Analysis of high-molecular-weight proteins using MALDI-TOF MS and Machine Learning for the differentiation of clinically relevant *Clostridioides difficile* ribotypes

Ana Candela[1*], David Rodriguez-Temporal[2*], Mario Blázquez-Sánchez[2], Manuel J. Arroyo[3], Mercedes Marín[2,4], Luis Alcalá[2,4], Germán Bou[1,5], Belén Rodríguez-Sánchez[2†] and Marina Oviaño[1,5†]

[1]Clinical Microbiology Department, Complexo Hospitalario Universitario A Coruña, Institute of Biomedical Research A Coruña (INIBIC) A Coruña, Spain; [2]Clinical Microbiology and Infectious Diseases Department, Hospital General Universitario Gregorio Marañón and Institute of Health Research Gregorio Marañón (IiSGM) Madrid, Spain; [3]Clover Bioanalytical Software, Av. del Conocimiento, 41, 18016 Granada, Spain; [4]CIBER de Enfermedades Respiratorias (CIBERES CB06/06/0058), Madrid 28007, Spain; [5]CIBER de Enfermedades Infecciosas (CIBERINFEC CB21/13/00055).

*Correspondence: acandelagon@gmail.com and david.rodriguez@iisgm.com

†Both authors are the senior authors of this article

Abstract

Clostridioides difficile is the main cause of antibiotic related diarrhea and some ribotypes (RT), such as RT027, RT181 or RT078, are considered high risk clones. A fast and reliable approach for *C. difficile* ribotyping is needed for a correct clinical approach. This study analyses high-molecular-weight proteins for *C. difficile* ribotyping with MALDI-TOF MS. Sixty-nine isolates representative of the most common ribotypes in Europe were analyzed in the 17,000-65,000 *m/z* region and classified into 4 categories (RT027, RT181, RT078 and 'Other RTs'). Five supervised Machine Learning algorithms were tested for this purpose: K-Nearest Neighbors, Support Vector Machine, Partial Least Squares-Discriminant Analysis, Random Forest and Light-Gradient Boosting Machine. All algorithms yielded cross-validation results >70%, being RF and Light-GBM the best performing, with 88% of agreement. Area under the ROC curve of these two algorithms was >0.9. RT078 was correctly classified with 100% accuracy and isolates from the RT181 category could not be differentiated from RT027.

Keywords: *Clostridioides difficile*; MALDI-TOF MS; Machine Learning, Ribotyping

Keywords: MALDI-TOF, high-molecular-weight proteins, *Clostridioides difficile*, Machine Learning, ribotyping

Introduction

Clostridioides difficile is an anaerobic Gram-positive rod that can persist on surfaces and in the environment, and is resistant to most conventional disinfectants such as alcohol or chlorhexidine due to its ability to form spores. This makes it highly transmissible if good hygiene and infection control measures are not implemented (1). This microorganism is the main cause of antibiotic related diarrhea and represents a public health concern due to its high morbidity and mortality rates and its involvement in nosocomial outbreaks.

Some *C. difficile* ribotypes (RTs) have shown to be more virulent and/or involved in nosocomial outbreaks due to the production of toxins. *C. difficile* RT NAP1/B1/RT027 and the recently described RT181 (“027-like”) (2-4) also present a deletion of 1 bp at position 117 of the regulatory gene *tcdC*, a pathogenicity locus that downregulates the production of toxins. The consequence of this deletion is the hyperproduction of toxins, which makes these RTs more pathogenic and associated to more severe outcomes (5-7).

C. difficile characterization and ribotyping should be fast and reliable, to enable the correct antibiotic implementation and infection control measures. Gold Standard techniques in the USA and Europe are Pulsed Field Gel Electrophoresis (PFGE) and PCR Ribotyping, respectively (8). These techniques are laborious and require specialized personnel trained in molecular biology. Also, they are cumbersome and final results are obtained after several days. Among the commercially available molecular techniques, one of the most used is Xpert® *C. difficile* BT (Cepheid, Sunnyvale, CA, USA), which allows the

detection of toxin B, the binary toxin and the deletion at position 117 in gene *tcdC* (9). However, it has been shown that this test does not distinguish the different toxigenic RTs that host the deletion in gene *tcdC* (2, 10, 11).

Matrix-Assisted Laser Desorption/Ionization Time of Flight Mass Spectrometry (MALDI-TOF MS) represents an alternative to the previously described reference methods since it is a technology available in most microbiology laboratories nowadays, easy to use and with a time-around time of only some minutes. Apart from the initial investment for the acquisition of the instrument, the cost per sample is lower than that of molecular techniques. Besides identification, MALDI-TOF MS has been applied for bacterial typing and antibiotic resistance detection in several species like *Pseudomonas aeruginosa*, *Staphylococcus aureus*, *Streptococcus pneumoniae* or *Escherichia coli* (12-15). The default region for spectra analysis is between 2,000 and 20,000 *m/z* -where most ribosomal proteins are found-, although it can be modified to analyze a higher molecular weight range.

For *C. difficile*, few studies have evaluated MALDI-TOF MS as a ribotyping alternative, and even fewer analyzed high molecular weight proteins. *C. difficile* external surface layer (S-layer) is composed of proteins with a molecular weight ranging from 40 to 200 kDa (16, 17). The study of this mass region increases the number of peaks available, expanding the chances for differentiation among different isolates. The aim of this study was to develop a rapid MALDI-TOF MS method for the differentiation of hypervirulent *C. difficile* RTs based the analysis of high molecular weight proteins.

Materials and Methods

Bacterial isolates and molecular characterization

A total of n=69 *C. difficile* isolates representing the most prevalent RTs in Spain and Europe were included in this study. All strains were isolated from clinical stool samples in Spain and ribotyped at Hospital General Universitario Gregorio Marañón, in Madrid, Spain. Clinical samples were directly analyzed upon reception with the commercial PCR Xpert® *C. difficile* BT (Cepheid, Sunnyvale, CA, USA) and cultured in CLO agar (Beckton Dickinson®, Franklin Lakes, NJ, US) for *C. difficile* isolation and ribotyping. After bacterial growth, isolates were identified by MALDI-TOF MS in an MBT Smart MALDI Biotyper (Bruker Daltonics, Bremen, Germany) with the updated database containing 11,096 mass spectra profiles. The standard on-plate protein extraction was applied with 1µl 100% formic acid followed by 1µl HCCA (α-Cyano-4-hydroxycinnamic acid) matrix solution.

Ribotyping was performed by PCR amplification of the intergenic spacer region between the 16S rRNA and the 23S rRNA followed by capillary electrophoresis. The results of this sequencing were interpreted with Bionumerics 5.0 software (bioMérieux, Marcy l'Etoile, France) (18-21). Isolates included in this study belonged to the following RTs (Table S1): RT027 (n=29), RT181 (n=7), RT078 (n=21) and n=12 strains from other less toxigenic RTs (RT001, RT106, RT207, RT014 and RT023).

MALDI-TOF MS spectra acquisition

For spectra acquisition, isolates were plated on Schaedler agar (Beckton Dickinson®, Franklin Lakes, NJ, US) and incubated at 37°C in anaerobic atmosphere for 48h. A few colonies were spotted on the MALDI target plate in duplicates and overlaid with 1µl of trans-ferulic acid matrix at a concentration of 15 mg/ml and dissolved in a solution of 33% acetonitrile, 17% formic acid and 50% water, as described previously (22, 23). Trans-ferulic acid matrix forms crystals upon drying (Figure 1) and allows the analysis of proteins in a higher molecular weight region, yielding larger mass peaks and higher signal-to-baseline ratios in comparison with other organic matrices (22).

Spectra were obtained manually in linear positive mode, reading over the formed crystals and acquired in the region 17,000-65,000 *m/z*, in pulses of 200 shots, for a summation of 800 spectra per strain and spot (Detector Gain 2.700V-2.800V, laser intensity 90%). Spectra were calibrated with the commercial calibrator PSII (Protein Standard II, Bruker Daltonics, Bremen, Germany).

MALDI-TOF MS spectra analysis and data modeling

High molecular weight spectra were analyzed with the commercial platform Clover MS Data Analysis Software (MSDAS, https://www.clovermsdataanalysis.com; Clover Bioanalytical Software, Granada, Spain). Spectra were processed by a pipeline of: a) baseline subtraction using Top-Hat filter (factor 0.02), b) smoothing via Savitzky-Golay filter (window length: 11; polynomial order: 3), c) peak alignment with constant mass tolerance of 3 Da and linear mass tolerance of 600 ppm and d) Total Ion

Current (TIC) normalization. Peaks were aligned in the 17,000-65,000 *m/z* region of the spectra and then merged in an average spectrum for each isolate. Full spectra were studied for this purpose.

Initial approach was to perform unsupervised algorithms to study the feasibility of using MALDI-TOF for the differentiation of *C. difficile* RTs analyzing the high molecular weight region. For that aim, hierarchical clustering with Principal Component Analysis (PCA) was performed. All isolates were included in this initial model and their high molecular weight spectra were compared.

After this initial study, several well-known supervised classification algorithms were applied: Random Forest (RF), Light-Gradient Boosting Machine (Light-GBM), Partial Least Squares-Discriminant Analysis (PLS-DA), Support Vector Machine (SVM) and K-Nearest Neighbor (KNN). These algorithms were trained by using a k-fold cross validation (k=5) as internal validation, meaning that 20% of the samples were randomly extracted from the model and blindly validated against it, on 5 different times. Hyperparameters applied for the construction of the models are summarized on Table S2.

Area under the receiver operating characteristic (AUROC) curve was obtained from these algorithms to evaluate their discrimination power for each category. The AUROC measures the ability of the model to discriminate among classes (i.e., higher values indicate greater ability) (24). A peak matrix with the full spectrum of each isolate (from 17,000 to 65,000 *m/z*) was employed as training data set. Biomarker analysis was applied to identify putative biomarkers capable of discriminating each category. For this purpose, ANOVA analysis and AUROC were calculated for all peaks found using a threshold of 1% of maximum intensity for peak selection.

Results

All isolates were identified as *C. difficile* by MALDI-TOF MS with a score >2.0. Initial study by unsupervised methods showed three main groups in hierarchical clustering. First, a clear clustering of RT078, a second one composed of mainly isolates from the "Other RTs" category, and a third cluster that grouped together most isolates from categories RT027 and RT181 (Figure 2).

All supervised algorithms evaluated yielded accuracy greater than 75%. The highest accuracy results for RT differentiation were obtained with the algorithms RF and Light-GBM. 5-fold cross validation of these two models yielded an accuracy of almost 90%. Differentiation of RT027 was possible with all algorithms with an accuracy >70% and a Positive Predictive Value (PPV) of 80.0% and 86.7% for RF and Light-GBM, respectively. RT078 could clearly be separated from the rest of RTs with an accuracy of 100% with all algorithms, and a PPV of 100% for RF and 95.5% for Light-GBM, with one RT027 isolate classified as RT078. Separation of RT181 from RT027 could not be achieved. Finally, "Other RTs" category also obtained an accuracy greater than 90% with RF and Light-GBM, and a PPV of 92.3% and 100%, respectively. Results of the 5-fold cross validation and distance plots of the algorithms are summarized in Table 1 and Figure 3, respectively.

AUROC was greater than 0.85 for all categories with all trained algorithms, except for RT181 because it could not be separated from RT027.

With the models that yielded best results (RF and Light-GBM) area under the ROC curves was greater than 0.9 for RT027, RT078 and the category “Other RTs” (Figure 4). These results for AUROC curves imply that all categories but RT181 can be differentiated with high accuracy from the rest of categories.

RT078 could be differentiated from all other RTs thanks to three putative biomarker peaks (19,222 *m/z*, 33,562 *m/z* and 41,253 *m/z*) present in these isolates and with higher intensities than in the rest (Figure 5), which AUROC = 1 and p-values in ANOVA analysis lower than 0.05 (4.85·e-25, 1.63·e-36 and 2.54·e-27, respectively). Whereas peak 19,222 *m/z* was the unique peak in this region, peaks 33,562 and 41,253 *m/z* appeared as a shift of peaks 33,840 and 40,722 *m/z*, respectively, which were present in other RTs.

Discussion

In this study, we evaluated the ability of MALDI-TOF MS to classify hypervirulent strains analyzing proteins with higher molecular mass than those usually evaluated for identification purposes. Our results showed that ribotyping with MALDI-TOF MS is feasible, differentiating relevant RTs like RT027 and RT078, and that this automated approach reduces the time to obtain results from several days to a few hours from the isolation of the microorganism in culture. Furthermore, the application of this technique is relatively simple, with sample preparation being identical to the preparation for routine identification using MALDI-TOF MS. Machine Learning algorithms such as RF and Light-GBM were the best to perform, both achieving a 5-fold cross validation of 88%.

External validation could not be carried out due to the limited number of isolates.

The results obtained in this study with MALDI-TOF MS can compare to the molecular assay Xpert® *C. difficile* BT performance, as it can separate RT027 from other clinically relevant RTs, although not from RT181. RT181 is a similar RT to RT027, as it produces toxin B, binary toxin and presents the same deletion in the regulatory gene *tcdC*. Differentiation between these two RTs can be only achievable with PCR ribotyping, which can take up to a week. The clinical implications of the differentiation between these two RTs are yet unknown as RT181 has been recently described and literature about this RT is still limited (2). Three biomarker peaks were found in our study for RT078 differentiation (19,222 *m/z*, 33,562 *m/z* and 41,253 *m/z*). Two of them, were described in a previous study (33,600±200 *m/z* and 41,375±125 *m/z*) as specific for RT078 (25). They appeared next to other peaks present in other RTs (33,840 and 40,722 *m/z*), with a shift of 300-500 *m/z*. They could correspond to different forms of the same protein, with variations between different RTs, although this was not confirmed. The peak at 19,222 *m/z* has not been described before to our knowledge. The identification of these three peaks could be applied as a fast tool for RT078 ribotyping. No other biomarkers of interest were found for the differentiation of the rest of categories.

Several studies have been published trying to ribotype *C. difficile* with MALDI-TOF MS, but the majority of them analyze the default region of identification, between 2,000 and 20,000 *m/z*, with variable results (8, 26-29). Two studies analyzed a higher molecular weight region with MALDI-TOF MS (up to 80,000 *m/z*). They achieved *C. difficile* typing by creating an internal

238 database according to their own mass spectra (what they define the “High
239 Molecular Weight Profile” -HMW Profile-), which then was compared to PCR
240 ribotyping (25, 30). They could not clearly correlate their HMW profiles with
241 conventional ribotyping, as some HMW Profiles include isolates from different
242 conventional RTs and at the same time, some conventional RTs include
243 isolates from different HMW profiles. They concluded that HMW ribotyping
244 could be a useful tool for *C. difficile* outbreak detection as they detected isolates
245 with the same HMW profile in different clinical outbreaks.

246 In this study *C. difficile* typing was achieved, directly correlating MALDI-
247 TOF MS results with what is considered the Gold Standard technique in
248 Europe, PCR ribotyping of the intergenic region between 16S rRNA and 23S
249 rRNA. This allows for extrapolation and comparison of results with other centers
250 that perform this technique and have a MALDI-TOF MS instrument available.

251 One of the flaws of this study is the limited number of isolates available,
252 although they belong to a multicentric collection and were isolated from several
253 hospitals in Spain. Strains were selected according to national and European
254 epidemiology, representing the most common RTs, but were isolated only from
255 the Spanish territory. A broader and more representative number of isolates
256 from Europe may be needed to validate this study and confirm the results. It is
257 possible that with an increase in the number of strains analyzed, separation of
258 RT181 from RT027 also improves.

259 The methodology developed in this study is a valuable tool for the rapid
260 discrimination of clinically relevant isolates of *C. difficile* and opens a new path
261 for future typing studies with MALDI-TOF MS and other clinically relevant
262 bacteria. The implementation of MALDI-TOF MS significantly reduces the time

and costs for *C. difficile* ribotyping in comparison with reference methods, allowing a better optimization of hospital resources and prompt initiation of treatment according to the characterized RT, as well as more efficient and cost-effective control of the infection.

Conflicts of Interests

MJA is employed by Clover Bioanalytical Software (Granada, Spain) but had no role in the design of the study or methodology. The rest of authors declare no conflicts of interest.

Acknowledgments

This work is partially supported by the project PI18/00997 and PI20/00686 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 de-sign, 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.

Figure Legends

Figure 1. a) Example of crystallization of Trans-ferulic acid matrix on the MALDI plate; b) Zoomed-in portion of a MALDI spot during spectra acquisition.

Figure 2. Dendrogram built with all isolates using Euclidean distance and Ward metric. Three main clusters could be observed. Dimensionality reduction was automatically applied to reach 95% of variance. In this experiment, there were automatically included 56 principal components.

Figure 3. Distance plots of the algorithms studied: a) Random Forest; b) K-Nearest Neighbor; c) Partial Least Squares-Discriminant Analysis; d) Support Vector Machine.

Figure 4. Area Under the Receiver Operating Characteristic (AUROC) curve for: a) Light-Gradient Boosting Machine algorithm; and b) Random Forest algorithm.

Figure 5. Specific biomarker peaks for the differentiation of RT078: 19,222, 22,562 and 41,253 *m/z*.

Author Contributions: Conceptualization, AC, DRT, MO and BRS; Methodology, AC, MM and LA; software, MJA; formal analysis, AC and MJA; writing—original draft preparation, AC and DRT; writing—review and editing, GB, MO and BRS; supervision, MO and BRS; project administration, MO and BRS; funding acquisition, MO and BRS. All authors have read and agreed to the published version of the manuscript.

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. European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology 38:1211-1221.

2. 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.
3. 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.
4. Baktash A, Corver J, Harmanus C, Smits WK, Fawley W, Wilcox MH, Kumar N, Eyre DW, Indra A, Mellmann A, Kuijper EJ. 2022. Comparison of Whole-Genome Sequence-Based Methods and PCR Ribotyping for Subtyping of *Clostridioides difficile*. Journal of clinical microbiology 60:e0173721.
5. Markovska R, Dimitrov G, Gergova R, Boyanova L. 2023. *Clostridioides difficile*, a New "Superbug". Microorganisms 11.
6. Burnham CA, Carroll KC. 2013. Diagnosis of *Clostridium difficile* infection: an ongoing conundrum for clinicians and for clinical laboratories. Clinical microbiology reviews 26:604-30.
7. Wolff D, Bruning T, Gerritzen A. 2009. Rapid detection of the *Clostridium difficile* ribotype 027 tcdC gene frame shift mutation at position 117 by real-time PCR and melt curve analysis. European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology 28:959-62.
8. Calderaro A, Buttrini M, Martinelli M, Farina B, Moro T, Montecchini S, Arcangeletti MC, Chezzi C, De Conto F. 2021. Rapid Classification of *Clostridioides difficile* Strains Using MALDI-TOF MS Peak-Based Assay in Comparison with PCR-Ribotyping. Microorganisms 9.
9. 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. Brazilian journal of microbiology : [publication of the Brazilian Society for Microbiology] 52:1937-1949.
10. 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. Journal of clinical medicine 9.
11. Chapin KC, Dickenson RA, Wu F, Andrea SB. 2011. Comparison of five assays for detection of *Clostridium difficile* toxin. The Journal of molecular diagnostics : JMD 13:395-400.
12. Mulet X, Garcia R, Gaya M, Oliver A. 2019. O-antigen serotyping and MALDI-TOF, potentially useful tools for optimizing semi-empiric antipseudomonal treatments through the early detection of high-risk clones. European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology 38:541-544.
13. Sauget M, van der Mee-Marquet N, Bertrand X, Hocquet D. 2016. Matrix-assisted laser desorption ionization-time of flight Mass spectrometry can detect *Staphylococcus aureus* clonal complex 398. Journal of microbiological methods 127:20-23.
14. Pinto TC, Costa NS, Castro LF, Ribeiro RL, Botelho AC, Neves FP, Peralta JM, Teixeira LM. 2017. Potential of MALDI-TOF MS as an alternative approach for capsular typing *Streptococcus pneumoniae* isolates. Scientific reports 7:45572.
15. Chui H, Chan M, Hernandez D, Chong P, McCorrister S, Robinson A, Walker M, Peterson LA, Ratnam S, Haldane DJ, Bekal S, Wylie J, Chui L, Westmacott G, Xu B, Drebot M, Nadon C, Knox JD, Wang G, Cheng K. 2015. Rapid, Sensitive, and Specific *Escherichia coli* H Antigen Typing by Matrix-Assisted Laser Desorption Ionization-Time of Flight-Based Peptide Mass Fingerprinting. Journal of clinical microbiology 53:2480-5.
16. Mauri PL, Pietta PG, Maggioni A, Cerquetti M, Sebastianelli A, Mastrantonio P. 1999. Characterization of surface layer proteins from *Clostridium difficile* by liquid

chromatography/electrospray ionization mass spectrometry. Rapid communications in mass spectrometry : RCM 13:695-703.

17. Qazi O, Hitchen P, Tissot B, Panico M, Morris HR, Dell A, Fairweather N. 2009. Mass spectrometric analysis of the S-layer proteins from *Clostridium difficile* demonstrates the absence of glycosylation. Journal of mass spectrometry : JMS 44:368-74.
18. 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. Antimicrobial agents and chemotherapy 59:586-9.
19. Stubbs SL, Brazier JS, O'Neill GL, Duerden BI. 1999. PCR targeted to the 16S-23S rRNA gene intergenic spacer region of *Clostridium difficile* and construction of a library consisting of 116 different PCR ribotypes. Journal of clinical microbiology 37:461-3.
20. 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.
21. Reigadas E, Alcala L, Gomez J, Marin M, Martin A, Onori R, Munoz P, Bouza E. 2018. Breakthrough *Clostridium difficile* Infection in Cirrhotic Patients Receiving Rifaximin. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 66:1086-1091.
22. Madonna AJ, Basile F, Ferrer I, Meetani MA, Rees JC, Voorhees KJ. 2000. On-probe sample pretreatment for the detection of proteins above 15 KDa from whole cell bacteria by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Rapid communications in mass spectrometry : RCM 14:2220-9.
23. Meetani MA, Voorhees KJ. 2005. MALDI mass spectrometry analysis of high molecular weight proteins from whole bacterial cells: pretreatment of samples with surfactants. Journal of the American Society for Mass Spectrometry 16:1422-1426.
24. Gato E, Arroyo MJ, Mendez G, Candela A, Rodino-Janeiro BK, Fernandez J, Rodriguez-Sanchez B, Mancera L, Arca-Suarez J, Beceiro A, Bou G, Oviano M. 2023. Direct Detection of Carbapenemase-Producing *Klebsiella pneumoniae* by MALDI-TOF Analysis of Full Spectra Applying Machine Learning. Journal of clinical microbiology 61:e0175122.
25. Rizzardi K, Akerlund T. 2015. High Molecular Weight Typing with MALDI-TOF MS - A Novel Method for Rapid Typing of *Clostridium difficile*. PloS one 10:e0122457.
26. 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.
27. Carneiro LG, Pinto TCA, Moura H, Barr J, Domingues R, Ferreira EO. 2021. MALDI-TOF MS: An alternative approach for ribotyping *Clostridioides difficile* isolates in Brazil. Anaerobe 69:102351.
28. Cheng JW, Liu C, Kudinha T, Xiao M, Yu SY, Yang CX, Wei M, Liang GW, Shao DH, Kong F, Tong ZH, Xu YC. 2018. Use of matrix-assisted laser desorption ionization-time of flight mass spectrometry to identify MLST clade 4 *Clostridium difficile* isolates. Diagnostic microbiology and infectious disease 92:19-24.
29. Li R, Xiao D, Yang J, Sun S, Kaplan S, Li Z, Niu Y, Qiang C, Zhai Y, Wang X, Zhao X, Zhao B, Welker M, Pincus DH, Jin D, Kamboj M, Zheng G, Zhang G, Zhang J, Tang YW, Zhao J. 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.
30. Ortega L, Ryberg A, Johansson A. 2018. HMW-profiling using MALDI-TOF MS: A screening method for outbreaks of *Clostridioides difficile*. Anaerobe 54:254-259.

414

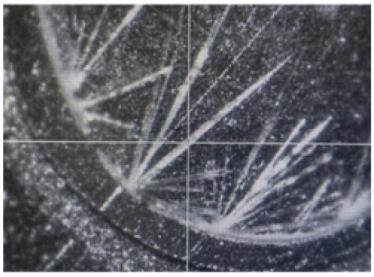
Table 1. 5-fold cross validation of the different algorithms studied, showing the accuracy for each one of the categories of the model and the total accuracy.

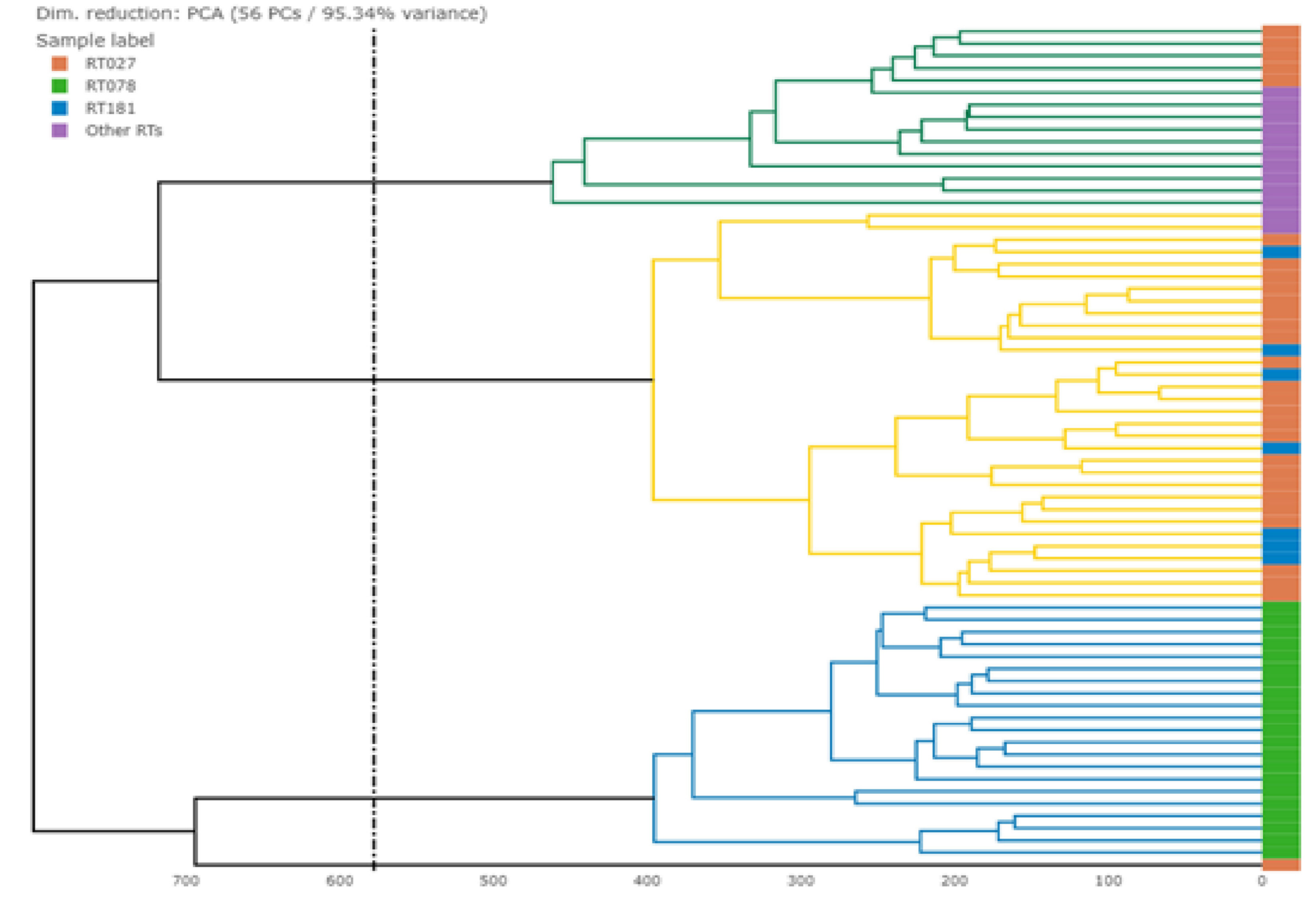
% Correct	Number of isolates	KNN	PCA-SVM	Light-GBM	PLS-DA	RF
RT027	n=29	79.3%	72.4%	89.7%	96.5%	96.5%
RT078	n=21	100%	100%	100%	100%	100%
RT181	n=7	0%	57.1%	42.9%	0%	0%
Other RTs	n=12	66.7%	83.3%	91.7%	83.3%	100%
Total % Correct	n=69	75.4%	81.2%	**88.4%**	85.5%	**88.4%**

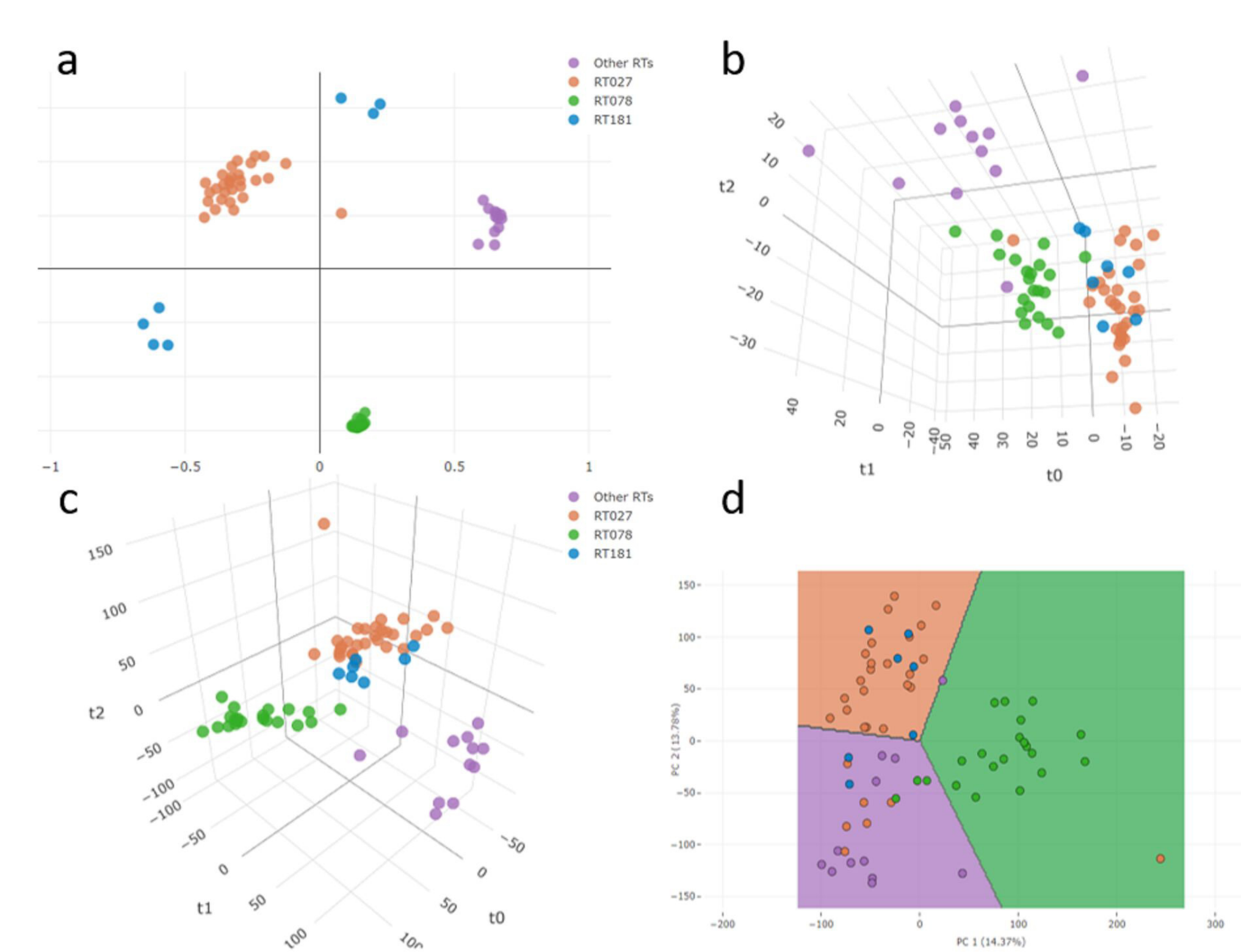
KNN: K-Nearest Neighbor; PCA-SVM: Principal Component Analysis-Support Vector Machine; Light-GBM: Light Gradient Boosting Machine; PLS-DA: Partial Least Squares Discriminant Analysis; RF: Random Forest.

a)

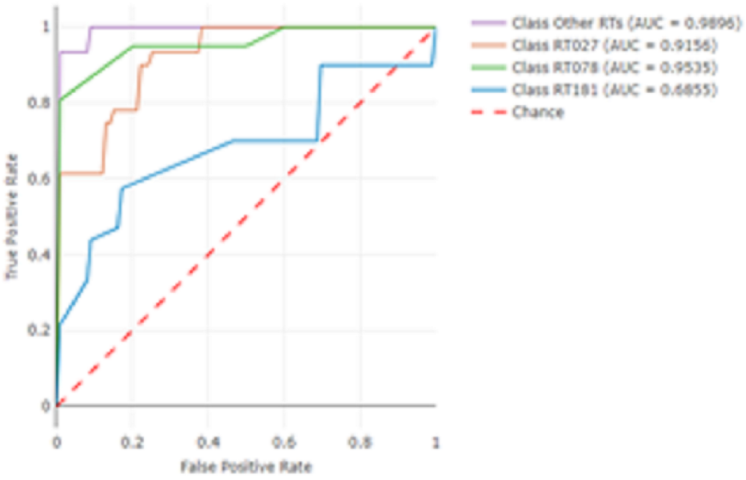
b)



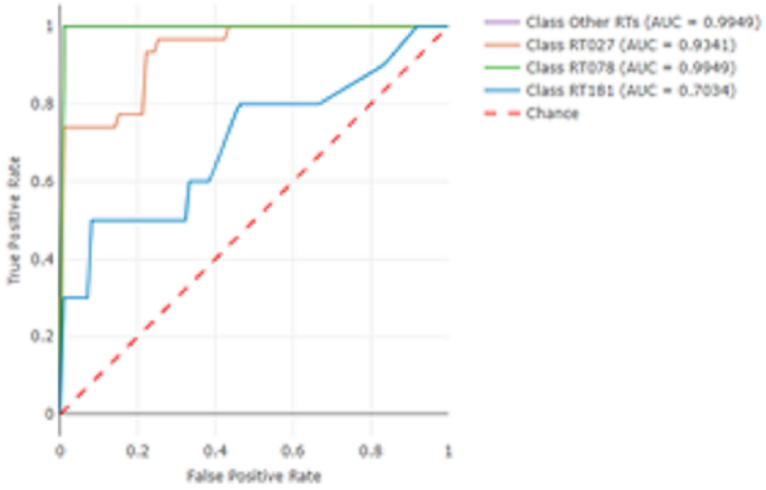

Dim. reduction: PCA (56 PCs / 95.34% variance)
Sample label
RT027
RT078
RT181
Other RTs
700
600
500
400
300
200
100
0

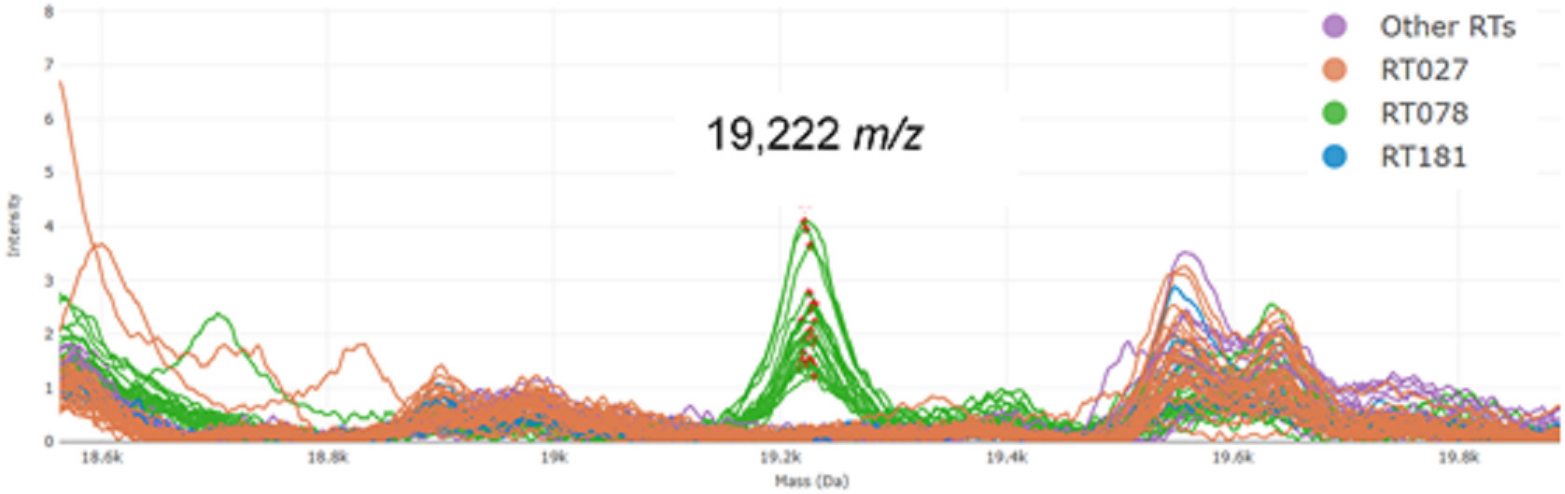

a
Other RTs
RT027
RT078
RT181
−1
−0.5
0
0.5
1
b
t2
t1
t0
c
Other RTs
RT027
RT078
RT181
t2
t1
t0
d
PC 2 (13.78%)
PC 1 (14.37%)

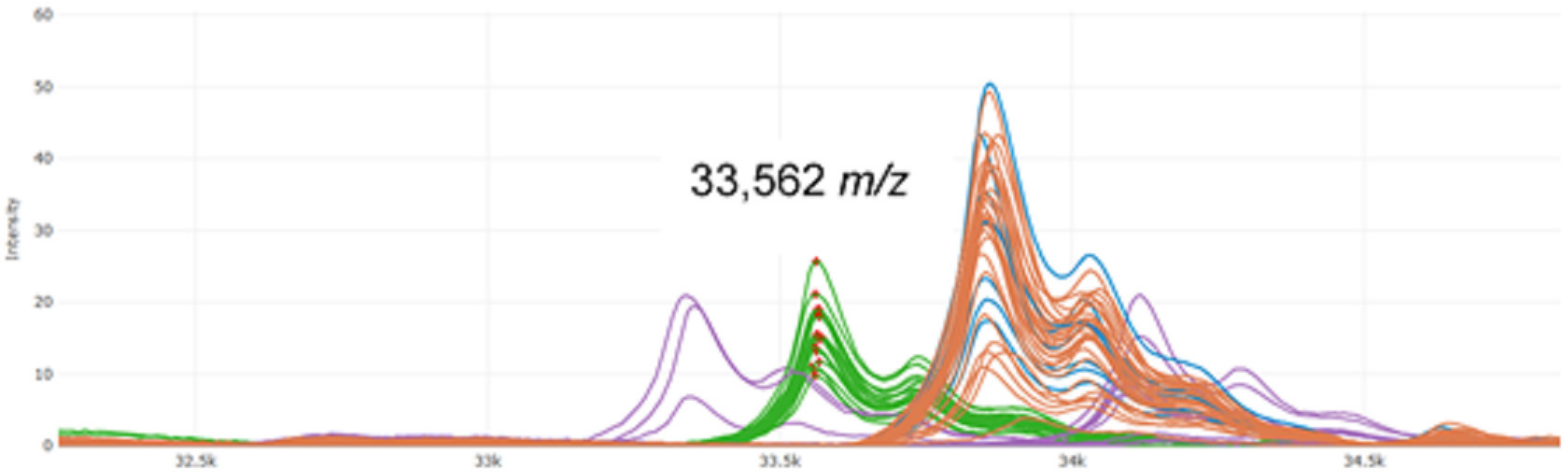
a)

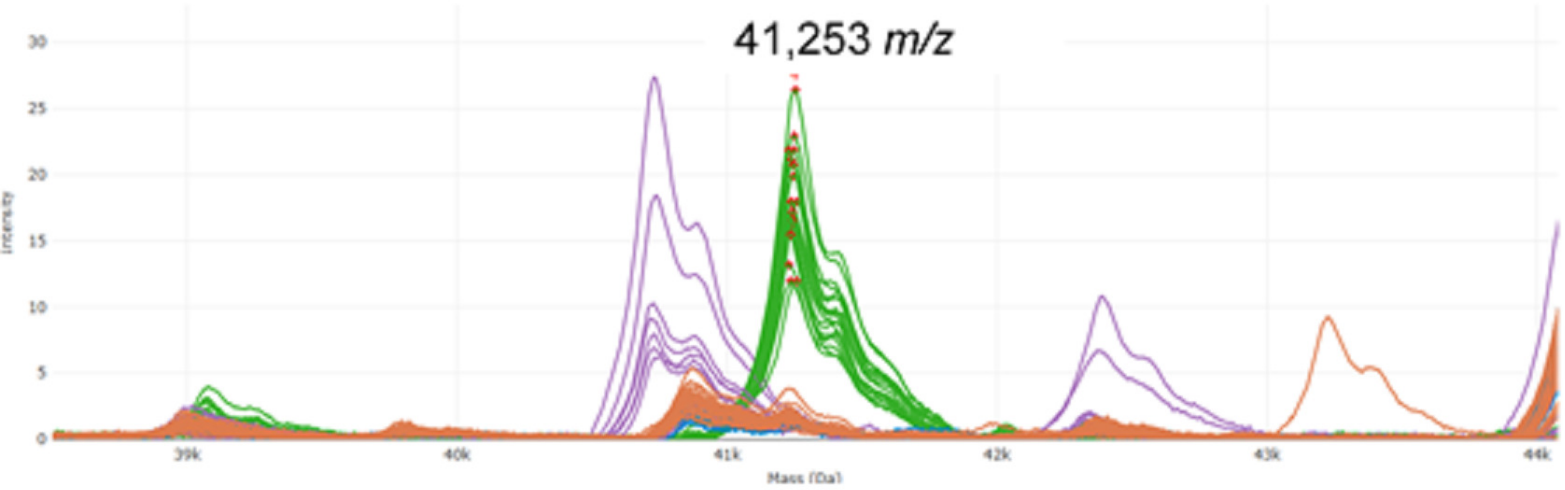

True Positive Rate
False Positive Rate
0
0.2
0.4
0.6
0.8
1
Class Other RTs (AUC = 0.9896)
Class RT027 (AUC = 0.9156)
Class RT078 (AUC = 0.9535)
Class RT181 (AUC = 0.6855)
Chance

b)


True Positive Rate
False Positive Rate
0
0.2
0.4
0.6
0.8
1
Class Other RTs (AUC = 0.9949)
Class RT027 (AUC = 0.9341)
Class RT078 (AUC = 0.9949)
Class RT181 (AUC = 0.7034)
Chance


Other RTs
RT027
RT078
RT181
19,222 m/z
Intensity
Mass (Da)


33,562 m/z
Intensity


41,253 m/z
Intensity
Mass (Da)