Rapid, automatic typing of *Clostridioides difficile* Ribotypes Using MALDI-TOF MS

Mario Blázquez-Sánchez[1,2*], Alejandro Guerrero-López[3*], Ana Candela[1,2#], Albert Belenguer-Llorens[4], José Miguel Moreno[5], Carlos Sevilla-Salcedo[4], María Sánchez-Cueto[1,2], Manuel J. Arroyo[6], Natacha Calama[1,2], Adoración Martín[1], Vanessa Gómez-Verdejo[4], Pablo M. Olmos[2,4], Luis Mancera[6], Patricia Muñoz[1,2,7,8], Mercedes Marín[1,2,7,8], Luis Alcalá[1,2], David Rodríguez-Temporal[1,2∞], Belén Rodríguez-Sánchez[1,2∞] and the AutoCdiff Study Group[†]

[1]Clinical Microbiology and Infectious Diseases Department, Hospital General Universitario Gregorio Marañón, Madrid, Spain. [2]Institute of Health Research Gregorio Marañón (IiSGM), Madrid, Spain. [3]Department of Signals, Systems and Radio communications, Universidad Politécnica de Madrid, Spain. [4]Department of Signal Processing and Communications, Universidad Carlos III de Madrid, Spain. [5]Department of Computer Science and Engineering, Universidad Carlos III de Madrid, Spain. [6]Clover Bioanalytical Software, Av. del Conocimiento, 41, 18016 Granada, Spain. [7]CIBER de Enfermedades Respiratorias (CIBERES CB06/06/0058), Madrid, Spain. [8]Medicine Department, Faculty of Medicine, Universidad Complutense de Madrid, Spain.

Running Title: Rapid *C. difficile* ribotyping with MALDI-TOF MS

[#] **Corresponding author**: Ana Candela, PhD.

Clinical Microbiology and Infectious Diseases Department, Hospital General Universitario Gregorio Marañón, Madrid, Spain. Dr. Esquerdo 46. 28007 Madrid, Spain

Department of Signal Processing and Communications, Universidad Carlos III de Madrid, Spain. Phone: +34- 91- 426 7163, Fax: +34- 91- 426 9595

E-mail: acandelagon@gmail.com

*These authors contributed equally to the article as first author

[∞]Both authors are the senior authors of this article.

ABSTRACT

Clostridioides difficile is a major cause of hospital-acquired diarrhea, posing significant clinical challenges due to its high mortality rates and its involvement in nosocomial outbreaks. Detecting its toxigenic ribotypes (RTs) rapidly and accurately is crucial for effective management and preventing fatal outcomes. This research aimed to create a methodology based on MALDI-TOF MS and Machine Learning (ML) algorithms to differentiate *C. difficile* RTs. MALDI-TOF spectra were acquired from 363 clinical isolates sourcing from 10 Spanish hospitals and analysed using Clover MSDAS and AutoCdiff, an *ad hoc* software developed in this study. Experiments confirmed seven biomarker peaks differentiating RT027 and RT181 from other RTs. Automatic classification tools in Clover MSDAS and AutoCdiff showed up to 100% balanced accuracy, even for isolates from real-time outbreaks. The developed models, available on the AutoCdiff website -https://bacteria.id-, offer researchers a valuable tool for quick RT determination. This approach significantly reduces time, costs, and hands-on time.

INTRODUCTION

Clostridioides difficile, a gram-positive, anaerobic, spore-forming, faecal-oral transmitted rod, is the leading cause of infectious, antibiotic-related, nosocomial diarrhoea (Czepiel, Drozdz et al. 2019; Markovska, Dimitrov et al. 2023). Certain strains of *C. difficile* can produce toxins that are responsible for intestinal damage and play an important role in the pathogenesis and evolution of the disease (Leffler and Lamont 2015; Krutova, Kinross et al. 2018). Several *C. difficile* ribotypes (RTs) have been described, including those that show higher pathogenicity and/or involved in nosocomial outbreaks (Davies, Ashwin et al. 2016; ECDC 2022). Ribotype B1/NAP1/027 (RT027) is of great clinical interest and has been reportedly associated with hospital outbreaks (He, Miyajima et al. 2013; Viprey, Davis et al. 2022). Strains belonging to RT027 are considered as “hypervirulent” due to the hyperproduction of toxins (Markovska, Dimitrov et al. 2023). Other RTs, such as the hypervirulent RT078 and RT181, have recently been reported as highly prevalent (9.0%) and emerging in European countries, respectively (Kachrimanidou, Metallidis et al. 2022; Viprey, Davis et al. 2022).

Rapid and reliable characterization of *C. difficile* is required to tackle transmission and avoid fatal outcomes (Krutova, Kinross et al. 2018). PCR ribotyping, Multiple Locus Variant Analysis and Pulsed-Field Gel Electrophoresis (PFGE) are Europe and North America's most widely implemented genotyping methods, respectively (Calderaro, Buttrini et al. 2022). These methods are considered the gold standard to differentiate *C. difficile* RTs. They are highly informative, but also time-consuming (48-96h respectively from bacterial culture) and labor-intensive, requesting highly trained personnel and specific instruments, both factors lacking in most diagnostic laboratories (Bidet, Lalande, et al. 2000; Abad-Fau, Sevilla et al. 2023). The Xpert® *C. difficile* BT assay has been implemented worldwide for the rapid differentiation of toxigenic *C. difficile* strains (Bai, Hao et al. 2021). This test is fast (50 min from a direct sample) and simple to perform (Whang and Joo 2014). It allows the detection of toxin B (*tcdB*), binary toxin (*cdt*) and the deletion in position 117 of the regulatory gene *tcdC* associated to RT027;

however, the targeted genes do not allow the differentiation of RTs closely related to RT027, such as RT181 or RT176 (Kachrimanidou, Baktash, et al. 2020; Novakova, Kotlebova et al. 2020). Furthermore, the cost per sample is still high (≥ 54 $, approximately 50€) (Chapin, Dickenson et al. 2011).

Matrix-Assisted Laser Desorption/Ionization Time of Flight (MALDI-TOF) Mass Spectrometry (MS), coupled with Machine Learning (ML) algorithms for spectra analysis, holds promise as an efficient method for distinguishing various *C. difficile* RTs rapidly (1h from bacterial culture) (Calderaro, Buttrini et al. 2022), with a cost per isolate less than 1$ (Clark, Kaleta et al. 2013). However, the current state-of-the-art of this approach has notable limitations, lacking a standard procedure for spectra analysis, an accessible code or software, and remaining exploratory in nature.

The objective of this study was to develop an efficient methodology for specific typing of the hypervirulent strains RT027 and RT181 utilizing MALDI-TOF MS and ML algorithms. For this purpose, we analysed two independent tools: Clover Mass Spectrometry Data Analysis Software (MSDAS), a commercial *ad hoc* software for MALDI-TOF MS spectra analysis (https://cloverbiosoft.com) considered state-of-the-art, and AutoCdiff (https://bacteria.id), an open-access software specifically developed by the authors in this study.

RESULTS

Biomarker Peak Analysis

A total of 7 potential biomarker peaks were identified (Table 1) using Clover univariate analysis. Specific peaks for RT027 differentiation were found at 2463 *m/z* (Fig. 1a and 1b) and 4933 *m/z* (Fig. 1c and 1d). A peak at 4993 *m/z* (Fig. 1c and 1e) was absent in RT027 but present in RT181 and other RTs. Finally, peaks at 3353 *m/z* (Fig. 1f and 1g), 6187 *m/z* (Fig. 1h and 1i), 6651 *m/z* (Fig. 1j and 1k) and 6710 *m/z* (Fig. 1j and 1l) were shown to differentiate both RT027 and RT181 from Other RTs. Two of these peaks, at 6651 and 6710 *m/z*, were

correlated with two isoforms of the methionine-cleaved ribosomal protein L28 using the Ribopeaks database (https://lcad.deinfo.uepg.br/~ppgca/ribopeaks/). This was also suggested by the information about protein L28 for *C. difficile* available in the Uniprot database (https://www.uniprot.org/). The difference between these two isoforms is the aminoacidic substitution G9D (Table 2), according to Uniprot database. Overall, these biomarker peaks were used as input for all Clover MSDAS algorithms and as a prior Expert Knowledge EK in the DBLR-FS algorithm of AutoCdiff.

Differentiation of *C. difficile* ribotypes

Experiment 1: RT027 Specific Typing

The aim of this experiment was to differentiate *C. difficile* RT027 isolates from other related RTs classified as "presumptive RT027" by Xpert® *C. difficile* BT assay using the commercial software Clover MSDAS and the AutoCdiff software, developed within this study.

The Clover MSDAS algorithms used as input the peaks at 2463, 4933, 4993, and 6651 *m/z*. The remaining biomarker peaks were excluded since they are shared between RT027 and RT181. Among the four selected peaks, 4933 and 4993 *m/z* were the most different between both categories in terms of peak intensity (Fig. S1a). When Principal Component Analysis (PCA) was performed, spectra of both categories were grouped in different clusters (Fig. S1b). After applying classification algorithms, Light-GBM obtained the highest results with a balanced accuracy of 96.6% and positive predictive values (PPV) of 91.4% and 98.7% for RT027 and "Other RTs", respectively. Algorithms with lower accuracy were RF and KNN with 95.1% and 93.7% of balanced accuracy, respectively, followed by SVM and PLS-DA (Table 3).

In AutoCdiff, the algorithms used full spectra as input, incorporating EK-enhancement for DBLR-FS, as outlined in the methodology. DBLR-FS, both with and without EK-enhancement, achieved a balanced accuracy of 100%, ensuring precision for both classes. While the addition of EK did not enhance performance in this context, it significantly sped up

the training convergence of the DBLR-FS model. Moreover, RF also obtained a 100% balanced accuracy, while FA-VAE and DT exhibited slightly lower accuracies of 99.1% and 97.2%, respectively.

Experiment 2: RT027 and RT181 Specific Typing

We tested the capacity of the described methodology to differentiate RT027 from RT181 (a clinically relevant ribotype identified as "presumptive RT027" by Xpert® *C. difficile* BT) and other clinically relevant RTs or "presumptive RT027" RTs.

The Clover MSDAS algorithms used the 7 biomarkers listed in Table 1. Adding 3 more peaks to those included in Experiment 1, allowed the differentiation of RT181 from the previous categories, maintaining the peaks 4933 and 4993 *m/z*, the most significant ones in terms of intensity (Fig. S2a). After applying a PCA, three distinct clusters were obtained, correlated to the three defined categories -RT027, RT181, "Other RTs"- (Fig. S2b). Regarding ML algorithms, RF yielded a balanced accuracy of 99.7%. The PPV of this model was 98.5% for RT027, 96.2% for RT181, and 97.3% for "Other RTs". The balanced accuracy of RF model was followed by KNN with 96.7% and SVM with 95.6% (Table 3).

In AutoCdiff, the algorithms used full spectra as input, incorporating EK-enhancement for DBLR-FS, as described in the methodology. Along the full spectra, DBLR-FS automatically detected 3 main peaks with high importance for RT differentiation: the presence of 4993 *m/z* for RT027 (Fig. S3a), the absence of 4993 and 6651 *m/z* for RT181 (Fig. S3b) and the presence of 6651 *m/z* combined with the absence of 6710 *m/z* for Other RTs (Fig. S3c). These peaks correlated with those found by implementing the Clover MSDAS Biomarker Analysis. Algorithm DBLR-FS, with EK-enhancement, achieved a balanced accuracy of 99.9%, with PPV of 94.1% for RT027, 100% for RT181, and 100% for Others. The inclusion of EK increased results from 99.5% to 99.9% while significantly speeding up training convergence. Additionally, RF attained a balanced accuracy of 99.9%, while DT and FA-VAE exhibited slightly lower accuracies of 94.5% and 92.0%, respectively (Table 3).

Experiment 3: Application of the algorithms to real-time cases

In this experiment, the available algorithms were re-trained with the full dataset of spectra from our *C. difficile* collection (n=348). Subsequently, they were tested during suspected nosocomial outbreaks at Hospital Central de la Defensa Gómez Ulla -CDGU- (n=12) and Hospital General Universitario Gregorio Marañón -HGUGM- (n=3) in real-time. All algorithms from both software packages consistently classified HCDGU isolates as RT181 and HGUGM isolates as "Other RTs" within 1 hour after culture positivity. PCR-ribotyping results validated the ML classifications 48h later, confirming that the outbreak *C. difficile* isolates from HCDGU belonged to RT181 and those from HGUGM corresponded to RT651.

In summary, the results from the three experiments underscore the efficacy of Machine Learning models in rapidly and accurately categorizing *C. difficile* ribotypes. Besides, all algorithms in AutoCdiff are available at https://bacteria.id for other researchers to test (Fig. S4).

DISCUSSION

In this study, it was demonstrated that the implementation of ML methods for the differentiation of *C. difficile* RTs based on MALDI-TOF MS spectra is a cost-effective, easy-to-apply, and reliable tool with great potential for rapid screening of these isolates, offering a real alternative to complex typing methods such as PCR ribotyping and Pulsed-Field Gel Electrophoresis (PFGE). Its implementation could reduce the turn-around time for definitive characterization of the isolates to 24-48 hours from the reception of the clinical sample in the laboratory. The hands-on time of the method is approximately of 1 hour once the *C. difficile* isolates are grown on solid agar media.

ML-based classification methods were applied to the analysis of protein spectra obtained by MALDI-TOF MS for 1) the rapid differentiation of RT027 from other *C. difficile* isolates classified as "presumptive RT027" by the rapid Xpert® *C. difficile* BT assay and, 2) the rapid discrimination of RT027 and RT181 from RT027-like RTs and other RTs prevalent

in Europe, both on independent test sets and real-time outbreak scenarios. Different ML algorithms and pre-processing pipelines consistently achieved this demonstration of versatility. Firstly, using a variety of algorithms trained and tested within two different packages: Clover MSDAS, a state-of-the art, proprietary software; and secondly, through AutoCdiff, an open-source package specifically developed for differentiation of *C. difficile* RTs within this study.

A total of 4 biomarker peaks (2463, 4933, 4993 and 6651 *m/z*) were found with Clover MSDAS software to differentiate the categories "RT027" and "Other RTs" in Experiment 1, and a total of 7 peaks (2463, 3353, 4933, 4993, 6187, 6651, 6710 *m/z*) when RT027 was differentiated from RT181 and Other RTs in Experiments 2 and 3. Three of these biomarkers correlated with those automatically found by DBLR-FS (4993, 6651, and 6710 *m/z*) for the same purpose. Some of these biomarkers have already been described in previous studies: protein peaks between 6647-6654 *m/z* and between 6707-6712 *m/z* (6651 *m/z* and 6710 *m/z* in this study, respectively) have been reportedly found in RT027 and RT176 (Emele, Joppe et al. 2019; Flores-Trevino, Garza-Gonzalez et al. 2019). These peaks could correspond to two isoforms of the ribosomal protein L28. In this study, the isoform 6710 *m/z* was found in RT027 and RT181, suggesting the great similarity between these two RTs, while the isoform 6651 *m/z* was specifically found in the rest of the analysed RTs, including those classified as “presumptive RT027”. Therefore, the only RT176 isolate included in this study showed the 6651 *m/z* peak, contrary to the results showed by Emele et al., 2019 (Emele, Joppe et al. 2019). Further, *C. difficile* RT176 isolates should be analysed to confirm this discrepancy. Furthermore, the peak at 3353 *m/z* was found to be specific for RT027 and RT181, as well as the peak at 6710 *m/z*, suggesting the possibility that it is the same isoform of the L28 with double charge ions. However, no biomarker peak with double charge was found for the other isoform of the L28 protein in addition to peak 6651 *m/z*.

The peak at 4928 *m/z* (4933 *m/z* in this study) has been described before as an uncharacterized protein present in RT027 (Corver, Sen et al. 2019). Our results correlated with this finding and showed that the 4933 *m/z* peak is specific to RT027 and the 2463 *m/z*

peak, which we suggest may be the same protein with double charge ions. However, the peak at 4993 *m/z*, absent in RT027 and present in the other two remaining categories ("RT181" and "Other RTs") had not been previously described, probably due to the limited literature about the RT181 available so far.

A recent paper using Clover MSDAS reported a balanced accuracy >95.0% when differentiating hypervirulent *C. difficile* RTs (Abdrabou, Sy et al. 2023). However, RT027 and RT176 could not be differentiated, although two separation steps were implemented. Despite not having established a specific category for RT176, this study validates a methodology to differentiate RTs with characteristics like RT176 ("presumptive RT027" by Xpert® *C. difficile* BT and cause of nosocomial outbreaks), as is the case of RT181. Our web application tool serves a dual purpose in this context. It acts as a practical platform for researchers to apply this methodology for their own data analysis and, importantly, functions as a data collection hub. By aggregating user-contributed data, the application becomes a vital resource for continually enhancing and refining our ML models. This evolving database is expected to significantly improve our models' capability to distinguish an expanding array of RTs with similar characteristics, paving the way for more precise and comprehensive typing in the near future.

In this study, Light-GBM and RF, operating within Clover MSDAS, exhibited superior performance in distinguishing RT027 from other "presumptive RT027" isolates. This pattern was consistent across all algorithms, excluding PLS-DA, when differentiating RT027 and RT181 from the rest of the analyzed RTs in Experiment 2. Furthermore, all algorithms achieved 100% accuracy in classifying the outbreak *C. difficile* isolates.

In AutoCdiff, the DBLR-FS with EK and RF algorithms showed the highest accuracy in differentiating RT027 from other "presumptive RT027" isolates. The Bayesian methods in AutoCdiff, like DBLR-FS, do not need cross-validation and provide valuable uncertainty measurements, enhancing their practical utility. In our second experiment, only one isolate, RT173 from the “Other RTs” group, was misclassified as RT027, likely due to the presence of the 4933 m/z peak, commonly associated with RT027. This highlights the need for further

analysis with a larger collection of this RT. Significantly, all algorithms in AutoCdiff correctly classified the isolates in the *C. difficile* outbreak scenarios, underscoring their effectiveness and ease of use in real-world applications.

Therefore, our results highlight the efficacy of the described methodology for rapid screening of *C. difficile* isolates. The implementation of the developed methodology would allow efficient control of the detected cases and early treatment of the affected patients, optimizing hospital resources and representing a substantial improvement in the workflow of clinical microbiology laboratories. In our commitment with the wide diffusion of MALDI-TOF MS-based bacterial typing, we have released the developed classification models for both Clover MSDAS and AutoCdiff, alongside the dataset used to train all algorithms. While access to Clover MSDAS offers a free plan with limited functionality and a paid version for full access, the owner generously offers access to the free full version for three months, and its use demands minimal knowledge about ML. In contrast, all classification algorithms crafted within AutoCdiff are freely accessible at www.bacteria.id. Users simply need to register and upload their MALDI-TOF MS spectra to the website for the AutoCdiff algorithms to classify them into RT027, RT181, or another RT. Besides, the dataset containing the MALDI-TOF spectra from the isolates analyzed in this study has been made accessible (https://doi.org/10.5281/zenodo.10370872).

Although this study represents a step forward in automating MALDI-TOF MS data for bacterial typing, it still has limitations. The first one is the number of isolates analyzed, sourced exclusively from Spain but sourcing from different hospitals of the country. Through the use of the web-based classification models at www.bacteria.id, we expect to contact with other research groups interested in the validation and expansion of the collection of *C. difficile* isolates. Due to the importance of RT027 and RT181, in this study we focussed mainly in these two RTs but we plan to differentiate other important RTs in future collaborative studies.

Besides, the need for a bacterial culture for MALDI-TOF analysis prolongs the detection period for 24-48h until colony growth, whereas the Xpert® *C. difficile* BT can be performed from direct samples in a shorter time. However, the reduction in costs is significant

compared to this technique, and while many microbiology laboratories do not have PCR ribotyping or Xpert® *C. difficile* BT available, MALDI-TOF is ubiquitous to almost all of them.

In conclusion, the integration of MALDI-TOF MS with ML methods presents a promising approach for effectively distinguishing clinically significant *C. difficile* RTs, offering a streamlined and time and effort-effective alternative to traditional molecular methods. A proposal for implementation of this methodology in the clinical laboratory is showed in Figure 2. Notably, the ML methods described in this study showed a high accuracy for the differentiation of RT027 and RT181 from other *C. difficile* RTs and their implementation significantly reduces the time to results from 120 hours (PCR-ribotyping) to 24-48h, providing a rapid screening solution, especially in the case of a suspected outbreak caused by hypervirulent RTs. The availability of the developed methods could empower laboratories with limited molecular resources to enhance their diagnostic capabilities. Finally, the findings underscore the transformative potential of ML applied to MALDI-TOF MS data, showcasing its capacity to revolutionize the landscape of *C. difficile* strain typing in clinical microbiology.

METHODS

Bacterial isolates

A total of 363 clinical isolates (Table 4) belonging to the most prevalent *C. difficile* RTs in Spain and Europe were included in the study (ECDC 2022). A subgroup of 348 isolates were sampled in 10 different hospitals in Spain from faeces of patients with diarrheal symptoms between 2010-2022. The remaining 15 *C. difficile* isolates sourced from two nosocomial outbreaks that took place in Madrid during May-June 2023 in the Hospital Central de la Defensa Gómez Ulla (HCDGU; n=12) and Hospital General Universitario Gregorio Marañón (HGUGM; n=3). All isolates were analysed by Xpert® *C. difficile* BT assay -Cepheid, USA- (Bai, Hao et al. 2021). Subsequently, they were also characterized by PCR-ribotyping and analysis of the amplicons in capillary electrophoresis (Stubbs, Brazier et al. 1999; Indra, Huhulescu et al. 2008; Marin, Martin et al. 2015). Briefly, the intergenic spacer region located between the 16S rRNA and 23S rRNA encoding regions was amplified by PCR. The DNA

fragments were subsequently analysed by capillary electrophoresis. Phylogenetic analysis of the ribotyping profiles was performed using the Bionumerics 5.0 software (bioMérieux, Marcy l'Etoile, France), as described before (Reigadas, Alcala et al. 2018).

Prior to MALDI-TOF MS analysis, the isolates were plated on Brucella Blood Agar (Beckton Dickinson®, Franklin Lakes, NJ, US) and incubated under anaerobic conditions at 37ºC. Incubation lasted 48h to allow sufficient time for *C. difficile* growth. Before analysis, all isolates were re-identified using MALDI-TOF MS technology in an MBT Smart MALDI Biotyper (Bruker Daltonics, Bremen, Germany) with the updated database containing 11,096 mass spectra profiles, applying the standard protocol for rapid, on-plate protein extraction with 100% formic acid followed by HCCA (α-Cyano-4-hydroxycinnamic acid) matrix solution.

MALDI-TOF MS spectra acquisition

For the acquisition of MALDI-TOF MS spectra, a small number of colonies of each *C. difficile* isolate was transferred to the MALDI plate in duplicate, overlaid with 1µl of 100% formic acid, allowed to dry and covered with 1µl of organic HCCA matrix. Two spectra were acquired from each spot on the MALDI plate in the positive ion mode within the 2,000 to 20,000 Da range. Consequently, for each bacterial isolate, a minimum of four spectra were available, consistent with the conditions applied for spectra analysis (Candela, Arroyo et al. 2022). With a subset of 275 isolates, spectra from three consecutive days were collected to test the repeatability of the method. Excluding the outliers and flat lines, each isolate was represented by an average of 5.26 spectra, with a standard deviation of 1.32. This cumulative approach involved a total of 1,879 spectra in the context of this study.

Spectra preprocessing

Each software tool (Clover MSDAS and AutoCdiff) uses a specific preprocessing. Clover MSDAS applied the following pre-processing pipeline: 1) Square root variance stabilization; 2) Smoothing via Savitzky-Golay filter (window length: 11; polynomial order: 3);

3) Baseline subtraction using Top-Hat filter (factor 0.02); 4) Peak alignment with constant mass tolerance of 2 Da and linear mass tolerance of 600 ppm; and 5) TIC normalization. On the other hand, AutoCdiff applied this pre-processing pipeline (Weis, Horn et al. 2020; Weis, Cuénod et al. 2022; Guerrero-López, Sevilla-Salcedo et al. 2023): 1) Square root variance stabilization; 2) Smoothing via Savitzky-Golay filter (window length: 11; polynomial order: 3); 3) Baseline subtraction using Top-Hat filter (factor 0.02); and 4) TIC normalization. Replicates were not merged but treated as unique spectrum to mimic a real-time scenario in both software.

Automatic ML bacteria typing

Different ML algorithms were employed to process MALDI-TOF MS spectra. Two development approaches were adopted: (i) utilizing commercial ad-hoc software tailored for bioanalysis, and (ii) employing open-source algorithms, developed in Python and integrated into a free, user-friendly web application tool created for this study.

The first was Clover MSDAS, a commercial *ad hoc* software developed by Clover Bioanalytical Software (Granada, Spain, https://cloverbiosoft.com) for pre-processing of the protein spectra, biomarker detection and automatic typing with well-known ML algorithms: Random Forest (RF), K-Nearest Neighbor (KNN), Partial Least Squares Discriminant Analysis (PLS-DA), Support Vector Machine (SVM), and Light Gradient Boosting Machine (Light-GBM). In Clover MSDAS, all algorithms underwent training using biomarkers identified by the software, as described in previous studies (Hamidi, Bagheri Nejad et al. 2022; Busby, Doyle et al. 2023; Candela, Arroyo et al. 2023). The use of these biomarkers for differentiation of specific categories is referred to as Expert-Knowledge (EK). The extraction of EK involved conducting a univariate analysis in Clover MSDAS, utilizing peak intensities to assign binary labels for Area Under the Curve (AUC) calculations. Peaks yielding an AUC greater than 0.7 were subsequently selected as potential biomarkers.

On the other hand, AutoCdiff is a novel, open-source tool developed in our study, designed to streamline the process of automatic bacteria typing. This tool has been allocated

in a webpage (https://bacteria.id) to allow MALDI users to efficiently analyse MALDI-TOF MS spectra by simply uploading raw data. The tool features a comprehensive array of pre-built Bayesian ML models, including sophisticated algorithms like the Factor Analysis-Variational AutoEncoder (FA-VAE) (Guerrero-López, Sevilla-Salcedo et al. 2022) and Dual Bayesian Logistic Regression with Feature Selection (DBLR-FS) (Belenguer-Llorens, Sevilla-Salcedo et al. 2022). In addition, AutoCdiff includes foundational models such as Random Forest (RF) and Decision Trees (DT), to meet a variety of analytical needs. In AutoCdiff, algorithms employed the complete spectrum for input, unrestricted to specific biomarkers. Specifically, for DBLR-FS, an additional approach allows the incorporation of EK as input to augment the data to speed up the model convergence. Consequently, biomarkers extracted by Clover MSDAS were utilized as Bayesian priors for the model's weights, thereby enriching the model with EK.

Regarding cross-validation Clover MSDAS utilized fixed parameters: RF with a minimum of 2 isolates per split, 1 leaf, 100 estimators, and sqrt maximum features; SVM with a linear kernel; Light-GBM with 32 leaves, a 0.1 learning rate, and 100 estimators; KNN with 5 neighbors; and PLS with 2 components. In AutoCdiff, a 10-fold cross-validation was applied for DT and RF algorithms, experimenting with various parameters like maximum depth (2, 4, 6, 8), minimum isolates per split (2, 4, 6), minimum leaves (1, 2, 4), and maximum features (auto, sqrt, log2). The FA-VAE and DBLR-FS models, due to their probabilistic nature, bypassed the need for cross-validation, enabling direct estimation of model hyperparameters and providing as output probabilities and uncertainties.

Experiment design

To evaluate the accuracy of the available ML algorithms, three experiments were designed: i) an initial experiment was created to specifically differentiate RT027 among the isolates labelled as "Presumptive RT027" by the Xpert® *C. difficile* BT assay (Figure 3). To facilitate this evaluation, a subset of isolates (n=100) that were categorized as "presumptive RT027" was analysed. These isolates belonged to RT027 (n=42) and other toxigenic RTs

(n=58). The classification models were tested with a set of 31 independent isolates belonging to the same categories (Table 4). ii) The second experiment aims to differentiate the main toxigenic ribotypes of our setting, namely RT027 and RT181, directly from the MALDI-TOF MS spectra obtained. To carry out this evaluation, a new dataset (n=269) was built where the most prevalent RTs in Spain and Europe (ECDC 2022; Viprey, Davis et al. 2022), such as RT001, RT014, RT106, RT207, RT023, RT002, and RT017 were represented (Table 4). Isolates from this experiment were classified into three distinct categories: i) RT027 (n=44), ii) RT181 (n=43) and iii) "Other RTs" (n=182). To validate this model, an independent subset of strains (n=79) from the same 3 categories was used. The distribution of *C. difficile* strains in the training and testing sets is represented in Table 4.iii) Finally, a third experiment was carried out in real-time with isolates from two nosocomial outbreaks to evaluate the efficacy of ML algorithms in a real-world scenario. All algorithms underwent training with the three different categories of the previous experiment using the whole bacterial collection, encompassing 348 *C. difficile* isolates (Table 4). Subsequently, these algorithms were put to the test with isolates sourcing from two outbreak scenarios belonging to HCDGU (n=12) and HGUGM (n=3) in real time. (Fig. 3).

CONFLICTS OF INTEREST

The authors declare no conflict of interests. MJA and LM are employees of Clover Bioanalytical Software, S.L.

FUNDING

This work is partially supported by grants PID2020-115363RB-I00, PID2021-123182OB-I00 and TED2021-132366B-I00 funded by MCIN/AEI/10.13039/501100011033, 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. AC (Rio Hortega CM21-00165), DRT (Sara Borrell CD22-00014) and BRS (Miguel Servet CPII19/00002) are funded by ISCIII. AGL and MBS are the recipients of the Intramural predoctoral contracts 2020 and 2022, respectively, from the Health Research Center of the Hospital Gregorio Marañón -IISGM-.

Data Availability

Aligned with our dedication to the FAIR (Findable, Accessible, Interoperable, and Reusable) principles, we have made the dataset from this study publicly accessible on Zenodo. It can be found at https://doi.org/10.5281/zenodo.10370872.

Code Availability

The code for all models implemented in AutoCdiff, which are open-source, is freely accessible at our GitHub repository: https://github.com/aguerrerolopez/Clostridium. This repository includes comprehensive documentation and usage notes to facilitate replication and further exploration of our models. For additional information or specific inquiries regarding the code, the authors are available to respond to reasonable requests.

Table 1. Biomarker peaks selected for the construction of classification models. Area Under the Curve values are represented in this table.

m/z	**2463**	**3353**	**4933**	**4993**	**6187**	**6651**	**6710**
RT027	0.93	0.85	0.98		0.88		0.96
RT181		0.90		0.98	0.89		0.98
Other RTs				0.77		0.97	

RT: Ribotype.

Table 2. Two isoforms of ribosomal protein L28 in *Clostridioides difficile*, according to the ribotypes (RTs) included in this study. L28-M indicates the mass of the protein excluding the N-terminal methionine.

Ribotypes	**Mass L28 (Da)**	**Mass L28-M (Da)**	**Aminoacidic sequence of Uniprot database**
RT027 and RT181	6837	6710	MAKVCSVCDKGKVSGNQVSHSNKHNKRTWS ANLRSVRAIIDGAPKRVKVCTRCLRSGKIERA

Other RTs	6779	6651	MAKVCSVCGKGKVSGNQVSHSNKHNKRTWS ANLRSVRAIIDGAPKRVKVCTRCLRSGKIERA

Table 3. Balanced accuracy results on test sets for various experiments, comparing the performance of algorithms across Clover Mass Spectrometry Data Analysis Software (MSDAS) and AutoCdiff software packages. The Dual Bayesian Logistic Regression with Feature Selection (DBLR-FS) algorithm was evaluated with and without Expert Knowledge (EK) and all results are compared against the gold standard method for ribotype identification. In addition, time to obtaining results from positive culture is indicated.

Method	Algorithm	Exp 1	Exp 2	Exp 3	Time
PCR ribotyping	Gold standard	100%	100%	100%	48-96h
Clover MSDAS	**RF**	95.1%	**99.7%**	**100%**	1 h
	LGBM	96.6%	95.6%	100%	
	KNN	93.7%	96.7%	100%	
	SVM	92.9%	95.6%	100%	
	PLS-DA	88.4%	32.1%	100%	
AutoCdiff	**DBLR-FS (EK)**	**100%**	**99.9%**	**100%**	1 h
	DBLR-FS	100%	99.5%	100%	
	RF	**100%**	**99.9%**	**100%**	
	DT	97.2%	94.5%	100%	
	FA-VAE	99.1%	92.0%	100%	

Exp: Experiment.

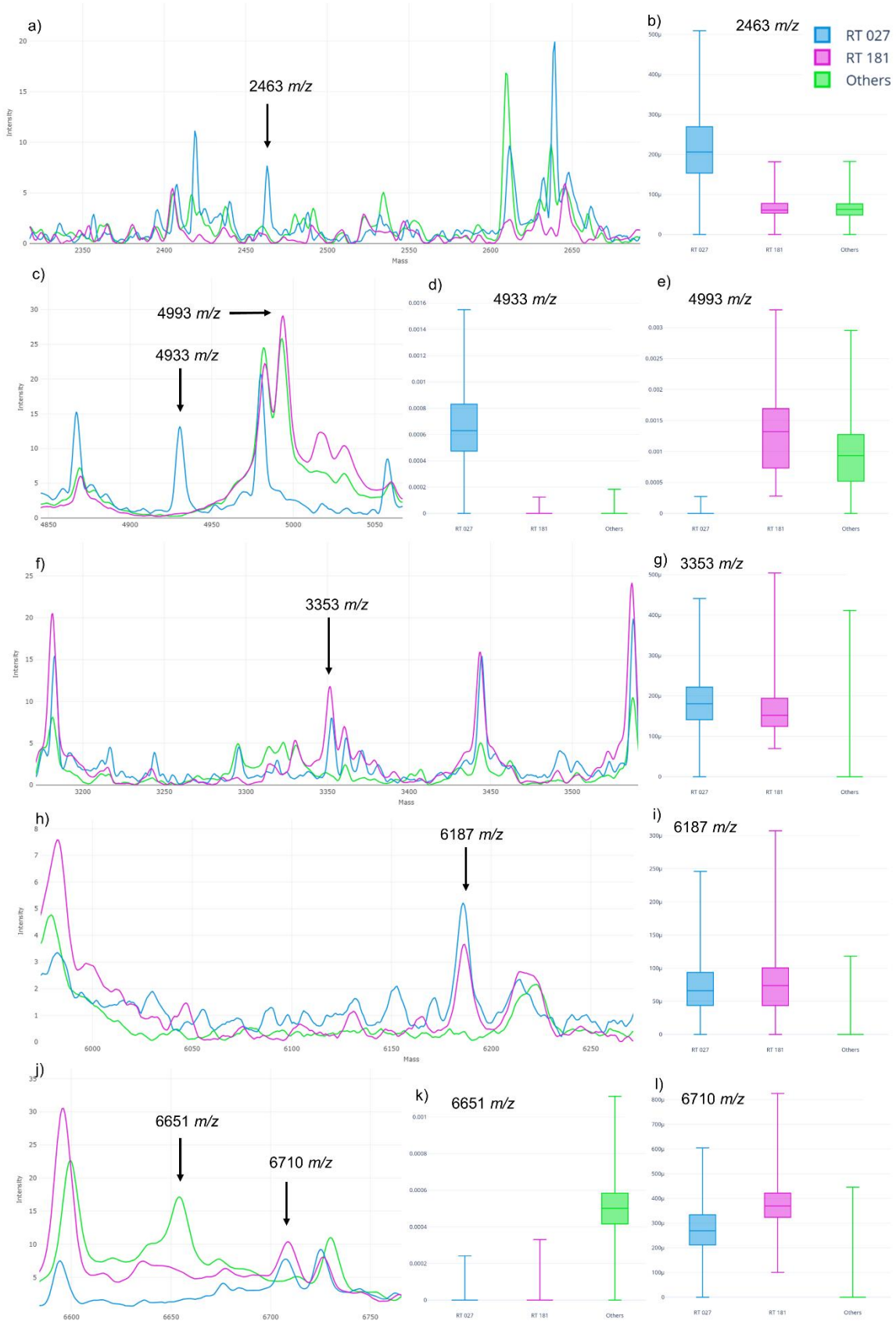
Table 4. List of *C. difficile* isolates, classified according to their respective ribotype (RT) and the classification model in which they were included.

	Total	**Exp 1**		**Exp 2**		**Exp 3**	
		Training	Test	Training	Test	Training	Test
RT181	67	45	10	44	11	55	12
RT027	53	42	11	42	11	53	0
RT106	31	0	0	25	6	31	0
RT001	30	0	0	24	6	30	0
RT017	30	0	0	24	6	30	0
RT078	30	0	0	24	6	30	0
RT207	30	0	0	24	6	30	0
RT014	29	0	0	23	6	29	0
RT023	19	0	0	15	4	19	0
RT002	18	0	0	14	4	18	0
RT165	6	4	2	4	2	6	0
RT166	4	1	3	3	1	4	0
RT651	3	0	0	0	0	0	3
RT250	2	1	1	1	1	2	0
RT170	1	0	1	0	1	1	0
RT171	1	0	1	0	1	1	0
RT173	1	1	0	0	1	1	0

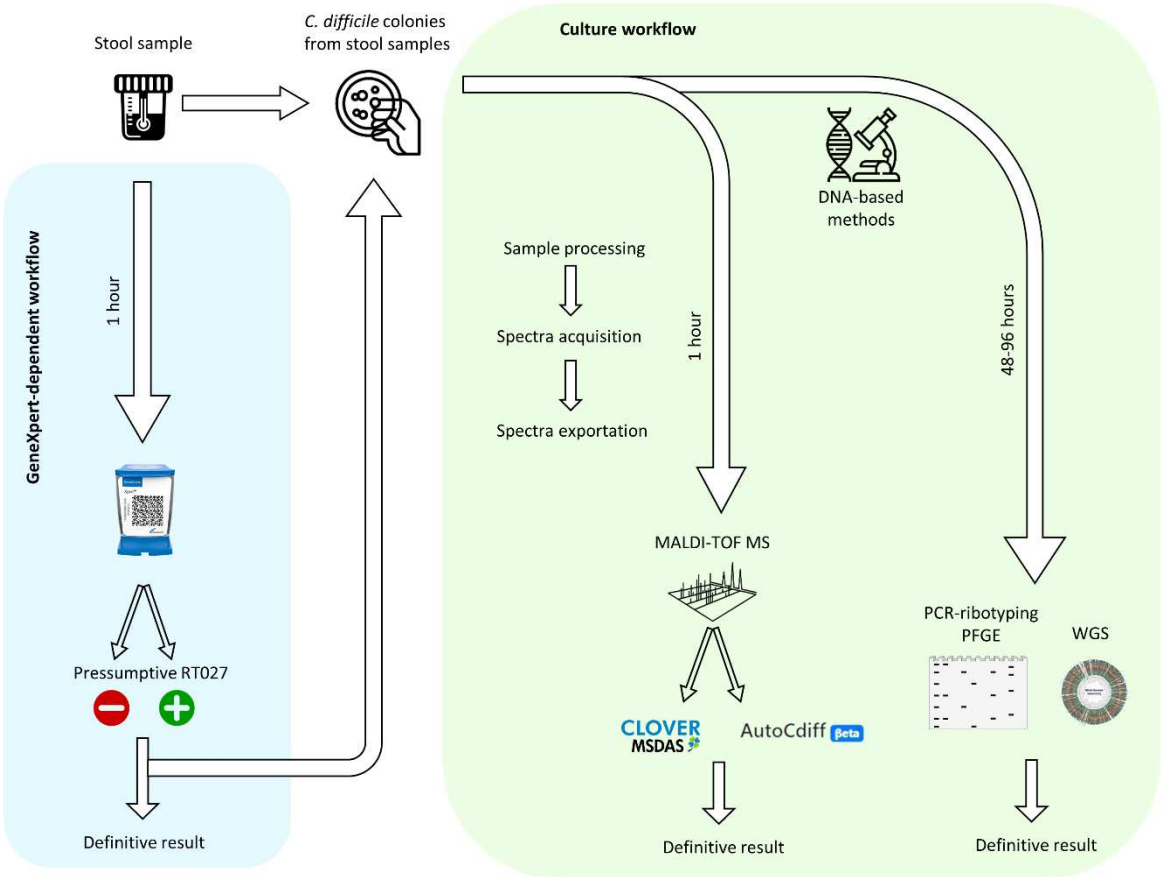
RT174	1	1	0	0	1	1	0
RT175	1	0	1	0	1	1	0
RT176	1	1	0	0	1	1	0
RT187	1	1	0	0	1	1	0
RT578	1	1	0	0	1	1	0
UNK	3	2	1	2	1	3	0
Total	363	100	31	269	79	348	15

Exp: Experiment. UNK: Unknown ribotype.

FIGURE LEGENDS



458 **Figure 1.** Peaks used for creation of predictive models for Clover Mass Spectrometry Data
459 Analysis Software. a) Peak 2463 *m/z*. b) Intensities of 2463 *m/z* according to model categories.
460 c) Peak 3353 *m/z*. d) Intensities of 3353 *m/z*. e) Peaks 4933 and 4993 *m/z*. f) Intensities of
461 4933 *m/z*. g) Intensities of 4993 *m/z*. h) Peak 6187 *m/z*. i) Intensities of 6187 *m/z*. j) Peaks
462 6651 and 6710 *m/z*. k) Intensities of 6651 *m/z*. l) Intensities of 6710 *m/z*.
463
464
465



466
467 **Figure 2.** Proposed laboratory workflow with the implementation of *Clostridioides difficie*
468 typing by MALDI-TOF MS methodology presented in this paper. The workflow shows a
469 GeneXpert-dependent option and a direct culture option for those laboratories with no
470 GeneXpert system available. PFGE: Pulsed-Field Gel Electrophoresis. WGS: Whole-Genome
471 Sequencing.

Figure 3. Description of the experiments developed for the classification of *C. difficile* protein spectra.

REFERENCES

Abad-Fau, A., E. Sevilla, et al. (2023). "Update on Commonly Used Molecular Typing Methods for Clostridioides difficile." Microorganisms **11**(7).

Abdrabou, A. M. M., I. Sy, et al. (2023). "Discrimination between hypervirulent and non-hypervirulent ribotypes of Clostridioides difficile by MALDI-TOF mass spectrometry and machine learning." Eur J Clin Microbiol Infect Dis.

Bai, Y., Y. Hao, et al. (2021). "Evaluation of the Cepheid Xpert C. difficile diagnostic assay: an update meta-analysis." Braz J Microbiol **52**(4): 1937-1949.

Belenguer-Llorens, A., C. Sevilla-Salcedo, et al. (2022). "A Novel Bayesian Linear Regression Model for the Analysis of Neuroimaging Data." Applied Sciences **12**(5).

Bidet, P., V. Lalande, et al. (2000). "Comparison of PCR-ribotyping, arbitrarily primed PCR, and pulsed-field gel electrophoresis for typing Clostridium difficile." Journal of clinical microbiology **38**(7): 2484-2487.

Busby, E. J., R. M. Doyle, et al. (2023). "Evaluation of Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry for Molecular Typing of Acinetobacter baumannii in Comparison with Orthogonal Methods." Microbiology Spectrum **11**(3).

Calderaro, A., M. Buttrini, et al. (2022). "Characterization of Clostridioides difficile Strains from an Outbreak Using MALDI-TOF Mass Spectrometry." Microorganisms **10**(7).

Candela, A., M. J. Arroyo, et al. (2023). "Rapid Discrimination of Pseudomonas aeruginosa ST175 Isolates Involved in a Nosocomial Outbreak Using MALDI-TOF Mass Spectrometry and FTIR Spectroscopy Coupled with Machine Learning." Transboundary and Emerging Diseases **2023**: 1-11.

Candela, A., M. J. Arroyo, et al. (2022). "Rapid and Reproducible MALDI-TOF-Based Method for the Detection of Vancomycin-Resistant Enterococcus faecium Using Classifying Algorithms." Diagnostics (Basel) **12**(2).

Clark, A. E., E. J. Kaleta, et al. (2013). "Matrix-assisted laser desorption ionization-time of flight mass spectrometry: a fundamental shift in the routine practice of clinical microbiology." Clinical microbiology reviews **26**(3): 547-603.

Corver, J., J. Sen, et al. (2019). "Identification and validation of two peptide markers for the recognition of Clostridioides difficile MLST-1 and MLST-11 by MALDI-MS." Clin Microbiol Infect **25**(7): 904 e901-904 e907.

Czepiel, J., M. Drozdz, et al. (2019). "Clostridium difficile infection: review." Eur J Clin Microbiol Infect Dis **38**(7): 1211-1221.

Chapin, K. C., R. A. Dickenson, et al. (2011). "Comparison of five assays for detection of Clostridium difficile toxin." The Journal of molecular diagnostics : JMD **13**(4): 395-400.

Davies, K. A., H. Ashwin, et al. (2016). "Diversity of Clostridium difficile PCR ribotypes in Europe: results from the European, multicentre, prospective, biannual, point-prevalence study of Clostridium difficile infection in hospitalised patients with diarrhoea (EUCLID), 2012 and 2013." Euro Surveill **21**(29).

ECDC (2022). "European Centre for Disease Prevention and Control. Clostridioides (Clostridium) difficile infections. Annual epidemiological report for 2016–2017." ECDC.

Emele, M. F., F. M. Joppe, et al. (2019). "Proteotyping of Clostridioides difficile as Alternate Typing Method to Ribotyping Is Able to Distinguish the Ribotypes RT027 and RT176 From Other Ribotypes." Frontiers in microbiology **10**: 2087.

Flores-Trevino, S., E. Garza-Gonzalez, et al. (2019). "Screening of biomarkers of drug resistance or virulence in ESCAPE pathogens by MALDI-TOF mass spectrometry." Scientific reports **9**(1): 18945.

Guerrero-López, A., C. Sevilla-Salcedo, et al. (2023). "Automatic antibiotic resistance prediction in Klebsiella pneumoniae based on MALDI-TOF mass spectra." Engineering Applications of Artificial Intelligence **118**.

Guerrero-López, A., C. Sevilla-Salcedo, et al. (2022). "Multimodal hierarchical Variational AutoEncoders with Factor Analysis latent space." arXiv.

Hamidi, H., R. Bagheri Nejad, et al. (2022). "A Combination of MALDI-TOF MS Proteomics and Species-Unique Biomarkers' Discovery for Rapid Screening of Brucellosis." Journal of the American Society for Mass Spectrometry **33**(8): 1530-1540.

He, M., F. Miyajima, et al. (2013). "Emergence and global spread of epidemic healthcare-associated Clostridium difficile." Nat Genet **45**(1): 109-113.

Indra, A., S. Huhulescu, et al. (2008). "Characterization of Clostridium difficile isolates using capillary gel electrophoresis-based PCR ribotyping." Journal of Medical Microbiology **57**(11): 1377-1382.

Kachrimanidou, M., A. Baktash, et al. (2020). "An outbreak of Clostridioides difficile infections due to a 027-like PCR ribotype 181 in a rehabilitation centre: Epidemiological and microbiological characteristics." Anaerobe **65**: 102252.

Kachrimanidou, M., S. Metallidis, et al. (2022). "Predominance of Clostridioides difficile PCR ribotype 181 in northern Greece, 2016-2019." Anaerobe **76**: 102601.

Krutova, M., P. Kinross, et al. (2018). "How to: Surveillance of Clostridium difficile infections." Clin Microbiol Infect **24**(5): 469-475.

Leffler, D. A. and J. T. Lamont (2015). "Clostridium difficile infection." N Engl J Med **372**(16): 1539-1548.

Marin, M., A. Martin, et al. (2015). "Clostridium difficile isolates with high linezolid MICs harbor the multiresistance gene cfr." Antimicrob Agents Chemother **59**(1): 586-589.

Markovska, R., G. Dimitrov, et al. (2023). "Clostridioides difficile, a New "Superbug"." Microorganisms **11**(4).

Novakova, E., N. Kotlebova, et al. (2020). "An Outbreak of Clostridium (Clostridioides) difficile Infections within an Acute and Long-Term Care Wards Due to Moxifloxacin-Resistant PCR Ribotype 176 Genotyped as PCR Ribotype 027 by a Commercial Assay." J Clin Med **9**(11).

Reigadas, E., L. Alcala, et al. (2018). "Breakthrough Clostridium difficile Infection in Cirrhotic Patients Receiving Rifaximin." Clin Infect Dis **66**(7): 1086-1091.

Stubbs, S. L., J. S. Brazier, et al. (1999). "PCR targeted to the 16S-23S rRNA gene intergenic spacer region of Clostridium difficile and construction of a library consisting of 116 different PCR ribotypes." J Clin Microbiol **37**(2): 461-463.

Viprey, V. F., G. L. Davis, et al. (2022). "A point-prevalence study on community and inpatient Clostridioides difficile infections (CDI): results from Combatting Bacterial Resistance in Europe CDI (COMBACTE-CDI), July to November 2018." Euro Surveill **27**(26).

Weis, C., A. Cuénod, et al. (2022). "Direct antimicrobial resistance prediction from clinical MALDI-TOF mass spectra using machine learning." Nature Medicine **28**(1): 164-174.

Weis, C., M. Horn, et al. (2020). "Topological and kernel-based microbial phenotype prediction from MALDI-TOF mass spectra." Bioinformatics **36**(Suppl_1): i30-i38.

Whang, D. H. and S. Y. Joo (2014). "Evaluation of the diagnostic performance of the xpert Clostridium difficile assay and its comparison with the toxin A/B enzyme-linked fluorescent assay and in-house real-time PCR assay used for the detection of toxigenic C. difficile." Journal of clinical laboratory analysis **28**(2): 124-129.