



**HAL**  
open science

# Contribution of MALDI-TOF mass spectrometry and machine learning including deep learning techniques for the detection of virulence factors of *Clostridioides difficile* strains

Alexandre Godmer, Quentin Gai Gianetto, Killian Le Neindre, Valentine Latapy, Mathilda Bastide, Muriel Ehmig, Valérie Lalande, Nicolas Veziris, Alexandra Aubry, Frédéric Barbut, et al.

## ► To cite this version:

Alexandre Godmer, Quentin Gai Gianetto, Killian Le Neindre, Valentine Latapy, Mathilda Bastide, 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*, 2024, 17 (6), pp.e14478. 10.1111/1751-7915.14478 . pasteur-04854508

**HAL Id: pasteur-04854508**

**<https://pasteur.hal.science/pasteur-04854508v1>**

Submitted on 23 Dec 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.


L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

## RESEARCH ARTICLE

# Contribution of MALDI-TOF mass spectrometry and machine learning including deep learning techniques for the detection of virulence factors of *Clostridioides difficile* strains

Alexandre Godmer<sup>1,2</sup>  | Quentin Gai Gianetto<sup>3,4</sup> | Killian Le Neindre<sup>5</sup> |  
Valentine Latapy<sup>2</sup> | Mathilda Bastide<sup>2</sup> | Muriel Ehmig<sup>5</sup> | Valérie Lalande<sup>2,5</sup> |  
Nicolas Veziris<sup>1,2</sup> | Alexandra Aubry<sup>1,6</sup> | Frédéric Barbut<sup>5,7,8</sup> |  
Catherine Eckert<sup>1,2,5,8</sup>

<sup>1</sup>U1135, Centre d'Immunologie et Des Maladies Infectieuses (Cimi-Paris), Sorbonne Université, Paris, France

<sup>2</sup>Département de Bactériologie, AP-HP, Sorbonne Université (Assistance Publique Hôpitaux de Paris), Groupe Hospitalier Universitaire, Sorbonne Université, Hôpital, Saint-Antoine, Paris, France

<sup>3</sup>Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics HUB, Paris, France

<sup>4</sup>Institut Pasteur, Université Paris Cité, Proteomics Platform, Mass Spectrometry for Biology Unit, UAR CNRS 2024, Paris, France

<sup>5</sup>AP-HP, Sorbonne Université (Assistance Publique Hôpitaux de Paris), National Reference Laboratory for *Clostridioides Difficile*, Paris, France

<sup>6</sup>Centre National de Référence Des Mycobactéries et de la Résistance Des Mycobactéries Aux Antituberculeux, AP-HP, Sorbonne Université (Assistance Publique Hôpitaux de Paris), Hôpital Pitié Salpêtrière, Paris, France

<sup>7</sup>INSERM 1139, Université Paris Cité, Paris, France

<sup>8</sup>Paris Center for Microbiome Medicine (PaCeMM) FHU, Paris, France

## Correspondence

Catherine Eckert, U1135, Centre d'Immunologie et Des Maladies Infectieuses (Cimi-Paris), Sorbonne Université, Paris, France.  
Email: [catherine.eckert@aphp.fr](mailto:catherine.eckert@aphp.fr)

## Funding information

Santé Publique France (Public Health France)

## Abstract

*Clostridioides difficile* (CD) infections are defined by toxins A (TcdA) and B (TcdB) along with the binary toxin (CDT). The emergence of the 'hyper-virulent' (Hv) strain PR 027, along with PR 176 and 181, two decades ago, reshaped CD infection epidemiology in Europe. This study assessed MALDI-TOF mass spectrometry (MALDI-TOF MS) combined with machine learning (ML) and Deep Learning (DL) to identify toxigenic strains (producing TcdA, TcdB with or without CDT) and Hv strains. In total, 201 CD strains were analysed, comprising 151 toxigenic (24 ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup>, 22 ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup> Hv<sup>+</sup> and 105 ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>-</sup>) and 50 non-toxigenic (ToxA<sup>-</sup>B<sup>-</sup>) strains. The DL-based classifier exhibited a 0.95 negative predictive value for excluding ToxA<sup>-</sup>B<sup>-</sup> strains, showcasing accuracy in identifying this strain category. Sensitivity in correctly identifying ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>-</sup> strains ranged from 0.68 to 0.91. Additionally, all classifiers consistently demonstrated high specificity (>0.96) in detecting ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup> strains. The classifiers' performances for Hv strain detection were linked to high specificity (≥0.96). This study highlights MALDI-TOF MS enhanced by ML techniques as a rapid and cost-effective tool for identifying CD strain virulence factors. Our results brought a proof-of-concept

Alexandre Godmer, Quentin Gai Gianetto, Frédéric Barbut and Catherine Eckert contributed equally to this article

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Authors. *Microbial Biotechnology* published by John Wiley & Sons Ltd.

concerning the ability of MALDI-TOF MS coupled with ML techniques to detect virulence factor and potentially improve the outbreak's management.

## INTRODUCTION

*Clostridioides difficile* (CD) is a major cause of healthcare-associated diarrhoea and can cause outbreaks in healthcare settings. The latest European survey performed in 2017 showed that the median incidence of *C. difficile* infection (CDI) cases in hospital facilities was 4.1 per 10,000 bed-days in Europe and 2.3 in France (Viprey et al., 2023).

The main virulence factors of this anaerobic bacteria are toxins A (TcdA) and B (TcdB) encoded by *tcdA* and *tcdB* genes, and a third toxin, the binary toxin (CDT) encoded by *cdtA* and *cdtB* genes is considered to be an additional virulence factor (Eckert et al., 2018). Current diagnosis of CDI relies on both clinical (i.e. diarrhoea or pseudomembranous colitis) and microbiological criteria (i.e. presence of TcdA and TcdB in stools detected by immuno-enzymatic tests or presence of a toxigenic strain detected by toxigenic culture or molecular methods) (Crobach et al., 2016).

Twenty years ago, the epidemiology of CDI changed dramatically. Large outbreaks of severe CDI were described first in North America and then in Europe. These outbreaks were attributed to a specific clone, the epidemic PCR-ribotype (PR) 027 strain, also called NAP-1 or BI, according to the typing method used. This so-called 'hypervirulent' (Hv) PR 027 strain is CDT-positive and characterized by an overproduction of TcdA, TcdB and CDT. A deletion in position 117, leading to a stop codon in the negative regulator *tcdC* gene of the transcription of *tcdA* and *tcdB*, could be responsible for the increased virulence of this strain (Warny et al., 2005). Based on a European multi-centre point prevalence survey performed in 2018, PR 027 represents 11% of toxigenic strains isolated in Europe. This strain and the closely related PR 181 and PR 176 strains are the predominant ribotypes isolated in Eastern Europe (Viprey et al., 2022).

Epidemiologic surveillance of circulating clones in Europe is based on PCR ribotyping. Currently, there is a lack of rapid and inexpensive methods to allow recognition of Hv clones or to detect clusters in hospital facilities for the timely implementation of infection control measures.

Some molecular methods are able to give a presumptive identification of this PR027 strain by detecting the deletion in position 117 in *tcdC* (e.g. Xpert® Cdiff/Epi, Cepheid; VERIGENE® *Clostridium difficile* nucleic acid test [CDF], Luminex) or a putative conjugative transposon *pct* (e.g. Amplidiag *C. difficile*+027®, Mobidiag Ltd.) found in these Hv strains. However, these methods are expensive, and the result has to

be confirmed by the reference PCR ribotyping method which is time-consuming and not routinely performed by non-reference laboratories.

Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) is a reliable method for identifying microorganisms in medical microbiological laboratories (Croxatto et al., 2012). Besides identification, other applications of MALDI-TOF MS, such as identifying strains associated with virulence factors, are increasingly being evaluated (Flores-Treviño et al., 2019; Huang et al., 2015). Furthermore, artificial intelligence techniques have shown promise in analysing and designing algorithms (also called classifiers), enabling the results generated by MALDI-TOF MS to be improved, especially for bacterial identification (Garrigos et al., 2021; Godmer et al., 2021; Rodríguez-Temporal et al., 2023), detection of antimicrobial resistance, and identification of clonal strains associated with outbreaks (Flores-Treviño et al., 2019; Mohammad et al., 2023; Weis et al., 2022). Of the artificial intelligence techniques, machine learning (ML) focuses on developing algorithms that learn from a training data set to optimize a performance criterion, also known as a cost function (e.g. accuracy, Cohen's kappa and cross-entropy loss) with the goal of solving a given problem, for example the most precise classification prediction possible. A field of ML that uses neural networks, deep learning (DL), solves complex tasks using artificial neural architectures composed of many hidden layers. Therefore, ML, including DL, paves the way for the development of inexpensive and easy-to-use alternative methods enabling MALDI-TOF MS to be used for the surveillance of epidemic clones such as PR-027 or for the detection of virulence factors.

The objectives of our study were to evaluate the ability of MALDI-TOF MS coupled with ML-based classifiers to identify (i) toxigenic CD strains (producing TcdA, TcdB, so-called A<sup>+</sup>B<sup>+</sup>, with or without CDT) and (ii) Hv CD strains (PR 027 and closely related PR 176 and PR 181 strains).

## EXPERIMENTAL PROCEDURES

### Bacterial isolates

A total of 201 toxigenic and non-toxigenic CD strains representative of the diversity of those circulating in France were used in this study. These strains included reference strains (Brazier–Kuijper–Wilcox collection) and strains collected from the National Reference Laboratory for CD (Saint-Antoine Hospital, Paris,

France). Strains were previously identified by MALDI-TOF MS and characterized by multiplex PCR for the detection of the main virulence factors (*tcdA*, *tcdB* and *cdt* genes), and PR was determined by capillary gel-based electrophoresis PCR ribotyping as previously described (Eckert et al., 2018).

The 201 CD strains included 50 non-toxigenic strains (*tcdA*<sup>-</sup> *tcdB*<sup>-</sup>) (designated ToxA<sup>-</sup>B<sup>-</sup>) belonging to 19 different PRs, and 151 toxigenic strains harbouring toxins A and B genes (ToxA<sup>+</sup>B<sup>+</sup>). Among the 151 ToxA<sup>+</sup>B<sup>+</sup> strains, 46 (8 different PRs) also harboured the binary toxin genes (ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup>) and 105 (23 different PRs) did not (ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>-</sup>). Finally, among the 46 ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup> strains, 22 belonged to the Hv strains, that is, PR 027 (*n*=13), PR 176 (*n*=5) and PR 181 (*n*=4) strains (ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup>Hv) (Table S5).

The strains were divided into two datasets: (i) the first corresponded to 50% of the strains (*n*=102) (the 'Optimization dataset') and was used to train the classifiers and to identify the most efficient pipelines and (ii) the second corresponded to the rest of the strains (*n*=99) the 'Test dataset' and was used to estimate the performance of the pipelines previously identified with the 'Optimization dataset'.

## Sample preparation

Each isolate was stored at -80°C (Microbank, Inc., Canada) and thawed and cultivated on CBA (bioMérieux SA, Marcy l'Etoile, France) incubated in an anaerobic atmosphere at 37°C for 48 h. A subculture was performed in the same conditions. Chemical protein extraction was then carried out. Briefly, a single colony was suspended in 200 µL water and vortexed. After adding 900 µL ethanol, samples were vortexed and centrifuged at 13,000 × *g* for 2 min. The supernatant was removed, and the remaining ethanol was evaporated at room temperature. Next, 25 µL of 70% formic acid was added and mixed with the pellet, then 25 µL of acetonitrile was added. After centrifugation at 13,000 × *g* for 2 min, the supernatant was ready for analysis. Each isolate was extracted once, and eight deposits were made for each extract (eight technical replicates). The dried spots were coated with MALDI matrix (10 mg/mL of α-cyano-4-hydroxy-cinnamic acid [α-HCCA] in 50% acetonitrile-2.5% trifluoroacetic acid; Bruker® Daltonics, Bremen, Germany) and each spot was analysed three times by MALDI-TOF MS. A total of 17 target plates were necessary to produce spectra at a rate of one plate per day.

## MALDI-TOF MS acquisition and analysis

Mass spectra were acquired using a Microflex LT instrument (Bruker® Daltonics, Bremen, Germany).

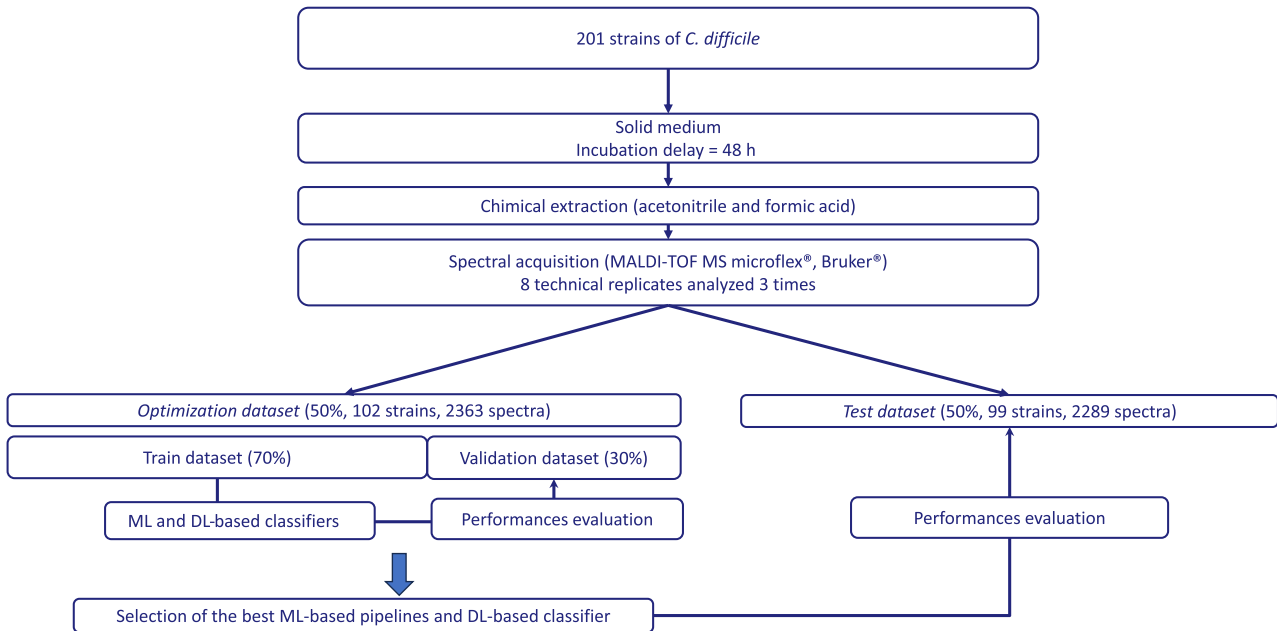
The standard parameters of the CE-IVD (Conformité Européenne—In Vitro Diagnostic) method recommended by the manufacturer were used. As recommended by the Bruker® manufacturer for database creation (Fergusson et al., 2020), spectra from the same strain were visually analysed to discard those of poor quality according to the following criteria: (i) presence of outlier peaks compared to the other spectra, (ii) presence of spectra with flat peaks compared to the other spectra and (iii) spectra with a mass shift greater than 500 ppm compared to the other spectra. A minimum of 20 spectra were selected per strain and imported in R using the MALDIForeign package (Gibb & Franceschi, 2022). The different stages of post-acquisition signal processing were performed in the R environment using the MsclassifR and MALDIquant packages (Gibb & Strimmer, 2012; Godmer et al., 2022) according to the following pipeline: (i) square root intensity transformation, (ii) spectrum smoothing (Undecimated Wavelet Transform [UDWT] algorithm), (iii) baseline processing (SNIP for statistics-sensitive non-linear iterative peak-clipping algorithm), (iv) intensity calibration (TIC for Total Ion Current algorithm) and (v) spectrum alignment (500 ppm) and selection of peaks with signal-to-noise ratio (S/N) greater than 3.

## Design of the machine learning analysis

Spectra from the 'optimization dataset' were divided into two sub-datasets: the first—the 'training dataset' (70% of the spectra)—was used to train the classifiers, and the second—the 'validation dataset' (30% of the spectra)—was used to estimate the performances of the generated classifiers of ML algorithms (Figure 1). Spectra from the same strain could only be part of one of the two datasets.

Due to the training characteristics of each ML-based algorithm to build classifiers, two classification approaches were used to achieve the objectives of the study:

- (i) Multi-class classification, where the ML-based classifiers were trained to classify the spectra into a single exclusive class. Each spectrum was assigned to one of the possible classes (a spectrum belongs to a category with the highest probability). In this study, the classes were (i) ToxA<sup>-</sup>B<sup>-</sup> strains, (ii) ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>-</sup> strains, (iii) ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup> strains and (iv) ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup>Hv strains. Each spectrum was classified into one of these classes based on its spectral characteristics.
- (ii) Multilabel classification, where the classifiers were trained to predict multiple classes for the same spectrum at the same time. This means that a single spectrum can be associated with more than one class using a score between 0 and 1 with a cut-off of 0.5 for each category independently.



**FIGURE 1** Performances of the 12 ML-based pipelines on the ‘optimization dataset’ to distinguish  $ToxA^-B^-$ ,  $ToxA^+B^+CDT^-$ ,  $ToxA^+B^+CDT^+$ , and  $ToxA^+B^+CDT^+Hv$  strains using traditional ML-based algorithms. mda, Mean Decrease Accuracy; multinom, linear regression; nnet, single-hidden-layer neural network; RF, random forests; SelectionVarStat, ANOVA test; sPLSDA, Sparse Partial Least Squares Discriminant Analysis; svm, linear Support Vector;  $ToxA^-B^-$ , spectra from non-toxicogenic strains;  $ToxA^+B^+$ , spectra from strains harbouring toxins A and B genes;  $ToxA^+B^+CDT^+$ , spectra from strains harbouring toxins A, B and binary toxin genes;  $ToxA^+B^+CDT^+Hv$ , strains harbouring toxins A, B and binary toxin genes and hypervirulent strains (associated with PR 027, PR 176, PR 181).

(iii) For the implementation of these approaches, the ML-based classifier approach was used for multi-class classification, where traditional ML techniques were applied. In addition, another ML-based classifier was performed using DL algorithms (Ctox-classifier) (Figure 1).

## Design of the machine learning-based classifiers approach

Twelve traditional ML-pipelines including two steps were used to create various ML-based classifiers using the MSclassifR R package (Godmer et al., 2022).

The first step eliminated non-informative mass-over-charge values that mainly lead to intensity peaks corresponding to background noise. Three methods using ML algorithms were used:

- (i) ‘Mean decrease accuracy’ (mda), which consisted of selecting variables using the distribution of mda values previously determined by measuring the effect of a random variable permutation using a random forests (RF) algorithm.
- (ii) ‘SelectionVarStat’ consisted of selection of informative peaks using analysis of variance (ANOVA) to determine if the variables were significantly different across the different groups.
- (iii) ‘Sparse partial least squares discriminant analysis’ (sPLS-DA) with k-folds cross-validation ( $k=5$ ).

In this process, in the first iteration, the first fold ( $k=1$ ) was used as a test for the algorithm, while the others ( $k=4$ ) were used for training; this process is then repeated until every fold has been used as a test set.

The second step estimated classification models using ML algorithms from the intensity peaks measured at mass-over-charge values shortlisted from the first step. Four ML algorithms were used: (i) linear regression (multinom), (ii) single-hidden-layer neural network (nnet), (iii) RF and (iv) linear support vector machine (svm). These classifiers were trained with k-fold cross-validation ( $k=5$ ) on the training dataset. The various hyperparameters for each algorithm and each training session were researched using a random search grid. For a mass spectrum, each classifier provides probabilities that it belongs to any of the considered categories. Finally, we considered a mass spectrum to belong to a particular category when this category was associated with the maximum of these probabilities measured on all categories. According to the predictions of each ML algorithm from the different pipelines, performance criteria (Cohen's kappa ( $\kappa$ ) coefficients) were estimated for each ‘validation dataset’. The optimization dataset was randomly split 10 times into a training dataset (70%) and a validation dataset (30%) while preserving the proportion of each category in each set. This allowed variations of the performance criteria to be estimated.

In addition, a DL-based classifier (the Ctox-classifier) was performed using a different approach, that is, the open-source library Keras using an artificial neuronal network organized in several layers (Keras: deep learning for humans, [n.d.](#)). Because this Ctox classifier took longer to train, only one classifier was generated after determining the appropriate learning rate using a random search grid. We used an architecture consisting of a fully connected network structure with six layers: (i) an input layer containing the variables, (ii) a first hidden layer with 256 nodes, (iii) a second hidden layer with 64 nodes, (iv) a third hidden layer with 256 nodes (v) a fourth hidden layer with 512 nodes and (vi) an output layer with four nodes and which used the sigmoid activation function. The activation function of a fully connected network was 'rectified linear units'. The binary cross-entropy was selected as the loss function and was selected as the optimizer 'Root Mean Square Propagation' (RMSprop). The learning rate was set to 0.001 and the number of epochs was set to 200. This C-tox-classifier was trained with the same methodology as previously described using the 'optimization dataset'. Due to the imbalanced data, the metric to train the MPL classifier was an area under the precision-recall curve (AUC-PR), and a weighting of the loss function or bias was applied during the training process. Of note, the classifier may have been biased in favour of the majority class in unbalanced datasets where some groups were under-sampled relative to others. Therefore, by giving more weight to the minority class, the weighting of the loss function mitigates the effect of class imbalance.

### Determination of the most efficient pipeline

Among the 12 standard ML-based pipelines, the three with the highest median kappa coefficient combined with the lowest standard deviation were considered the most efficient. In addition, due to imbalanced data, the classifiers from the three best ML-based pipelines were retrained using the following resampling methods: (i) down-sampling, randomly removing majority class spectra; (ii) up-sampling, randomly replicating minority class spectra and (iii) synthetic minority oversampling technique, using a machine learning algorithm (K-nearest-neighbours) to generate new minority class spectra. Since one classifier was employed using DL algorithms (the Ctox-classifier), we included it in the study.

### Machine learning-based classifiers evaluation on the *test dataset*

The performance of each classifier was evaluated on the different 'test datasets' using Cohen's kappa coefficient. The classifiers trained with or without resampling

methods from the three best ML-based pipelines determined by the high mean kappa coefficient and the Ctox-classifier were selected to estimate the performance on the 'test dataset'. The performance of the classifier for each pipeline with the best kappa coefficient and that of the Ctox-classifier were then reported.

### Peak analysis

Discriminant peaks with a frequency  $\geq 90\%$  were identified in one group of spectra and with a frequency  $\leq 10\%$  in the other three groups. This was done using an in-house programme in the R environment. The visualization tool is available at <https://agodmer.github.io/Clostri/>.

### Statistical analysis

To evaluate the ML-based classifiers comprehensively, several metrics were used to estimate performance on the 'test dataset': accuracy (proportion of correctly classified spectra out of the total), sensitivity (corresponding to the proportion of well-identified spectra in this study per subspecies), the reliability of identification representing the percentage of certainty of correct identification (corresponding to the positive predictive value [PPV]), and the reliability of identification indicating the percentage of certainty in correct identifications of negative results (corresponding to the negative predictive value [NPV]). In addition, the kappa coefficient (which measures inter-rater reliability on a scale from  $-1$  [complete disagreement] to  $1$  [complete agreement]) was also reported. All these metrics were calculated using the caret package in the R software (version 4.2.2) (Kuhn, 2008). Statistical comparisons were performed using the unilateral Wilcoxon rank sum tests with the Benjamini–Hochberg correction, and a  $p$ -value  $< 0.05$  was considered as statistically significant.

### Data storage

The spectra generated during this study are available only for medical research in accordance with the FAIR principle (Godmer et al., 2023).

## RESULTS

### Determination of the ML-based most efficient pipelines

A total of 4659 spectra were generated. The 'optimization dataset' consisted of 2363 spectra from 102 strains across 41 different PR and served as the training set for ML-based classifiers. These classifiers have been

developed from a variety of combinations of ML-based algorithms through various ML pipelines (Figure 2). During the training process, resampling methods were applied to account for the significantly different numbers of strains in each group. Subsequently, the 'test dataset' encompassed 2296 spectra from 99 strains across 42 PR, and it was employed to assess the performance of the ML-based classifiers derived from ML pipelines.

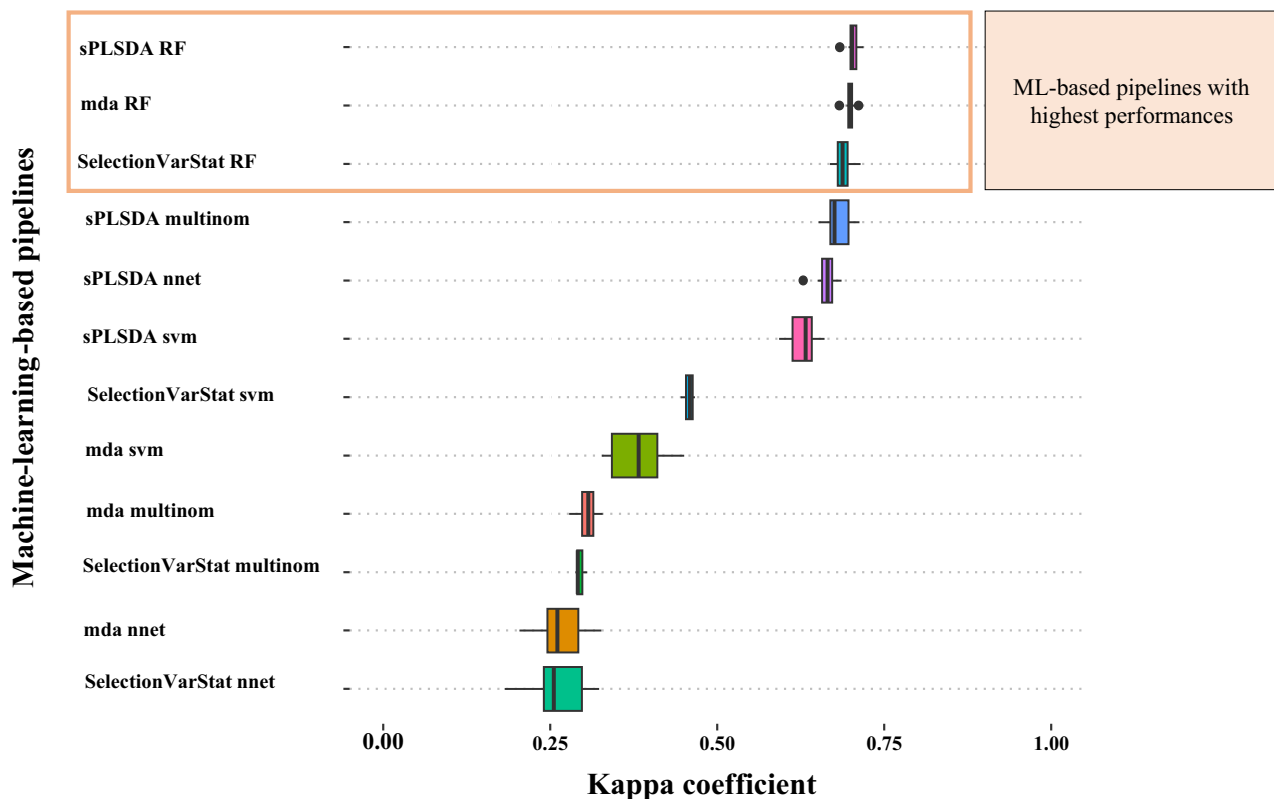
The 12 ML pipelines designed from the 'optimization dataset' led to the conception of 120 ML-based classifiers, enabling predictions to be made according to the study objectives (each pipeline was estimated on 10 pairs of training and validation sets). A preliminary assessment of the 'optimization dataset' showed that the global performance (accuracy, mean  $\pm$  standard deviation [SD]) to identify ToxA<sup>-</sup>B<sup>-</sup>, ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>-</sup>, ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup>, and ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup> Hv strains translated to a global rate of correct identification of 0.65  $\pm$  0.16 (Table S1). The three best pipelines were sPLSDA-RF, mda-RF, and SelectionVarStat-RF with mean  $\pm$  SD kappa coefficients of 0.70  $\pm$  0.01, 0.70  $\pm$  0.01, and 0.69  $\pm$  0.01, respectively (Figure 2; Table S1). In addition, Random Forest (RF) corresponding to the algorithm selected in these three ML-based pipelines demonstrated statistically better performance versus all the other ML-based algorithms ( $p < 10^{-7}$ ) (Table S2).

## Detection of toxigenic strains by MALDI Biotyper® (Bruker®) coupled with ML- and DL-based approaches (Table 1)

The maximum performance of the classifiers from the ML-based approaches for detecting toxigenic strains with the highest kappa coefficient values is presented in Table 1. The global performance of all the traditional ML-based classifiers (excluding the Ctox-classifier obtained with the DL method) from the 12 ML-based are listed in Table S3. Regarding the traditional ML-based classifiers, no resampling method statistically outperformed the others (Table S4).

Regarding the ability to identify strains producing no toxins (ToxA<sup>-</sup>B<sup>-</sup>) with the ML-based classifiers, sensitivities and specificities ranged from 61% to 86% and 90% to 94%, respectively, which also reflected an increased ability to exclude a ToxA<sup>-</sup>B<sup>-</sup> spectrum (i.e. positive predictive value (PPV) or negative predictive value (NPV), respectively), rather than to correctly identify a ToxA<sup>-</sup>B<sup>-</sup> spectrum, considering the prevalence of this group in our setting.

Regarding the ability to identify strains producing toxins (ToxA<sup>+</sup>B<sup>+</sup>) with the ML-based classifiers, sensitivities and specificities ranged from 65% to 91% and 68% to 99%, respectively. More specifically, the ML-based classifiers generally performed better at classifying spectra for ToxA<sup>+</sup>B<sup>+</sup> CDT<sup>-</sup> strains in comparison



**FIGURE 2** Workflow of machine learning (ML)-based approach used in this study. \*The Ctox classifier was performed using Deep Learning algorithms, whereas the other classifiers were obtained using traditional ML-based algorithms.

to those for ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup> strains (mean sensitivity and standard deviation: 0.86±0.06 for the ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>-</sup> group versus 0.72±0.05 for the ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup> group). Conversely, there was a reversal of this trend when considering mean specificity, with the values of 0.78±0.07 for the ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>-</sup> group and 0.98±0.01 for the ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup> group.

### Detection of Hv strains (ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup> Hv) by MALDI Biotyper® (Bruker®) coupled with ML-DL approaches (Table 1)

The maximum performances of the classifiers from the ML-based approaches for detecting Hv strains (ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup> Hv) with the highest kappa coefficient values are presented in Table 1. Regarding the traditional ML-based classifiers, no resampling method statistically outperformed the others (Table S4). In general, the ML-based classifiers exhibited similar performance with better specificity (96% to 100%) than sensitivity (65% to 76%) which translated to high PPV or NPV, considering the prevalence of Hv strains in our setting.

### Analysis of discriminant peaks

Among the 2127 peaks analysed, we found two specific peaks that were associated with non-Hv strains: (i)  $m/z=6650$  was found in 94.4% of spectra of non-Hv strains versus 5.6% of Hv spectra (ii)  $m/z=3354$  was found in 90.5% of spectra of non-Hv strains versus 9.5% of Hv spectra. Discriminating peaks capable of separating the other groups were not found with our method. A visualization tool is available at <https://agodmer.github.io/Clostri/>.

## DISCUSSION

MALDI-TOF MS is a widely used, fast, and cost-saving laboratory method for the identification of microorganisms (Dekker & Branda, 2011). MALDI-TOF coupled with ML techniques has been used successfully to identify genetically similar species, to detect antibiotic resistance or virulence factors, and to detect epidemic clones (Elbehiry et al., 2022; Flores-Treviño et al., 2019; Li et al., 2022; Weis et al., 2022). For example, the *Escherichia coli* B2 phylogroup, which is more virulent than the other phylogroups, can be distinguished using MALDI-TOF MS (Sauget et al., 2014), as well as enterohemorrhagic *E. coli* pathotypes (Christner et al., 2014; Clark et al., 2013; Fagerquist et al., 2010; Mazzeo et al., 2006). In this work, we assessed the potential of MALDI-TOF MS to distinguish between different groups of toxin-producing CD strains, including some Hv strains.

Previously, several methods to detect virulence factors by MALDI-TOF MS have been used. One of these is based on the detection of a discriminant peak corresponding to the virulence factor. In this method, the peak should belong to the mass range studied by commercial equipment for routine use (2–20 kDa), such as the Delta-Toxin from *Staphylococcus aureus* (Gagnaire et al., 2012). However, the toxins often have molecular weights outside the mass range studied by commercial equipment, such as the Shiga toxin 1 and Shiga toxin 2 harboured by some *E. coli* strains (32–33 kDa) (Kubo et al., 2021). As the toxins produced by CD fall into this latter category (~308 and ~270 kDa for TcdA and TcdB, respectively), we chose to use ML-based algorithms to detect spectral patterns associated with strains producing virulence factors.

To the best of our knowledge, the ML-based approaches used in the literature for CD have mostly been used to distinguish specific PRs (Calderaro et al., 2021, 2022). Unfortunately, some of the most predominant PRs were absent in these studies (Calderaro et al., 2021, 2022). For instance, Calderaro et al. used ML-based models based on the Genetic Algorithm, QuickClassifier and Supervised Neural Network algorithms to create ML-based classifiers for the identification of 61 CD strains from 10 PR (arbitrarily noted PR1-10, PR1 and PR2 corresponding to PR 018 and PR 126, respectively, while other PRs remained unknown) with recognition close to 100% for PR1-5 (55 strains) whereas the six strains from PR6-10 could not be classified (Calderaro et al., 2021). Also, some authors have also used a statistical approach to identify a particular PR, such as PR017, which is particularly widespread in China (Li et al., 2018). The Brazilian study used MALDI-TOF MS to distinguish 19 PR of CD. Using an approach involving clustering of spectral data, 73% of spectra were correctly classified in a double-blind validation using 13 biomarker profiles. More specifically, the method proved highly effective for epidemic PR 027 and Brazilian-specific PR, with an accuracy of between 94% and 100% (Carneiro et al., 2021). To date, only one study has successfully evaluated (with greater than 95% accuracy) the use of MALDI-TOF MS in conjunction with ML-based algorithms (excluding DL-based methods) to distinguish strains producing binary toxin from strains that do not produce binary toxin (classified as Hv or non-Hv in their study) (Abdrabou et al., 2023). Unlike this study, our objective was not only to separate binary toxin-producing strains from the others but also to separate several distinct populations such as ToxA<sup>-</sup>B<sup>-</sup>, ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>-</sup>, ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup> and ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup>Hv. Despite different objectives compared to the Abdrabou et al. study, we also highlighted the effectiveness of the RF algorithm for spectral data analysis. In our research, we found that, among the ML-based classifiers we performed, RF demonstrated superior performance (Table 1). These observations



**TABLE 1** Performances of the machine learning-based classifiers on the 'test dataset' for the detection ToxA<sup>-</sup>B<sup>-</sup>, ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>-</sup>, ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup> and ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup>Hv strains.

Machine learning-based pipeline	Strain characteristics	Resampling	Prevalence on the 'test dataset'	Metric performance (mean ± SD)				
				Sensitivity	Specificity	NPV		
Mda and rf	ToxA <sup>-</sup> B <sup>-</sup>	None	0.26	0.67	0.90	0.71	0.89	
		Down		0.79	0.92	0.77	0.93	
		Smote		0.75	0.91	0.75	0.91	
	ToxA <sup>+</sup> B <sup>+</sup> CDT <sup>-</sup>	Up		0.65	0.93	0.76	0.88	
		None	0.52	0.87	0.72	0.77	0.84	
		Down		0.85	0.83	0.85	0.84	
	ToxA <sup>+</sup> B <sup>+</sup> CDT <sup>+</sup>	Smote		0.86	0.80	0.83	0.84	
		Up		0.90	0.70	0.77	0.86	
		None	0.12	0.67	0.98	0.85	0.96	
	ToxA <sup>+</sup> B <sup>+</sup> CDT <sup>+</sup> Hv	Down		0.75	0.95	0.67	0.97	
		Smote		0.78	0.97	0.76	0.97	
		Up		0.72	0.99	0.88	0.96	
	SelectionVarStat and rf	ToxA <sup>-</sup> B <sup>-</sup>	None	0.10	0.71	0.99	0.99	0.97
			Down		0.75	1.00	1.00	0.97
			Smote		0.75	1.00	1.00	0.97
ToxA <sup>+</sup> B <sup>+</sup> CDT <sup>-</sup>	Up		0.74	1.00	1.00	0.97		
	None	0.26	0.67	0.90	0.70	0.88		
	Down		0.78	0.92	0.77	0.92		
ToxA <sup>+</sup> B <sup>+</sup> CDT <sup>+</sup>	Smote		0.74	0.92	0.76	0.91		
	Up		0.67	0.93	0.76	0.89		
	None	0.52	0.87	0.72	0.77	0.83		
ToxA <sup>+</sup> B <sup>+</sup> CDT <sup>+</sup> Hv	Down		0.86	0.82	0.84	0.84		
	Smote		0.88	0.77	0.81	0.85		
	Up		0.9	0.7	0.77	0.87		
ToxA <sup>+</sup> B <sup>+</sup> CDT <sup>+</sup> Hv	None	0.12	0.67	0.98	0.85	0.96		
	Down		0.75	0.96	0.73	0.97		
	Smote		0.75	0.98	0.86	0.97		
ToxA <sup>+</sup> B <sup>+</sup> CDT <sup>+</sup> Hv	Up		0.65	0.99	0.89	0.96		
	None	0.10	0.71	1.00	0.98	0.97		
	Down		0.75	0.96	0.73	0.97		
ToxA <sup>+</sup> B <sup>+</sup> CDT <sup>+</sup> Hv	Smote		0.75	0.98	0.86	0.97		
	Up		0.65	0.99	0.89	0.96		
	None	0.10	0.71	1.00	0.98	0.97		

TABLE 1 (Continued)

Machine learning-based pipeline	Strain characteristics	Resampling	Prevalence on the 'test dataset'	Metric performance (mean $\pm$ SD)				
				Sensitivity	Specificity	PPV	NPV	
sPLSDA and rf	ToxA <sup>-</sup> B <sup>-</sup>	None	0.26	0.62	0.94	0.77	0.87	
		Down		0.81	0.91	0.76	0.93	
		Smote		0.77	0.93	0.79	0.92	
	ToxA <sup>+</sup> B <sup>+</sup> CDT <sup>-</sup>	Up		0.61	0.94	0.79	0.87	
		None	0.52	0.68	0.90	0.76	0.87	
		Down		0.86	0.83	0.85	0.85	
	ToxA <sup>+</sup> B <sup>+</sup> CDT <sup>+</sup>	Smote		0.89	0.79	0.83	0.87	
		Up		0.91	0.68	0.76	0.87	
		None	0.12	0.66	0.98	0.86	0.99	
	ToxA <sup>+</sup> B <sup>+</sup> CDT <sup>+</sup> Hv	Down		0.73	0.97	0.76	0.96	
		Smote		0.75	0.98	0.83	0.97	
		Up		0.71	0.98	0.82	0.96	
	Ctox (Deep Learning-based classifier)	ToxA <sup>+</sup> B <sup>+</sup> CDT <sup>+</sup> Hv	None	0.10	0.7	1.00	0.99	0.97
			Down		0.76	1.00	1.00	0.97
			Smote		0.74	1.00	1.00	0.97
Up				0.72	1.00	1.00	0.97	
ToxA <sup>-</sup> B <sup>-</sup>	ToxA <sup>+</sup> B <sup>+</sup> CDT <sup>-</sup>	NA	0.26	0.86	0.90	0.75	0.95	
		None	0.52	0.86	0.89	0.90	0.85	
		Down	0.12	0.79	0.96	0.70	0.97	
		Up	0.1	0.69	1.00	1.00	0.97	

Abbreviations: NA, not available; NPV, observed negative predictive value according to the prevalence of this study; PPV, observed Positive Predictive Value according to the prevalence of this study; RF, random forests; SelectionVarStat, ANOVA test; smote synthetic minority oversampling technique; sPLSDA, Sparse Partial Least; ToxA<sup>-</sup>B<sup>-</sup>, spectra from non-toxicogenic strains; ToxA<sup>+</sup>B<sup>+</sup>, spectra from strains harbouring toxins A and B genes; ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup>, spectra from strains harbouring toxins A, B and binary toxin genes; ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup>Hv, strains harbouring toxins A, B and binary toxin genes and hypervirulent strains (PR 027, PR 176 and PR 181).

confirm that the RF algorithm is particularly suitable for a variety of spectral data classification tasks, including distinguishing genetically related species, according to the results reported by Rodríguez-Temporal et al. (2023) and Candela et al. (2023).

Concerning DL methods, we showed that MALDI-TOF coupled with the Ctox-classifier could accurately exclude ToxA<sup>-</sup>B<sup>-</sup> spectra in 95% of cases. Regarding the identification of the ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup> strains, the specificity and observed NPV were more than 96% for all classifiers performed, indicating that this approach could be useful to exclude this type of strain. It should be noted that the observed NPV and PPV calculated in this study depend on the prevalence of each group of strains included and do not reflect that of the general population. The integration of the peak selection step directly into the neural architecture makes the use of the Ctox-classifier with DL algorithms an attractive choice. This approach eliminates the need for additional ML algorithms, simplifying the otherwise time-consuming process of testing multiple ML-based algorithms. In addition, with this method, the Ctox-classifier achieves performance that is comparable to other ML-based classifiers (Table 1).

In the present study, we were not able to correctly identify Hv strains (ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup> Hv) by MALDI-TOF combined with the ML-based classifier, but we could reliably exclude this type of strain. Interestingly, these results can be explained by the absence of two discriminating peaks in ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup> Hv strains ( $m/z=6650$ ,  $3354$ ). The first discriminant peak ( $m/z=6650-6654$ ) was previously described by Flores-Treviño et al. (2019). The authors used the absence of peak  $m/z=6654$  to successfully distinguish PR 027 strains from other strains with a sensitivity of 100%, specificity of 91.7%, and a PPV of 95%. Of note, in the same study, the peak at  $m/z=6654$  was also absent in PR 176, which is aligned with our findings. In addition, this peak was not found in the third PR (PR 181) which is closely related to PR 027 and PR 176 that we included in the ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup>Hv.

Moreover, the second discriminant peak found in our study ( $m/z=3354$ ), with a frequency of less than 10% in ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup>Hv but higher (90.5%) in non-Hv strains, also appears to be of interest but to our knowledge has not been described in the literature. These two peaks effectively exclude PR 027 strains, in agreement with the classifiers developed with specificities close to 1. However, the presence of the peak with an estimated variation of 6Da between studies  $m/z=6706-6712$ , which was considered to be specific for PR 027 strains in other studies (Emele et al., 2019; Flores-Treviño et al., 2019; Reil et al., 2011), was less discriminatory in our study. The latter ( $m/z=6709$  for our study) was associated with a frequency of 93.2% in ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup>Hv strains but was also present with a significant frequency in non-Hv strains (18.2%).

Our study has some limitations. We used the Columbia blood agar (CBA) medium, and the medium used can have an effect on the spectra and, notably, can affect the presence of discriminant peaks (Popović et al., 2023). It is therefore necessary to evaluate the ML-based classifiers with strains cultivated on other types of media, especially selective media that are frequently used in laboratories, to confirm our findings. We used a chemical extraction technique to get better-quality spectra. It would be useful to validate this methodology on deposits directly placed on the identification plate. This could result in a gain of time and make the technique simpler. Also, to validate the generalization and reproducibility of ML and DL-based classifiers, it would have been necessary to repeat the extractions. Another limitation of the study is that it was performed in a single centre. It would be interesting to test this algorithm on a much broader range of isolates with different MALDI-TOF SM instruments in different laboratories. This study is a proof-of-concept that demonstrates MALDI-TOF MS to be an easy and inexpensive technique to detect TcdA, TcdB and CDT and to presumptively exclude PR 027 strains.

While this methodology shows promise, it is currently facing challenges in being routinely integrated into microbiological laboratory diagnostics. Contrary to molecular PCR and immunochromatographic tests that are directly applicable to stool samples, our technique necessitates pre-processing involving both a culturing phase and an extraction procedure prior to MALDI-TOF MS analysis. These prerequisites are time-consuming and may hinder the effectiveness of the method for immediate diagnostic needs. However, for epidemiological surveillance and outbreak management, our approach offers complementary advantages over molecular methods. It is worth noting that MALDI-TOF MS coupled with algorithms based on ML including DL provides high specificity, which has the potential to rapidly identify and rule out some types of strains such as ToxA<sup>+</sup>B<sup>+</sup>CDT<sup>+</sup>Hv. For instance, this could eliminate the need for more expensive and time-consuming molecular typing methods such as PCR ribotyping or whole genome sequencing. A targeted application could considerably improve response strategies to epidemics and contribute to the rapid implementation of infection control measures. Targeted applications could significantly improve epidemic response strategies and contribute to the rapid implementation of infection control measures. For example, the high NPV observed in our study for CDT<sup>+</sup> strains, such as PR 078 strains known to cause epidemics of severe diarrhoea in communities and particularly in young people, means that this type of strain can be immediately excluded in an epidemic context.

## AUTHOR CONTRIBUTIONS

**Alexandre Godmer:** Conceptualization; data curation; formal analysis; investigation; methodology; software;

supervision; validation; visualization; writing – original draft; writing – review and editing. **Quentin Gai Gianetto**: Conceptualization; formal analysis; methodology; writing – review and editing. **Killian Le Neindre**: Data curation; methodology; resources; writing – review and editing. **Valentine Latapy**: Methodology; writing – review and editing. **Mathilda Bastide**: Data curation; methodology. **Muriel Ehmig**: Writing – review and editing. **Valérie Lalande**: Methodology; writing – review and editing. **Nicolas Veziris**: Methodology; writing – review and editing. **Alexandra Aubry**: Writing – review and editing. **Frédéric Barbut**: Investigation; methodology; project administration; resources; supervision; writing – review and editing. **Catherine Eckert**: Investigation; methodology; resources; supervision; validation; visualization; writing – original draft; writing – review and editing.

## ACKNOWLEDGEMENTS

The National Reference Laboratory for *Clostridioides difficile* is financially supported by Santé Publique France (Public Health France). The funder had no role in the study design, data collection and interpretation or the decision to submit the work for publication. The authors would like to thank Dr Andrew Lane (Lane Medical Writing) for editorial support for English language and grammar.

## CONFLICT OF INTEREST STATEMENT

All authors report no potential conflicts of interest with respect to the research, authorship and publication of this article.

## DATA AVAILABILITY STATEMENT

Data are available on request Alexandre Godmer, Quentin Gai Gianetto, Killian Le Neindre, Valentine Latapy, Mathilda Bastide, Muriel Ehmig, Valérie Lalande, Nicolas Veziris, Alexandra Aubry, Frédéric Barbut, & Catherine Eckert. (2023). Supplemental Data from “Contribution of MALDI-TOF Mass Spectrometry and Machine Learning Including Deep Learning Techniques for the Detection of Virulence Factors of *Clostridioides difficile* Strains” (Version 1). Zenodo. <https://doi.org/10.5281/zenodo.8424800>.

## ETHICS STATEMENT

All specimens were processed and anonymized in accordance with ethical and legal standards, and patients were not physically involved in this study. Informed consent was not needed for this study.

## ORCID

Alexandre Godmer  <https://orcid.org/0000-0002-5211-5796>

## REFERENCES

Abdrabou, A.M.M., Sy, I., Bischoff, M., Arroyo, M.J., Becker, S.L., Mellmann, A. et al. (2023) 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*, 42, 1373–1381.
- Calderaro, A., Buttrini, M., Farina, B., Montecchini, S., Martinelli, M., Arcangeletti, M.C. et al. (2022) Characterization of *Clostridioides difficile* strains from an outbreak using MALDI-TOF mass spectrometry. *Microorganisms*, 10, 1477.
- Calderaro, A., Buttrini, M., Martinelli, M., Farina, B., Moro, T., Montecchini, S. et al. (2021) Rapid classification of *Clostridioides difficile* strains using MALDI-TOF MS peak-based assay in comparison with PCR-ribotyping. *Microorganisms*, 9, 661.
- Candela, A., Guerrero-López, A., Mateos, M., Gómez-Asenjo, A., Arroyo, M.J., Hernandez-García, M. et al. (2023) 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.
- Carneiro, L.G., Pinto, T.C.A., Moura, H., Barr, J., Domingues, R.M.C.P., Ferreira, E. et al. (2021) MALDI-TOF MS: an alternative approach for ribotyping *Clostridioides difficile* isolates in Brazil. *Anaerobe*, 69, 102351.
- Christner, M., Trusch, M., Rohde, H., Kwiatkowski, M., Schlüter, H., Wolters, M. et al. (2014) Rapid MALDI-TOF mass spectrometry strain typing during a large outbreak of Shiga-toxigenic *Escherichia coli*. *PLoS ONE*, 9, e101924.
- Clark, C.G., Kruczkiewicz, P., Guan, C., McCorrister, S.J., Chong, P., Wylie, J. et al. (2013) Evaluation of MALDI-TOF mass spectroscopy methods for determination of *Escherichia coli* pathotypes. *Journal of Microbiological Methods*, 94, 180–191.
- Crobach, M.J.T., Planche, T., Eckert, C., Barbut, F., Terveer, E.M., Dekkers, O.M. et al. (2016) European Society of Clinical Microbiology and Infectious Diseases: update of the diagnostic guidance document for *Clostridium difficile* infection. *Clinical Microbiology and Infection*, 22(Suppl 4), S63–S81.
- Croxatto, A., Prod'hom, G. & Greub, G. (2012) Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology. *FEMS Microbiology Reviews*, 36, 380–407.
- Dekker, J.P. & Branda, J.A. (2011) MALDI-TOF mass spectrometry in the clinical microbiology laboratory. *Clinical Microbiology Newsletter*, 33, 87–93.
- Eckert, C., Devallière, T., Syed-Zaidi, R., Lalande, V. & Barbut, F. (2018) Evaluation of a novel molecular assay to diagnose toxigenic strains of *Clostridium difficile*. *Anaerobe*, 52, 111–114.
- Elbehry, A., Aldubaib, M., Abalkhail, A., Marzouk, E., Albeloushi, A., Moussa, I. et al. (2022) How MALDI-TOF mass spectrometry technology contributes to microbial infection control in healthcare settings. *Vaccine*, 10, 1881.
- Emele, M.F., Joppe, F.M., Riedel, T., Overmann, J., Rupnik, M., Cooper, P. 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.
- Fagerquist, C.K., Garbus, B.R., Miller, W.G., Williams, K.E., Yee, E., Bates, A.H. et al. (2010) Rapid identification of protein biomarkers of *Escherichia coli* O157:H7 by matrix-assisted laser desorption ionization–time-of-flight–time-of-flight mass spectrometry and top-down proteomics. *Analytical Chemistry*, 82, 2717–2725.
- Fergusson, C.H., Coloma, J.M.F., Valentine, M.C., Haeckl, F.P.J. & Lington, R.G. (2020) Custom matrix-assisted laser desorption ionization–time of flight mass spectrometric database for identification of environmental isolates of the Genus *Burkholderia* and related genera. *Applied and Environmental Microbiology*, 86, e00354-20.
- Flores-Treviño, S., Garza-González, E., Mendoza-Olazarán, S., Morfín-Otero, R., Camacho-Ortiz, A., Rodríguez-Noriega, E. et al. (2019) Screening of biomarkers of drug resistance or

- virulence in ESCAPE pathogens by MALDI-TOF mass spectrometry. *Scientific Reports*, 9, 18945.
- Gagnaire, J., Dauwalder, O., Boisset, S., Khau, D., Freyrière, A.M., Ader, F. et al. (2012) Detection of *Staphylococcus aureus* delta-toxin production by whole-cell MALDI-TOF mass spectrometry. *PLoS ONE*, 7, e40660.
- Garrigos, T., Neuwirth, C., Chapuis, A., Bador, J., Amoureux, L., Andre, E. et al. (2021) Development of a database for the rapid and accurate routine identification of *Achromobacter* species by matrix-assisted laser desorption/ionization-time-of-flight mass spectrometry (MALDI-TOF MS). *Clinical Microbiology and Infection*, 27, 126.e1–e5.
- Gibb, S. & Franceschi, P. (2022) *MALDIquantForeign: Import/Export Routines for "MALDIquant"*.
- Gibb, S. & Strimmer, K. (2012) Maldiquant: a versatile R package for the analysis of mass spectrometry data. *Bioinformatics*, 28, 2270–2271.
- Godmer, A., Benzerara, Y., Normand, A.C., Veziris, N., Gallah, S., Eckert, C. et al. (2021) Revisiting species identification within the *Enterobacter cloacae* Complex by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *Microbiology Spectrum*, 9, e0066121.
- Godmer, A., Benzerara, Y., Veziris, N., Matondo, M., Aubry, A. & Gianetto, Q.G. (2022) MSclassifR: An R package for supervised classification of mass spectra with machine learning methods. *bioRxiv*. Available from: <https://doi.org/10.1101/2022.03.14.484252>
- Godmer, A., Gianetto, Q.G., Neindre, K.L., Aubry, A., Veziris, N., Barbut, F. et al. (2023) *Spectra from "Contribution of MALDI-TOF mass spectrometry and Machine Learning including Deep Learning techniques for the detection of virulence factors of Clostridioides difficile strains"*.
- Huang, Y., Li, J., Gu, D., Fang, Y., Chan, E.W., Chen, S. et al. (2015) Rapid detection of K1 hypervirulent *Klebsiella pneumoniae* by MALDI-TOF MS. *Frontiers in Microbiology*, 6, 1435.
- Keras: Deep Learning for humans. (n.d.) *Keras: Deep Learning for humans*. Available from: <https://keras.io/>
- Kubo, Y., Ueda, O., Nagamitsu, S., Yamanishi, H., Nakamura, A. & Komatsu, M. (2021) Novel strategy of rapid typing of Shiga toxin-producing *Escherichia coli* using MALDI Biotyper and ClinProTools analysis. *Journal of Infection and Chemotherapy*, 27, 1137–1142.
- Kuhn, M. (2008) Building predictive models in R using the caret package. *Journal of Statistical Software*, 28, 1–26.
- Li, D., Yi, J., Han, G. & Qiao, L. (2022) MALDI-TOF mass spectrometry in clinical analysis and research. *ACS Measurement Science Au*, 2, 385–404.
- Li, R., Xiao, D., Yang, J., Sun, S., Kaplan, S., Li, Z. et al. (2018) Identification and characterization of *Clostridium difficile* sequence type 37 genotype by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *Journal of Clinical Microbiology*, 56, e01990-17.
- Mazzeo, M.F., Sorrentino, A., Gaita, M., Cacace, G., Di Stasio, M., Facchiano, A. et al. (2006) Matrix-assisted laser desorption ionization-time of flight mass spectrometry for the discrimination of food-borne microorganisms. *Applied and Environmental Microbiology*, 72, 1180–1189.
- Mohammad, N., Normand, A.-C., Nabet, C., Godmer, A., Brossas, J.-Y., Blaize, M. et al. (2023) Improving the detection of epidemic clones in *Candida parapsilosis* outbreaks by combining MALDI-TOF mass spectrometry and deep learning approaches. *Microorganisms*, 11, 1071.
- Popović, N.T., Kazazić, S.P., Bojanić, K., Strunjak-Perović, I. & Čož-Rakovac, R. (2023) Sample preparation and culture condition effects on MALDI-TOF MS identification of bacteria: a review. *Mass Spectrometry Reviews*, 42(5), 1589–1603.
- Reil, M., Erhard, M., Kuijper, E.J., Kist, M., Zaiss, H., Witte, W. et al. (2011) Recognition of *Clostridium difficile* PCR-ribotypes 001, 027 and 126/078 using an extended MALDI-TOF MS system. *European Journal of Clinical Microbiology & Infectious Diseases*, 30, 1431–1436.
- Rodríguez-Temporal, D., Herrera, L., Alcaide, F., Domingo, D., Héry-Arnaud, G., van Ingen, J. et al. (2023) Identification of mycobacterium abscessus subspecies by MALDI-TOF mass spectrometry and machine learning. *Journal of Clinical Microbiology*, 61, e01110-22.
- Sauget, M., Nicolas-Chanoine, M.-H., Cabrolier, N., Bertrand, X. & Hocquet, D. (2014) Matrix-assisted laser desorption ionization-time of flight mass spectrometry assigns *Escherichia coli* to the phylogroups a, B1, B2 and D. *International Journal of Medical Microbiology*, 304, 977–983.
- Viprey, V.F., Davis, G.L., Benson, A.D., Ewin, D., Spittal, W., Vernon, J.J. 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 Surveillance*, 27, 2100704.
- Viprey, V.F., Granata, G., Vendrik, K.E.W., Davis, G.L., Petrosillo, N., Kuijper, E.J. et al. (2023) European survey on the current surveillance practices, management guidelines, treatment pathways and heterogeneity of testing of *Clostridioides difficile*, 2018–2019: results from the combatting bacterial resistance in Europe CDI (COMBACTE-CDI). *The Journal of Hospital Infection*, 131, 213–220.
- Warny, M., Pepin, J., Fang, A., Killgore, G., Thompson, A., Brazier, J. et al. (2005) Toxin production by an emerging strain of *Clostridium difficile* associated with outbreaks of severe disease in North America and Europe. *Lancet*, 366, 1079–1084.
- Weis, C., Cuénod, A., Rieck, B., Dubuis, O., Graf, S., Lang, C. et al. (2022) Direct antimicrobial resistance prediction from clinical MALDI-TOF mass spectra using machine learning. *Nature Medicine*, 28, 164–174.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Godmer, A., Gai Gianetto, Q., Le Neindre, K., Latapy, V., Bastide, M., Ehmig, M. et al. (2024) 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. Available from: <https://doi.org/10.1111/1751-7915.14478>