevant for predicting the AR label, and all of them are associated with the MALDI-TOF view. However, only 3 of these 14 features are capable of capturing all the available information simultaneously. Additionally, it is worth mentioning that the MALDI-TOF view requires 51 private latent features, which capture the behavior of this view alone, much like a PCA would do.

In Fig. 5.4, it can be observed that there is a correlation between the AR of each strain and its original domain, as indicated by the presence of 3 shared latent features. Additionally, the domain label is utilised to explain the projection of the

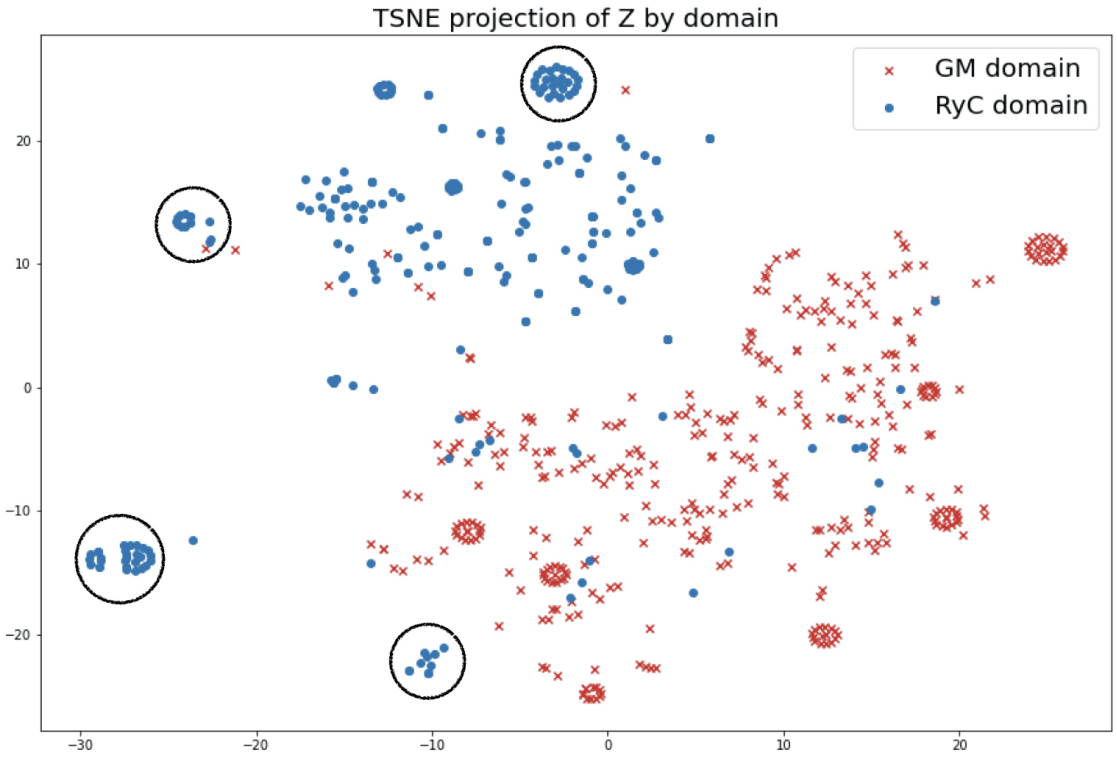


Figure 5.5: t-SNE 2-dimensional representation of the 14 latent variables of $\mathbf{G}$ that are relevant for the domain view. Red crosses stand for a $\mathbf{g}_{n,:}$ whose observation comes from the HGUGM domain, while every blue dot stands for a $\mathbf{g}_{n,:}$ whose observation comes from the HURyC domain.

MS, thereby demonstrating that the distributions of MALDI-TOF differ depending on the epidemiology.

In terms of low-dimensional projection, Fig. 5.5 illustrates the use of T-distributed Stochastic Neighbor Embedding (t-SNE) to map the 14 relevant latent features that explain the domain view onto a 2-dimensional space. The two domains exhibit separate distributions, which can be easily distinguished using a simple linear classifier. The HGUGM domain is more compact due to its samples being derived from a single hospital, whereas the HURyC domain, which consists of samples from 19 different hospitals, displays a sparser distribution with distinct clusters of data points. The four black circles in Fig. 5.5 correspond to four different hospitals within the HURyC domain collection. Therefore, our model can cluster the data by both domain and hospital, without being explicitly informed of this information.

By reducing the initial 10,000 MALDI-TOF features to only 76 latent features, KSSHIBA is able to produce a low-dimensional latent space representation. To-

gether with domain characterisation, this accounts for the superior results achieved by KSSHIBA over all SOTA methods in the inter-domain scenario, since it can identify the differences between MALDI-TOF data depending on their domain, facilitating improved generalisation of predictions.

5.4 Conclusions

In this study, we present a novel method for predicting antibiotic resistance of *K. pneumoniae* to ESBL and CP production that utilises both MALDI-TOF spectra and epidemiological information. Our approach is based on an enhanced version of the KSSHIBA algorithm that achieves superior performance in terms of AUC when compared to state-of-the-art algorithms such as XGBoost, LightGBM, MLP, SVMs, and GP. Notably, our method is the first to process raw MALDI-TOF data without requiring external preprocessing using MQ. It also offers dimensionality reduction by simultaneously integrating MALDI-TOF and epidemiological data into a low-dimensional latent space. Moreover, the KSSHIBA algorithm provides interpretable results by leveraging epidemiological information through its multi-view architecture. Furthermore, it is capable of automatically tuning the model hyperparameters using Bayesian inference.

In this chapter, we investigated the advantages of our approach in two distinct bacterial domains: (1) using data from a single hospital (GM) and (2) grouping strains from 18 hospitals across different geographic locations, selected based on their phenotypic and genotypic resistance to beta-lactams (RyC). We found that all current non-heterogeneous models, such as GPs or SVMs, exhibited overfitting to one of the domains and performed poorly in the smaller domain, as shown in Table 6. Therefore, heterogeneous models that can analyze epidemiological information are necessary to predict AR in a fair and unbiased manner between domains. Our experiments demonstrated that it is critical to adjust for different data distributions when working with two domains simultaneously. In fact, the inclusion of domain information improved the learning process of KSSHIBA, enabling it to properly model the different data distributions by overcoming the bias introduced by the data itself, which can lead to overfitting, particularly if there is domain imbalance. However, the process of learning a probabilistic model involves learning a full distribution of the data, which adds complexity to the training phase.

Our work contributes towards the important goal of reducing ineffective antibiotic prescribing by enabling the prediction of possible resistance mechanisms in *K. pneumoniae*. The implementation of our method in microbiological laboratories has the potential to improve the detection and treatment of multidrug-resistant infections, as well as significantly reduce the time required to obtain resistance results compared to traditional manual methods. This could have a substantial impact on global public health by improving patient outcomes.

As a next step, we propose a longitudinal study over real test samples in clinical settings, where KSSHIBA and other baseline models can be used to automatically predict AR and assess its viability in a real-world scenario. After validation, we plan to develop a web server for KSSHIBA as a rapid AR detection method in laboratories.

CHAPTER 6

AUTOMATIC RIBOTYPING BASED ON PROBABILISTIC TECHNIQUES

Following the collaboration with the Instituto de Investigación Sanitaria Gregorio Marañón, we customise and implement the technical contributions explained in Chapters 3 and 4 to a pilot study of the automatic ribotyping of *Clostridium difficile* (*C. diff*).

C. diff is a pathogen that has gained widespread recognition for its significant economic burden in the United States [100] and its rapid spread in Europe [101]. This pathogen is particularly prevalent among patients who have recently been treated with antibiotics [102] and its infection can cause life-threatening complications, such as sepsis [103] or colitis [104], particularly in severe cases. Currently, hyper-virulent *C. diff* strains, such as RT027 [107, 108], are associated with severe colitis and higher mortality rates [109, 110, 111] due to a deletion in the tcdC gene, which regulates toxins production. Current laboratories use real-time PCR assays, for example the GeneXpert *C. diff*, to detect the presence of toxins B and binary, and the deletion in the tcdC gene in *C. diff* strains [112]. However, new ribotypes, such as RT181, which share those characteristics with hyper-virulent RT027 but differs in symptoms, present challenges for accurate detection using PCR methods.

Therefore, in this chapter we propose a solution using ML techniques following the philosophy established in Chapter 5. To do so, we apply and tailor KSSHIBA [122] (Chapter 3) and FA-VAE [123] (Chapter 4) to the automatic ribotyping of *C. diff* based on their MALDI-TOF MS. We follow the same approach as in Chapter 5 using KSSHIBA with RBF, Linear and PIKE kernels. Then, we compare their performance in time, ribotyping accuracy, and latent space dimensionality to FA-VAE. We adapt FA-VAE, presented in Chapter 4, by implementing different VAE architectures, such as 1D-CNN and MLP encoder-decoder. Both models are analysed over a dataset of 275 *C. diff* from the Hospital General Universitario Gregorio Marañón (GM) containing 10 different ribotypes (RT) that are grouped into three groups: hyper-virulent RT027, new RT181, and nonvirulent RTs. In this dataset, a train-test split is performed to evaluate the performance of the different models. Finally, best models are evaluated in a real outbreak involving 3 *C. diff* isolates ocurred on January 24th. Specifically, FA-VAE perfectly predicted the RT the first day of the outbreak, while current laboratories methodologies lasted 6 days on obtaining results. Our objective is not only to demonstrate the effectiveness of

probabilistic models, such as FA-VAE, in automating microbiological laboratories and rapid ribotyping *C. diff*, but to drastically reduce the time required for this process by up to 6 days, ultimately enabling more efficient and effective management of outbreaks and control of infectious diseases.

The results of the applications presented in this chapter have been submitted as a preliminary study to the XXVI Congreso Nacional de la Sociedad Española de Enfermedades Infecciosas y Microbiología Clínica. In accordance with the open science philosophy upheld in this thesis, all associated experiments detailed in this chapter are readily accessible through a public repository on GitHub, under the link [1], developed in collaboration with the team led by Belén Rodríguez-Sánchez of the IISGM.

The organisation of this chapter is as follows: Section 6.1 reviews the state-of-the-art in the automatic ribotyping of *C. diff* and motivates the use of ML techniques. Section 6.2 describes the *C. diff* collection, including the pre-processing steps and technical adaptations to both KSSHIBA and Factor Analysis VAE (FA-VAE). Section 6.3 presents the results of two experiments: a preliminary study to test the feasibility of the proposal and a study of a real outbreak of *C. diff*. Finally, in Section 6.4, the chapter concludes with a summary of the main findings and suggestions for future work.

6.1 State-of-the-art in *C. diff* ribotyping

In the last decade, more than 70 different strains of *C. diff* have been discovered [106]. In particular, the RT027 strain is often referred to as hyper-virulent [107, 108] due to its association with severe colitis and higher mortality rates [109, 110, 111]. Recent studies [113, 114] discovered an outbreak of a new pathogenic ribotype, called RT181, which is detected as false RT027 by PCR methods as it also presents a deletion in the tcdC gene. Hence, it presents a challenge to an accurate detection in clinics.

As noted by Cuénod et al. [115] in their 2021 literature review, MALDI-TOF MS can be used for ribotyping of *C. diff*. This was previously demonstrated by Reil et al. [116] in 2011 through the analysis of 355 *C. diff* samples using MALDI-TOF MS, where biomarkers were discovered for manual identification of RT001, RT027, and RT078/126 in the mass range between $3K - 13K$ Da. Moreover, Rizzardi et al. [117] in 2015 also observed that extended MALDI-TOF MS, with mass range of $30K - 50K$ Da, was able to identify biomarkers for various ribotypes, including RT010, RT011, RT012, RT015, RT017 and RT020, among others.

To date, no studies have properly applied ML methods to perform automatic ribotyping of *C. diff*. In Calderaro et al. [121] MALDI-TOF MS was used on a

[1] https://github.com/aguerrerolopez/Clostridium

set of 29 *C. diff* samples to classify them as epidemic or non-epidemic, but were unable to achieve conclusive results regarding automatic ribotyping. Additionally, they did not make their code or data publicly available exhibiting the same trend as in *K. pneumoniae* depicted in Chapter 5. Hence, as stated before, there is a lack of reproducibility in the current state-of-the-art. In light of this gap in the literature, we propose performing automatic ribotyping of *C. diff* using the technical advancements presented in this dissertation, such as KSSHIBA [122] (Chapter 3) and FA-VAE [123] (Chapter 4).

6.2 Materials and Methods

6.2.1 Bacterial isolates

Two data sets are included in this pilot study, comprising of an initial dataset of 275 *C. diff* isolates and a collention of a real clinical scenario consisted of 3 *C. diff* isolates obtained. The isolates from the first data set were obtained during the first two quarters of 2022 whereas the outbreak isolates are from January 24th of 2023.

Before processing, the isolates were grown on Brucella blood agar (Beckton Dickinson®) and incubated under stringent anaerobic conditions at $37°C$. The incubation period was extended to 48 hours to provide sufficient growth time for *C. diff*. Prior to analysis, all isolates were reconfirmed using MALDI-TOF MS technology on an MBT Smart MALDI Biotyper (Bruker Daltonics, Bremen, Germany) with updated database containing 9957 mass spectra profiles (MSP), following the standard protocol of Formic Acid followed by HCCA (α-Cyano-4-hydroxycinnamic acid) matrix solution.

For the acquisition of spectra, a small amount of each *C. diff* colony was placed on the MALDI metal plate in duplicate, overlaid with $1\mu l$ of 100% formic acid, allowed to dry and spotted with 1μ l of organic HCCA matrix. Each spot on the MALDI plate was read twice. MALDI-TOF spectra were acquired after a 48h incubation period for each isolate, obtaining 4 spectra per strain. Spectra were acquired in positive mode in the region of $2K$ to $20K$ Da.

The study is carried out on microbiological samples, not human products, so informed consent from the patient is not required.

6.2.2 Isolates ribotyping

All *C. diff* strains were ribotyped at the HGUGM facility and Table 6.1 shows the relation of each RT of the database. For prediction purposes, the isolates were grouped in three categories: hyper-virulent RT027, 027-like RT181 and others, presenting the relation shown in Table 6.2.

Table 6.1: Relation of number of samples per ribotype and per dataset.

Ribotype (RT)	Training samples	Outbreak
RT001	30	
RT002	38	
RT014	29	
RT017	30	
RT023	19	
RT027	30	
RT078	30	
RT106	30	
RT181	30	
RT207	30	
RT651		3
Total	275	3

Table 6.2: Manual labelled samples for ML purposes in three classes: RT027, RT181 and Others.

Ribotype (RT)	Training samples	Outbreak
RT027	30	
RT181	30	
Others	216	3
Total	275	3

6.2.3 MALDI-TOF MS preprocessing and methodology

As shown in the paper by Guerrero et al. [45] and in Chapter 5 of this thesis, it has been demonstrated that no preprocessing is required for MALDI-TOF MS. As a result, the standard processing, consisting of the four steps (i) variance stabilising, (ii) smoothing, (iii) baseline removal, and (iv) intensity calibration, are avoided in this study.

Regarding the methodology, a preliminary study is conducted in which the training dataset is partitioned using a 60-40 train-test split. To address the issue of label imbalance, random oversampling of RT027 and RT181 is carried out resulting in stratified train partitions. The performance of each model is evaluated using balanced accuracy, taking into account the label imbalance in the test set, and in macro-average AUC with respect to ribotyping classification. Secondly, the best models are selected by performance, namely FA-VAE MLP and KSSHIBA Linear,

and they are retrained with full training dataset, i.e., the 275 samples. Finally, these retrained models are tested on the real outbreak scenario where they are employed in conjunction with traditional PCR techniques to ribotype.

As means of comparison, KSHHIBA is tested with three different kernels: RBF, Linear and PIKE, where kernel parameters (σ in RBF and t smooth in PIKE) are cross-validated. Then, in FA-VAE two different encoder-decoder architectures are tested: 1 layer MLP and 3 layer 1D-CNN.

6.2.4 KSSHIBA adapted to *C. diff*

In this context, we adapt the KSSHIBA architecture to tackle MALDI-TOF MS data and efficiently ribotype *C. diff* strains. In particualr, a two-view KSSHIBA is proposed, as shown in Fig. 6.1. The first view deals with the kernelised MALDI-TOF MS data, with different kernels such as Linear, RBF or PIKE. That means that instead of working with $\mathbf{X} \in \mathbb{R}^{275\times 18000}$ we work with the kernelised version $\mathbf{K} \in \mathbb{R}^{275\times 275}$. The second view models the RT information containing a one-hot encoder version of the RT that identifies each isolate. Using KSSHIBA, we can efficiently deal, in terms of computational cost, with the high-dimensional MALDI-TOF data at the same time that we exploit non-linear relationships.

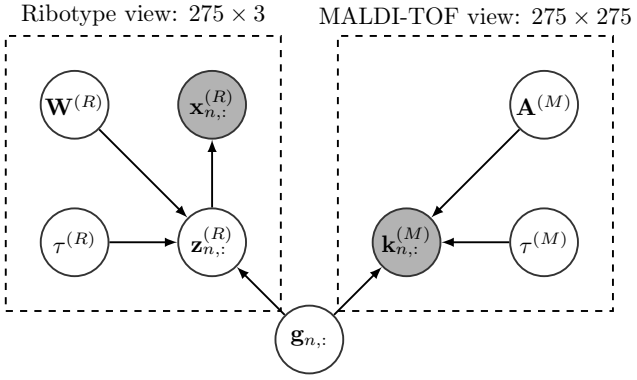


Figure 6.1: Probabilistic graphical model for the evaluated data set: view **M** corresponds to the VAE that handles MALDI-TOF MS, and view **R** corresponds to the RT (RT027, RT181, Others). The white circles represent random variables that the model learns, while the grey circles represent the observations.

6.2.5 FA-VAE adapted to *C. diff*

Furthermore, we suggest the FA-VAE model presented in Chapter 4 as a generative approach to process MALDI-TOF MS data for automatic ribotyping of *C. diff* isolates. To tailor the FA-VAE model to the specific characteristics of the data,

we propose the implementation of two distinct methods: a static approach and a mass-temporal approach.

For the static approach, a FA-VAE with two views is proposed. As first view, the MALDI-TOF MS data is processed using a VAE with a MLP encoder-decoder structure. This involves using an encoder-decoder architecture with linear layers to handle input data represented by $\mathbf{X}^{(1)} \in \mathbb{R}^{275\times 18000}$, and projecting it into a latent space represented by $\mathbf{Z}^{(1)} \in \mathbb{R}^{275\times H}$, where in the experimentation we tested $H = 20, 100$. The second view, the RT view, comprises a one-hot encoder version of the RT that identifies each isolate as proposed in KSSHIBA.

In the context of our study, we would like to emphasise that the m/z axis of the MALDI-TOF MS instrument captures the masses arriving at different time steps, as described in Chapter 1. Due to the fact that lighter masses reach the detector earlier, they appear at lower m/z values, which leads to the interpretation of the m/z axis as a temporal axis. To take advantage of this temporal dependence, we propose using a one-dimensional convolutional neural network (1D-CNN) as the encoder-decoder structure for the first view of our multi-view model. This involves the exploitation of 1D-CNNs to handle the input data represented by $\mathbf{X}^{(1)} \in \mathbb{R}^{275\times 18000}$, and project it into a latent space represented by $\mathbf{Z}^{(1)} \in \mathbb{R}^{275\times K}$, where in the experimentation we tested $H = 20, 100$, taking into account the temporal relationships between the peaks. This temporal concept is motivated by the behaviour of PIKE, as described by Weis et al. (2020) [11], where peaks that are closer in m/z coordinates are more correlated as they may be peptides of the same protein. The employment of a 1D-CNN is henceforth justified on the grounds of its potential to exploit the correlation between adjacent peaks while exhibiting relatively weaker association with distal peaks along the m/z axis.

Both approaches share the same graphical model shown in Fig. 6.2. The key difference between both approaches is which architecture deals with the encoding-decoding phase of MALDI-TOF MS data.

6.3 Results

In this section, the results obtained using the proposed models are presented. Two different scenarios are discussed: a preliminar study and an actual outbreak scenario.

In the first case, we investigate the potential benefits of applying ML techniques to automate ribotyping of *C. diff*. We evaluate the performance of ML models based on various metrics, including dimensionality reduction, computational efficiency, balanced accuracy, and AUC. Furthermore, we analyse the latent spaces generated by the KSSHIBA and FA-VAE methods.

Ultimately, to analyse its implementation in laboratories, a real outbreak is

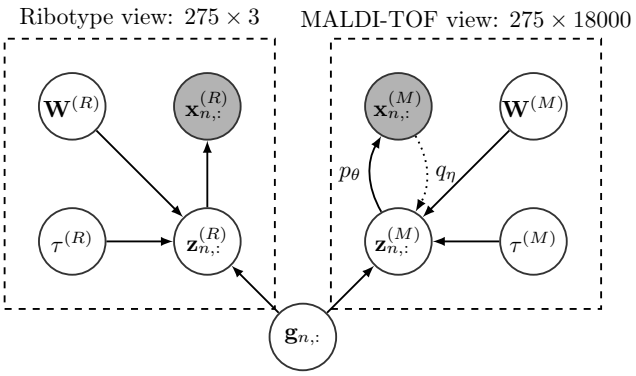


Figure 6.2: Probabilistic graphical model for the evaluated data set: view **M** corresponds to the VAE that handles MALDI-TOF MS, and view **R** corresponds to the RT (RT027, RT181, Others). The white circles represent random variables that the model learns, while the grey circles represent the observations.

studied. Best models selected in the first study are used in real time along traditional PCR techniques to check the feasibility of using ML techniques in laboratories.

6.3.1 Preliminar study

As motivated in Section 6.1, a gap in literature exists with regard to the utilisation of ML techniques for the automated ribotyping of *C. diff.* In light of this, an exploratory study is undertaken to assess the feasibility of such an approach.

Table 6.3: Results of KSSHIBA and FA-VAE approaches in AUC and Balanced Accuracy for a fixed 40% random test samples. First reduction represents the dimension of the first technique used, in case of FA-VAE the H output dimension of the encoder and in case of KSSHIBA the N number of samples used for the kernelisation. The best results for each case is shown in bold.

Model	First reduction (H or N)	K	Iterations	Time	Balanced Accuracy	AUC
FA-VAE MLP	20	7	**544**	**42**	0.97	0.98
FA-VAE MLP	100	9	622	57	0.94	0.96
FA-VAE 1D-CNN	20	**5**	721	207	0.84	0.88
FA-VAE 1D-CNN	100	8	787	234	0.83	0.97
KSSHIBA RBF	388	87	10000	290	0.95	0.98
KSSHIBA LINEAR	388	100	10000	332	**0.99**	**0.99**
KSSHIBA PIKE	388	100	10000	334	**0.99**	**0.99**

The results of testing the different approaches under study on 40% of the data are presented in Table 6.3. All models employed techniques for reducing the dimensionality of the data and we analyse their performance for different latent

dimensions. For instance, when the FA-VAE approach is utilised, different dimension of the VAE latent space, H, are analysed and when the KSSHIBA approach is used, the number of relevant vectors, N, used in the kernel for projecting the data is indicated together to the dimension of the FA latent space, K, which is automatically optimised. Additionally, the time cost required for the convergence of the model is measured by iterations and time.

The evaluation results, measured by the balanced accuracy and AUC scores, indicate that all models achieve high accurate results, with none of the approaches having a score lower than 0.83 in balanced accuracy or 0.88 in AUC. The highest scores have been obtained by the KSSHIBA models in both the PIKE and LINEAR versions, with values of 0.99 in both metrics, suggesting near-perfect results. The stationary version of FA-VAE, known as FA-VAE MLP, closely follows the performance of KSSHIBA, with scores of 0.97 and 0.98 in balanced accuracy and AUC, respectively. In terms of time cost, it is noted that the KSSHIBA implementations require 10000 iterations to converge, while the FA-VAE MLP achieve comparable results in significantly less time, requiring 18 times fewer iterations and 8 times less time. Moreover, the interpretability of the models is also taken into account, providing FA-VAE a more compact latent representation, that is, lower K value. However, the results obtained from the 1D-CNN approaches, which employed a temporal approach, were found to be comparatively lower and not as suitable for the given task.

In conclusion, the results of this preliminary study suggest that KSSHIBA (LINEAR or PIKE) achieves superior performance in terms of accuracy, while FA-VAE MLP performs comparably with the added benefit of a more compact latent space and lower time cost. The findings of this study indicates the viability of using MALDI-TOF data for automatic ribotyping of *C. diff* strains.

6.3.2 Real outbreak on January 24th

On January 24th, a real outbreak of *C. diff* was reported in the HGUGM, which involved three distinct strains and MALDI-TOF spectra were obtained. In order to evaluate the feasibility of using FA-VAE and KSSHIBA models in a laboratory context, both models were applied to the collected spectra. Since it was a real scenario, no label was available for the affected strains, so two control samples (both RT027) were tested as a reference. Table 6.4 shows the prediction made by 5 different approaches and the true label obtained by traditional rybotyping.

With regards to performance, the FA-VAE MLP 20 model was superior to all other methods, as it achieved a complete classification of the three outbreak samples and both control samples. The KSSHIBA RBF model also showed a satisfactory performance, correctly identifying the outbreak and one control sample. However, both the FA-VAE 1D-CNN and KSSHIBA PIKE models were able to identify

Table 6.4: Results of KSSHIBA and FA-VAE approaches in terms of Accuracy for the outbreak samples and two control samples. The control samples are denoted as c_n whereas outbreak samples are anotated as o_n. The best model is shown in bold.

Model	c_1	c_2	o_1	o_2	o_3	Accuracy	Time
PCR ribotyping	RT027	RT027	Others	Others	Others	-	7 days
FA-VAE MLP 20	RT027	RT027	Others	Others	Others	**1.00**	∅
KSSHIBA RBF	Others	RT027	Others	Others	Others	0.80	∅
FA-VAE 1D-CNN 20	RT181	RT181	Others	Others	Others	0.60	∅
KSSHIBA PIKE	Others	Others	Others	Others	Others	0.60	∅
KSSHIBA Linear	RT181	RT027	Others	RT181	RT181	0.40	∅

the outbreak but failed to correctly classify the control samples. Conversely, the KSSHIBA LINEAR model demonstrated poor performance, misclassifying almost all samples.

The results obtained in this real scenario suggest that both KSSHIBA PIKE and LINEAR may have suffered from over-fitting in the preliminary study, while models with lower accuracy such as KSSHIBA RBF and FA-VAE MLP 20 seem to have demonstrated better generalisation capabilities.

6.3.3 Latent space analysis

To gain a deeper understanding of the results presented in Table 6.4 and the possible over-fitting presented in Table 6.3, we perform a latent space analysis. As Table 6.3 shows, each model uses a different size of latent space, K. To make these projections comparable, we use t-SNE to reduce all models into a 2D space. Fig. 6.3 show the 2D t-SNE projections for the previous models. In these projections, predictions of the test samples are depicted by a circle and the true value is indicated by a cross. If the circle and cross have the same colour, the model made a correct prediction, and if the colours differ, it is a missclassification.

A closer examination of the performance of the KSSHIBA models is conducted. As we saw in Table 6.4, KSSHIBA models lacked generalisation power, hence we analysed their latent space projections, represented in Fig 6.3. If we focus on KSSHIBA models (first row of Fig 6.3) theirs projections are not well-clustered, with no clear separation between the different data points, which might indicate a potential over-fitting issue.

However, the FA-VAE approaches (second row of Fig 6.3) show a more clustered latent projection compared to the KSSHIBA models. It is worth noting that both FA-VAE models have a lower dimensional latent space, making the projection to 2D easier to perform. In the case of FA-VAE MLP, which performed perfectly in the real outbreak scenario, three clear clusters can be observed, each containing

one of the categories RT027, RT181 and Others. However, there is also a fourth cluster that groups some RT027 and RT181 indicating that the problem is not straight-forward. Moreover, RT027 and RT181 are the ones that the PCR technique cannot differ due to their similarities.

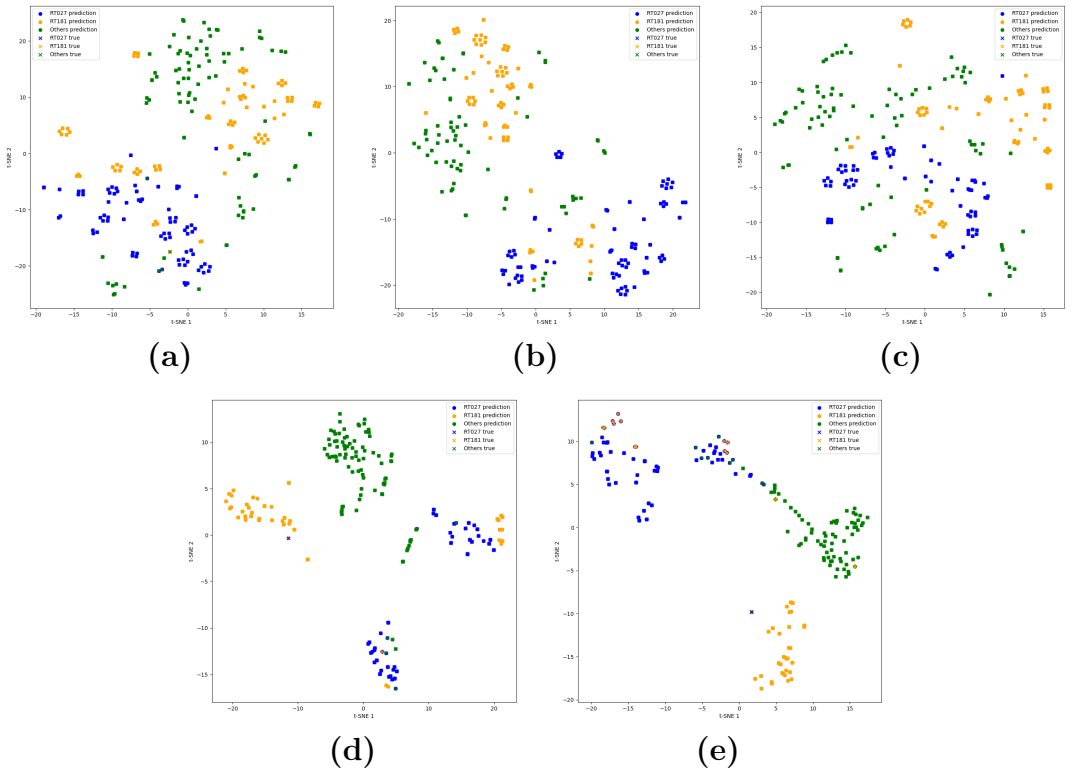


Figure 6.3: **(a)** KSSHIBA RBF **(b)** KSSHIBA PIKE **(c)** KSSHIBA LINEAR **(d)** FA-VAE MLP **(e)** FA-VAE 1D-CNN

6.4 Conclusions

In this preliminary study, we present a new approach for the automatic ribotyping of *C. diff* by harnessing the potential of probabilistic deep learning techniques using MALDI-TOF data. We investigate the practical viability of the proposed Bayesian FA models, namely KSSHIBA and FA-VAE, as a solution to this issue. To the best of our knowledge, this is the first demonstration of the feasibility of utilising probabilistic models to perform ribotyping of *C. diff*. To assess the viability of our approach, we conducted experiments on 275 samples from the HGUGM and achieved accuracy rates above 80% for all models, where particular configurations of KSSHIBA even reached perfect accuracy. Additionally, we tested KSSHIBA and FA-VAE in a real-life outbreak scenario in the HGUGM where FA-VAE performed a successfully classification. Our results not only exhibit high accuracy in predicting the ribotype of each strain, but also reveal an interpretable latent space, which represents a crucial advancement in the field. Additionally, the traditional ribotyping methods typically took 7 days to provide results, while our proposed methods were able to produce results on the same day, offering a significant reduction in the time required to take action.

It is important to note the limitations of this study, as they present opportunities for future research. For example, it would be beneficial to analyse strains with geographical differences in order to determine the generalisability of the findings. Moreover, a wider testing over time should be performed, along with PCR techniques, to keep evaluating the system. Despite these limitations, this preliminary study successfully demonstrated the potential of using MALDI-TOF-based probabilistic deep learning for automating bacterial ribotyping. The promising results obtained in a real outbreak provide a solid foundation for further advancements in this field.

The ultimate goal of this study would be to establish the viability of probabilistic models based on MALDI-TOF for clinical use and to demonstrate its superiority over traditional methods by reducing time cost. The next step is to conduct a rigorous longitudinal study, which will use the models, FA-VAE and KSSHIBA, in real-world laboratory procedures and compare them to traditional PCR techniques.

Moreover, the scope of future studies should be expanded to include a diverse range of sample origins, in order to fully understand the impact of epidemiological characteristics on bacterial ribotyping. This will not only deepen our understanding of bacterial ribotyping but also inform more effective public health measures.

In conclusion, this study is a crucial stepping stone towards realising the full potential of MALDI-TOF for bacterial ribotyping and advancing our ability to tackle bacterial outbreaks.

CHAPTER 7

CONCLUSIONS

In this dissertation, we have presented two technical innovations that address the challenge of incorporating diverse and multiple data sources. Specifically, we have tailored these solutions to handle microbiological data, working in collaboration with the Instituto de Investigación Sanitaria Gregorio Marañón (IISGM), and implemented them in real-world microbiology laboratories. We have developed a novel approach that merges advanced Factor Analysis (FA) techniques with kernel-based methods (as outlined in Chapter 3) and powerful generative models such as Variational AutoEncoder (VAE)s (as presented in Chapter 4). The result has been a set of robust, modular, and easily interpretable models that have been applied to two important microbiological scenarios, including the prediction of antibiotic resistance in Chapter 5 and the automation of the ribotyping procedure in Chapter 6. These contributions represent an advancement in the field and pave the way for future breakthroughs in microbiological research.

7.1 Summary of Methods and Contributions

The first theoretical model presented in this dissertation, Kernelised Semi-Supervised Heterogeneous Inter-Battery Analysis (KSSHIBA), has addressed semi-supervised, multi-view, high-dimensional data challenges by integrating kernel-based strategies and diverse, semi-supervised data perspectives. The model has generated condensed representations through the automatic selection of RVs and feature selection by executing ARD over the kernel. KSSHIBA has offered a Bayesian interpretation of FA that can seamlessly incorporate high-dimensional data while exploiting non-linearities by using kernel methods, along with categorical, binary, positive, and real data. The feasibility of KSSHIBA has been proven by outperforming state-of-the-art proposals in multidimensional regression tasks, high-dimensional image databases, and interpretability in high-dimensional classification tasks.

The second model presented in this dissertation, Factor Analysis-Variational AutoEncoder (FA-VAE), combines a powerful generative model, VAE, with the FA framework established in KSSHIBA. This model is the first deep hierarchical VAE for mixed and diverse data using an interpretable FA latent space. Specifically, FA-VAE has been able to adapt multiple VAEs, creating a modular model that can operate with a wider range of data domains. Exploiting the VAE architectural

choice, the model can handle categorical, binary, positive, real, image, or even temporal data. Each VAE has been conditioned by a deep hierarchical structure using FA, learning a disentangled and explicable latent space. The effectiveness of FA-VAE has been demonstrated by its superior performance compared to existing state-of-the-art approaches in (i) conditioning generative models, (ii) performing domain adaptation between datasets, and (iii) performing transfer learning between generative models.

As part of the partnership with the IISGM, the first model, KSSHIBA, has been customised and applied to detect antibiotic resistance in *K. pneumoniae* in the IISGM and in the IRyCIS. Specifically, we have adapted KSSHIBA to work with Linear, RBF and PIKE kernels, being the last one particularly suited for handling MALDI-TOF MS data. Additionally, we have made use of the multi-view capability of KSSHIBA to integrate epidemiological data into the model. We have also taken advantage of the semi-supervised approach to incorporate all relevant data, including MALDI-TOF MS and epidemiological information, into the prediction model. This led to KSSHIBA performing better than state-of-the-art models, achieving 0.88 and 0.77 AUC scores in ESBL+CP and WT antibiotic resistance prediction. We have deduced that, based on the results obtained, heterogeneous models are superior to traditional methods by incorporating epidemiological information about the bacteria. Its implementation in clinical laboratories could accelerate the rapid identification of multidrug-resistant isolates, thereby reducing the time for a therapeutic decision from 96 hours to 24 hours and thus having a significant impact on global public health improving patient outcomes.

The second model, FA-VAE, has been also tailored and applied in the microbiological laboratory of IISGM. FA-VAE has been presented as a novel method for automatic ribotyping of *C. diff* using MALDI-TOF data and probabilistic DL techniques is introduced. The feasibility of both, KSSHIBA and FA-VAE, has been evaluated as a potential solution for this problem. To the best of our knowledge, this is the first demonstration of using probabilistic DL for *C. diff* ribotyping. To determine the viability of the proposed approach, experiments have been conducted on 275 samples from the HGUGM. The results showed high accuracy rates, with KSSHIBA even achieving perfect accuracy. The models have been also tested in a real-life outbreak scenario in the HGUGM, where FA-VAE proved successful classification. The results not only demonstrated high accuracy in predicting each strain's RT but also revealed an interpretable latent space. Additionally, the traditional ribotyping methods typically took 7 days, while the proposed methods produced results on the same day, offering a significant reduction in response time.

In conclusion, this dissertation has made a significant contribution to the field of Bayesian FA by addressing its limitations in processing diverse data types through the development of the novel models KSSHIBA and FA-VAE. Furthermore, a comprehensive examination of the limitations in automating laboratory procedures

within the field of microbiology has been conducted, and the efficacy of the newly developed models has been demonstrated through their successful application to important problems, such as antibiotic resistance prediction and automatic ribotyping. Therefore, KSSHIBA and FA-VAE, both in terms of their technical and applied contributions, represent major advancements in the field of Bayesian statistics and pave the way for future breakthroughs in automating microbiological laboratories.

7.2 Proposals for future research

In light of the results presented in this dissertation there is immense potential for the utilisation of probabilistic models to revolutionise the field of microbiology and fully automate laboratory procedures. Consequently, the established lines of research developed in this thesis can be expanded to encompass more comprehensive and far-reaching objectives.

7.2.1 Enhance predictivity of FA-VAE in unbalanced multi-view problems

An exciting direction for extending the FA-VAE model would be to improve predictivity in the embedded latent space generated by the VAE in unbalanced multi-view settings. Our proposal currently faces a limitation whereby certain non-VAE views may be inadequately represented in the global $\mathbf{g}_{\mathrm{n},:}$ latent variables in some multi-view scenarios. This is due to the fact that these variables are constructed by combining all embedded latent variables $\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})}$ and observed variables $\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})}$. As a result, a $\mathbf{z}_{\mathrm{n},:}^{(\mathrm{m})}$ that represents a MALDI-TOF may be over-represented against a multi-label $\mathbf{x}_{\mathrm{n},:}^{(\mathrm{m})}$ that determines epidemiological information. Thus, a promising avenue for future work is to modify the regularisation term in each v-VAE by incorporating the posterior predictive distribution of $\mathbf{g}_{\mathrm{n},:}$, denoted as $\mathbf{g}_{n,:}^{*}$. This predictive version would be constructed using all m-views except for the one handled by the v-VAE. By implementing this approach, the regularisation term could induce the $\mathbf{z}_{n,:}^{(v)}$ embedded latent space to contain predictive information about the other views, thereby mitigating the bias stemming from an over-reliance on the specific view itself.

7.2.2 Widely epidemiological study over *K. pneumoniae*

An additional application of KSSHIBA and FA-VAE would be to broaden the experiment conducted in Chapter 5 to include samples from additional locations. Future research could take the method outlined by Weis et al. [97] to the next

level by integrating their extensive dataset of 5000 *K. pneumoniae* samples from Switzerland with our own study presented in [45] to thoroughly investigate variations in bacterial epidemiology on a European scale. We aim to push the boundaries of both articles by investigating if it is possible to predict antibiotic resistance of Iberian bacteria using a model trained only with Swiss bacteria, and then further enhance the model by fine-tuning it with a small number of Iberian bacteria. This research could lead to two potential hypotheses: (i) it is possible to create a highly accurate and versatile model that automatically distinguishes antibiotic resistance regardless of the country, or (ii) transfer learning is necessary to achieve optimal results when transitioning between countries. Both cases, suggest the creation of a comprehensive European database of *K. pneumoniae* to advance the field of medical microbiology and improve patient outcomes.

7.2.3 Multi-view unsupervised *E. coli* spread analysis

An interesting application of our FA-VAE model would be to conduct a comprehensive, multi-view unsupervised analysis of *Escherichia coli* at the Hospital General Universitario Gregorio Marañón (HGUGM). *E. coli* is a prevalent bacteria within the *Enterobacter* family that causes diarrhoea, stomach cramps, and fever, commonly found in hospitals. In 2022, more than 8000 MALDI-TOF MS samples have been analysed in HGUGM regarding *E. coli*. Medical doctors hypothesise that *E. coli* spreads in the hospital, causing the infection of other patients, even healthcare professionals, resulting in nosocomial outbreaks in various parts of the hospital. Therefore, a future study could use the proposed FA-VAE model to conduct a thorough, unsupervised multi-view analysis of the spread of *E. coli* at the HGUGM. To achieve this, we propose using FA-VAE in the following way: (i) the first view will handle MALDI-TOF MS data with a VAE; (ii) the second view will deal with categorical information on which hospital service the *E. coli* has been found; and (iii) the third view will manage demographic data about patients. This approach would enable us to identify key patterns and trends in the spread of *E. coli* in the HGUGM, potentially helping to prevent future nosocomial outbreaks.

The application of time-based spectral clustering to track the spread of hospital infections would be a valuable tool for health specialists. By implementing a real-time clustering system that updates daily or weekly with the current outbreaks at the hospital, specialists can gain insights into the evolution of the spread of infections over time. An interesting technical contribution would be to apply spectral clustering using the PIKE kernel, which is specifically designed for analysing MALDIs, to conduct daily or weekly clustering and then incorporate previous time steps, such as the preceding week, to create a more comprehensive view of the infection's progression.

7.2.4 Longitudinal and international study of *C. diff*

The current method of bacterial ribotyping, which relies on traditional PCR techniques, has proven to be effective in many ways. However, the limitations of this method have become increasingly evident, particularly in terms of accuracy, as discussed in Chapter 6. In light of these limitations, the exploration of new and innovative methods of bacterial ribotyping is of utmost importance. The preliminary study described in this thesis has provided compelling evidence of the potential of probabilistic DL and MALDI-TOF for bacterial ribotyping. The results of this study have indicated that MALDI-TOF-based probabilistic models have the potential to outperform traditional PCR techniques in terms of time cost and accuracy.

However, an international longitudinal study is necessary to fully validate these findings and determine the viability of probabilistic DL for ribotyping in real-world laboratory procedures. By considering a broader range of sample origins, the study would provide valuable insights into the impact of epidemiological factors on bacterial ribotyping. Moreover, the validation of the models during time would result into a confidence measure about how accurate probabilistic DL models can be performing this job. This will not only deepen our understanding of bacterial ribotyping but also inform more effective public health measures.

BIBLIOGRAPHY

[1] M. Pavlovic, I. Huber, R. Konrad, and U. Busch, "Application of maldi-tof ms for the identification of food borne bacteria," *The open microbiology journal*, vol. 7, p. 135, 2013.

[2] C. Sevilla-Salcedo, V. Gómez-Verdejo, and P. M. Olmos, "Sparse semi-supervised heterogeneous interbattery bayesian analysis," *Pattern Recognition*, vol. 120, p. 108141, 2021.

[3] A. Nazabal, P. M. Olmos, Z. Ghahramani, and I. Valera, "Handling incomplete heterogeneous data using vaes," *Pattern Recognition*, vol. 107, p. 107501, 2020.

[4] Y. Gong, H. Hajimirsadeghi, J. He, T. Durand, and G. Mori, "Variational selective autoencoder: Learning from partially-observed heterogeneous data," in *International Conference on Artificial Intelligence and Statistics*, pp. 2377–2385, PMLR, 2021.

[5] M. Wu and N. Goodman, "Multimodal generative models for scalable weakly-supervised learning," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[6] M. Lee and V. Pavlovic, "Private-shared disentangled multimodal vae for learning of latent representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1692–1700, 2021.

[7] J. Xu, Y. Ren, H. Tang, X. Pu, X. Zhu, M. Zeng, and L. He, "Multi-vae: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9234–9243, October 2021.

[8] W. Youpeng, L. Hongxiang, G. Yiju, and Z. Liang, "Amvae: Asymmetric multimodal variational autoencoder for multi-view representation," in *International Conference on Artificial Neural Networks*, pp. 391–402, Springer, 2021.

[9] C. Ma, S. Tschiatschek, R. Turner, J. M. Hernández-Lobato, and C. Zhang, "Vaem: a deep generative model for heterogeneous mixed type data," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11237–11247, 2020.

[10] I. Peis, C. Ma, and J. M. Hernández-Lobato, "Missing data imputation and acquisition with deep hierarchical models and hamiltonian monte carlo," in *Advances in Neural Information Processing Systems 35*, 2022.

[11] 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.

[12] P. Hewage, M. Trovati, E. Pereira, and A. Behera, "Deep learning-based effective fine-grained weather forecasting model," *Pattern Analysis and Applications*, vol. 24, no. 1, pp. 343–366, 2021.

[13] Y. Duan, J. W. Goodell, H. Li, and X. Li, "Assessing machine learning for forecasting economic risk: Evidence from an expanded chinese financial information set," *Finance Research Letters*, vol. 46, p. 102273, 2022.

[14] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[15] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, *et al.*, "Evaluating large language models trained on code," *arXiv preprint arXiv:2107.03374*, 2021.

[16] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[17] Ö. Aydın and E. Karaarslan, "Openai chatgpt generated literature review: Digital twin in healthcare," *Available at SSRN 4308687*, 2022.

[18] T. H. Kung, M. Cheatham, A. Medinilla, C. Sillos, L. De Leon, C. Elepano, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, *et al.*, "Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models," *medRxiv*, 2022.

[19] J. Kim, "Search for medical information and treatment options for musculoskeletal disorders through an artificial intelligence chatbot: Focusing on shoulder impingement syndrome," *medRxiv*, 2022.

[20] W. Zhang, S. Yan, J. Li, X. Tian, and T. Yoshida, "Credit risk prediction of smes in supply chain finance by fusing demographic and behavioral data," *Transportation Research Part E: Logistics and Transportation Review*, vol. 158, p. 102611, 2022.

[21] J. Wu, F. Orlandi, M. S. Pathan, D. O'Sullivan, and S. Dev, "Augmenting weather sensor data with remote knowledge graphs," in *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1264–1267, IEEE, 2022.

[22] L. A. Castro, N. Generous, W. Luo, A. Pastore y Piontti, K. Martinez, M. F. Gomes, D. Osthus, G. Fairchild, A. Ziemann, A. Vespignani, *et al.*, "Using heterogeneous data to identify signatures of dengue outbreaks at fine spatio-temporal scales across brazil," *PLoS neglected tropical diseases*, vol. 15, no. 5, p. e0009392, 2021.

[23] B. E. Stein and M. A. Meredith, *The merging of the senses.* The MIT press, 1993.

[24] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[25] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE signal processing magazine*, vol. 34, no. 6, pp. 96–108, 2017.

[26] D. Carneiro, J. C. Castillo, P. Novais, A. Fernández-Caballero, and J. Neves, "Multimodal behavioral analysis for non-invasive stress detection," *Expert Systems with Applications*, vol. 39, no. 18, pp. 13376–13389, 2012.

[27] O. M. Mozos, V. Sandulescu, S. Andrews, D. Ellis, N. Bellotto, R. Dobrescu, and J. M. Ferrandez, "Stress detection using wearable physiological and sociometric sensors," *International journal of neural systems*, vol. 27, no. 02, p. 1650041, 2017.

[28] I. Peis, P. M. Olmos, C. Vera-Varela, M. L. Barrigón, P. Courtet, E. Baca-Garcia, and A. Artes-Rodriguez, "Deep sequential models for suicidal ideation from multiple source data," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 6, pp. 2286–2293, 2019.

[29] L. Sun, S. Zhang, H. Chen, and L. Luo, "Brain tumor segmentation and survival prediction using multimodal mri scans with deep learning," *Frontiers in neuroscience*, vol. 13, p. 810, 2019.

[30] Y. Wang, G. Ma, L. An, F. Shi, P. Zhang, D. S. Lalush, X. Wu, Y. Pu, J. Zhou, and D. Shen, "Semisupervised tripled dictionary learning for standard-dose pet image prediction using low-dose pet and multimodal mri," *IEEE transactions on biomedical engineering*, vol. 64, no. 3, pp. 569–579, 2016.

[31] I. Shomorony, E. T. Cirulli, L. Huang, L. A. Napier, R. R. Heister, M. Hicks, I. V. Cohen, H.-C. Yu, C. L. Swisher, N. M. Schenker-Ahmed, *et al.*, "An unsupervised learning approach to identify novel signatures of health and disease from multimodal data," *Genome medicine*, vol. 12, no. 1, pp. 1–14, 2020.

[32] P. A. Bradford, "Extended-spectrum β-lactamases in the 21st century: characterization, epidemiology, and detection of this important resistance threat," *Clinical microbiology reviews*, vol. 14, no. 4, pp. 933–951, 2001.

[33] D. M. Livermore and N. Woodford, "The β-lactamase threat in enterobacteriaceae, pseudomonas and acinetobacter," *Trends in microbiology*, vol. 14, no. 9, pp. 413–420, 2006.

[34] T. Czekaj, M. Ciszewski, and E. M. Szewczyk, "Staphylococcus haemolyticus–an emerging threat in the twilight of the antibiotics age," *Microbiology*, vol. 161, no. 11, pp. 2061–2068, 2015.

[35] C. Ayoub Moubareck and D. Hammoudi Halat, "Insights into acinetobacter baumannii: a review of microbiological, virulence, and resistance traits in a threatening nosocomial pathogen," *Antibiotics*, vol. 9, no. 3, p. 119, 2020.

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

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

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

[39] A. H. Arslan, F. U. Ciloglu, U. Yilmaz, E. Simsek, and O. Aydin, "Discrimination of waterborne pathogens, cryptosporidium parvum oocysts and bacteria using surface-enhanced raman spectroscopy coupled with principal component analysis and hierarchical clustering," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 267, p. 120475, 2022.

[40] D. Klein, R. Breuch, J. Reinmüller, C. Engelhard, and P. Kaul, "Rapid detection and discrimination of food-related bacteria using ir-microspectroscopy in combination with multivariate statistical analysis," *Talanta*, vol. 232, p. 122424, 2021.

[41] A. Y. Guh, Y. Mu, L. G. Winston, H. Johnston, D. Olson, M. M. Farley, L. E. Wilson, S. M. Holzbauer, E. C. Phipps, G. K. Dumyati, *et al.*, "Trends in us burden of clostridioides difficile infection and outcomes," *New England Journal of Medicine*, vol. 382, no. 14, pp. 1320–1330, 2020.

[42] H. S. Lee, K. Plechot, S. Gohil, and J. Le, "Clostridium difficile: Diagnosis and the consequence of over diagnosis," *Infectious diseases and therapy*, vol. 10, no. 2, pp. 687–697, 2021.

[43] 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.

[44] 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, pp. 318–327, Mar. 2018.

[45] 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, "Automatic antibiotic resistance prediction in klebsiella pneumoniae based on maldi-tof mass spectra," *Engineering Applications of Artificial Intelligence*, vol. 118, p. 105644, 2023.

[46] C. Weis, B. A. Rieck, S. Balzer, A. Cuenod, A. Egli, and K. Borgwardt, "Improved maldi-tof ms based antimicrobial resistance prediction through hierarchical stratification," *bioRxiv*, 2022.

[47] J. P. Anhalt and C. Fenselau, "Identification of bacteria using mass spectrometry," *Analytical chemistry*, vol. 47, no. 2, pp. 219–225, 1975.

[48] M. Karas and F. Hillenkamp, "Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons," *Analytical chemistry*, vol. 60, no. 20, pp. 2299–2301, 1988.

[49] S.-Y. Hsieh, C.-L. Tseng, Y.-S. Lee, A.-J. Kuo, C.-F. Sun, Y.-H. Lin, and J.-K. Chen, "Highly efficient classification and identification of human pathogenic bacteria by maldi-tof ms," *Molecular & cellular proteomics*, vol. 7, no. 2, pp. 448–456, 2008.

[50] R. Patel, "Maldi-tof ms for the diagnosis of infectious diseases," *Clinical chemistry*, vol. 61, no. 1, pp. 100–111, 2015.

[51] M. Y. Ashfaq, D. A. Da'na, and M. A. Al-Ghouti, "Application of maldi-tof ms for identification of environmental bacteria: A review," *Journal of Environmental Management*, vol. 305, p. 114359, 2022.

[52] E. Torres-Sangiao, C. Leal Rodriguez, and C. García-Riestra, "Application and perspectives of maldi–tof mass spectrometry in clinical microbiology laboratories," *Microorganisms*, vol. 9, no. 7, p. 1539, 2021.

[53] M. Martín Gómez and M. Ballesteros González, "Espectrometría de masas y análisis de biomarcadores," *Monografías de la Real Academia Nacional de Farmacia*, 2010.

[54] J. O. Lay and R. D. Holland, "Rapid identification of bacteria based on spectral patterns using maldi-tofms," *Mass Spectrometry of Proteins and Peptides*, pp. 461–487, 2000.

[55] C. G. Clark, P. Kruczkiewicz, C. Guan, S. J. McCorrister, P. Chong, J. Wylie, P. van Caeseele, H. A. Tabor, P. Snarr, M. W. Gilmour, *et al.*, "Evaluation of maldi-tof mass spectroscopy methods for determination of escherichia coli pathotypes," *Journal of microbiological methods*, vol. 94, no. 3, pp. 180–191, 2013.

[56] R. Giebel, C. Worden, S. Rust, G. Kleinheinz, M. Robbins, and T. Sandrin, "Microbial fingerprinting using matrix-assisted laser desorption ionization time-of-flight mass spectrometry (maldi-tof ms): applications and challenges," *Advances in applied microbiology*, vol. 71, pp. 149–184, 2010.

[57] C. Fenselau and P. A. Demirev, "Characterization of intact microorganisms by maldi mass spectrometry," *Mass spectrometry reviews*, vol. 20, no. 4, pp. 157–171, 2001.

[58] R. Holland, J. Wilkes, F. Rafii, J. Sutherland, C. Persons, K. Voorhees, and J. Lay Jr, "Rapid identification of intact whole bacteria based on spectral patterns using matrix-assisted laser desorption/ionization with time-of-flight mass spectrometry," *Rapid Communications in Mass Spectrometry*, vol. 10, no. 10, pp. 1227–1232, 1996.

[59] J. Abián, M. Carrascal, and M. Gay, "Introducción a la espectrometría de masas para la caracterización de péptidos y proteínas en proteómica," *Proteómica*, 2008.

[60] E. Jordana-Lluch, E. M. Català, and V. A. Ruiz, "La espectrometría de masas en el laboratorio de microbiología clínica," *Enfermedades Infecciosas y Microbiología Clínica*, vol. 30, no. 10, pp. 635–644, 2012.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] Z. Liu, D. Elashoff, and S. Piantadosi, "Sparse support vector machines with l0 approximation for ultra-high dimensional omics data," *Artificial intelligence in medicine*, vol. 96, pp. 134–141, 2019.

[68] H.-Y. Wang, C.-H. Kuo, C.-R. Chung, T.-W. Lin, J.-R. Yu, J.-J. Lu, and T.-S. Wu, "Rapid and accurate discrimination of mycobacterium abscessus subspecies based on matrix-assisted laser desorption ionization-time of flight spectrum and machine learning," *bioRxiv*, 2022.

[69] B. Zhou, L. Sun, T. Fang, H. Li, R. Zhang, and A. Ye, "Rapid and accurate identification of pathogenic bacteria at the single-cell level using laser tweezers raman spectroscopy and deep learning," *Journal of Biophotonics*, p. e202100312, 2022.

[70] S. Yu, X. Li, W. Lu, H. Li, Y. V. Fu, and F. Liu, "Analysis of raman spectra by using deep learning methods in the identification of marine pathogens," *Analytical Chemistry*, vol. 93, no. 32, pp. 11089–11098, 2021.

[71] B. L. Thomsen, J. B. Christensen, O. Rodenko, I. Usenov, R. B. Grønnemose, T. E. Andersen, and M. Lassen, "Accurate and fast identification of minimally prepared bacteria phenotypes using raman spectroscopy assisted by machine learning," *Scientific reports*, vol. 12, no. 1, pp. 1–12, 2022.

[72] A. Candela, A. Guerrero-López, M. Mateos, A. Gómez-Asenjo, M. J. Arroyo, M. Hernandez-García, R. del Campo, E. Cercenado, A. Cuénod, G. Méndez, L. Mancera, J. de Dios Caballero, L. Martínez-García, D. Gijón, M. I. Morosini, P. Ruiz-Garbajosa, A. Egli, R. Cantón, P. Muñoz, D. Rodríguez-Temporal, and B. Rodríguez-Sánchez, "Automatic discrimination of species within the enterobacter cloacae complex using maldi-tof mass spectrometry and supervised algorithms," *bioRxiv*, 2022.

[73] M. L. Mezzatesta, F. Gona, and S. Stefani, "Enterobacter cloacae complex: clinical impact and emerging antibiotic resistance," *Future microbiology*, vol. 7, no. 7, pp. 887–902, 2012.

[74] A. Kremer and H. Hoffmann, "Prevalences of the enterobacter cloacae complex and its phylogenetic derivatives in the nosocomial environment," *European journal of clinical microbiology & infectious diseases*, vol. 31, no. 11, pp. 2951–2955, 2012.

[75] M. Akbari, B. Bakhshi, and S. N. Peerayeh, "Particular distribution of enterobacter cloacae strains isolated from urinary tract infection within clonal complexes," *Iranian biomedical journal*, vol. 20, no. 1, p. 49, 2016.

[76] A. Davin-Regli, J.-P. Lavigne, and J.-M. Pagès, "Enterobacter spp.: update on taxonomy, clinical aspects, and emerging antimicrobial resistance," *Clinical microbiology reviews*, vol. 32, no. 4, pp. e00002–19, 2019.

[77] H. Hoffmann and A. Roggenkamp, "Population genetics of the nomenspecies enterobacter cloacae," *Applied and environmental microbiology*, vol. 69, no. 9, pp. 5306–5318, 2003.

[78] N. K. Singh, D. Bezdan, A. Checinska Sielaff, K. Wheeler, C. E. Mason, and K. Venkateswaran, "Multi-drug resistant enterobacter bugandensis species isolated from the international space station and comparative genomic analyses with human pathogenic strains," *BMC microbiology*, vol. 18, no. 1, pp. 1–13, 2018.

[79] R. Podschun and U. Ullmann, "Klebsiella spp. as nosocomial pathogens: epidemiology, taxonomy, typing methods, and pathogenicity factors," *Clinical microbiology reviews*, vol. 11, no. 4, pp. 589–603, 1998.

[80] H. W. Boucher, G. H. Talbot, J. S. Bradley, J. E. Edwards, D. Gilbert, L. B. Rice, M. Scheld, B. Spellberg, and J. Bartlett, "Bad bugs, no drugs: no eskape! an update from the infectious diseases society of america," *Clinical infectious diseases*, vol. 48, no. 1, pp. 1–12, 2009.

[81] D. J. Doorduijn, S. H. Rooijakkers, W. van Schaik, and B. W. Bardoel, "Complement resistance mechanisms of klebsiella pneumoniae," *Immunobiology*, vol. 221, no. 10, pp. 1102–1109, 2016.

[82] M. K. Paczosa and J. Mecsas, "Klebsiella pneumoniae: going on the offense with a strong defense," *Microbiology and Molecular Biology Reviews*, vol. 80, no. 3, pp. 629–661, 2016.

[83] H. Yigit, A. M. Queenan, G. J. Anderson, A. Domenech-Sanchez, J. W. Biddle, C. D. Steward, S. Alberti, K. Bush, and F. C. Tenover, "Novel carbapenem-hydrolyzing β-lactamase, kpc-1, from a carbapenem-resistant strain of klebsiella pneumoniae," *Antimicrobial agents and chemotherapy*, vol. 45, no. 4, pp. 1151–1161, 2001.

[84] 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, pp. e00037–18, Nov. 2018.

[85] 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, pp. 4155–4162, Dec. 2014. Publisher: American Society for Microbiology.

[86] L. Wang, C. Han, W. Sui, M. Wang, and X. Lu, "Maldi-tof ms applied to indirect carbapenemase detection: a validated procedure to clearly distinguish between carbapenemase-positive and carbapenemase-negative bacterial strains," *Analytical and bioanalytical chemistry*, vol. 405, no. 15, pp. 5259–5266, 2013.

[87] B. Li, T. Guo, F. Qu, B. Li, H. Wang, Z. Sun, X. Li, Z. Gao, C. Bao, C. Zhang, *et al.*, "Matrix-assisted laser desorption ionization: time of flight mass spectrometry-identified models for detection of esbl-producing bacterial strains," *Medical Science Monitor Basic Research*, vol. 20, p. 176, 2014.

[88] S. Angeletti, G. Dicuonzo, A. Lo Presti, E. Cella, F. Crea, A. Avola, M. A. Vitali, M. Fagioni, and L. De Florio, "Maldi-tof mass spectrometry and blakpc gene phylogenetic analysis of an outbreak of carbapenem-resistant k. pneumoniae strains," *New Microbiol*, vol. 38, no. 4, pp. 541–550, 2015.

[89] Y. Huang, J. Li, D. Gu, Y. Fang, E. W. Chan, S. Chen, and R. Zhang, "Rapid detection of k1 hypervirulent klebsiella pneumoniae by maldi-tof ms," *Frontiers in microbiology*, vol. 6, p. 1435, 2015.

[90] K. Z. Vardakas, D. K. Matthaiou, M. E. Falagas, E. Antypa, A. Koteli, and E. Antoniadou, "Characteristics, risk factors and outcomes of carbapenem-resistant klebsiella pneumoniae infections in the intensive care unit," *Journal of Infection*, vol. 70, no. 6, pp. 592–599, 2015.

[91] C. Giordano and S. Barnini, "Rapid detection of colistin-resistant klebsiella pneumoniae using maldi-tof ms peak-based assay," *Journal of microbiological methods*, vol. 155, pp. 27–33, 2018.

[92] Y. Huang, J. Li, Q. Wang, K. Tang, and C. Li, "Rapid detection of kpc-producing klebsiella pneumoniae in china based on maldi-tof ms," *Journal of Microbiological Methods*, vol. 192, p. 106385, 2022.

[93] E. Gato, I. P. Constanso, A. Candela, F. Galán, B. K. Rodiño-Janeiro, M. J. Arroyo, G. Méndez, L. Mancera, T. Alioto, M. Gut, *et al.*, "An improved matrix-assisted laser desorption ionization–time of flight mass spectrometry data analysis pipeline for the identification of carbapenemase-producing klebsiella pneumoniae," *Journal of clinical microbiology*, vol. 59, no. 7, pp. e00800–21, 2021.

[94] 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, 2020.

[95] H. López-Fernández, H. M. Santos, J. L. Capelo, F. Fdez-Riverola, D. Glez-Peña, and M. Reboiro-Jato, "Mass-up: an all-in-one open software application for maldi-tof mass spectrometry knowledge discovery," *BMC bioinformatics*, vol. 16, no. 1, pp. 1–12, 2015.

[96] C. Wang, Z. Wang, H.-Y. Wang, C.-R. Chung, J.-T. Horng, J.-J. Lu, and T.-Y. Lee, "Large-scale samples based rapid detection of ciprofloxacin resistance in klebsiella pneumoniae using machine learning methods," *Frontiers in microbiology*, p. 186, 2022.

[97] C. Weis, A. Cuénod, B. Rieck, O. Dubuis, S. Graf, C. Lang, M. Oberle, M. Brackmann, K. K. Søgaard, M. Osthoff, *et al.*, "Direct antimicrobial resistance prediction from clinical maldi-tof mass spectra using machine learning," *Nature Medicine*, vol. 28, no. 1, pp. 164–174, 2022.

[98] S. Kalkan, "Multimodal analysis of south-eastern black sea sediment bacterial population diversity," *Marine Pollution Bulletin*, vol. 183, p. 114063, 2022.

[99] C. M. Tressler, V. Ayyappan, S. Nakuchima, E. Yang, K. Sonkar, Z. Tan, and K. Glunde, "A multimodal pipeline using nmr spectroscopy and maldi-tof mass spectrometry imaging from the same tissue sample," *NMR in Biomedicine*, 2022.

[100] E. R. Dubberke and M. A. Olsen, "Burden of clostridium difficile on the healthcare system," *Clinical infectious diseases*, vol. 55, no. suppl_2, pp. S88–S92, 2012.

[101] T. van der Kooi, A. Lepape, P. Astagneau, C. Suetens, M. A. Nicolaie, S. de Greeff, I. Lozoraitiene, J. Czepiel, M. Patyi, D. Plachouras, *et al.*, "Mortality review as a tool to assess the contribution of healthcare-associated infections to death: results of a multicentre validity and reproducibility study, 11 european union countries, 2017 to 2018," *Eurosurveillance*, vol. 26, no. 23, p. 2000052, 2021.

[102] S. B. Debast, M. P. Bauer, E. J. Kuijper, and Committee, "European society of clinical microbiology and infectious diseases: update of the treatment guidance document for clostridium difficile infection," *Clinical microbiology and infection*, vol. 20, pp. 1–26, 2014.

[103] S.-L. Amaya, E.-S. Rosa, G.-F. Sergio, A. Noelia, R.-R. Lourdes, G. Eduardo, R. G. Patricia, and C. Javier, "Extraintestinal clostridioides difficile infection: septic arthritis 12 months after colitis," *Anaerobe*, vol. 69, p. 102318, 2021.

[104] L. C. McDonald, D. N. Gerding, S. Johnson, J. S. Bakken, K. C. Carroll, S. E. Coffin, E. R. Dubberke, K. W. Garey, C. V. Gould, C. Kelly, *et al.*, "Clinical practice guidelines for clostridium difficile infection in adults and children: 2017 update by the infectious diseases society of america (idsa) and society for healthcare epidemiology of america (shea)," *Clinical infectious diseases*, vol. 66, no. 7, pp. e1–e48, 2018.

[105] I. C. Hall and E. O'toole, "Intestinal flora in new-born infants: with a description of a new pathogenic anaerobe, bacillus difficilis," *American journal of diseases of children*, vol. 49, no. 2, pp. 390–402, 1935.

[106] K. E. Dingle, D. Griffiths, X. Didelot, J. Evans, A. Vaughan, M. Kachrimanidou, N. Stoesser, K. A. Jolley, T. Golubchik, R. M. Harding, *et al.*, "Clinical clostridium

difficile: clonality and pathogenicity locus diversity," *PloS one*, vol. 6, no. 5, p. e19993, 2011.

[107] M. He, F. Miyajima, P. Roberts, L. Ellison, D. J. Pickard, M. J. Martin, T. R. Connor, S. R. Harris, D. Fairley, K. B. Bamford, *et al.*, "Emergence and global spread of epidemic healthcare-associated clostridium difficile," *Nature genetics*, vol. 45, no. 1, pp. 109–113, 2013.

[108] R. Fatima and M. Aziz, "The hypervirulent strain of clostridium difficile: Nap1/b1/027-a brief overview," *Cureus*, vol. 11, no. 1, 2019.

[109] V. G. Loo, L. Poirier, M. A. Miller, M. Oughton, M. D. Libman, S. Michaud, A.-M. Bourgault, T. Nguyen, C. Frenette, M. Kelly, *et al.*, "A predominantly clonal multi-institutional outbreak of clostridium difficile–associated diarrhea with high morbidity and mortality," *New England Journal of Medicine*, vol. 353, no. 23, pp. 2442–2449, 2005.

[110] P. Wiegand, D. Nathwani, M. Wilcox, J. Stephens, A. Shelbaya, and S. Haider, "Clinical and economic burden of clostridium difficile infection in europe: a systematic review of healthcare-facility-acquired infection," *Journal of Hospital Infection*, vol. 81, no. 1, pp. 1–14, 2012.

[111] E. Kuijper, F. Barbut, J. Brazier, N. Kleinkauf, T. Eckmanns, M. Lambert, D. Drudy, F. Fitzpatrick, C. Wiuff, D. Brown, *et al.*, "Update of clostridium difficile infection due to pcr ribotype 027 in europe, 2008," *Eurosurveillance*, vol. 13, no. 31, p. 18942, 2008.

[112] W. Jamal, E. M. Pauline, and V. O. Rotimi, "Comparative performance of the genexpert c. difficile pcr assay and c. diff quik chek complete kit assay for detection of clostridium difficile antigen and toxins in symptomatic community-onset infections," *International Journal of Infectious Diseases*, vol. 29, pp. 244–248, 2014.

[113] M. Kachrimanidou, A. Baktash, S. Metallidis, O. Tsachouridou, F. Netsika, D. Dimoglou, A. Kassomenaki, E. Mouza, M. Haritonidou, and E. Kuijper, "An outbreak of clostridioides difficile infections due to a 027-like pcr ribotype 181 in a rehabilitation centre: Epidemiological and microbiological characteristics," *Anaerobe*, vol. 65, p. 102252, 2020.

[114] M. Kachrimanidou, S. Metallidis, O. Tsachouridou, C. Harmanus, V. Lola, E. Protonotariou, L. Skoura, and E. Kuijper, "Predominance of clostridioides difficile pcr ribotype 181 in northern greece, 2016–2019," *Anaerobe*, p. 102601, 2022.

[115] A. Cuénod and A. Egli, "Advanced applications of maldi-tof ms–typing and beyond," in *Application and Integration of Omics-powered Diagnostics in Clinical and Public Health Microbiology*, pp. 153–173, Springer, 2021.

[116] M. Reil, M. Erhard, E. J. Kuijper, M. Kist, H. Zaiß, W. Witte, H. Gruber, and S. Borgmann, "Recognition of clostridium difficile pcr-ribotypes 001, 027 and 126/078 using an extended maldi-tof ms system," *European journal of clinical microbiology & infectious diseases*, vol. 30, no. 11, pp. 1431–1436, 2011.

[117] K. Rizzardi and T. Åkerlund, "High molecular weight typing with maldi-tof ms-a novel method for rapid typing of clostridium difficile," *Plos one*, vol. 10, no. 4, p. e0122457, 2015.

[118] J. Skraban, S. Dzeroski, B. Zenko, D. Mongus, S. Gangl, and M. Rupnik, "Gut microbiota patterns associated with colonization of different clostridium difficile ribotypes," *PloS one*, vol. 8, no. 2, p. e58005, 2013.

[119] B. Y. Li, J. Oh, V. B. Young, K. Rao, and J. Wiens, "Using machine learning and the electronic health record to predict complicated clostridium difficile infection," in *Open forum infectious diseases*, vol. 6, p. ofz186, Oxford University Press US, 2019.

[120] D. Ruzicka, T. Kondo, G. Fujimoto, A. P. Craig, S.-W. Kim, and H. Mikamo, "Development of a clinical prediction model for recurrence and mortality outcomes after clostridioides difficile infection using a machine learning approach," *Anaerobe*, vol. 77, p. 102628, 2022.

[121] A. Calderaro, M. Buttrini, B. Farina, S. Montecchini, M. Martinelli, M. C. Arcangeletti, C. Chezzi, and F. De Conto, "Characterization of clostridioides difficile strains from an outbreak using maldi-tof mass spectrometry," *Microorganisms*, vol. 10, no. 7, p. 1477, 2022.

[122] C. Sevilla-Salcedo, A. Guerrero-López, P. M. Olmos, and V. Gómez-Verdejo, "Bayesian sparse factor analysis with kernelized observations," *Neurocomputing*, vol. 490, pp. 66–78, 2022.

[123] A. Guerrero-López, C. Sevilla-Salcedo, V. Gómez-Verdejo, and P. M. Olmos, "Multi-view hierarchical variational autoencoders with factor analysis latent space," *arXiv preprint arXiv:2207.09185*, 2022.

[124] C. Bishop, "Bayesian pca," in *Advances in Neural Information Processing Systems* (M. Kearns, S. Solla, and D. Cohn, eds.), vol. 11, MIT Press, 1998.

[125] A. Klami, S. Virtanen, and S. Kaski, "Bayesian canonical correlation analysis," *Journal of Machine Learning Research*, vol. 14, no. Apr, pp. 965–1003, 2013.

[126] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[127] 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.

[128] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[129] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4. Springer, 2006.

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

[131] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*, vol. 2. MIT press Cambridge, MA, 2006.

[132] T. Bui, D. Hernández-Lobato, J. Hernandez-Lobato, Y. Li, and R. Turner, "Deep gaussian processes for regression using approximate expectation propagation," in *International conference on machine learning*, pp. 1472–1481, PMLR, 2016.

[133] J. Fei, J. Zhao, S. Sun, and Y. Liu, "Active learning methods with deep gaussian processes," in *International Conference on Neural Information Processing*, pp. 473–483, Springer, 2018.

[134] V. A. Vuyyuru, G. A. Rao, and Y. Murthy, "A novel weather prediction model using a hybrid mechanism based on mlp and vae with fire-fly optimization algorithm," *Evolutionary Intelligence*, vol. 14, no. 2, pp. 1173–1185, 2021.

[135] K. K. Santhosh, D. P. Dogra, P. P. Roy, and A. Mitra, "Vehicular trajectory classification and traffic anomaly detection in videos using a hybrid cnn-vae architecture," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[136] S. Lin, R. Clark, R. Birke, S. Schönborn, N. Trigoni, and S. Roberts, "Anomaly detection for time series using vae-lstm hybrid model," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4322–4326, Ieee, 2020.

[137] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *International Conference on Learning Representations*, 2014.

[138] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," *ICLR 2017*, 2016.

[139] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in β-vae," *arXiv preprint arXiv:1804.03599*, 2018.

[140] M. Suzuki, K. Nakayama, and Y. Matsuo, "Joint multimodal learning with deep generative models," *arXiv preprint arXiv:1611.01891*, 2016.

[141] D. Barrejón, P. M. Olmos, and A. Artés-Rodríguez, "Medical data wrangling with sequential variational autoencoders," *IEEE Journal of Biomedical and Health Informatics*, 2021.

[142] A. Javaloy, M. Meghdadi, and I. Valera, "Boosting heterogeneous vaes via multi-objective optimization," in *NeurIPS 2021 Workshop on Your Model is Wrong: Robustness and misspecification in probabilistic modeling*, 2021.

[143] R. Vedantam, I. Fischer, J. Huang, and K. Murphy, "Generative models of visually grounded imagination," *arXiv preprint arXiv:1705.10762*, 2017.

[144] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[145] C. K. Williams, C. Nash, and A. Nazábal, "Autoencoders and probabilistic inference with missing data: An exact solution for the factor analysis case," *arXiv preprint arXiv:1801.03851*, 2018.

[146] E. J. Gumbel, "Statistical theory of extreme values and some practical applications," *NBS Applied Mathematics Series*, vol. 33, 1954.

[147] Y. Shi, B. Paige, P. Torr, *et al.*, "Variational mixture-of-experts autoencoders for multi-modal deep generative models," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[148] Y. Burda, R. Grosse, and R. Salakhutdinov, "Importance weighted autoencoders," *arXiv preprint arXiv:1509.00519*, 2015.

[149] J. Tomczak and M. Welling, "Vae with a vampprior," in *International Conference on Artificial Intelligence and Statistics*, pp. 1214–1223, PMLR, 2018.

[150] A. Belenguer-Llorens, C. Sevilla-Salcedo, M. Desco, M. L. Soto-Montenegro, and V. Gómez-Verdejo, "A novel bayesian linear regression model for the analysis of neuroimaging data," *Applied Sciences*, vol. 12, no. 5, p. 2571, 2022.

[151] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of machine learning research*, vol. 1, no. Jun, pp. 211–244, 2001.

[152] I. M. de Diego, A. Muñoz, and J. M. Moguerza, "Methods for the combination of kernel matrices within a support vector framework," *Machine learning*, vol. 78, no. 1-2, p. 137, 2010.

[153] S. Qiu and T. Lane, "A framework for multiple kernel support vector regression and its applications to sirna efficacy prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 6, no. 2, pp. 190–199, 2008.

[154] G. Fung, M. Dundar, J. Bi, and B. Rao, "A fast iterative algorithm for fisher discriminant using heterogeneous kernels," in *Proceedings of the twenty-first international conference on Machine learning*, p. 40, 2004.

[155] E. Spyromitros-Xioufis, G. Tsoumakas, W. Groves, and I. Vlahavas, "Multi-target regression via input space expansion: treating targets as inputs," *Machine Learning*, vol. 104, no. 1, pp. 55–98, 2016.

[156] A. Karalič and I. Bratko, "First order regression," *Machine learning*, vol. 26, no. 2-3, pp. 147–176, 1997.

[157] S. Džeroski, D. Demšar, and J. Grbović, "Predicting chemical parameters of river water quality from bioindicator data," *Applied Intelligence*, vol. 13, no. 1, pp. 7–17, 2000.

[158] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, V. Vapnik, *et al.*, "Support vector regression machines," *Advances in neural information processing systems*, vol. 9, pp. 155–161, 1997.

[159] F. Murtagh, "Multilayer perceptrons for classification and regression," *Neurocomputing*, vol. 2, no. 5-6, pp. 183–197, 1991.

[160] A. Damianou, N. D. Lawrence, and C. H. Ek, "Multi-view learning as a nonparametric nonlinear inter-battery factor analysis," *Journal of Machine Learning Research*, vol. 22, no. 86, pp. 1–51, 2021.

[161] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep. 07-49, University of Massachusetts, Amherst, 2007.

[162] L. Wolf, T. Hassner, and Y. Taigman, "Effective unconstrained face recognition by combining multiple descriptors and learned background statistics," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 10, pp. 1978–1990, 2010.

[163] C. M. Bishop and G. D. James, "Analysis of multiphase flows using dual-energy gamma densitometry and neural networks," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 327, no. 2-3, pp. 580–593, 1993.

[164] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[165] A. Royer, K. Bousmalis, S. Gouws, F. Bertsch, I. Mosseri, F. Cole, and K. Murphy, "Xgan: Unsupervised image-to-image translation for many-to-many mappings," in *Domain Adaptation for Visual Understanding*, pp. 33–49, Springer, 2020.

[166] 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, p. 106259, Feb. 2021.

[167] 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, p. dlaa084, Oct. 2020.

[168] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[169] 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, pp. 1–6, May 2014.

[170] 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.

[171] 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*, June 2021.

[172] R. Schaback, "A unified theory of radial basis functions: Native hilbert spaces for radial basis functions ii," *Journal of computational and applied mathematics*, vol. 121, no. 1-2, pp. 165–177, 2000.

[173] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.