MARISMa: a routine MALDI-TOF MS database from 2018 to 2024

Lucía Schmidt-Santiago[1] Inés López-Mareca[1] Mario Blázquez-Sánchez[2]
José Miguel Moreno[3] Carlos Sevilla-Salcedo[1] Vanessa Gómez-Verdejo[2]
Belén Rodríguez-Sánchez[2] David Rodríguez-Temporal[2,†,*]
Alejandro Guerrero-López[1,† *]

[1]Department of Signal and Communication Theory, Universidad Carlos III de Madrid, 28911 Spain
[2]Department of Clinical Microbiology and Infectious Diseases, Gregorio Marañón Health Research Institute, Madrid, 28040 Spain
[3]Computer Science and Engineering Department, Universidad Carlos III de Madrid, 28911 Spain
†these authors contributed equally to this work

Abstract

Clinical microbiology laboratories play a crucial role in identifying pathogens, guiding antibiotic treatment, and managing antimicrobial resistance (AMR). Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF MS) has become essential for rapid, accurate, and cost-effective microbial identification. Recent advances in integrating MALDI-TOF MS with Artificial Intelligence (AI) show promise in improving microbial detection and prediction of AMR. However, progress is limited by the lack of comprehensive and openly accessible databases that restrict the validation, reproducibility, and applicability of the model.

To address this gap, we introduce a publicly available, extensively documented MALDI-TOF MS database comprising *207,950* unique spectra from isolates collected between 2018 and 2024 at the Hospital General Universitario Gregorio Marañón, Spain. This dataset includes *191,048* bacteria, *16,537* fungal, and *365* mycobacterial isolates, all provided in raw format with detailed metadata and AMR annotations for *13,114* bacterial isolates. In contrast to many existing datasets, this resource is openly and freely shared, rigorously curated, and designed to support a wide range of machine learning and epidemiological applications. By ensuring unrestricted access to high-quality, standardized data, this database aims to promote transparency, reproducibility, comparative benchmarking, and collaborative progress in AI-driven clinical microbiology.

*Corresponding author: david.rodriguez@iisgm.es

Background & Summary

The accurate and timely identification of microorganisms is crucial in clinical microbiology for effective patient management and infection control. Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF MS) has revolutionized this field by offering rapid, precise, and cost-effective identification of a broad spectrum of bacteria and fungi. This technology generates unique spectral fingerprints by ionizing proteins from microbial isolates. Its high reproducibility and efficiency have made MALDI-TOF MS indispensable in clinical diagnostics, proteomics, and biomarker discovery.

Despite these advances, significant limitations persist, particularly in regard to the identification of filamentous fungi and less common bacterial species [1, 2, 3]. Commercial databases often lack comprehensive spectral data for these organisms, leading to challenges in accurate identification and potential misdiagnosis. For instance, the primary limitation in identifying filamentous fungi using MALDI-TOF MS is the limited reference spectra available, with only one the Database of Resistance Information on Antimicrobials and MALDI-TOF Mass Spectra (DRIAMS) available offering 5,203 fungi MALDI-TOF samples [4]. This gap underscores the need for more inclusive and representative spectral databases to enhance the diagnostic capabilities of MALDI-TOF MS in clinical settings.

The integration of artificial intelligence (AI) techniques with MALDI-TOF MS data has significantly advanced microbial identification and characterization. Recent studies utilizing publicly accessible databases, such as DRIAMS[4, 5] and the Robert Koch Institute (RKI) database[6], have demonstrated considerable improvements in microbial detection and antimicrobial resistance (AMR) prediction through the application of machine learning techniques[7, 8, 9, 10, 11, 12, 13]. For example, [7] use DRIAMS along with a private database to train MLP to predict AMR in *Pseudomona aeruginosa*. Similarly, in [8] they combined DRIAMS public *Staphylococcus aureus* with a private Taiwan database to train a LightGBM model to predict AMR. Other researchers, such as [9], propose to use DRIAMS as a pretraining database to then finetune to their own private database for AMR prediction. Despite these advancements, current publicly available datasets remain limited in number and scope, significantly hindering further developments in this rapidly growing research domain.

While certain studies have shared datasets alongside their research for reproducibility, these datasets often lack comprehensive database papers or thorough metadata descriptions, reducing their practical utility for novel research initiatives. Furthermore, some publicly available data have already been preprocessed, thus constraining researchers solely to reproducing published findings and limiting their potential to explore novel hypotheses [14, 15, 16, 17, 18, 19, 20]. In contrast, numerous investigations rely on proprietary databases or small datasets with limited representativeness [21, 22, 23, 24, 25, 26, 27], severely restricting reproducibility and generalizability.

Moreover, the shortage of openly accessible, well-characterized MALDI-TOF datasets critically impedes robust evaluation of AI models, particularly with respect to out-of-distribution (OOD) validation which is often lacked in the literature [12, 22, 21, 24, 20, 25, 18, 28, 29, 30]. Without open datasets, researchers are unable to rigorously test and demonstrate the transferability and clinical applicability of their models to external, heterogeneous data.

The open availability of both raw MALDI-TOF MS spectral data and corresponding code is essential for advancing reproducible research and fostering innovation in machine learning applications. However, our review [31] revealed that only a limited number of studies (n=13) provide open access to their datasets, and even fewer (n=7) accompany these with publicly shared machine learning implementations or raw spectral files in the original Bruker Daltonics XMASS format, or *BrukerRaw* for short [32, 33]. Moreover, incon-

sistent file formats and frequent omission of critical metadata (e.g. acquisition time, device information) in such datasets further constrain their utility for standardization and reproducibility. This scenario underscores a persistent gap in the field: the lack of a large, standardized, and fully open-access repository that covers the full spectrum of microorganisms, including bacteria, fungi, and mycobacteria. Creating and openly sharing these resources will facilitate: (i) development of innovative and robust AI-driven methodologies, (ii) rigorous and comprehensive OOD validation studies, (iii) enhanced transparency in method reporting, (iv) standardized comparative analysis across models and research groups, and ultimately, (v) acceleration of collaborative advancements in AI-driven clinical microbiology research.

In light of these limitations, we present here the most extensive MALDI-TOF MS dataset to date, comprising **N=207,950** unique spectra, which include data from **191,048** bacterial, **16,537** fungal, and **365** mycobacterial isolates. All spectra are provided in the original *BrukerRaw* format, ensuring that complete metadata regarding acquisition time, device information, and calibration details are available to researchers. This resource is intended to support a wide range of research applications and to set a new benchmark for open-access resources in microbiology.

Methods

A description of several aspects of the corpus is presented next, including the ethics declaration, a description of the participants, the speech protocol tasks, and the audio registration process.

Ethics declaration

The Ethics Committee of the Gregorio Marañón Health Research Institute (CEIm) evaluated this project and considered that all the conditions for waiving informed consent were met, since the study was conducted with protein spectra from microbiological samples and not with human products.

Patients' information was completely removed, and a unique identifier per bacteria was kept that is completely different from the one used in the Hospital's electronic records. No personal data was exchanged with external researchers who had access to the corpus.

MALDI-TOF spectra included

This dataset includes all routine MALDI-TOF MS spectra acquired at Hospital General Universitario Gregorio Marañón between May 1, 2018 and December 31, 2024. The year 2018 was selected as the starting point to ensure consistency in acquisition instrumentation, as this marks the installation of the current MALDI-TOF device at the hospital. To maintain complete annual coverage, data from 2025 onwards is not included in this release. Spectra were included if they originated from clinical routine workflows and were successfully assigned to a genus or species by the clinical microbiology laboratory. Exclusion criteria comprised (i) spectra flagged as low-quality (score below 1.7) by internal quality control systems or visual inspection, and (ii) spectra from unidentified or unclassified isolates.

This release represents the first installment of an ongoing initiative. Future versions of the database are planned and may include additional years or retrospective corrections to improve coverage and metadata quality.

To ensure the taxonomic reliability of species annotations assigned to each spectrum, a post-processing validation step was applied to the taxonomic identification outputs generated by the MALDI Biotyper system

(Bruker Daltonics). Each processed sample yields both the raw spectrum in the *BrukerRaw* format and an associated `.xml` file containing the taxonomic results for each of the analytes, individual spectra acquired from different wells of the MALDI target plate.

Within each XML file, the Biotyper software returns a ranked list of up to 10 potential identifications per analyte, ordered by their `globalMatchValue`, a numerical score reflecting the similarity between the measured spectrum and entries in the reference library. Each result includes a genus and species-level classification, reported under the `referencePatternName` field.

The following filtering strategy was applied to retain only high-confidence and taxonomically consistent identifications:

1. **Confidence threshold check:**
   The first match (Rank 1) was evaluated. If its `globalMatchValue` was below a threshold of **1.7**, the analyte was **discarded**, as the classification was considered unreliable.
2. **Genus consistency check:**
   The genus of the second match (Rank 2) was compared with that of the first. If they differed, the analyte was **discarded**, indicating taxonomic inconsistency at the genus level.
3. **Species concordance:**
   - If the genus matched but the species differed, the `globalMatchValue` of the second match was evaluated:
     - If it was **greater than 1.7**, the analyte was **discarded**, due to conflicting high-confidence species-level identifications.
     - If it was **less than 1.7**, the analyte was **retained**, assuming the conflict was minor and not strongly supported.
4. **Third match validation:**
   - The genus must match that of the first two entries. If not, the analyte was **discarded**.
   - If the species differed:
     - A `globalMatchValue` above the threshold led to **rejection**.
     - A value below the threshold allowed the analyte to be **accepted**.

This multi-stage filtering process was systematically applied to all spectra in the dataset to maximize the accuracy and consistency of genus and species labels. By enforcing these criteria, we ensured that only spectra with coherent, high-confidence taxonomic annotations were retained in the final database.

It is important to note that each well-level spectrum (analyte) is individually annotated following this procedure. However, since multiple wells (analytes) correspond to the same original sample, discrepancies in their assigned genus or species can arise. In cases where wells from the same sample were classified into different taxonomic groups, the entire spectrum (i.e., the complete sample) was removed from the database to maintain label consistency and ensure data integrity.

An overview of this validation process is illustrated in Figure 1.

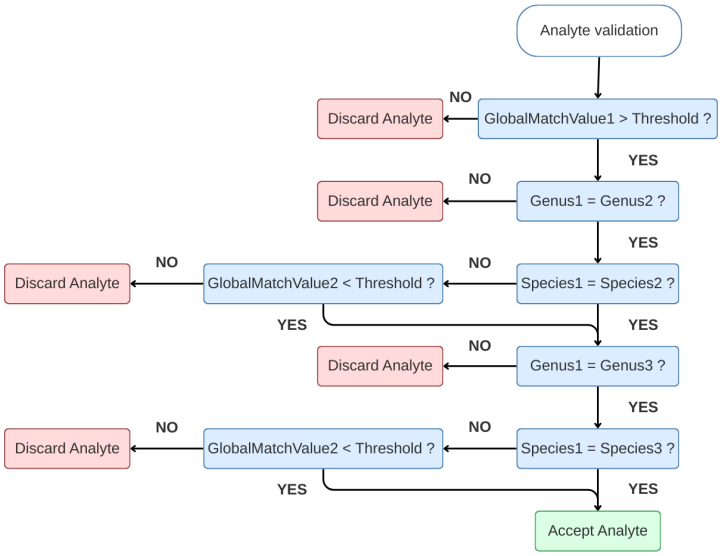


Figure 1: Flowchart describing the analyte validation procedure applied during the preprocessing of the MARISMa database. Only analytes with consistent genus and species predictions among the top-3 classification matches, and with high-confidence global match values (threshold ≥ 1.7), are retained.

The final dataset comprises *191,118* bacterial, *16,537* fungal, and *365* mycobacterial isolates, all shared in their raw spectral form. Each isolate may have multiple spectral replicates, generated as part of routine reproducibility testing; in total, the dataset includes *N* MALDI-TOF spectra. Figure 2 highlights the most represented microorganisms in the collection, which predominantly belong to bacterial species classified as WHO critical or high-priority pathogens [34]. These include *Escherichia coli* and *Klebsiella pneumoniae* (critical priority), as well as *Enterococcus faecalis* and *Staphylococcus aureus* (high priority).

For bacterial species, we reference the 2024 WHO Bacterial Priority Pathogens List (BPPL) [34], which categorizes pathogens into three priority tiers—critical, high, and medium—based on their public health relevance and the urgency for research and antimicrobial development.

Although we do not provide antibiotic susceptibility profiles for all samples, we include isolates identified as belonging to these species regardless of resistance phenotype. Our dataset thus contains representatives from all bacterial families listed in the WHO classification, as detailed in Figure 3.

For the fungal species, we follow the WHO Fungal Priority Pathogens List (FPPL) published in 2022 [35], which identifies 19 fungal pathogens of concern and similarly groups them into critical, high, and medium priority categories. Our database includes representative isolates from all fungal families listed in this classification as detailed in Figure 4.

By encompassing all WHO-designated bacterial and fungal priority pathogens, this dataset offers an exceptionally valuable resource for the development, benchmarking, and evaluation of AI-driven tools in clinical microbiology, antimicrobial resistance surveillance, and pathogen detection.

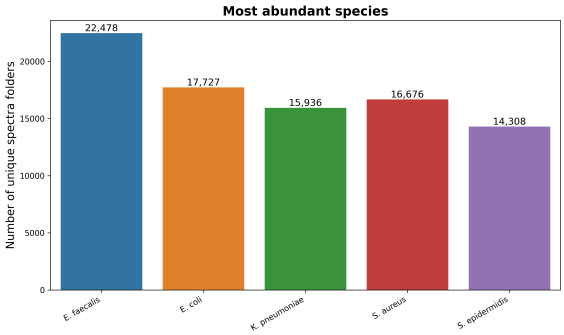


Figure 2: The top 5 most populated microorganisms of our database. Four of the five are part of the top priority pathogen list defined by the WHO [34]

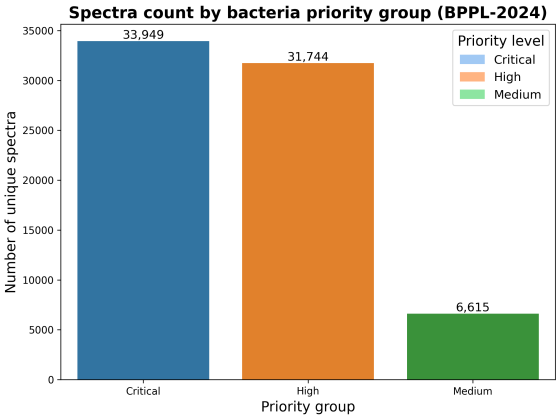


Figure 3: Number of bacterial samples available in our database for the three groups—critical, high, and medium— based on their impact on health defined by the 2024 WHO BPPL.

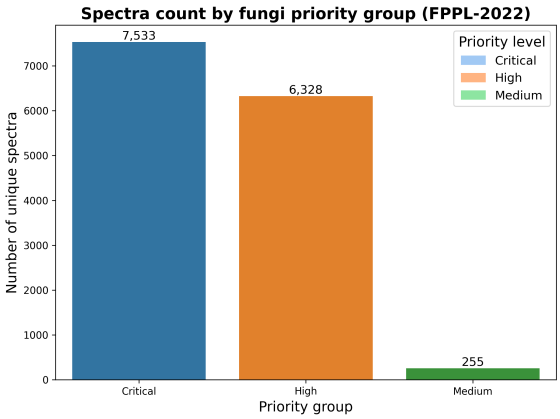


Figure 4: Number of fungal samples available in our database for the three groups—critical, high, and medium— based on their impact on health defined by the 2022 WHO FPPL.

Sample protocol

Protein spectra included in the MARISMa database were acquired from isolates sent for routine identification to the Clinical Microbiology and Infectious Diseases of the Hospital General Universitario Gregorio Marañón. One spectrum was acquired per microbial isolate using a mass spectrometer MBT Smart MALDI Biotyper (Bruker Daltonics, Bremen, Germany). The instrument was calibrated daily using a commercial bacterial test standard (Bruker Daltonics), prior to spectra acquisition. Isolates were cultured on standard media for bacterial and fungal isolates. The spectra from microorganisms analysed directly from blood cultures were obtained by using a commercial kit (Sepsityper kit, Bruker Daltonics) that concentrates the microorganisms present in 1ml of blood culture and removes blood components using a washing buffer.

Metadata

BrukerRaw spectra are composed of several files, mainly `fid`, `acqu` and `proc`. The first is a binary file with the intensity values of the spectrogram, and the last two are plain text files containing metadata that characterize the acquired sample and the instrument used. Among the hundreds of variables found between this pair of files alone, we highlight a dozen that are useful to researchers:

- **Sample ID (`$ID_raw`).** Random, unique identifier automatically generated during acquisition.
- **Target ID (`$TgIDS`).** Same as sample ID, but identifies the whole group of samples acquired at once in the same batch.
- **Sample position (`$PATCHNO`).** Acquired position inside the plate (e.g., A1).
- **Acquisition date (`$AQ_DATE`).** Timestamp for when the sample was acquired.
- **Calibration date (`$CLDATE`).** Timestamp of the last calibration of the acquisition instrument.
- **Instrument serial number (`$InstrID`).** Unique identifier of the acquisition instrument set by the manufacturer.

- **Instrument type (`$InstTyp`).** Numeric code determining the type of instrument (e.g., type 9 is microflex). As with most of the *BrukerRaw* format, these constants are not officially documented, but the research community has already taken on this task [36].
- **Digitizer type (`$DIGTYP`).** Same as instrument type, but for the digitizer (e.g., type 19 is Bruker BD0G5).
- **flexControl version (`$FCVer`).** Acquisition software version.
- **AIDA version (`$Acquver`).** Calibration software version.
- **Spectrum size (`$TD`).** Number of spectrum peaks (intensity time periods).
- **Spectrum delay (`$DELAY`).** Time in nanoseconds until the first measured intensity.
- **Sample rate (`$DW`).** Nanoseconds between each measured intensity.
- **Mass calibration constants (`$ML1`, `$ML2`, `$ML3`).** Also referred to as dwell times. Parameters needed to convert measured intensities from the `fid` file to pairs of m/z using the formula described in [37].

Antimicrobial resistance report

We provide antimicrobial resistance (AMR) annotations for 13,114 *Klebsiella pneumoniae* isolates, including susceptibility results for up to 41 antibiotics. Table 1 shows the AST results, showing counts of susceptible (S) and resistant (R) phenotypes based on routine clinical testing.

The most commonly tested antibiotics include beta-lactams, carbapenems, and fluoroquinolones. As expected, certain antibiotics such as ampicillin show nearly universal resistance, whereas others such as colistin and imipenem retain high susceptibility rates. These data illustrate the clinical relevance and variability present in the dataset and underscore its utility for AMR prediction tasks.

Table 1: Full antibiotic susceptibility breakdown for *Klebsiella pneumoniae* isolates with AMR metadata (n = 13,114), including all tested antibiotics.

Antibiotic	Susceptible (S)	Resistant (R)	Intermediate (I)	Tested (Total)
Ampicillin	13,110	4	0	13,114
Cefepime	3,085	9,991	37	13,113
Cefuroxime	4,419	7,737	957	13,113
Ertapenem	1,136	11,974	3	13,113
Gentamicin	1,715	11,397	1	13,113
Imipenem	268	12,699	146	13,113
Piperacillin-Tazobactam	3,646	9,465	2	13,113
Trimethoprim-Sulfamethoxazole	3,264	9,831	18	13,113
Tobramycin	2,383	10,726	2	13,111
Ciprofloxacin	3,155	9,624	323	13,102
Fosfomycin	3,351	9,658	0	13,009
Amikacin	809	12,163	0	12,972
Colistin	183	12,745	0	12,928
Levofloxacin	2,944	9,926	48	12,918
Meropenem	245	12,511	85	12,841
Piperacillin	12,584	125	5	12,714
Aztreonam	2,398	9,822	15	12,235
Cefotaxime	850	9,928	14	10,792
Ceftazidime	849	9,821	122	10,792
Ticarcillin	10,279	1	0	10,280
Cefixime	2,303	7,976	0	10,279
Ceftazidime-Avibactam	88	10,103	0	10,191
Ceftolozane-Tazobactam	936	9,255	0	10,191
Norfloxacin	2,959	6,867	3	9,829
Nitrofurantoin	2,310	7,438	0	9,748
Ampicillin-Sulbactam	3,406	4,741	277	8,424
Amoxicillin-Clavulanic Acid E	2,114	5,514	0	7,628
Amoxicillin-Clavulanic Acid	1,659	3,766	0	5,425
Cephalexin	1,650	3,531	0	5,181
Cefoxitin	246	2,566	50	2,862
Minocycline	522	2,125	3	2,650
Tigecycline	83	2,140	0	2,223
Trimethoprim	550	1,509	26	2,085
Cefalotin	499	1,322	0	1,821
Cefazolin	982	93	0	1,075
Mecillinam	1	645	0	646
Chloramphenicol	111	394	0	505
Nalidixic Acid	297	114	0	411
Doripenem	0	22	0	22
Cefiderocol	6	3	1	10
Meropenem/Vaborbactam	0	1	0	1

Data Record

Data access is facilitated through both the Zenodo repository [38]. The structural representation of the dataset is illustrated in Fig. 5 and Fig. 6. The dataset comprises **207,950** unique MALDI-TOF MS spectra acquired from bacterial and fungal isolates, all provided in the original *BrukerRaw* format. This format ensures that complete metadata—such as acquisition time, device information, and calibration details—is available to researchers, thereby facilitating robust reproducibility and further data-driven investigations.

The database is systematically organized to facilitate efficient data retrieval and analysis. The database is

first divided by year of data acquisition (2018–2024). Within each year, data is grouped by bacterial genera, and inside each genus, by species. For each species, the database contains the studies in which that species was analyzed, identified by an anonymized external ID. Inside each study folder, data is organized following *BrukerRaw* format, i.e., subfolders named according to the target position on the MALDI target plate (e.g., `0_A1`, `0_A2`, `0_D5`), which correspond to biological replicates. Within each target position folder, there is always a folder named `1`, representing the first acquisition of the sample; when experiments are repeated on the same sample, additional folders named `2`, `3`, etc., appear, corresponding to technical replicates. Each acquisition folder contains the standard *BrukerRaw* structure: a `1SLin` folder with acquisition files (`acqu`, `acqus`, `fid`, `sptype`) and a `pddata` folder with additional metadata. Additionally, at the year level, a statistics file is included summarizing the number of genera, species, studies, and spectra available in total. Similarly, each year folder contains a statistics file reporting the number of genera, species, studies, and spectra available within that year.

Table 2: Folder structure of the *MARISMa* database.

Level	**Description**
`Year`	Year of data acquisition (e.g., `2018`, `2019`, `2020`...)
`Genus`	Bacterial genus folder
`Species`	Bacterial species folder
`Study (extern_ID)`	Study folder identified by external ID
`Target_Position`	Target plate position folder (e.g., `0_A1`) = Biological replicate
`1`, `2`, ...	Acquisition folders (`1` = first acquisition, others = technical replicates)
`1SLin`	Raw data folder (BrukerRaw structure)
`acqu`, `acqus`, `fid`, `sptype`	Acquisition files
`pddata`	Metadata folder

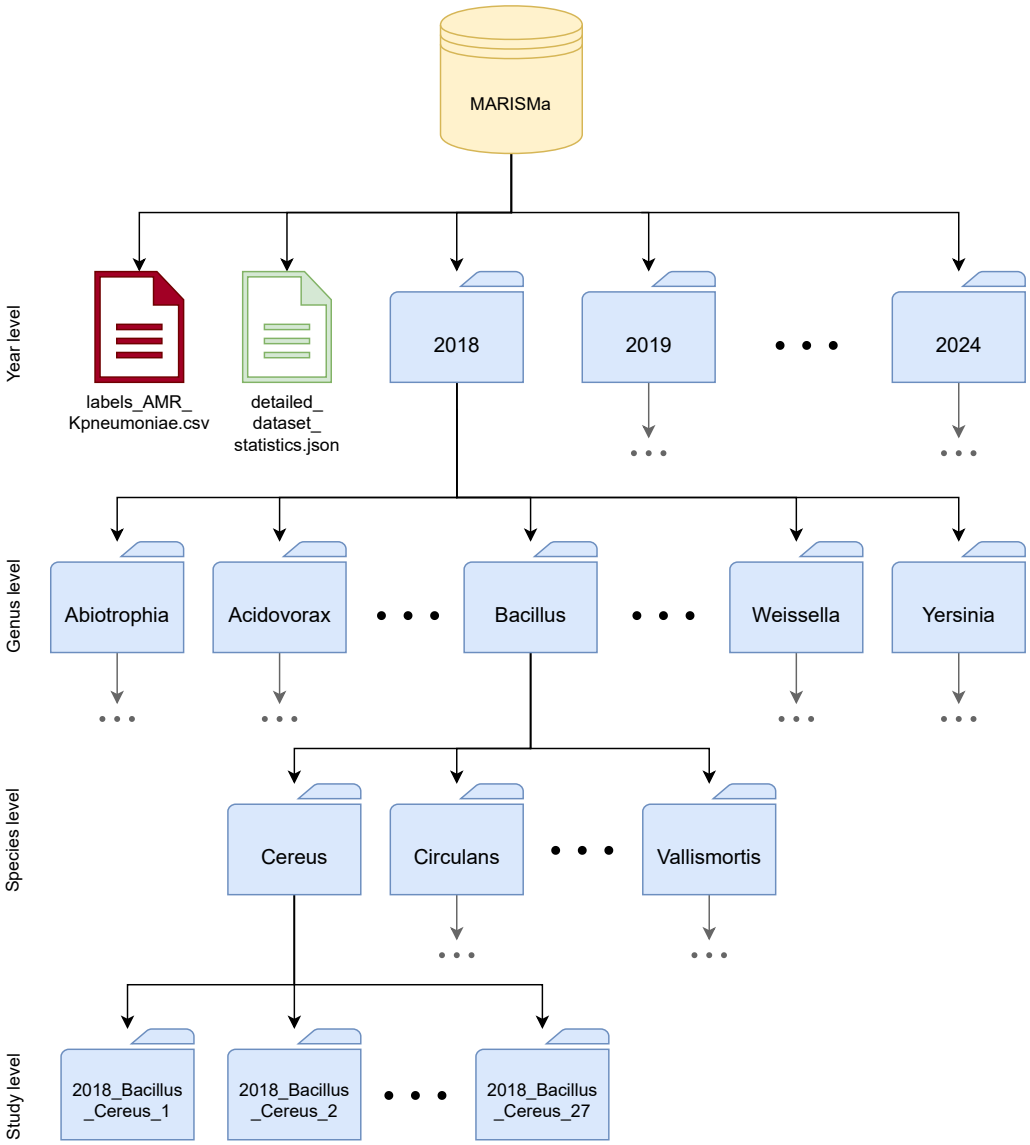


Figure 5: Hierarchical folder structure of the *MARISMa* database. The dataset is organized chronologically, with one folder per acquisition year (`2018` to `2024`). Within each year, data is grouped by bacterial genus, then by species, and finally by study. Study folders follow the format `year_genus_species_index`. At the root level of the dataset, two files — `detailed_dataset_statistics.json` and `labels_AMR_Kpneumoniae.csv` — provide global summary statistics of spectra distribution per genus and species across the dataset and the labels of AMR for *Klebsiella pneumoniae* isolates, respectively.

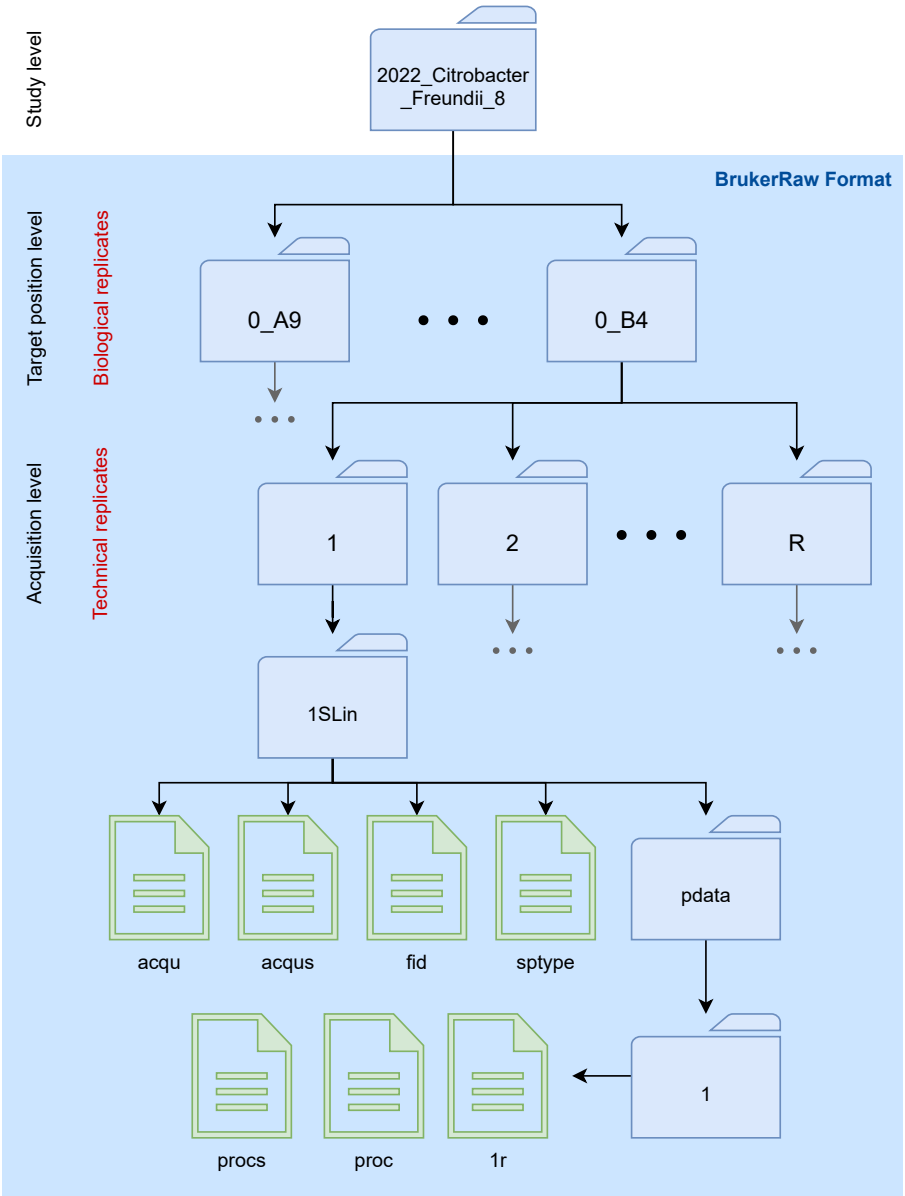


Figure 6: Structure of the raw data format used in the *MARISMa* database, following the BrukerRaw standard. Study folders follow the format `year_genus_species_index` and contain subfolders representing MALDI target plate positions (e.g., `0_A1`, `0_B8`), which correspond to biological replicates. Each target position folder may include multiple acquisition folders (`1`, `2`, ...) representing technical replicates. These replicate folders contain raw spectra in Bruker format (within a `1SLin` structure) with raw acquisition files and a `pdata` folder. Relevant files are the `fid` file, which stores the mass spectrum in binary format, and the `acqu` file, which contains essential metadata such as the acquisition and calibration dates. Other files like `acqus` and `sptype` contain complementary acquisition settings. This structure ensures reproducibility and traceability of each spectral acquisition.

Metadata

Each MALDI-TOF spectrum in our dataset is accompanied by a metadata file in XML format, following Bruker's *MspMatchResult* schema. This file stores detailed acquisition and classification information, including one or more `<Analyte>` entries, each corresponding to a well on the MALDI plate (e.g., `F9`, `F10`). For each analyte, the metadata includes acquisition details (instrument, operator, method), a peak list with mass and intensity values, and up to ten identification matches sorted by descending confidence. Each match contains the predicted genus and species (`referencePatternName`) and its corresponding similarity score (`globalMatchValue`).

AMR evaluation

Antimicrobial resistance (AMR) data is provided in a CSV file, where each row corresponds to a unique biological sample, identified by a spectrum identifier in the first column. This identifier does not refer to an individual well on the MALDI target plate, but to the sample as a whole—meaning that all analytes (wells) derived from the same sample share the same AMR profile. In the folder structure, this identifier corresponds to a folder located under each `Species/` directory, containing all subfolders for the related target wells.

The CSV includes several metadata fields: the sample identifier, the date of collection, the isolation status date, the clinical service of origin, the anatomical source of the sample, and the type of isolation. Following this, antimicrobial susceptibility results are organized in column pairs: each pair includes the Minimum Inhibitory Concentration (MIC) for a given antibiotic and its corresponding clinical interpretation (e.g., susceptible, intermediate, resistant). Resistance profiles are available for up to 31 antibiotics.

Importantly, resistance data is only available for samples identified as *Klebsiella pneumoniae*, and not all entries contain full susceptibility results for all 31 antibiotics—i.e., the dataset includes missing values as detailed in Fig. 7.

Technical Validation

The MARISMa corpus has been partially examined in several studies, establishing its usefulness for different studies such as antibiotic resistance prediction in *K. pneumoniae* [39], automatic ribotyping *C. difficile* [40], data augmentation techniques development for microbiology [41], automatic surveillance of *E. coli*[42], among others.

Demonstration of dataset utility and reproducibility

To facilitate immediate use and reproducibility, we provide a demonstration notebook, `simple_classifier`, available in the associated GitHub repository. This notebook exemplifies how to query, preprocess, and classify spectra using our MALDI-TOF MS dataset.

The provided code includes a flexible function that allows users to construct training datasets by selecting specific acquisition years, genera, and species. For instance, by specifying years `[2018, 2019]` and genus `Klebsiella`, the function retrieves all matching samples and assembles a corresponding dataset, which can be used for downstream machine learning tasks.

Each spectrum is then preprocessed using a series of standard signal processing techniques widely adopted in MALDI-TOF analysis, whose hyperparameters can be tuned by the user. For instance, the example given

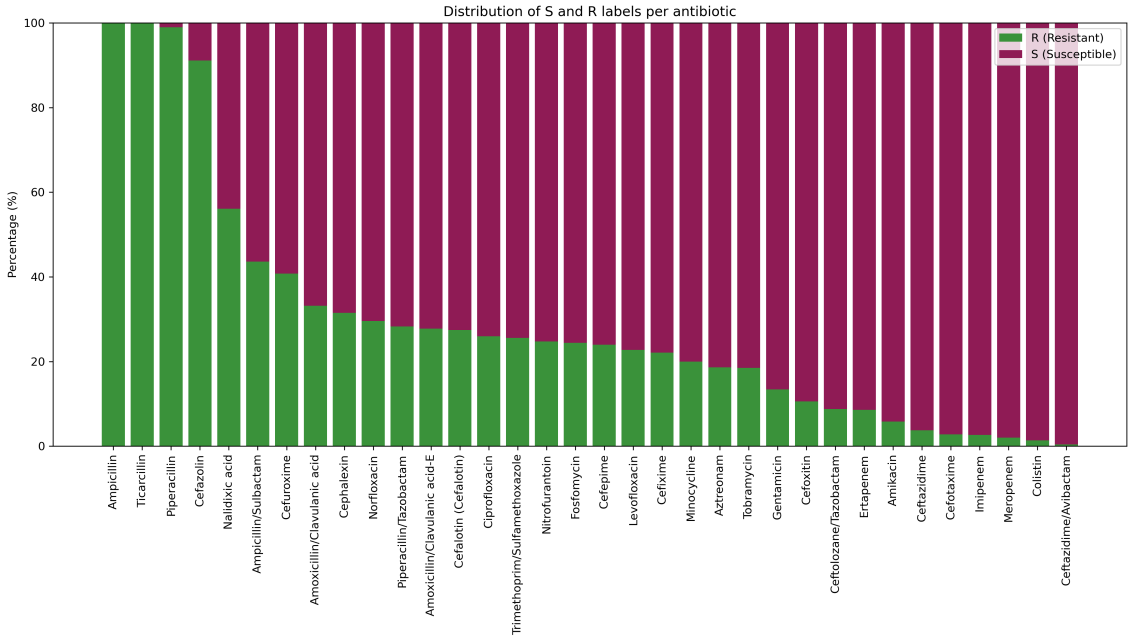


Figure 7: Missingness distribution per antibiotic

follows:

```
bin_step = 9
spec = SpectrumObject.from_bruker(acqu_file, fid_file)
spec = VarStabilizer(method='sqrt')(spec)
spec = Smoother(halfwindow=10, polyorder=3)(spec)
spec = BaselineCorrecter(method='SNIP', snip_n_iter=10)(spec)
spec = Normalizer()(spec)
spec = Trimmer(min=2000, max=20000)(spec)
spec = Binner(step=bin_step)(spec)
```

A simple Random Forest classifier is then trained to discriminate between all available *Klebsiella* subspecies within the selected timeframe. This example showcases not only the high quality and diversity of the dataset, but also its ease of use and adaptability. Users are encouraged to modify the notebook to select different time periods, genera, or species, enabling rapid prototyping and testing of new machine learning approaches.

Impact of preprocessing on spectral clustering To evaluate the structure and interpretability of the MALDI-TOF MS spectra, we performed a t-SNE analysis of the top ten most represented organisms in the dataset, including bacterial and fungal species. Figure 8 illustrates the resulting embeddings before and after applying standard preprocessing steps.

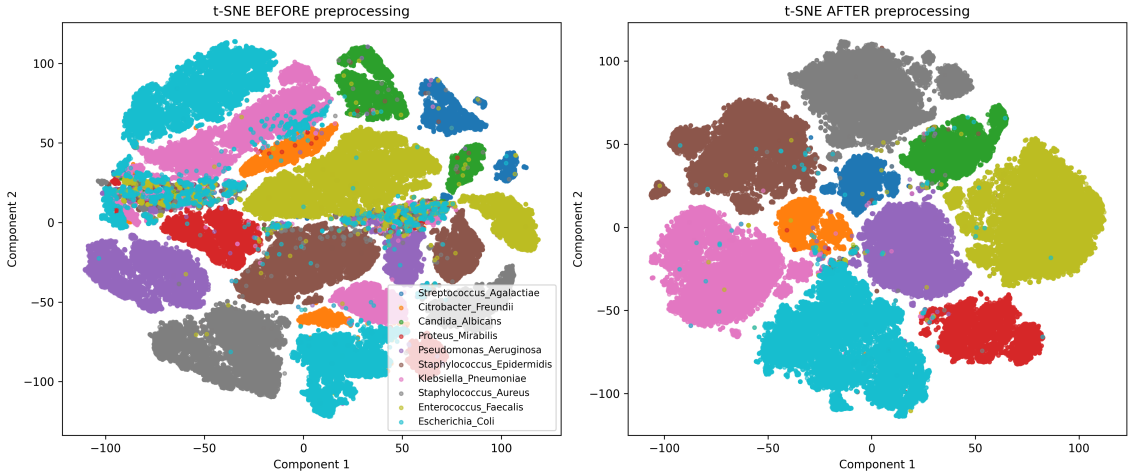


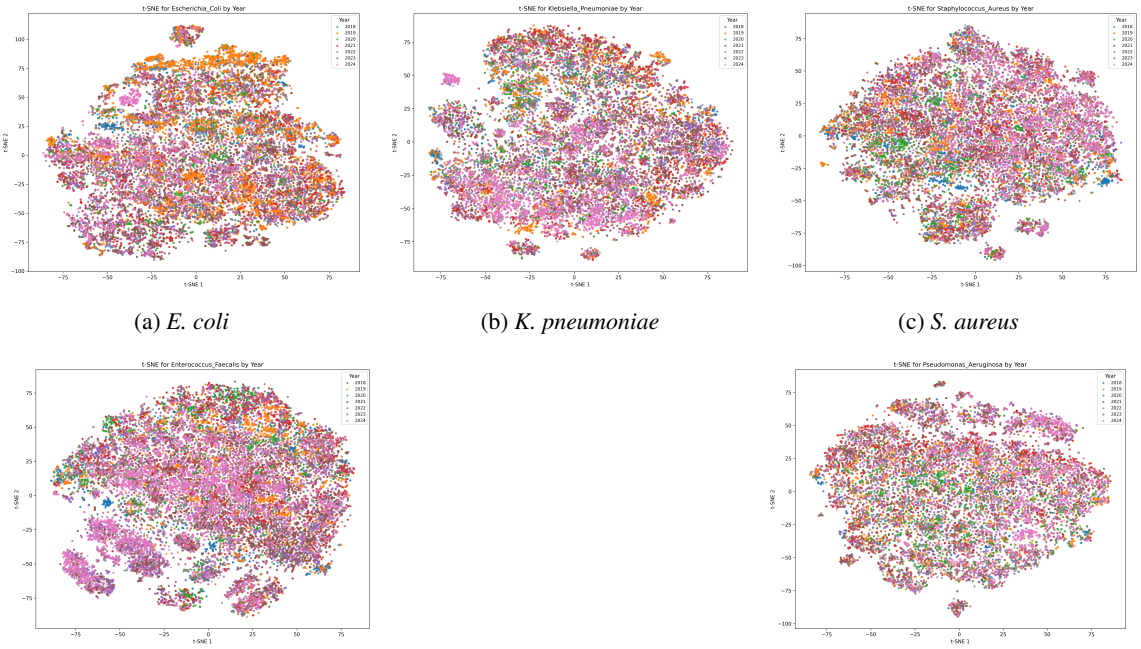
Figure 8: t-SNE visualizations of the top ten most represented microorganisms in the dataset. **Left**: Raw spectra as acquired, interpolated to equal length. Some overlap and misclassification are visible. **Right**: The same data after preprocessing, including variance stabilization, smoothing, baseline correction, normalization, and trimming. Well-separated clusters indicate improved inter-class structure and highlight the utility of preprocessing.

In the left panel, the t-SNE was computed on the raw spectra (interpolated to a common length) as they are originally acquired. While some clustering is evident, several species exhibit overlapping or poorly defined boundaries, with certain samples scattered across clusters. This highlights the challenges of working directly with unprocessed spectra, which may contain high variance, noise, or baseline drift.

In the right panel, we show the same dataset after applying a commonly adopted preprocessing pipeline in MALDI-TOF analysis. Specifically, each spectrum underwent variance stabilization, smoothing, baseline correction, normalization, and trimming. The resulting t-SNE reveals clearly defined clusters with minimal overlap, reflecting a more coherent biological structure in the data.

Although we release all spectra in raw form to ensure full flexibility and transparency, this comparison underscores the importance of preprocessing in downstream machine learning tasks and supports the use of established preprocessing pipelines for spectral feature extraction and classification.

Temporal consistency of spectra acquisition To assess the robustness and temporal consistency of the MALDI-TOF MS data, we conducted a t-SNE 2D analysis for five of the most important bacterial species in the dataset. For each species, we visualized the t-SNE embedding of their spectra, color-coded by the year of acquisition (2018–2024).



(a) *E. coli* (b) *K. pneumoniae* (c) *S. aureus*

(d) *E. faecalis* (e) *P. aeruginosa*

Figure 9: t-SNE visualization of MALDI-TOF spectra for five key bacterial species, color-coded by year of acquisition. The absence of temporal clustering indicates consistency in spectral quality across the years.

Preservation of taxonomic separability across years To further evaluate the diversity and consistency of the MALDI-TOF MS data across the study period, we performed an annual t-SNE analysis from 2018 to 2024. For each year, we extracted spectra corresponding to the ten most frequently represented microorganisms (including both bacteria and fungi) and applied dimensionality reduction to visualize their latent structure.

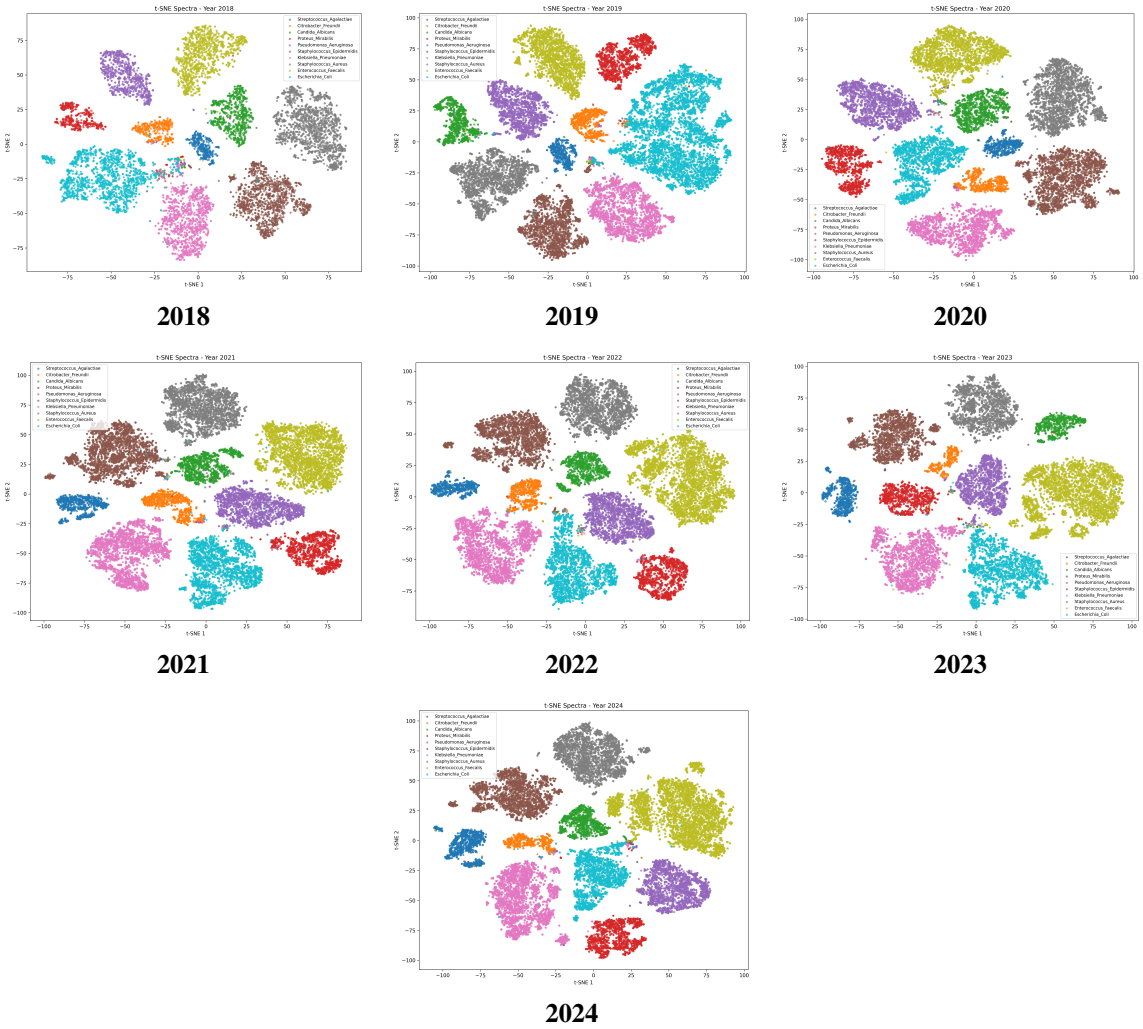


Figure 10: t-SNE embeddings of the ten most represented microorganisms in the dataset, plotted separately for each year from 2018 to 2024. While the t-SNEs are computed independently and thus not aligned spatially, all years consistently exhibit ten separable clusters, reflecting robust inter-class separability and data quality across time.

Figure 10 displays the resulting t-SNE plots for each year. Although each embedding was computed independently and is thus not directly comparable in position or scale, each year consistently yields ten well-separated clusters. This repeated clustering pattern across years strongly supports the taxonomic distinguishability of the major organisms in the dataset and reflects both the biological diversity of the samples and the stable acquisition quality over time.

Despite previous research conducted with partial subsets of this dataset, the complete database presented here has not yet been fully utilized or explored to its maximum potential. Similar to the impact of pub-

licly available resources like the DRIAMS and RKI databases [4, 6], we anticipate that this comprehensive dataset will significantly advance the development and application of more sophisticated deep learning models, including Transformer-based approaches [9, 23], thereby fostering innovation in clinical microbiology research.

Usage Notes

The MARISMa dataset is complemented with a set of Python scripts for data management and analysis. These scripts were designed to provide a practical introduction to working with the dataset and to facilitate initial experiments for bacterial identification within this corpus.

The script "`simple_classifier.ipynb`" offers a straightforward demonstration on processing and analyzing MALDI-TOF spectra, followed by a simple Random Forest classifier trained to discriminate between all available *Klebsiella* subspecies within the selected timeframe. This script is intended to inspire further analytical explorations.

Code Availability

The codes used for data preprocessing and cleaning can be accessed in a specific GitHub® repository (`https://github.com/ilmareca/MaldiTof-BacteriaID-GregorioMaranon`). Additionally, the referenced scripts used for classification purposes are included in the same repository.

Acknowledgements

This work is partially supported by grant PID2023-146684NB-I00 funded by MCIN/AEI/10.13039/501100011033 and ERDF/EU and by the project TEC-2024/COM-89, by the project PI18/00997 from the Health Research Fund (FIS. Instituto de Salud Carlos III. Plan Nacional de I+D+I 2013-2016) of the Carlos III Health Institute (ISCIII, Madrid, Spain) partially financed by the European Regional Development Fund (FEDER) 'A way of making Europe'. The funders had no role in the study design, data collection, analysis, decision to publish, or preparation/content of the manuscript. DRT (Sara Borrell CD22-00014) and BRS (Miguel Servet CPII19/00002) are funded by ISCIII. MBS is the recipients of the Intramural predoctoral contract 2022 from the Health Research Center of the Hospital Gregorio Marañón -IISGM-.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

L.S.S., I.L.M., and J.M.M. processed, cleaned, and curated the dataset. D.R.T. and M.B.S. collected the clinical isolates from routine diagnostics. I.L.M. developed the software used for data analysis. D.R.T., B.R.S., and V.G.V. provided resources and acquired funding. A.G.L. wrote the initial manuscript draft. A.G.L. and C.S.S. conducted code validation and addressed technical issues. All authors critically reviewed, edited, and approved the final version of the manuscript.

References

[1] Stein, M. *et al.* Evaluation of three maldi-tof mass spectrometry libraries for the identification of filamentous fungi in three clinical microbiology laboratories in manitoba, canada. *Mycoses* **61**, 743–753 (2018).

[2] Zvezdanova, M. E. *et al.* Detection of azole resistance in aspergillus fumigatus complex isolates using maldi-tof mass spectrometry. *Clinical Microbiology and Infection* **28**, 260–266 (2022).

[3] Wan, T.-W. *et al.* Performance assessment of the bruker biotyper maldi-tof ms for the identification of difficult-to-identify viridans group streptococci. *Journal of Clinical Microbiology* **61**, e01143–23 (2023).

[4] Weis, C., Cuénod, A., Rieck, B. *et al.* DRIAMS: Database of resistance information on antimicrobials and MALDI-TOF mass spectra (2022). Dryad https://doi.org/10.5061/dryad.bzkh1899q.

[5] Weis, C. *et al.* Direct antimicrobial resistance prediction from clinical MALDI-TOF mass spectra using machine learning. *Nature Medicine* **28**, 164–174, 10.1038/s41591-021-01619-9 (2022).

[6] Lasch, P., Stämmler, M. & Schneider, A. A MALDI-TOF mass spectrometry database for identification and classification of highly pathogenic microorganisms from the Robert Koch-Institute (RKI) (2016). Zenodo https://doi.org/10.5281/zenodo.163517.

[7] Fu, J. *et al.* Rapid AMR prediction in Pseudomonas aeruginosa combining MALDI-TOF MS with DNN model. *Journal of Applied Microbiology* **134**, lxad248, 10.1093/jambio/lxad248 (2023).

[8] Wang, W.-Y., Chiu, C.-F., Tsao, S.-M., Lee, Y.-L. & Chen, Y.-H. Accurate prediction of antimicrobial resistance and genetic marker of Staphylococcus aureus clinical isolates using MALDI-TOF MS and machine learning: Across DRIAMS and Taiwan database. *International Journal of Antimicrobial Agents* **64**, 107329, 10.1016/j.ijantimicag.2024.107329 (2024).

[9] De Waele, G., Menschaert, G. & Waegeman, W. An antimicrobial drug recommender system using MALDI-TOF MS and dual-branch neural networks. *eLife* **13**, RP93242, 10.7554/eLife.RP93242 (2024).

[10] López, X., Manriquez, J., Hernández García, R. & Peralta, D. MSDeepAMR: Antimicrobial resistance prediction based on deep neural networks and transfer learning. *Frontiers in Microbiology* **15**, 1361795, 10.3389/fmicb.2024.1361795 (2024).

[11] Astudillo, C. A., López-Cortés, X. A., Ocque, E. & Manríquez-Troncoso, J. M. Multi-label classification to predict antibiotic resistance from raw clinical MALDI-TOF mass spectrometry data. *Scientific Reports* **14**, 31283, 10.1038/s41598-024-31283-6 (2024).

[12] Macaya Mejias, V. *et al.* Predicting bacterial antibiotic resistance using MALDI-TOF mass spectrometry databases with ELM applications. *Journal of Computer Science & Technology* **24** (2024).

[13] Ren, M., Chen, Q. & Zhang, J. Repurposing MALDI-TOF MS for effective antibiotic resistance screening in Staphylococcus epidermidis using machine learning. *Scientific Reports* **14**, 24139, 10.1038/s41598-024-24139-8 (2024).

[14] Asakura, K. *et al.* Rapid and easy detection of low-level resistance to vancomycin in methicillin-resistant Staphylococcus aureus by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *PLOS ONE* **13**, e0194212, 10.1371/journal.pone.0194212 (2018).

[15] Yu, J. *et al.* Rapid identification of methicillin-resistant Staphylococcus aureus using MALDI-TOF MS and machine learning from over 20,000 clinical isolates. *Microbiology Spectrum* **10**, e00483–22, 10.1128/spectrum.00483-22 (2022).

[16] Wang, Z. *et al.* A risk assessment framework for multidrug-resistant Staphylococcus aureus using machine learning and mass spectrometry technology. *Briefings in Bioinformatics* **24**, bbad330, 10.1093/bib/bbad330 (2023).

[17] Chung, C.-R. *et al.* Data-driven two-stage framework for identification and characterization of different antibiotic-resistant Escherichia coli isolates based on mass spectrometry data. *Microbiology Spectrum* **11**, e03479–22, 10.1128/spectrum.03479-22 (2023).

[18] Yu, H. *et al.* Direct prediction of carbapenem-resistant, carbapenemase-producing, and colistin-resistant Klebsiella pneumoniae isolates from routine MALDI-TOF mass spectra using machine learning and outcome evaluation. *International Journal of Antimicrobial Agents* **61**, 106849, 10.1016/j.ijantimicag.2023.106849 (2023).

[19] Wang, Z. *et al.* Promising potential of machine learning-assisted MALDI-TOF MS as an effective detector for Streptococcus suis serotype 2 and virulence thereof. *Applied and Environmental Microbiology* **89**, e01284–23, 10.1128/aem.01284-23 (2023).

[20] Mao, Q. *et al.* Identification of Escherichia coli strains using MALDI-TOF MS combined with long short-term memory neural networks. *Aging* **16**, 205995, 10.18632/aging.205995 (2024).

[21] Xu, X. Modelling the rapid detection of carbapenemase-resistant Klebsiella pneumoniae based on machine learning and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Diagnostic Microbiology and Infectious Disease* **110**, 116467, 10.1016/j.diagmicrobio.2024.116467 (2024).

[22] Jian, M.-J. *et al.* Artificial intelligence-clinical decision support system in infectious disease control: Combatting multidrug-resistant Klebsiella pneumoniae with machine learning. *Infection and Drug Resistance* 2899–2912, 10.2147/IDR.S460295 (2024).

[23] Nguyen, H.-A. *et al.* Predicting Pseudomonas aeruginosa drug resistance using artificial intelligence and clinical MALDI-TOF mass spectra. *mSystems* **9**, e00789–24, 10.1128/msystems.00789-24 (2024).

[24] Lin, T.-H. *et al.* Artificial intelligence-clinical decision support system for enhanced infectious disease management: Accelerating ceftazidime-avibactam resistance detection in Klebsiella pneumoniae. *Journal of Infection and Public Health* **17**, 102541, 10.1016/j.jiph.2024.102541 (2024).

[25] Godmer, A. *et al.* Contribution of MALDI-TOF mass spectrometry and machine learning including deep learning techniques for the detection of virulence factors of Clostridioides difficile strains. *Microbial Biotechnology* **17**, e14478, 10.1111/1751-7915.14478 (2024).

[26] Huang, L. *et al.* Automated classification of group B Streptococcus into different clonal complexes using MALDI-TOF mass spectrometry. *Frontiers in Molecular Biosciences* **11**, 1355448, 10.3389/fmolb.2024.1355448 (2024).

[27] Jian, M.-J. *et al.* Pioneering Klebsiella pneumoniae antibiotic resistance prediction with artificial intelligence-clinical decision support system–enhanced matrix-assisted laser desorption/ionization time-of-flight mass spectrometry: Retrospective study. *Journal of Medical Internet Research* **26**, e58039, 10.2196/58039 (2024).

[28] Abdrabou, A. *et al.* Discrimination between hypervirulent and non-hypervirulent ribotypes of Clostridioides difficile by MALDI-TOF mass spectrometry and machine learning. *European Journal of Clinical Microbiology & Infectious Diseases* 10.1007/s10096-023-04665-y (2023).

[29] Zeng, Y. *et al.* Machine learning model of imipenem-resistant Klebsiella pneumoniae based on MALDI-TOF-MS platform: An observational study. *Health Science Reports* **6**, e1108, 10.1002/hsr2.1108 (2023).

[30] Iskender, S. *et al.* Rapid Determination of Colistin Resistance in Klebsiella pneumoniae by MALDI-TOF Peak Based Machine Learning Algorithm with MATLAB. *Diagnostic Microbiology and Infectious Disease* 10.1016/j.diagmicrobio.2023.116052 (2023).

[31] Schmidt-Santiago, L. *et al.* Applied machine learning for human bacteria maldi-tof mass spectrometry: a systematic review. *bioRxiv* 2025–01 (2025).

[32] Candela, A. *et al.* Automatic discrimination of species within the Enterobacter cloacae complex using matrix-assisted laser desorption ionization-time of flight mass spectrometry and supervised algorithms. *Journal of Clinical Microbiology* **61**, e01049–22, 10.1128/jcm.01049-22 (2023).

[33] Zintgraff, J. *et al.* Exploring Streptococcus pneumoniae capsular typing through MALDI-TOF mass spectrometry and machine-learning algorithms in Argentina: Identifying prevalent NON PCV13 serotypes alongside PCV13 serotypes. *Journal of Mass Spectrometry and Advances in the Clinical Lab* **30**, 100093, 10.1016/j.jmsacl.2023.11.003 (2023).

[34] Organization, W. H. *WHO bacterial priority pathogens list, 2024: bacterial pathogens of public health importance, to guide research, development, and strategies to prevent and control antimicrobial resistance* (World Health Organization, 2024).

[35] Organization, W. H. *WHO fungal priority pathogens list to guide research, development and public health action* (World Health Organization, 2022).

[36] Gibb, S. acqu file (2016). https://github.com/sgibb/readBrukerFlexData/wiki/acqu-file.

[37] Titulaer, M. K. *et al.* A database application for pre-processing, storage and comparison of mass spectra derived from patients and controls. *BMC Bioinformatics* **7**, 1–16, 10.1186/1471-2105-7-403 (2006).

[38] Schmidt-Santiago, L. *et al.* Marisma: Maldi-tof acquisition repository of infectious species at marañón, 10.5281/zenodo.14218003 (2025).

[39] Guerrero-López, A. *et al.* Automatic antibiotic resistance prediction in klebsiella pneumoniae based on maldi-tof mass spectra. *Engineering Applications of Artificial Intelligence* **118**, 105644 (2023).

[40] Blázquez-Sánchez, M. *et al.* Automatic, web-based typing of clostridioides difficile ribotypes using maldi-tof ms. *bioRxiv* 2024–01 (2024).

[41] Guerrero-López, A. *et al.* Overcoming variability challenges for clostridioides difficile via data augmentation techniques. *bioRxiv* 2024–10 (2024).

[42] Rodriguez Palomo, R. *et al.* Automatic surveillance of escherichia coli bacteriological strains within clinical settings. *bioRxiv* 2024–11 (2024).