# Automatic, web-based 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], Mark Gutiérrez-Pareja[1,2], 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[1,2.7], 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:** Web-based typing of *C. difficile* by MALDI-TOF MS

[#]**Corresponding author**: Ana Candela, PhD. (https://orcid.org/0000-0003-0009-7116)

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

Phone: +34-91- 426 9595, Fax: +34-91- 426 7163

E-mail: acandelagon@gmail.com

*These authors contributed equally to the article as first authors

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

[†]The AutoCdiff Study Group: Mariela Martínez[9], Cristina Verdú[10], Juan Soliveri[10], Juan Cuadros[11], Felipe Pérez[11], Alberto Nieto[12]; Luís Moisés Ruiz[13], María Simón[14], Joan López[15], Óscar Cuevas[16], Carmen Ezpeleta[17], Soledad Illescas[18].

[9]Hospital Universitario de Guadalajara, Spain; [10]Microbiology Department, Universidad de Alcalá de Henares, Spain; [11]Hospital Universitario Príncipe de Asturias, Alcalá de Henares, Spain; [12]Hospital Universitario Puerta de Hierro, Majadahonda, Spain; [13]Hospital Central de la Cruz Roja, Madrid, Spain; [14]Hospital Central de la Defensa Gómez Ulla, Madrid, Spain; [15]Hospital Sant Joan de Déu, Manresa, Spain; [16]Hospital Universitario Severo Ochoa, Getafe, Spain; [17]Hospital Universitario de Navarra, Pamplona, Spain; Hospital General Universitario de Ciudad Real, Ciudad Real, Spain.

## ABSTRACT

*Clostridioides difficile* is a major cause of hospital-acquired diarrhea, posing significant clinical challenges due to its high morbi-mortality rates and its involvement in nosocomial outbreaks. Detecting its toxigenic ribotypes rapidly and accurately is crucial for effective outbreak control. This study aimed to create a rapid diagnostic methodology based on MALDI-TOF MS and Machine Learning algorithms to differentiate toxigenic *C. difficile* RTs.

MALDI-TOF spectra were acquired from 379 clinical isolates sourcing from 10 Spanish hospitals and analysed using Clover MSDAS, a specific software for MALDI-TOF spectra analysis, considered as the state-of-the-art tool for this purpose, and AutoCdiff, an *ad hoc* software developed in this study.

Seven biomarker peaks were found to differentiate epidemic RT027 and RT181 strains from other RTs (2463, 3353, 4933, 4993, 6187, 6651 and 6710 *m/z*). Two peaks (2463 and 4993 *m/z*) were specifically found in RT027 isolates while combinations of the other 5 peaks allowed the differentiation of RT181 from other ribotypes. Automatic classification tools developed in Clover MSDAS and AutoCdiff using the specific peaks and the entire protein spectra, respectively, showed up to 100% balanced accuracy. Both methods allowed correct ribotype assignment for isolates sourcing from real-time outbreaks.

The developed models, available from Clover MSDAS and the AutoCdiff website -https://bacteria.id- offer researchers a valuable tool for quick *C. difficile* ribotype determination based on MALDI-TOF spectra analysis. Although further validation of the models is still required, they represent rapid and cost-effective methods for standardized *C. difficile* ribotype assignment.

**Keywords:** MALDI-TOF Mass Spectrometry, *Clostridioides difficile*, ribotype, nosocomial outbreak

## INTRODUCTION

*Clostridioides difficile*, a Gram-positive, anaerobic, spore-forming, faecal-oral transmitted rod, is the leading cause of infectious, antibiotic-related, nosocomial and community-acquired diarrhoea (1, 2). Ribotype 027 (RT027) is of great clinical interest and has been reportedly associated with hospital outbreaks (3, 4). Strains belonging to RT027 have been considered as "hypervirulent" due to the hyperproduction of toxins (2). Other ribotypes (RT), such as the hypervirulent RT078 and RT181, have recently been reported as highly prevalent (9.0%) and emerging in European countries, respectively (4, 5).

The Xpert® *C. difficile* BT assay has been implemented worldwide for the rapid diagnostic of toxigenic *C. difficile* strains from faeces (6). This test is fast -50 min from a direct sample- and simple to perform (7). It allows the detection of toxin B (*tcdB*), binary toxin (*cdtA*) and the deletion in position 117 of the regulatory gene *tcdC* associated to RT027, but present in other *C. difficile* RTs. Therefore, the targeted genes do not allow the differentiation of RTs closely related to RT027, such as RT181 or RT176, belonging to clade 2 ST1. The Xpert® *C. difficile* BT assay classifies these RTs as "presumptive RT027" (8, 9).

Matrix-Assisted Laser Desorption/Ionization Time of Flight (MALDI-TOF) Mass Spectrometry (MS), coupled with Machine Learning (ML) algorithms for spectra analysis, represents an efficient alternative to DNA-based methods -Pulse Field Gel Electrophoresis and PCR ribotyping- for distinguishing *C. difficile* RTs rapidly -1h from bacterial culture- (10), with a cost per isolate below 1$ (11). However, the current state-of-the-art for this approach has notable limitations, lacking a standard procedure for spectra analysis and an accessible code or software.

The goal of this study was to develop an efficient methodology for specific typing of the hypervirulent *C. difficile* strains RT027 and RT181 utilizing MALDI-TOF MS and ML algorithms.

## METHODS

### Bacterial isolates

A total of 379 clinical isolates (Table 1) belonging to the most prevalent *C. difficile* RTs in Spain and Europe were included in this study (12). A subgroup of 348 isolates were obtained from faeces of patients with diarrheal symptoms sourcing from 10 Spanish hospitals. The samples were submitted to Hospital General Universitario Gregorio Marañón -HGUGM- for further characterization of the *C. difficile* RTs. The remaining 31 *C. difficile* isolates sourced from two suspected nosocomial outbreaks detected in Madrid during May-June 2023 in the Hospital Central de la Defensa Gómez Ulla -HCDGU- (n=28) and HGUGM (n=3). All isolates were analyzed by Xpert® *C. difficile* BT assay -Cepheid, USA- (6). Subsequently, they were characterized by PCR-ribotyping (13-15). Briefly, the intergenic spacer region located between the 16S rRNA and 23S rRNA encoding regions was amplified by PCR. The DNA fragments were analysed by capillary electrophoresis. Phylogenetic analysis of the ribotyping profiles was performed using the Bionumerics 7.2 software (bioMérieux, Marcy l'Etoile, France) and WebRibo (https://webribo.ages.at/) (13).

Prior to MALDI-TOF MS analysis, samples were plated on Brucella Blood Agar (Beckton Dickinson®, Franklin Lakes, NJ, US) and incubated under anaerobic conditions at 37ºC during 48h. Before analysis, all isolates were re-identified using MALDI-TOF MS in an MBT Smart MALDI Biotyper (Bruker Daltonics, Bremen, Germany).

### 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 duplicates, overlaid with 1µl of 100% formic acid, allowed to dry and covered with 1µl of organic HCCA (α-Cyano-4-hydroxycinnamic acid) matrix. Two spectra were acquired from each spot in the positive ion

mode within the 2,000 to 20,000 Da range. Consequently, for each bacterial isolate, four spectra were available, consistent with the conditions applied for spectra analysis (16).

**Table 1.** List of *C. difficile* isolates, classified according to their respective ribotype (RT) and the classification model in which they were included.

| | Categories | Experiment 1 | | Experiment 2 | | Total |
|---|---|---|---|---|---|---|
| | | Training | Test | Training | Test | |
| **RT181** | **RT181** | 44 | 11 | 55 | 28 | 83 |
| **RT027** | **RT027** | 42 | 11 | 53 | 0 | 53 |
| **RT106** | **Other RTs** | 25 | 6 | 31 | 0 | 31 |
| **RT001** | | 24 | 6 | 30 | 0 | 30 |
| **RT017** | | 24 | 6 | 30 | 0 | 30 |
| **RT078** | | 24 | 6 | 30 | 0 | 30 |
| **RT207** | | 24 | 6 | 30 | 0 | 30 |
| **RT014** | | 23 | 6 | 29 | 0 | 29 |
| **RT023** | | 15 | 4 | 19 | 0 | 19 |
| **RT002** | | 14 | 4 | 18 | 0 | 18 |
| **RT651** | | 0 | 0 | 0 | 3 | 3 |
| **RT250** | | 1 | 1 | 2 | 0 | 2 |
| **RT578** | | 0 | 1 | 1 | 0 | 1 |
| **HGUGM RTs*** | | | | | | |
| **165** | **Other RTs** | 4 | 2 | 6 | 0 | 6 |
| **166** | | 3 | 1 | 4 | 0 | 4 |
| **170** | | 0 | 1 | 1 | 0 | 1 |
| **171** | | 0 | 1 | 1 | 0 | 1 |
| **173** | | 0 | 1 | 1 | 0 | 1 |
| **174** | | 0 | 1 | 1 | 0 | 1 |
| **175** | | 0 | 1 | 1 | 0 | 1 |
| **176** | | 0 | 1 | 1 | 0 | 1 |
| **187** | | 0 | 1 | 1 | 0 | 1 |
| **UNK** | | 2 | 1 | 3 | 0 | 3 |
| **Total** | | 269 | 79 | 348 | 31 | 379 |

UNK: Unknown ribotype. *These RTs were analysed by MLVA and classified as uncharacterized RTs (unpublished data). They were assigned correlative numbers that do not correlate with those described in the literature. Thus, ribotype HGUGM 176 does not correspond to the RT176 described by Novakova et al. 2020. All were classified as "presumptive RT027" by the Xpert® *C. difficile* BT assay.

## Spectra 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: 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 (17-19). Replicates were not merged but treated as unique spectra in both software packages, to mimic a real-time scenario.

## Automatic ML bacteria typing

Clover MSDAS, a commercial *ad hoc* software developed by Clover Bioanalytical Software (Granada, Spain, https://cloverbiosoft.com) was applied for biomarker detection and automatic typing with well-known ML algorithms: Random Forest (RF), K-Nearest Neighbor (KNN), Support Vector Machine (SVM) and Light Gradient Boosting Machine (Light-GBM). In Clover MSDAS, all algorithms underwent training using biomarker peaks identified by the software, as described in previous studies (20-22). 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 this study, designed to streamline the process of automatic bacteria typing. It features pre-built Bayesian

ML models, including algorithms like the Factor Analysis-Variational AutoEncoder (FA-VAE) (23) and Dual Bayesian Logistic Regression with Feature Selection (DBLR-FS) (24), as well as foundational models such as RF and Decision Trees (DT). In AutoCdiff, algorithms use the complete spectrum for input, unrestricted to specific biomarkers. An additional approach for DBLR-FS allows the incorporation of EK as input 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. AutoCdiff has been allocated in a webpage (https://bacteria.id) to allow MALDI users to efficiently analyze MALDI-TOF MS spectra by simply uploading raw data.

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 and KNN with 5 neighbors. 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-g, 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, two experiments were designed: i) to differentiate the main toxigenic RTs in the study collection, namely RT027 and RT181, directly from the MALDI-TOF MS spectra. To carry out this evaluation, a dataset (n=269) was built with *C. difficile* isolates from the most prevalent RTs in Spain and Europe (4, 12), such as RT001, RT014, RT106, RT207, RT023, RT002, and RT017 (Table 1). Isolates

from this experiment were classified into three distinct categories: RT027 (n=44), RT181 (n=43) and "Other RTs" (n=182). To validate this model, an independent subset of strains (n=79) from the same 3 categories was used; ii) Finally, the accuracy of the ML algorithms was evaluated in a real-world scenario with isolates from two nosocomial outbreaks. All algorithms underwent training using the whole bacterial collection, encompassing 348 *C. difficile* isolates (Table 1). Subsequently, these algorithms were tested with isolates sourcing from two outbreaks detected in HCDGU (n=28) and HGUGM (n=3) in real time (Figure S1).

**Ethics Statement**

This study was approved by the University Hospital Gregorio Marañon Ethics Committee (CEIm) -Study number MICRO.HGUGM.2021-025-. The Committee considered that all the conditions for waiving informed consent were met since the study was conducted with microbiological samples and not with human products.

## RESULTS

**Biomarker Peak Analysis**

A total of 7 potential biomarker peaks were identified (Table 2) using Clover univariate analysis. Specific peaks for RT027 differentiation were found at 2463 and 4933 *m/z* (Figure 1). A peak at 4993 *m/z* was absent in RT027 but present in RT181 and other RTs while peaks at 3353, 6187, 6651 and 6710 *m/z* were shown to differentiate both RT027 and RT181 from Other RTs (Figure 1). Three of these peaks, at 3353 -the same peak as 6710 *m/z* with double charge-, 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 according to Uniprot database (Table 3). Overall, these biomarker peaks were used as input for all Clover MSDAS algorithms and as a prior EK in the DBLR-FS algorithm of AutoCdiff.

**Table 2.** 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 3.** 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 | MAKVCSVC**D**KGKVSGNQVSHSNKHNKRTWSA NLRSVRAIIDGAPKRVKVCTRCLRSGKIERA |
| **Other RTs** | 6779 | 6651 | MAKVCSVC**G**KGKVSGNQVSHSNKHNKRTWSA NLRSVRAIIDGAPKRVKVCTRCLRSGKIERA |

## Differentiation of *C. difficile* ribotypes

Experiment 1: 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 "presumptive RT027" RTs.

Clover MSDAS algorithms used as input the 7 biomarkers listed in Table 2. Peaks at 4933 and 4993 *m/z* were the most significant ones in terms of intensity (Figure S2a). After applying PCA, three distinct clusters were obtained, correlated to the three defined categories

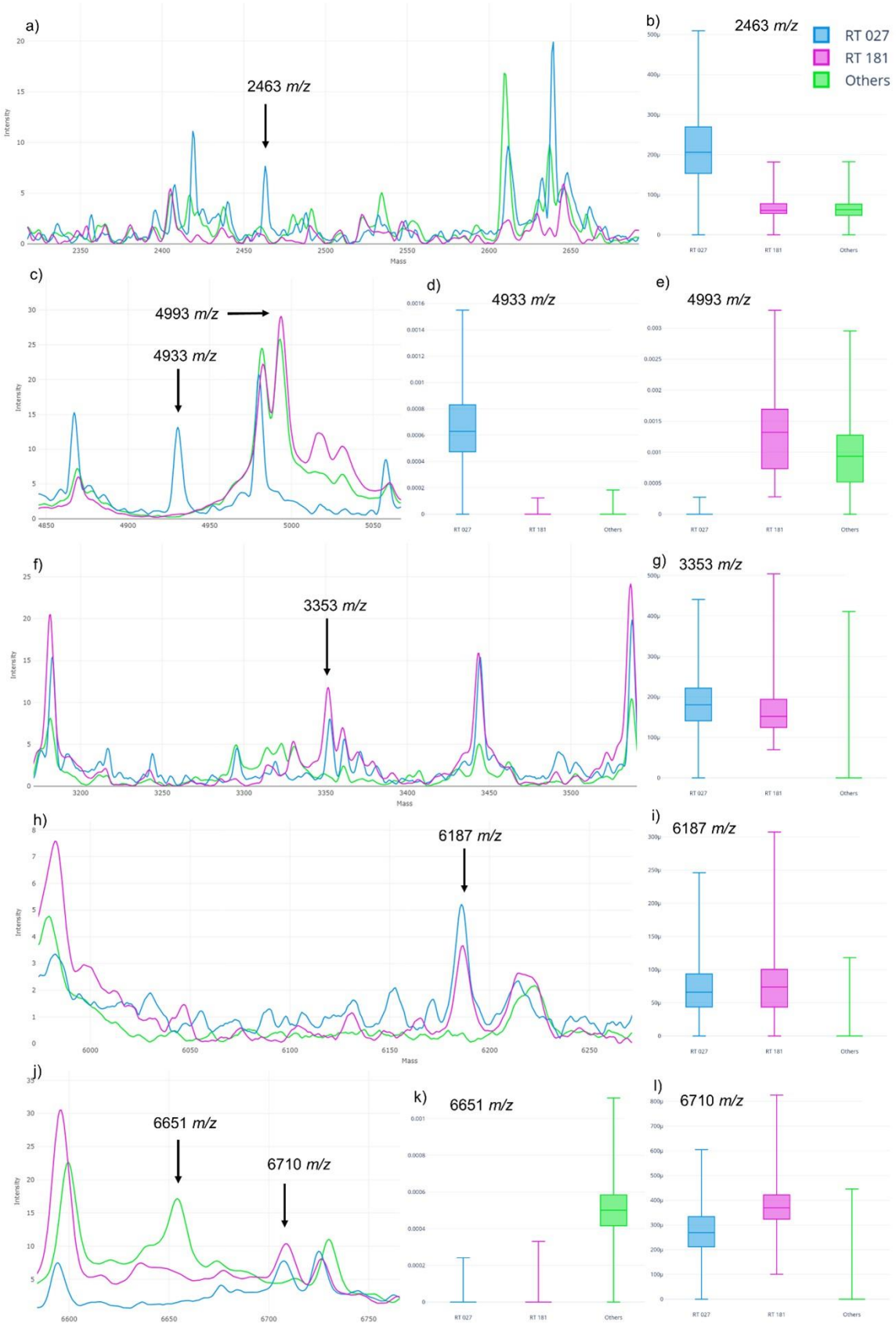
-RT027, RT181, “Other RTs”- (Figure S2b). Regarding ML algorithms, RF yielded a balanced accuracy of 99.7%. The PPV of this model was 98.5% for RT027, 96.2% for RT181, and 97.3% for “Other RTs”. The balanced accuracy of RF model was followed by KNN with 96.7% and SVM with 95.6% (Table 4).

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 4933 *m/z* for RT027 (Figure S3a), the absence of 4933 and 6651 *m/z* for RT181 (Figure S3b) and the presence of 6651 *m/z* combined with the absence of 6710 *m/z* for Other RTs (Figure 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. 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.The inclusion of EK increased results from 99.5% to 99.9% while significantly speeding up training convergence.

**Table 4.** 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 |
| | **L-GBM** | **96.6%** | 95.6% | 100% | |
| | KNN | 93.7% | 96.7% | 100% | |
| | SVM | 92.9% | 95.6% | 100% | |
| AutoCdiff | **DBLR-FS (EK)** | **100%** | **99.9%** | **100%** | |
| | DBLR-FS | 100% | 99.5% | 100% | 1 h |
| | **RF** | **100%** | **99.9%** | **100%** | |
| | DT | 97.2% | 94.5% | 100% | |
| | FA-VAE | 99.1% | 92.0% | 100% | |

Exp: Experiment.


a)
2463 m/z
Intensity
Mass
b)
2463 m/z
RT 027
RT 181
Others
c)
4993 m/z
4933 m/z
d)
4933 m/z
e)
4993 m/z
f)
3353 m/z
g)
3353 m/z
h)
6187 m/z
i)
6187 m/z
j)
6651 m/z
6710 m/z
k)
6651 m/z
l)
6710 m/z

265
266

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

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 4).

Experiment 2: Application of the algorithms to real-time cases

The available algorithms were re-trained with the full dataset of spectra from our *C. difficile* collection (n=348) and tested during suspected nosocomial outbreaks at HCDGU (n=28) and HGUGM (n=3) in real-time. A proposal for implementation of this methodology in the clinical laboratory is showed in Figure 2.

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 according to WebRibo database.

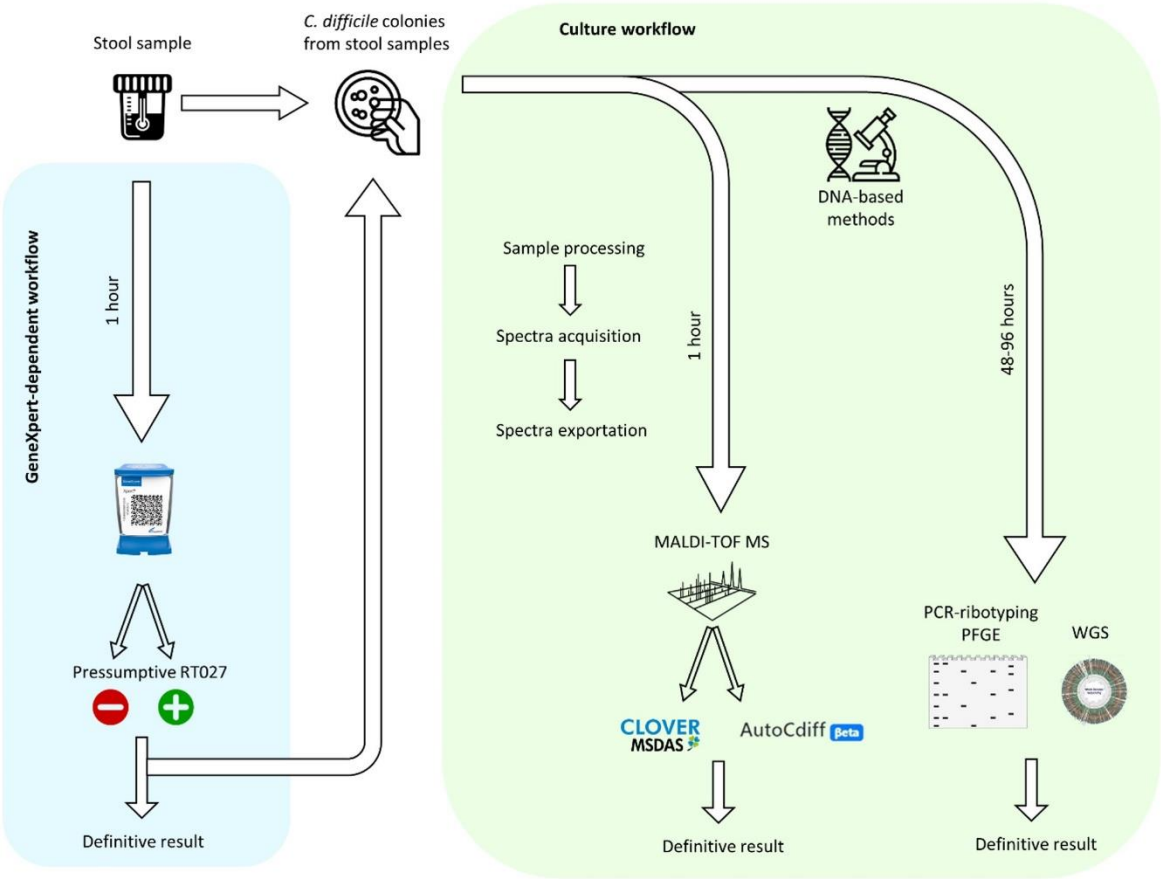



**Figure 2.** Proposed laboratory workflow with the implementation of *Clostridioides difficile* typing by MALDI-TOF MS methodology presented in this paper. The workflow shows a GeneXpert-dependent option and a direct culture option for those laboratories with no GeneXpert system available. PFGE: Pulsed-Field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

## 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 PFGE. Its implementation could reduce the turn-around time for definitive characterization of the isolates to 48-72h 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.

A total of 7 biomarker peaks (2463, 3353, 4933, 4993, 6187, 6651, 6710 *m/z*) where found to differentiate RT027 from RT181 and Other RTs. Three of these biomarkers correlated with those automatically found by DBLR-FS (4993, 6651 and 6710 *m/z*) for the same purpose. Some of them 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 (25, 26). 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 detected in the rest of the analyzed RTs, including those classified as "presumptive RT027". 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 6651 *m/z* isoform of the L28 protein.

The peak at 4928 *m/z* (4933 *m/z* in this study) has been described before as an uncharacterized protein present in RT027 (27). 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. 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 applying Clover MSDAS reported a balanced accuracy >95.0% when differentiating hypervirulent *C. difficile* RTs (28). However, RT027 and RT176 could not be differentiated. 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.

In this study, the algorithms Light-GBM and RF, available in Clover MSDAS, exhibited superior performance in distinguishing RT027 from other "presumptive RT027" isolates. This pattern was consistent across all algorithms, when differentiating RT027 and RT181 from the rest of the analyzed RTs. 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. Therefore, the developed methodology has proven to be very useful to characterize C. difficile ribotypes in real time. It requires very limited expertise and means and favors the rapid implementation of control measures.

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, sourcing exclusively from Spain although from geographically distant hospitals. This limitation is expected to be overcome soon by the collaborations established with different research groups who have started to test the web-based classification models at www.bacteria.id (Figure S4). Besides, the need for a bacterial culture for MALDI-TOF MS 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.

In conclusion, the integration of MALDI-TOF MS with ML methods presents a promising approach to effectively distinguish clinically significant *C. difficile* RTs, offering a time- and effort-effective screening solution, especially when an outbreak caused by

hypervirulent *C. difficile* RTs is suspected. The availability of the developed methods could empower laboratories with limited molecular resources to enhance their diagnostic capabilities.

**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.

## CONFLICTS OF INTEREST

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

## REFERENCES


1. Czepiel J, Drozdz M, Pituch H, Kuijper EJ, Perucki W, Mielimonka A, Goldman S, Wultanska D, Garlicki A, Biesiada G. 2019. Clostridium difficile infection: review. Eur J Clin Microbiol Infect Dis 38:1211-1221.
2. Markovska R, Dimitrov G, Gergova R, Boyanova L. 2023. Clostridioides difficile, a New "Superbug". Microorganisms 11.
3. He M, Miyajima F, Roberts P, Ellison L, Pickard DJ, Martin MJ, Connor TR, Harris SR, Fairley D, Bamford KB, D'Arc S, Brazier J, Brown D, Coia JE, Douce G, Gerding D, Kim HJ, Koh TH, Kato H, Senoh M, Louie T, Michell S, Butt E, Peacock SJ, Brown NM, Riley T, Songer G, Wilcox M, Pirmohamed M, Kuijper E, Hawkey P, Wren BW, Dougan G, Parkhill J, Lawley TD. 2013. Emergence and global spread of epidemic healthcare-associated Clostridium difficile. Nat Genet 45:109-13.
4. Viprey VF, Davis GL, Benson AD, Ewin D, Spittal W, Vernon JJ, Rupnik M, Banz A, Allantaz F, Cleuziat P, Coordinators C-CN, Wilcox MH, Davies KA, consortium C-C, Members of the C-CDINc. 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.
5. Kachrimanidou M, Metallidis S, Tsachouridou O, Harmanus C, Lola V, Protonotariou E, Skoura L, Kuijper E. 2022. Predominance of Clostridioides difficile PCR ribotype 181 in northern Greece, 2016-2019. Anaerobe 76:102601.
6. Bai Y, Hao Y, Song Z, Chu W, Jin Y, Wang Y. 2021. Evaluation of the Cepheid Xpert C. difficile diagnostic assay: an update meta-analysis. Braz J Microbiol 52:1937-1949.
7. Whang DH, Joo SY. 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 28:124-9.
8. Novakova E, Kotlebova N, Gryndlerova A, Novak M, Vladarova M, Wilcox M, Kuijper E, Krutova M. 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.
9. Kachrimanidou M, Baktash A, Metallidis S, Tsachouridou O, Netsika F, Dimoglou D, Kassomenaki A, Mouza E, Haritonidou M, Kuijper E. 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.

10. Calderaro A, Buttrini M, Farina B, Montecchini S, Martinelli M, Arcangeletti MC, Chezzi C, De Conto F. 2022. Characterization of Clostridioides difficile Strains from an Outbreak Using MALDI-TOF Mass Spectrometry. Microorganisms 10.
11. Clark AE, Kaleta EJ, Arora A, Wolk DM. 2013. Matrix-assisted laser desorption ionization-time of flight mass spectrometry: a fundamental shift in the routine practice of clinical microbiology. Clinical microbiology reviews 26:547-603.
12. ECDC. 2022. European Centre for Disease Prevention and Control. Clostridioides (Clostridium) difficile infections. Annual epidemiological report for 2016–2017. ECDC.
13. Indra A, Huhulescu S, Schneeweis M, Hasenberger P, Kernbichler S, Fiedler A, Wewalka G, Allerberger F, Kuijper EJ. 2008. Characterization of Clostridium difficile isolates using capillary gel electrophoresis-based PCR ribotyping. Journal of Medical Microbiology 57:1377-1382.
14. Marin M, Martin A, Alcala L, Cercenado E, Iglesias C, Reigadas E, Bouza E. 2015. Clostridium difficile isolates with high linezolid MICs harbor the multiresistance gene cfr. Antimicrob Agents Chemother 59:586-9.
15. Bidet P, Lalande V, Salauze B, Burghoffer B, Avesani V, Delmee M, Rossier A, Barbut F, Petit JC. 2000. Comparison of PCR-ribotyping, arbitrarily primed PCR, and pulsed-field gel electrophoresis for typing Clostridium difficile. Journal of clinical microbiology 38:2484-7.
16. Candela A, Arroyo MJ, Sanchez-Molleda A, Mendez G, Quiroga L, Ruiz A, Cercenado E, Marin M, Munoz P, Mancera L, Rodriguez-Temporal D, Rodriguez-Sanchez B. 2022. Rapid and Reproducible MALDI-TOF-Based Method for the Detection of Vancomycin-Resistant Enterococcus faecium Using Classifying Algorithms. Diagnostics (Basel) 12.
17. Guerrero-López A, Sevilla-Salcedo C, Candela A, Hernández-García M, Cercenado E, Olmos PM, Cantón R, Muñoz P, Gómez-Verdejo V, del Campo R, Rodríguez-Sánchez B. 2023. Automatic antibiotic resistance prediction in Klebsiella pneumoniae based on MALDI-TOF mass spectra. Engineering Applications of Artificial Intelligence 118.
18. Weis C, Cuénod A, Rieck B, Dubuis O, Graf S, Lang C, Oberle M, Brackmann M, Søgaard KK, Osthoff M, Borgwardt K, Egli A. 2022. Direct antimicrobial resistance prediction from clinical MALDI-TOF mass spectra using machine learning. Nature Medicine 28:164-174.
19. Weis C, Horn M, Rieck B, Cuenod A, Egli A, Borgwardt K. 2020. Topological and kernel-based microbial phenotype prediction from MALDI-TOF mass spectra. Bioinformatics 36:i30-i38.
20. Candela A, Arroyo MJ, Sánchez-Cueto M, Marín M, Cercenado E, Méndez G, Muñoz P, Mancera L, Rodríguez-Temporal D, Rodríguez-Sánchez B, Khurshid M. 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.
21. Hamidi H, Bagheri Nejad R, Es-haghi A, Ghassempour A. 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 33:1530-1540.
22. Busby EJ, Doyle RM, Leboreiro Babe C, Harris KA, Mack D, Méndez-Cervantes G, O'Sullivan DM, Pang V, Sadouki Z, Solanki P, Huggett JF, McHugh TD, Wey EQ, Luethy PM, Aboulela A, Alexander X. 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.
23. Guerrero-López A, Sevilla-Salcedo C, Gómez-Verdejo V, M. Olmos P. 2022. Multimodal hierarchical Variational AutoEncoders with Factor Analysis latent space. arXiv doi:10.48550/arXiv.2207.09185.
24. Belenguer-Llorens A, Sevilla-Salcedo C, Desco M, Soto-Montenegro ML, Gómez-Verdejo V. 2022. A Novel Bayesian Linear Regression Model for the Analysis of Neuroimaging Data. Applied Sciences 12.

473 25. Flores-Trevino S, Garza-Gonzalez E, Mendoza-Olazaran S, Morfin-Otero R, Camacho-Ortiz A,
474 Rodriguez-Noriega E, Martinez-Melendez A, Bocanegra-Ibarias P. 2019. Screening of
475 biomarkers of drug resistance or virulence in ESCAPE pathogens by MALDI-TOF mass
476 spectrometry. Scientific reports 9:18945.
477 26. Emele MF, Joppe FM, Riedel T, Overmann J, Rupnik M, Cooper P, Kusumawati RL, Berger FK,
478 Laukien F, Zimmermann O, Bohne W, Gross U, Bader O, Zautner AE. 2019. Proteotyping of
479 Clostridioides difficile as Alternate Typing Method to Ribotyping Is Able to Distinguish the
480 Ribotypes RT027 and RT176 From Other Ribotypes. Frontiers in microbiology 10:2087.
481 27. Corver J, Sen J, Hornung BVH, Mertens BJ, Berssenbrugge EKL, Harmanus C, Sanders I, Kumar
482 N, Lawley TD, Kuijper EJ, Hensbergen PJ, Nicolardi S. 2019. Identification and validation of
483 two peptide markers for the recognition of Clostridioides difficile MLST-1 and MLST-11 by
484 MALDI-MS. Clin Microbiol Infect 25:904 e1-904 e7.
485 28. Abdrabou AMM, Sy I, Bischoff M, Arroyo MJ, Becker SL, Mellmann A, von Muller L, Gartner
486 B, Berger FK. 2023. Discrimination between hypervirulent and non-hypervirulent ribotypes
487 of Clostridioides difficile by MALDI-TOF mass spectrometry and machine learning. Eur J Clin
488 Microbiol Infect Dis doi:10.1007/s10096-023-04665-y.

489