MAESTRÍA EN MICROBIOLOGÍA MOLECULAR

11va. Cohorte-Año 2019

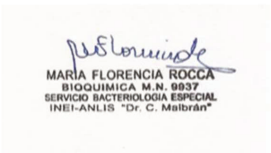
Tesis de Maestría:

Espectrometría de Masas MALDI-TOF MS como herramienta para el diagnóstico de enfermedades infecciosas.

Evaluación del potencial de MALDI-TOF MS acoplado al entrenamiento automatizado a través de un modelo predictivo que contribuya al diagnóstico de SARS-CoV-2.

Alumna: María Florencia Rocca

Directora: Rita Inés Armitano



Agradecimientos.

A mi familia, por acompañarme en este largo camino y esperarme en casa con una sonrisa.

A mis padres y hermanas, por su ayuda incondicional y por inculcarme el deseo y las ganas en cada proyecto.

A mi jefa Mónica Prieto, por darme la oportunidad de aprender y de llevar adelante este trabajo con total libertad y apoyo.

A Gisela Martínez y Gastón Dangiolo, por ser los mejores amigos y compañeros de vida, que me han dado fuerzas para continuar y me han sacado una sonrisa siempre.

A mi directora de tesis, Rita Armitano, por su acompañamiento desde el primer momento, por sus empujoncitos cuando flaqueaba en la escritura de esta tesis, por su presencia y valiosos aportes en cada corrección.

Este y todos mis logros son dedicados a Juana y Nina…

Indice general.

Contenido. **Página**

Indice de Tablas.

Contenido. **Página**

Indice de Figuras.

Contenido. **Página**

1. RESUMEN

La Espectrometría de Masas (EM) a través de la técnica MALDI-TOF MS, se ha utilizado desde hace varios años, para la detección de bacterias y hongos de relevancia clínica, además de sus diversas aplicaciones en proteómica, metabolómica, lipidómica, toxicología, endocrinología, genética y microbiología para caracterizar biomarcadores tales como proteínas, péptidos, lípidos, hormonas, metabolitos y nucleótidos. Aunque más recientemente, su uso se ha incrementado en el campo de la medicina y el diagnóstico clínico, siendo un enfoque que al día de hoy hace falta explorar.

Por otra parte, desde el mes de marzo del año 2020, el mundo se ha visto detenido por una nueva enfermedad, conocida como COVID-19, causada por el virus del síndrome respiratorio agudo severo coronavirus 2 (SARS-CoV-2). La detección temprana, sensible y específica del virus del SARS-CoV-2 es ampliamente reconocida como el punto crítico para responder al brote que preocupa al sistema de salud. Actualmente, el diagnóstico se basa fundamentalmente en técnicas moleculares de RT-PCR en tiempo real, aunque su implementación en los laboratorios, se ve amenazada por los altos costos y la extraordinaria demanda de insumos a nivel mundial. Es por eso, que el desarrollo de pruebas alternativas y / o técnicas complementarias se ha vuelto tan relevante.

En este proyecto, se explota el potencial de la tecnología de EM en combinación con algoritmos de aprendizaje automático, para la detección de perfiles proteicos característicos obtenidos a partir de muestras de hisopados nasofaríngeos de pacientes positivos y pacientes negativos de COVID-19. El procedimiento se propone de esta forma para simplificar la toma de muestra y el procesamiento de los datos, y para establecer una correlación directa entre el desempeño de los modelos propuestos basados en inteligencia artificial, con respecto a los resultados obtenidos mediante las técnicas de referencia actuales.

Según los resultados preliminares alcanzados a partir de este desarrollo (precisión = 67,66%, sensibilidad = 61,76%, especificidad = 71,72%, VPP =60,00%, VPN =73,20%), los métodos basados en espectrometría de masas junto con el análisis multivariado, demostraron que la proteómica es una herramienta interesante que merece ser explorada como un enfoque diagnóstico complementario de enfermedades infecciosas y no infecciosas, debido a su bajo costo y alto rendimiento. Sin embargo, se deben tomar en consideración pasos adicionales, como el análisis de un gran número de muestras y la implementación de técnicas de enriquecimiento y purificación, para evaluar la aplicabilidad del método desarrollado como técnica de tamizaje.

2. INTRODUCCIÓN

Generalidades.

Proteómica. El estudio de todos los genes de una persona (genoma), incluidas las interacciones de esos genes entre sí y con el entorno de la persona, se conoce como genómica. Esta ciencia ha proporcionado gran cantidad de información acerca de los genomas de un importante número de especies **(https://www.ncbi.nlm.nih.gov/genome/browse/)**. Sin embargo, aún se desconoce la función biológica de la mayoría de las proteínas codificadas por esos genes (**Westergren-Thorsson et al.**, 2006). Este hecho, junto con las mejoras en los métodos de separación proteica, como la cromatografía líquida y la electroforesis bidimensional, en combinación con el uso de la técnica de Espectrometría de Masas (EM), fueron el puntapié inicial para la aparición de la proteómica como tal (**Fenselau y Demirev,** 2001).

La palabra proteoma proviene de la fusión de **prote**ína y gen**oma**, acuñada por Marc Wilkins en la década del ´90 como una imagen dinámica. La proteómica es la ciencia que se basa en el estudio de las proteínas expresadas por un organismo o que están presentes en un medio biológico en un momento determinado y bajo condiciones concretas (**Wilkins et al.**, 1996). En general, una enfermedad no produce la alteración de una única proteína, por eso el estudio del conjunto puede resultar de utilidad para encontrar biomarcadores (BM) específicos que puedan ser usados como herramientas para el diagnóstico, pronóstico y seguimiento de enfermedades.

En el diseño de cualquier estudio de proteómica clínica se deben tener en cuenta algunos aspectos fundamentales tales como: 1) la selección correcta de los pacientes y los controles más adecuados; 2) la comparación de las muestras de pacientes con individuos sanos y con muestras procedentes de pacientes con perfiles similares al grupo de estudio, consideraciones necesarias para asegurar la especificidad de los BM potenciales hallados; 3) la selección del tipo de muestra, su recolección y almacenamiento; y 4) la clasificación del total de espectros adquiridos en los grupos de entrenamiento o descubrimiento y de validación posterior (**Righetti,** 2013).

Finalmente, pero no menos importante, se deberá llevar a cabo la comparación de los resultados surgidos de la proteómica con los del método patrón de oro (*gold standard*) para evaluar el rendimiento alcanzado y la posibilidad de automatizar al máximo el procesamiento y análisis de muestras futuras (**Mischak et al.**, 2007).

Espectrometría de Masas. La EM es una técnica analítica en la que los átomos o moléculas de una muestra son ionizados positivamente, separados por su relación masa/carga (*m/z*) y posteriormente detectados y registrados. Sus principales ventajas son, proporcionar una alta especificidad en la determinación del peso molecular debido a la posibilidad de medir exactamente su masa molecular, así como obtener información a partir de los fragmentos iónicos de un analito. Su sensibilidad es elevada y es muy versátil, ya que permite determinar la estructura de compuestos muy diversos. Es aplicable a todo tipo de muestras, volátiles, no volátiles, polares y apolares, sólidos, líquidos y gases (**Gonzalez de Buitrago y Ferreira**, 2006). En combinación con las denominadas técnicas de aprendizaje automatizado, es la más calificada para analizar los perfiles peptídicos de muestras complejas reales.

Tradicionalmente, la identificación de microorganismos se ha realizado por métodos basados en tinciones que permiten la clasificación de la morfología microscópica con el fin de apoyar decisiones diagnósticas y terapéuticas tempranas; así como también en pruebas *in vitro* apoyadas en reacciones bioquímicas mediante sistemas manuales o automatizados. Sin embargo, estos métodos fenotípicos tienen limitaciones asociadas a su dependencia de los procesos metabólicos de los microorganismos, ya que requieren de un cultivo con crecimiento adecuado y tiempos de incubación mínimos para alcanzar un resultado al cabo de varios días, lo cual impacta directamente sobre los aspectos terapéuticos y epidemiológicos. Otros métodos de identificación alternativos que superan las dificultades de los tradicionales han ganado espacio en el laboratorio de microbiología clínica; de este modo, las técnicas moleculares se han establecido como procedimientos complementarios a los métodos fenotípicos o incluso como referencia para la identificación de microorganismos, en especial la secuenciación del ácido ribonucleico ribosómico 16S (16S ARNr) o la detección de genes seleccionados por reacción en cadena de la polimerasa reversa en tiempo real (qRT-PCR, por sus siglas en inglés). Por su parte, el desarrollo de la tecnología de EM MALDI-TOF, ha permitido la utilización de la EM en la identificación de microorganismos completos mediante el análisis de proteínas, principalmente ribosomales, a través de la creación de un espectro de masas que es específico para cada género y especie (**Maldonado et al.,** 2018).

Su utilización se remonta a finales de la década del 80, cuando se observó que, mediante una fuente de láser ultravioleta y el sellado de la muestra con una matriz orgánica, se conseguía una ionización blanda para la detección de moléculas como proteínas, péptidos, azúcares y oligonucleótidos; este desarrollo mereció el Premio Nobel de Química en el año 2002, compartido por el químico estadounidense John Fenn y el ingeniero químico japonés Tanaka (**Tanaka et al.,** 1988).

Más adelante comenzó a ser aplicada para la identificación rápida de microorganismos intactos mediante el método de la huella peptídica a partir del desarrollo de bibliotecas proteicas que se integraron a programas informáticos con algoritmos de decisión (**Demirev et al,** 1999). Sin embargo, no fue hasta el año 2008 en que simultáneamente Mellman en Alemania y Degand en Francia, informan el desempeño de MALDI-TOF para la identificación de un gran número de Bacilos Gram Negativos No Fermentadores obtenidos de muestras clínicas (**Mellman et al.; Degand et al.**, 2008).
A partir de ese momento, aparecen numerosos trabajos describiendo que el rendimiento de MALDI-TOF es similar a realizar la secuenciación parcial del gen 16S ARNr, pero arrojando resultados confiables y más económicos en cuestión de minutos (**Seng et al.,** 2009).

MALDI-TOF MS. Por su sigla en inglés significa desorción ionización láser asistida por una matriz y *tof* alude al tiempo de vuelo de los iones hasta llegar al detector que está acoplado al equipo. Básicamente, es una tecnología que ofrece un patrón de proteínas de un organismo desconocido y compara el patrón generado, con una librería o base de datos, para finalmente presentar un valor de *score* o puntaje de esa comparación (**Claydon et al.,** 1996).

Plataformas comerciales. Actualmente existen dos sistemas disponibles en el mercado y ambos han sido aprobados por la Food and Drug Administration (FDA), a través de la norma 510k durante el año 2013, para su utilización en diagnóstico microbiológico clínico en humanos, sin embargo, esta validación solo incluyó los grupos taxonómicos más relevantes.
El sistema Vitek MS de Biomerieux (Francia), utiliza la base de datos Saramis desarrollada por AnagnosTec y comercializada por Shimadzu con los espectrómetros de masas Axima; en el año 2010 fue adquirida por Biomerieux para su incorporación a la plataforma comercial Vitek MS. En la actualidad, este sistema posee dos configuraciones: el modo IVD (*In Vitro Diagnostic*) y el modo RUO (*Research Use Only*) avalado para uso sólo en investigación y con una base de datos abierta, de este modo, el usuario puede incorporar perfiles de referencia. La plataforma Vitek MS cuenta con más de 15000 superespectros que equivalen a 200 géneros y 1046 especies.
Por otra parte, el sistema MALDI Biotyper fue creado y es comercializado por Bruker Daltonics (Alemania). El paquete del software Biotyper, acoplado a los espectrómetros de masas de la línea Flex, es una plataforma abierta que permite guardar los espectros creados por el usuario para expandir la base de datos comercial.
Cuenta con más de 9000 perfiles proteicos o MSPs (Mass Spectrum Profiles) que representan 540 géneros y 3000 especies.

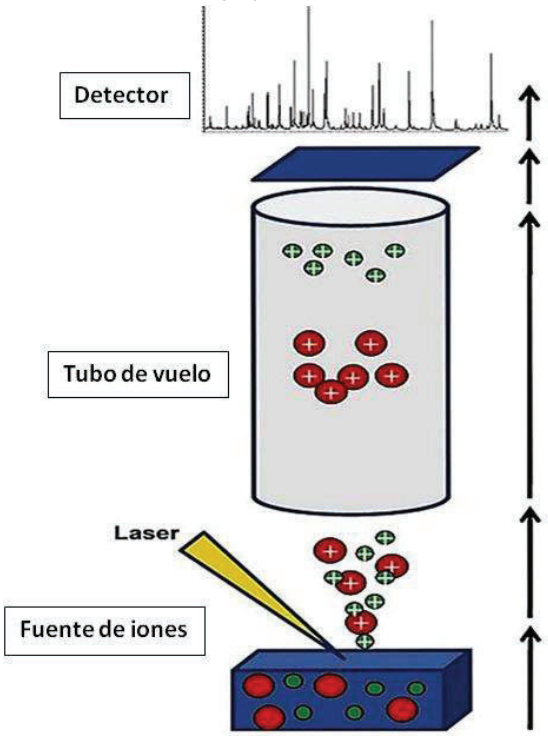
Aunque los principios técnicos de los sistemas Vitek MS y MALDI Biotyper son similares a grandes rasgos, existen diferencias en las bases de datos de referencia, en sus sistemas operativos y en los algoritmos empleados para la identificación, por lo que los resultados de ambos sistemas no son directamente comparables y las bases de datos no son transferibles entre plataformas (**Rocca et al.**, 2019).

Métodos de siembra. La biomasa microbiana depositada con un palillo de madera en la placa de acero debe cubrirse con la matriz recomendada por el fabricante, y se deja secar a temperatura ambiente.
Para algunas bacterias, la aplicación directa de la matriz a las células microbianas intactas (conocido como método directo de siembra) da como resultado una lisis celular adecuada, permitiendo el correcto análisis de las proteínas citoplasmáticas. Sin embargo, la lisis puede resultar insuficiente para organismos con paredes celulares gruesas, por ejemplo, levaduras, hongos filamentosos o micobacterias, donde se debe realizar un método de extracción. Este procedimiento puede ser *in situ* (agregando ácido fórmico al pocillo de muestra, previo al sellado con la matriz) o en tubo (agregando etanol-ácido fórmico-acetonitrilo y una lisis mecánica por vortex) y se realiza si no se obtuvieron resultados confiables por los otros métodos de siembra más sencillos (**Wayne,** 2017; **MALDI Biotyper 3.1 User Manual**).

Componentes del equipo. Los espectrómetros de masas están formados por tres elementos básicos. En primer lugar, la fuente de iones, donde a partir de la muestra cubierta con una matriz orgánica en un soporte adecuado, se forma un haz de iones en estado gaseoso. En segundo término, el separador de masas o tubo de vuelo, que separa los iones formados en función de su relación *m/z* y en tercera instancia el detector de esos iones (**Maldonado et al.**, 2018).

El esquema de las partes que componen un espectrómetro de masas se representa en la **Figura 1**.

Figura 1. Componentes principales de un Espectrómetro de Masas.
Adaptada de https://www.nanocell.org.br/maldi-tof-uma-ferramenta-revolucionaria-para-as-analises-clinicas-e-pesquisa-do-cancer/



Proceso de desorción-ionización. La muestra que se pretende identificar es depositada sobre el soporte metálico y mezclada con la matriz de baja masa que la cristaliza, estabilizándola; de esta forma las moléculas de la muestra quedan incorporadas dentro de la estructura de los cristales de la matriz.

Las matrices utilizadas, que deben contener en su estructura anillos aromáticos, son sólidos cristalinos con baja presión de vapor que pueden volatilizarse fácilmente para formar iones en el vacío. En el caso de MALDI-TOF MS que utiliza un láser UV, la molécula de la matriz también debe tener un cromóforo fuerte para ayudar a absorber energía y preservar la fragmentación proteica.

Existen diferentes matrices con diversas aplicaciones analíticas. La elección dependerá del tipo de muestra que se quiera analizar y del rango de lectura que se va a utilizar.

Las matrices más frecuentemente utilizadas son:

-Acido alfa-ciano-4-hidroxicinámico (HCCA): para péptidos, proteínas de 2000 a 20000 Da.

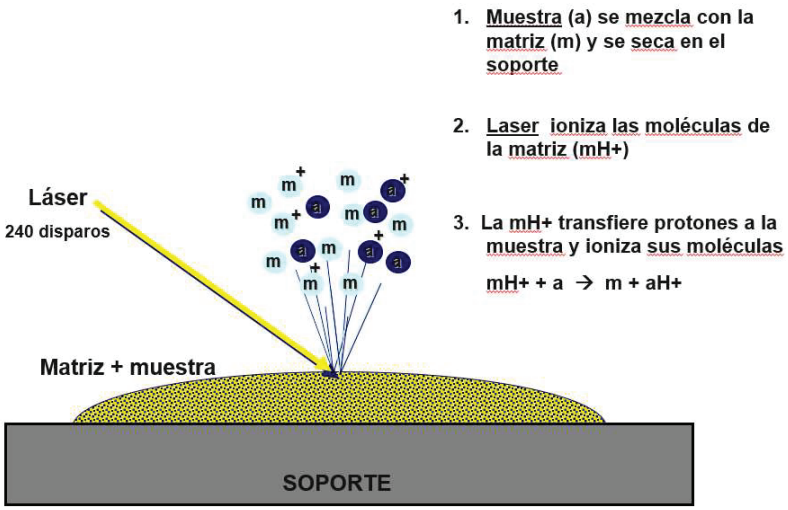
-Acido 2,5-dihidroxibenzoico (DHB): para moléculas pequeñas, productos de hidrólisis.

-Ácido sinapínico (SA): para proteínas enteras, hasta 150000 Da.

El HCCA es la matriz de elección para el diagnóstico de rutina y la que está disponible en los laboratorios clínicos, debido a que presenta el mejor rendimiento en la generación de perfiles para la identificación de microorganismos patógenos.

Esta etapa del procedimiento de co-cristalización resulta esencial para conseguir la posterior desorción-ionización suave y eficiente dada por el láser (**Marvin et al.,** 2003), cuyos pulsos deben ser de intensidad y duración determinadas para lograr una ionización blanda (energía entre 106 y 107 W/cm cuadrados y duración de 1 a 5 nanosegundos), en la longitud de onda del UV a 337nm, con lo que las moléculas orgánicas aromáticas de la matriz absorben una gran cantidad de energía por excitación de los electrones produciéndose la sublimación del analito y de la matriz (**Cortez et al.**). Ya en fase gaseosa, la estabilización de estas moléculas aromáticas tiene lugar por adición de protones que, en parte son captados por la muestra, generándose fragmentos de proteínas con carga positiva. El funcionamiento de la fuente de ionización se puede ver en la **Figura 2**.

Figura 2. Pasos del proceso de desorción-ionización de la muestra.
Adaptado de https://present5.com/maldi-tof-matrix-assisted-laser-desorption-ionization-time/



Analizador de tiempo de vuelo. El principio de los espectrómetros de tiempo de vuelo se basa en la relación entre la masa y la velocidad de los iones.

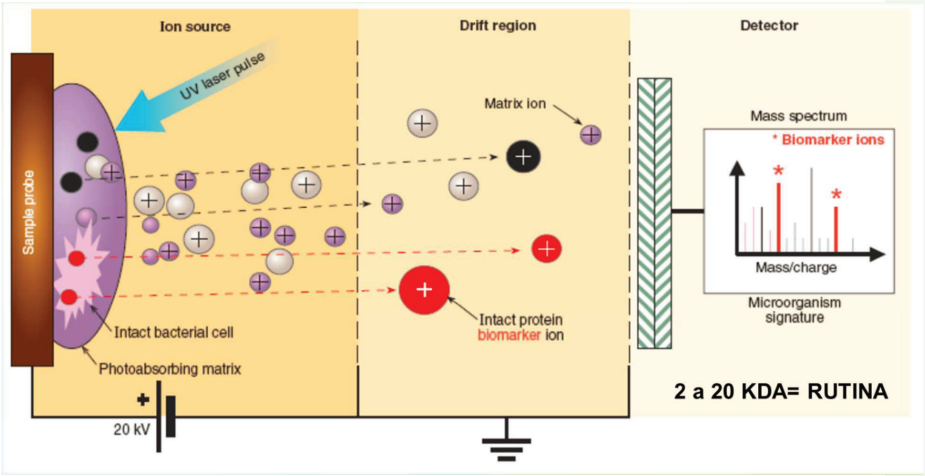
El analizador de tiempo de vuelo es básicamente un tubo de 1 a 4 metros donde entran los iones generados por el láser siguiendo una trayectoria lineal. Los analitos que salen del analizador van a impactar sobre el detector de iones que es una superficie de materiales semiconductores.

En estas condiciones, todos los iones que entran en el tubo de vuelo tienen la misma energía cinética, donde la relación *m/z* es proporcional al cuadrado del tiempo de vuelo, de esta manera, los iones más pequeños llegarán primero al detector y equivaldrán a un pico cuya intensidad está dada por la cantidad de iones con la misma masa que impactan en el mismo instante en el detector.

La representación de la intensidad frente a la relación *m/z* es lo q se conoce como espectro de masas o *mass fingerprint*, que es propio de cada microorganismo (**Figura 3**).

Si z=1, como es habitual en la ionización blanda de la técnica MALDI-TOF, en el espectro de masas se representa la intensidad en términos relativos, en escala de cero a cien, frente a la masa. Los rangos de lectura de masas se modifican de acuerdo con el tipo de ensayo que se pretenda realizar.

Figura 3. Principio de la generación de un espectro de masas en función del tiempo de vuelo. Adaptado de https://present5.com/maldi-tof-matrix-assisted-laser-desorption-ionization-time/

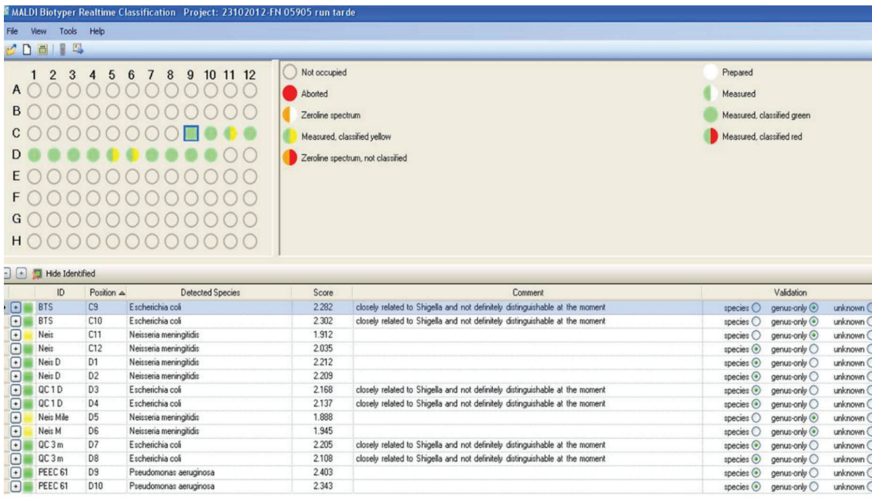


Entonces cada disparo es capaz de generar un espectro y la sumatoria de disparos, resulta un espectro promedio a partir de las proteínas ribosomales de la bacteria, el cual es referido a una base de datos donde se comparan los perfiles obtenidos con los de las cepas de referencia y así se diferencian los aislamientos bacterianos (**Maldonado et al.**, 2018).
Los resultados son revisados según un valor de puntuación (*Bruker*) o un nivel de confianza (*Biomérieux*).

Interpretación de resultados. En el sistema MALDI Biotyper, el resultado de la identificación junto con un puntaje también conocido como valor de *score*, se acompaña de un color y letras que equivalen a niveles de confianza, como se muestra en la **Figura 4**. Estos valores son calculados de forma automática por el equipo mediante un algoritmo estadístico en base a la cantidad de picos que comparte el microorganismo que se pretende identificar con los espectros de referencia presentes en la base de datos y les asigna un valor de acuerdo con un rango de tolerancia en el eje x (*m/z*) y en el eje y (concentración).
El sistema emplea un valor de puntuación que oscila entre 0.000 y 3.000, donde una puntuación entre 2.300 y 3.000 es interpretada como una identificación altamente probable a nivel de especie. Las puntuaciones entre 2.000 y 2.299 representan la identificación segura del género y la identificación probable a nivel de especie.

Las puntuaciones que van desde 1.700 a 1.999 representan una probable identificación de género, que requiere pruebas adicionales para una identificación positiva a nivel de la especie del microorganismo. Las puntuaciones que van desde 1.699 a 0.000 no se consideran una identificación fiable **(MALDI Biotyper 3.1 User Manual)**.

Figura 4. Resultados de la identificación en base a niveles de confianza que se acompañan de letras y valores de puntaje o scores (software MALDI Biotyper Realtime Classification).



Entonces, los perfiles proteicos obtenidos a partir de microorganismos desconocidos como bacterias y hongos de relevancia en salud pública son comparados con los perfiles proteicos de cepas de referencia de una Base de Datos y pueden ser identificados con altos niveles de confianza (**De Bel et al.**, 2011).

Sin embargo, la mayoría de las no identificaciones en MALDI-TOF, se deben a la ausencia del perfil de proteínas en la Base de Datos comercial, que son resueltas mediante la creación *in house* de nuevas librerías locales de espectros (**Cipolla et al.,** 2018).

Para incorporar un nuevo espectro de referencia a la base de datos, se seleccionan candidatos que no han sido identificados por la plataforma.

Esos candidatos son completamente caracterizados por métodos de referencia y sometidos al protocolo de extracción en tubo para la obtención de espectros más limpios reproducibles. Luego del análisis bioinformático, se incorporan a la librería complementaria como un MSP que estará compuesto por 20-30 espectros individuales de ese

microorganismo. El proceso de creación de un MSP requiere amplios conocimientos en softwares y entrenamiento previo en los métodos de preparación de las muestras a partir de cepas estándares para asegurar la calidad, empleando aislamientos frescos y puros en medios apropiados de crecimiento **(MALDI Biotyper 3.1 User Manual).**

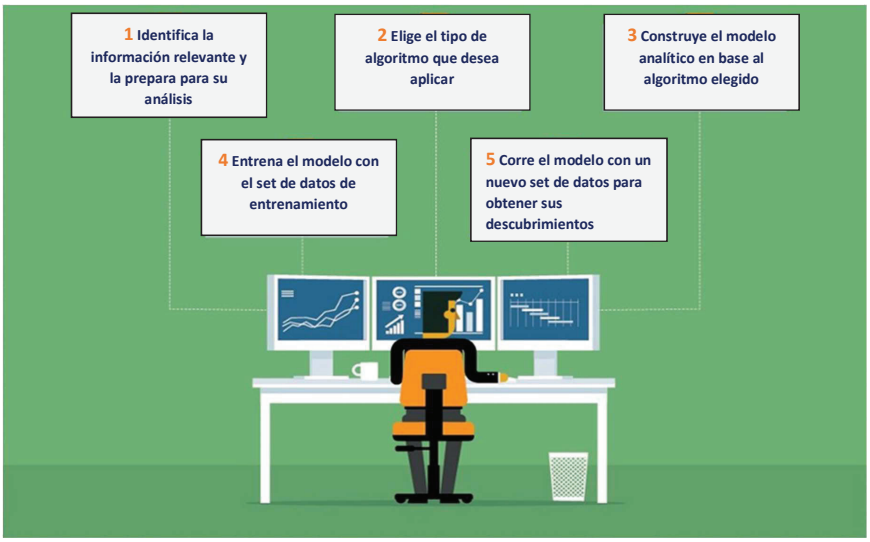
Aplicaciones. Más allá de la mera identificación convencional, se puede utilizar la EM para obtener perfiles peptídicos a partir de una gran variedad de muestras biológicas, como pueden ser el plasma, suero, saliva, orina, para estudios ómicos debido a sus capacidades específicas, no específicas y de alto rendimiento (**Lipi et al.,** 2020**; Zautner et al.**, 2016).
Para explotar estas novedosas aplicaciones, se requiere del procesamiento de un enorme número de muestras, incluso con varias repeticiones por muestra, generando grandes conjuntos de datos proteómicos que deberán ser sometidos al análisis mediante herramientas de entrenamiento automatizado o *machine learning (ML*) para poder interpretar los datos más relevantes del conjunto (**Schubert et al.,** 2017). Estos algoritmos de aprendizaje automático han permitido: detectar sublinajes, el perfil de resistencia a los antibióticos, la aparición de brotes o de cepas toxigénicas (**Feucherolles et al.**, 2019; **De Bruyne et al.**, 2011); por ejemplo, en la identificación preliminar de *Staphylococcus aureus* resistentes a la meticilina (SARM), evitando la utilización de métodos costosos, lentos e intensivos en mano de obra (**Yang et al.,** 2020) y para la tipificación de serovariedades de *Salmonella no-typhi* causantes de enfermedad en pollos, aplicado para la seguridad en la industria alimentaria a gran escala con costo-efectividad y tiempo mejorados (**Suthee et al.,** 2020).

Entrenamiento automatizado. El desarrollo de sistemas o algoritmos para procesar automáticamente grandes conjuntos de datos se conoce como ciencia de datos profunda o inteligencia artificial (**Granville**, 2017). El aprendizaje automático - *Machine Learning (ML)* es una disciplina científica del ámbito de la Inteligencia Artificial (IA) que se basa en crear sistemas que aprenden automáticamente; en este contexto, aprender quiere decir identificar patrones complejos en millones de datos biológicos. El algoritmo es entrenado con un enorme conjunto de datos y al revisarlos, es capaz de hacer predicciones sin que se lo haya programado explícitamente para esa tarea, mejorando su desempeño de forma autónoma con el tiempo. Cientos de ejemplos de IA rodean nuestra vida cotidiana, como los vehículos autónomos o las aplicaciones para hacer ejercicio, los buscadores de internet, entre otros. El principio básico es similar al que utilizan las aplicaciones del teléfono celular para escuchar música, que rápidamente aprenden sobre los gustos del usuario y sugieren música que le puede gustar. Esto constituye los algoritmos. Los datos de entrenamiento son las canciones que esa persona elige escuchar, el algoritmo aprende en base a esas

selecciones a lo largo del tiempo y luego puede elegir por sí mismo canciones que coinciden con los gustos del usuario. En los experimentos de investigación, se aplica la misma lógica. Se ingresan los datos biológicos, que se conocen como datos de entrenamiento, la máquina se entrena para que clasifique esos datos en grupos mediante modelos matemáticos y el sistema aprende a clasificar mejor a medida que se alimenta con más datos y así mejora su desempeño para separar las variables que el investigador desee (**Fitzpatrick et al.; Rhoads**, 2020).

El proceso del entrenamiento automatizado llevado a cabo por un científico de datos u otro especialista informático se resume en la **Figura 5**.

Figura 5. Etapas principales del procesamiento de datos en el aprendizaje automatizado-ML, aplicado por un científico de datos.



Para que los modelos de aprendizaje automático den resultados razonables, no solo se necesita alimentarlos con grandes cantidades de datos, sino que también se debe garantizar la calidad de estos. Los conjuntos de prueba y entrenamiento deficientes pueden producir efectos impredecibles en el resultado del modelo, pueden dar lugar a un ajuste excesivo o inadecuado de los datos y el modelo generado puede terminar arrojando resultados sesgados (**Granville**, 2017).

Idealmente, los datos se deben dividir en 3 conjuntos: 1)- ***conjunto de entrenamiento*** que contiene los datos que se introducirán en el modelo. En términos más simples, el modelo aprenderá de estos datos; 2)- ***conjunto de desarrollo o validación cruzada***, cuyos datos se utilizan para validar el modelo ya entrenado. Este es el escenario más importante, ya que, si la diferencia entre el error en el conjunto de entrenamiento y el error en el conjunto de desarrollo es muy grande, significa que el modelo tiene una alta varianza y, por lo tanto, un caso de sobreajuste durante su generación; 3)- ***conjunto de prueba*** que contiene los datos sobre los que se testea el modelo ya entrenado y validado. Este último dará una medida acerca de cuán eficiente es un modelo general y cuán probable es que prediga algo que no tiene sentido.

Existen una gran cantidad de parámetros estadísticos como precisión, exactitud, sensibilidad y especificidad que se pueden utilizar para medir el rendimiento teórico del modelo. Por otro lado, el conjunto de validación y el conjunto de prueba deben ser de la misma distribución; se debe tomar todo el conjunto de datos y mezclarlo para luego, dividirlo aleatoriamente en dos. En cambio, el conjunto de entrenamiento puede provenir de una distribución ligeramente diferente a la de los otros dos (**ClinPro Tools User Manual**, 2011).

El entrenamiento automatizado se clasifica según como aprende la computadora en:

- Entrenamiento supervisado: es el que más se utiliza en epidemiología para hacer regresiones lineales y donde el valor de la entrada de cada variable es conocido.

Los datos utilizados para construir el algoritmo contienen información conocida sobre la característica en estudio, que no está presente en los datos futuros. Por lo tanto, la información que se quiere predecir o por la que se quiere clasificar una población está disponible en los datos utilizados para construir el modelo.

Dentro del aprendizaje supervisado, los algoritmos pueden ser de clasificación o de regresión dependiendo de la naturaleza de la variable de respuesta. Un algoritmo de clasificación entrega resultados categóricos (por ejemplo, positivo/negativo, sano/enfermo) y uno de regresión da resultados continuos (por ejemplo, valores de índice de masa corporal IMC).

- Entrenamiento no supervisado: es el que trata de agrupar los datos, pero sin tener la respuesta correcta ya que no se dispone de la información de una variable que se

quiera predecir. Se aplica a desarrollos estadísticos que intentan identificar subgrupos con similares características (por ejemplo, *clustering k-means*).

Además, los métodos supervisados y no supervisados pueden ser:

-Discriminativos: estos algoritmos calculan directamente la probabilidad condicional de un resultado en un conjunto de datos observados. Por ejemplo, la probabilidad de que un individuo tenga diabetes dada una determinada masa corporal. La mayoría de los enfoques estadísticos epidemiológicos son discriminativos.

-Generativos: también calculan la probabilidad condicional de un resultado, pero indirectamente ya que primero modelan distribución de probabilidad conjunta. Es decir, todas las combinaciones posibles de masa corporal y resultados de diabetes. Estos algoritmos son generalmente más complejos, como es el caso de clasificadores bayesianos (*Naive Bayes*).

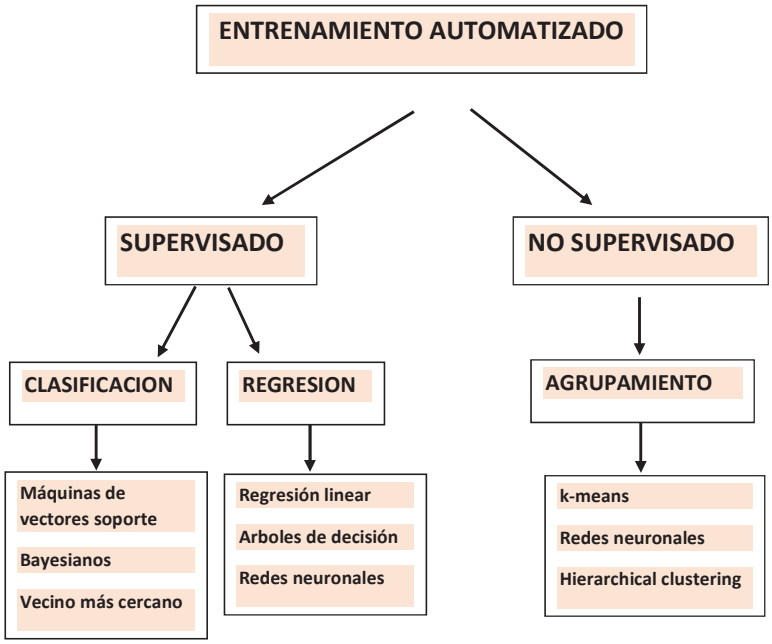
La elección de uno u otro método dependerá fundamentalmente del tipo de datos disponibles y del modelo de aprendizaje que se pretenda crear (**Shalev-Shwartz,** 2014).

A continuación, se mencionan algunos de los algoritmos matemáticos más comúnmente utilizados en EM (**Shalev-Shwartz** y **Shai**, 2014; **Fitzpatrick et al.; Rhoads**, 2020):

- Máquina de vectores de soporte (por su sigla en inglés, *SVM*)
- k- vecino más cercano (por su sigla en inglés, *k-NN*).
- Arboles de decisión (por su sigla en inglés, *DT*).
- Clasificadores bayesianos (en inglés, *Naïve Bayes*).
- Empaquetamiento (en inglés, boostrap).

Otros algoritmos de clasificación que también pueden aplicarse en entrenamiento automatizado son la regresión logística, el análisis discriminante lineal, bosques aleatorios y el análisis de conglomerados jerárquicos, dentro de los métodos no supervisados. Todos estos algoritmos se describen con mayor detalle en el glosario **(https://es.acervolima.com/2021/02/09/ventajas-y-desventajas-de-diferentes-modelos-de-clasificacion/).**

Resumiendo los conceptos:



Finalmente, los algoritmos preestablecidos en el software *ClinPro Tools* acoplado al equipo Microflex LT, son los más frecuentemente aplicados en este tipo de análisis de datos surgidos de la EM: algoritmo genético, red neuronal supervisada y clasificador rápido, cuyos fundamentos se detallan a continuación:

-El **algoritmo genético** (por su sigla en inglés, *GA*) es el más utilizado e imita la evolución en la naturaleza ya que depende de variaciones biológicas, como mutaciones y selección debidas a cambios evolutivos. Este algoritmo selecciona las combinaciones de picos que son más relevantes para la separación de clases. La ventaja de GA es que necesita poco tiempo de cálculo sin dejar de producir buenos resultados para alcanzar la separación con alta varianza entre las clases. El GA sólo se usa como una función de selección. La clasificación se realiza mediante el algoritmo *k-NN* basado en los picos seleccionados. Los valores predeterminados de *k-NN* son 1, 3, 5, 7; siendo 1 el más apropiado para bajo número de muestras y hasta 7 cuando el número de datos es de gran tamaño.

-El **clasificador rápido** (por su sigla en inglés, *QC*) calcula el área promedio de cada pico en el conjunto y proporciona un valor estadístico p por clase. Durante la clasificación, las áreas de los picos se ordenan mediante el algoritmo de clasificación univariante y se calcula un promedio de todos los picos que indica la pertenencia a esa clase. La clasificación permite determinar la membresía y también una probabilidad p para cada clase. Este algoritmo funciona mejor si hay pocas muestras disponibles para la generación del modelo.

-**Las redes neuronales supervisadas** (por su sigla en inglés, *SNN*) se inspiran en el comportamiento de señalización de las neuronas en redes biológicas y se emplean para problemas supervisados y no supervisados. Las neuronas en las capas de entrada y salida corresponden a las variables independientes y dependientes, respectivamente. Las neuronas en capas adyacentes se comunican entre sí a través de funciones de activación. El objetivo de este algoritmo es reducir la función de pérdida a cero, es decir, hacer que la salida prevista del *SNN* coincida con la verdad lo más fielmente posible. Puede requerir grandes conjuntos de datos para lograr un rendimiento óptimo del modelo, pero funciona mejor con múltiples clases.

Procesamiento de datos. Una vez que se genera una gran cantidad de información proteica existe un flujo de trabajo general para su procesamiento. En primera instancia se realiza un análisis exhaustivo de la calidad de los datos y se aplican diversos procesos de transformación que incluyen la suavización y el alineamiento de espectros, y la corrección de la línea de base. Luego se procede a la búsqueda de los "potenciales picos biomarcadores" de una determinada condición, por ejemplo, sano o enfermo, toxigénico o no toxigénico. A partir de estos biomarcadores será posible armar una matriz de entrenamiento computarizado basada en algoritmos matemáticos para crear modelos de clasificación (**Ressom et al**., 2005).
Una vez creado el modelo, aplicando alguno o varios de estos algoritmos, se deberá evaluar su solidez mediante una validación cruzada que puede realizarse en base a diferentes métodos estadísticos (aleatorio, plegado en *k-fold* y *leave one out*). En la validación cruzada, se divide un conjunto de datos para generar el modelo y un conjunto de prueba para evaluarlo y determinar la capacidad teórica de predicción (**Shalev-Shwartz** y **Shai**, 2014).
Este tipo de validación es realizada automáticamente por el software y únicamente se puede calcular si al menos 20 espectros no excluidos de todos los grupos están disponibles.

En cambio, para la validación externa, que debe ser realizada por el operador, se requiere de la carga de nuevos datos de espectros para cada clase que no se hayan empleado en la

generación del modelo. De estos nuevos datos se conoce a qué clase corresponden y deben ser procesados igual que los datos empleados en la generación del modelo.

A lo largo de nueve años de experiencia en MALDI-TOF MS, desde el Laboratorio Nacional de Referencia (LNR) de Argentina se han aplicado estas herramientas con objetivos muy diversos, entre los que se pueden mencionar: la creación de una Base de Datos *in house* a partir de patas de garrapatas para la identificación de vectores con gran impacto en salud pública e involucrados en la transmisión de bacterias del género *Rickettsia* (datos no publicados al momento); la subtipificación de clones hipervirulentos de *Streptococcus pyogenes* causantes de enfermedad invasiva en niños (**Rocca et al.,** 2018); la discriminación de *Escherichia coli o157:h7* de otras cepas diarreigénicas aplicando simples algoritmos predictivos en combinación con la detección rápida y fácil de biomarcadores específicos (**Manfredi et al.,** 2019); la detección de mecanismos de resistencia a los antimicrobianos (**Espinosa et al.,** 2018); incluso en la detección del estadío inflamatorio durante una sepsis a partir de las diferencias observadas en los perfiles del proteoma obtenido de sueros de pacientes (**Ledesma et al.,** 2020). Sin embargo, el potencial de la proteómica para la detección de virus a partir de muestras clínicas no ha sido evaluado hasta el momento.

Actualmente, la biotipificación de microorganismos utilizando MALDI-TOF MS es la aplicación más exitosa de la EM en los laboratorios clínicos. Desafortunadamente, este enfoque no es el más adecuado para identificar virus, por dos razones principales: en primer lugar, los virus se encuentran dentro de las células y no pueden aislarse por métodos simples como es el caso de las bacterias, por lo tanto la sensibilidad está muy limitada por el rango dinámico del instrumento; en segundo lugar los virus constan de muchas menos proteínas que las bacterias y, por ende, producen sólo unas pocas, si es que las hay, en el estrecho rango de masas de aproximadamente 2.000 a 12.000 Da, donde, además pueden estar cubiertas por la señal proteica de las células completas.

Actual pandemia por COVID-19. A fines del año 2019, fue reportado un tipo de coronavirus, denominado SARS-CoV-2. El mismo es considerado el agente causal de la nueva enfermedad por coronavirus 2019, conocida como COVID-19, que ha sido declarada pandemia por la Organización Mundial de la Salud el 12 de marzo del mismo año tras su aparición en Wuhan, China (**Yang et al.**, 2020). Al 4 de abril del 2020, ya se habían reportado más de 1.2 millones de casos confirmados de COVID-19 en 175 países, con más de 65,000 muertes (**COVID-19 Map - Johns Hopkins Coronavirus Resource Center**, 2020**)**. En el primer reporte de Emergencia por COVID-19 de la Sociedad Argentina de Virología del 26 de marzo del año 2020, se describe a los coronavirus como virus

envueltos cuyo genoma consiste en una única molécula de ARN simple cadena de sentido positivo. Estos virus pertenecen a la familia *Coronaviridae* que infecta aves y mamíferos, incluyendo camélidos, murciélagos, ciervos, ratas, ratones, perros, gatos y humanos. Ocasionalmente, los coronavirus pueden emerger como patógenos mediante un salto a una especie hospedadora diferente. En humanos, algunos miembros de la familia (229E, OC43, NL63, y HKU1) son conocidos desde hace décadas por causar los síntomas del resfrío común, pero luego del brote del Síndrome Respiratorio Agudo Severo (SARS) en la provincia de Guangdong, China, en el año 2002, los coronavirus han sido reconocidos como agentes causantes de graves infecciones respiratorias e intestinales. El SARS-CoV-2 es uno de los cuatro nuevos virus patógenos que han pasado de huéspedes animales a humanos en los últimos 20 años.

Dado que estamos frente a un virus emergente, hasta el momento es escasa la información específica que se tiene sobre el mecanismo de patogenia que presenta SARS-CoV-2. Por lo tanto, los datos que existen a nivel mundial se basan, en su mayoría, en la similitud con el SARS-CoV. La respuesta inmunitaria, tanto innata como adaptativa son necesarias para la eliminación viral, pero siempre bajo una regulación estricta, de lo contrario puede desencadenarse la inmunopatología asociada (**Guo et al.,** 2020).

Se estima que para limitar la propagación del SARS-CoV-2 sería necesario realizar pruebas de diagnóstico a gran escala para detectar y aislar a los individuos asintomáticos. Dado que este tipo de pruebas son extremadamente difíciles de realizar en la población general, el testeo periódico de individuos asintomáticos en los grupos más expuestos, como los trabajadores de la salud, sería de particular utilidad para controlar la circulación del virus, de forma tal de poder indicar el aislamiento de aquellos que resultasen positivos, así como de sus convivientes o contactos estrechos.

La carga viral en muestras respiratorias podría ser un marcador potencialmente útil para la evaluación de la gravedad, el pronóstico de la enfermedad y el monitoreo de la evolución de la infección en los pacientes. Las muestras de elección son principalmente el hisopado nasofaríngeo y orofaríngeo además del esputo y/o aspirado nasofaríngeo, aspirado endotraqueal o lavado broncoalveolar. Además de su presencia en muestras de origen respiratorio, el virus SARS-CoV-2 ha sido detectado en muestras fecales y sanguíneas (**Zhang et al**., 2020). Para ensayos serológicos complementarios las muestras de elección actuales son sueros obtenidos en la fase aguda y convaleciente.

El diagnóstico de la infección por SARS-CoV-2, se basa en la detección del genoma viral a través de técnicas de biología molecular como es la reacción en cadena de la polimerasa en tiempo real (**Chan et al.,** 2020). La sensibilidad es alrededor del 80% en la práctica, mientras que su especificidad es cercana al 95%.

A partir del conocimiento de la secuencia completa del genoma viral, se pudieron diseñar secuencias de iniciadores para los ensayos estandarizados en institutos de distintos países del mundo y cuyos protocolos están disponibles y publicados en la página de la OMS **(https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/laboratoryguidance).**

Dado el complejo contexto mundial por la pandemia se comenzó a evaluar fuertemente el rol potencial de MALDI-TOF MS acoplado al entrenamiento automatizado en la detección del virus, lo que se evidenció en el creciente número de publicaciones y preimpresiones que tratan sobre la detección del virus SARS-CoV-2 por esta metodología (**Bezstarosti et al.; Ihling et al.; Orsburn et al.; Gouveia et al.; Cardozo et al.; Nikolaev et al.; Nachtigall et al.,** 2020). Cabe señalar que, al momento del desarrollo de este proyecto de tesis, muchas de estas publicaciones no existían o solo se encontraban disponibles como preimpresiones y debieron evaluarse con suma precaución.

3. HIPOTESIS

En la actualidad, el diagnóstico de COVID-19 basado en técnicas moleculares se ve amenazado por la extraordinaria demanda de insumos a nivel mundial (**Antezack et al.**, 2020). Es por eso que el desarrollo de pruebas alternativas y / o técnicas complementarias resulta tan relevante (**Chin et al., 2020; Wenzhong y Hualan**, 2020). Entre las opciones metodológicas, la secuenciación de nueva generación (por su sigla en inglés, *NGS*) es un método que permite el análisis completo del genoma independientemente de la muestra (**Brown et al.**, 2018). La ventaja de *NGS* es que no es necesario ningún conocimiento previo del virus objetivo, pero los límites de detección pueden variar drásticamente y sus costos son muy elevados.

Por su parte, los métodos inmunoserológicos, basados en la detección de anticuerpos específicos antiproteínas SARS-CoV-2, dependen de la aparición de estos anticuerpos en el suero del paciente. Estas pruebas se realizan en diferentes formatos, como ensayos de flujo lateral, ELISA o *Western Blot*. El principal inconveniente en comparación con la qRT-PCR se refleja en la curva de la cinética de seroconversión. La menor sensibilidad durante la primera semana luego del inicio de los síntomas, los vuelve una opción poco apropiada para su uso en el diagnóstico de personas con sospecha de infección activa por SARS-CoV-2, sumado a problemas potenciales con la especificidad, ya que puede ocurrir reactividad cruzada contra especies de virus relacionadas, dependiendo del epítope detectado (**Van der Heide,** 2020).

Reportes previos sugieren que los métodos basados en la detección de proteínas virales podrían ser más beneficiosos, en comparación con la metodología basada en ácidos nucleicos, debido a la mayor estabilidad composicional de ciertas proteínas (**Buckley,** 2018). Se describe que las proteínas virales son menos propensas a degradarse durante el transporte de la muestra en comparación con el ARN viral, lo que podría conducir a menos falsos negativos, siendo, además, la proteómica basada en EM, el método actual de elección para la identificación de proteínas de forma sensible y selectiva. Sin embargo, SARS-CoV-2 ha sido analizado en este último tiempo a partir de una gran variedad muestras de pacientes, esto hace que los resultados de la EM varíen mucho y hasta ahora, no se alcancen sensibilidades similares a las obtenidas a través de la técnica de PCR que sigue siendo el método estándar de oro (**Grenga, Armengaud**, 2020).

Debido a la complejidad del proceso biológico que ocurre en las células respiratorias durante la infección, sumado a la dificultad en el procesamiento de este tipo de muestras, la presencia abundante de proteínas del huésped y a la peligrosidad de la manipulación viral en cultivo celular, se postula la detección de un perfil proteico diferencial entre muestras de pacientes positivos y negativos, es decir, la búsqueda de modificaciones en la presencia/

ausencia e intensidad de los picos debidos a la infección viral, más allá de la detección de un único biomarcador (**Prodan et a**l., 2016) directamente sobre una muestra de hisopado nasofaríngeo.

Las limitadas investigaciones previas realizadas sobre la patogénesis del SARS-CoV-2 se han centrado en la respuesta sistémica del huésped (**Rhoades et al**., 2021), por lo que la detección de picos estaría asociada directamente a modificaciones en las proteínas de la muestra en diferentes condiciones (sano o enfermo) y su variación individual.

La respuesta inflamatoria es muy compleja y para caracterizar las proteínas detectadas en mayor concentración, como propias del huésped o del virus, sería necesario contar con el virus inactivado y en cultivo celular como control positivo. Pero, además, de la dificultad para poder llevar a cabo este procedimiento complejo y peligroso, se conoce que el poder de resolución de MALDI-TOF MS no permitiría la cuantificación absoluta y específica de los péptidos hallados ya que solo puede detectar diferencias relativas entre perfiles de muestras para picos particulares y donde sólo pueden obtenerse identidades putativas haciendo coincidir los valores *m / z* (**Chivte et al**., 2021).

En base a todo lo anteriormente expuesto, sumado a la experiencia previa en el área, se postula la evaluación de la EM acoplada al ML, como una alternativa atractiva y complementaria a los métodos basados en PCR, por ejemplo, cuando los reactivos de biología molecular escasean o cuando las capacidades diagnósticas deben expandirse rápidamente y se requiere el análisis de un gran número de datos en el corto tiempo (**Grossegesse et al.**, 2020).

4. APORTES DEL PROYECTO

En el marco de crisis desatado por la actual pandemia (**COVID-19 Map - Johns Hopkins Coronavirus Resource Center**, 2020; **Li**; **Yang et al.; Zhou et al.,** 2020), investigadores de instituciones prestigiosas de todo el mundo han propuesto varias iniciativas que promueven compartir de manera rápida, los hallazgos, protocolos, ideas y resultados de los desarrollos relacionados a COVID-19 a través de diversas plataformas digitales, de forma tal de establecer redes de conocimiento para compartir experiencias y discusión. Hasta este momento no existían en la literatura científica antecedentes sobre este abordaje, por lo que resultó válido evaluar el potencial de MALDI-TOF MS debido a que es una técnica simple, económica y rápida que analiza los perfiles de proteínas de una muestra en cuestión de minutos y está ampliamente disponible en los laboratorios de microbiología clínica de todo el mundo (**Croxatto et al.,** 2012; **Rocca et al.,** 2020).

En nuestro país, en el marco del Sistema Nacional de Redes de Laboratorios, en el año 2015 se creó la RENAEM (disposición EX -2019-98413674-APN-ANLIS#MSYDS) como una Red

Nacional que incluye a laboratorios públicos y privados que han implementado la EM como herramienta de diagnóstico microbiológico. El objetivo principal de la red es generar homogeneidad y excelencia en diagnósticos oportunos, confiables y accesibles para mejorar la eficiencia y efectividad del sistema de vigilancia de salud (**Rocca et al.**, 2019).

Esta red ha crecido enormemente sumando nodos de referencia, nodos colaboradores y participantes alcanzando 31 instituciones al momento de redactar este escrito (20 equipos Microflex LT de la firma Bruker Daltonics, 11 equipos Vitek MS comercializados por Biomerieux).

Con el fin de transferir todos los hallazgos de los Laboratorios Nacionales de Referencia, se diseñó un sitio web de la RENAEM, **http://www.anlis.gov.ar/renaem/.** Del mismo modo, todos los manuales de procedimientos generados por los referentes de la red están depositados en el Sistema de Gestión del conocimiento del ANLIS Malbrán, **http://sgc.anlis.gob.ar/handle/123456789/614**, que ya cuenta con más de 1000 descargas en todo el mundo en sus dos versiones (inglés y español) desde su incorporación al sitio en el año 2019.

Por lo que cualquier descubrimiento surgido de este proyecto de tesis podría ser comunicado fácilmente al resto de la comunidad científica a través de estos medios, fomentando el desarrollo de nuevos temas de investigación entre los participantes de la Red Nacional de Espectrometría de Masas en todo el país.

5. OBJETIVO GENERAL

Evaluar el potencial de MALDI-TOF MS en la generación de espectros de masas característicos, obtenidos directamente a partir de hisopados nasofaríngeos provenientes de individuos con sintomatología compatible de COVID-19, con el fin de encontrar picos discriminatorios específicos que contribuyan al diagnóstico de SARS-CoV-2. Mediante el uso de la inteligencia artificial, a partir de los datos proteómicos obtenidos, se evaluará si estos potenciales picos biomarcadores podrían ser útiles para diferenciar muestras positivas de muestras negativas para COVID-19.

6. OBJETIVOS ESPECÍFICOS

- Obtener los perfiles peptídicos en forma manual en el equipo de EM, Microflex LT a partir de hisopados nasofaríngeos previamente caracterizadas mediante la técnica de referencia.

- Revisar, corregir y analizar en los diferentes softwares disponibles, los espectros obtenidos para evaluar la calidad de estos y la posibilidad de detectar señales propias de un determinado perfil (biomarcador).

- Analizar la diversidad proteómica empleando diversos algoritmos de ML y diseñar modelos predictivos de clasificación rápida mediante el uso de la EM acoplada a la Inteligencia Artificial que permitan predecir el agrupamiento de los pacientes.

- Comparar los resultados obtenidos a partir de la EM con el método molecular, para evaluar su potencial aplicación como técnica de tamizaje sencilla y económica en el diagnóstico, pronóstico y seguimiento de la enfermedad. Comparar además los valores de CT (*cycle treshold*) de muestras positivas con la técnica de referencia, con respecto a los resultados de ML para evaluar una posible correlación entre ellos.

- Determinar los parámetros de desempeño analítico, valor predictivo positivo (VPP) y valor predictivo negativo (VPN) de la metodología MALDI-TOF-MS en comparación con el método de q RT-PCR.

7. MATERIALES Y MÉTODOS

Muestras. Durante el período abril-agosto del año 2020, se analizaron **311** muestras de hisopado nasofaríngeo, provenientes de individuos con sintomatología compatible con COVID-19.

Del total de muestras, **123** fueron seleccionadas y provistas por el servicio de Virosis Respiratorias del INEI-ANLIS "Carlos G. Malbrán", Laboratorio Nacional de Referencia (LNR) (se adjunta Cronograma de trabajo y total de muestras procesadas en la **Tabla S1** del apartado Material Suplementario).

Las **188** muestras restantes provinieron de laboratorios colaboradores de la Red Nacional de Espectrometría de Masas (laboratorio de microbiología de la FFyB del Hospital de Clínicas José de San Martín y laboratorio del Hospital Naval). Estas últimas muestras fueron procesadas por personal de laboratorio de los hospitales mencionados y en nuestro servicio se recibió un panel con los espectros proteicos obtenidos (**Tabla S2** del apartado Material Suplementario).

Cabe aclarar que todas las muestras fueron procesadas por triplicado mediante EM y los datos de espectros se adquirieron siempre en las mismas condiciones. A partir de estas **311** muestras procesadas se obtuvieron un total de 933 espectros proteicos.

Características de las muestras. Se cumplieron las normas recomendadas por la Organización Mundial de la Salud para la recolección, conservación, embalaje, transporte y análisis de las muestras de pacientes que se ajustan a la definición de caso sospechoso de COVID-19 (**https://www.who.int/publications-detail/laboratory-testing-for-2019-novelcoronavirus-in-suspected-human-cases-20200117**).

Una vez realizado el hisopado nasofaríngeo, el hisopo de fibra sintética se colocó en un tubo estéril que contenía 2 a 3 ml de solución salina. Estas mismas muestras empleadas para realizar el diagnóstico por qRT-PCR, se utilizaron directamente para el análisis en MALDI-TOF MS sin otra preparación previa. Todo el procedimiento se propuso de este modo, para optimizar el rendimiento de las muestras disponibles en el corto plazo, simplificar el procesamiento de los datos, y para establecer una correlación directa del desempeño del desarrollo propuesto, con respecto a los resultados obtenidos mediante las técnicas de referencia actuales.

Declaración ética y de bioseguridad. El trabajo siguió los protocolos éticos ya que no se relacionó la identidad de ninguna muestra con el nombre del paciente u otra información que pudiera conducir a la identificación personal. No se obtuvo la información demográfica

específica de los pacientes, tales como gravedad de la enfermedad, edad y comorbilidades, durante este trabajo.

Diagnóstico de referencia mediante qRT-PCR. Las muestras fueron caracterizadas en el Servicio de Virosis Respiratorias-LNR mediante RT-PCR, según lo recomendado por **Corman** y colaboradores (2020).
Debido a que el estudio fue llevado a cabo al inicio de la actual pandemia, cuando el LNR recibía únicamente muestras de pacientes sintomáticos, se obtuvieron los hisopados nasofaríngeos de pacientes con los siguientes resultados de RT-PCR:
No detectable (ND) o negativo para COVID-19,
Detectable o positivo para COVID-19,
Detectable o positivo para otros virus relacionados: virus influenza (FLU), sarampión (SAR), coronavirus humano endémico (VH).

Una vez realizado el diagnóstico por el método de referencia, las muestras se almacenaron a -80°C hasta su uso en MALDI-TOF MS.

Consideraciones de bioseguridad. Para la protección del personal, se debió garantizar la inactivación del virus durante la preparación de la muestra (**Grossegesse et al.**, 2020). Hasta la actualidad no existe ningún estudio previo que haya investigado sistemáticamente la inactivación de virus para métodos proteómicos. Se han propuesto numerosos tratamientos tales como, acetona, acetona/metanol, metanol/cloroformo, exposición a la radiación UV a longitud de onda de 254 nm, tratamiento térmico de 65°C, condiciones alcalinas o acídicas o formalina (**Iles et al**, 2020). Sin embargo, en este estudio se debió explorar una técnica específica de preparación de muestras multiómica, ya que además de la inactivación viral, se debía conservar toda señal peptídica perteneciente al virus y al paciente; para ello, se consultaron especialistas de Argentina y Chile referentes en el tema, quienes aseguraron que SARS-CoV-2 es un virus lábil y que el ácido que conforma la matriz orgánica HCCA que se utilizó para el procedimiento en MALDI-TOF, era suficiente para lograr la correcta inactivación viral manteniendo las propiedades peptídicas de las muestras clínicas y asegurando de este modo, la mejor calidad para el análisis. Estas muestras tampoco representaban peligro biológico una vez salidas del equipo debido a que fueron sometidas al intenso desgaste del láser.

Todas las manipulaciones se realizaron bajo cabina de seguridad biológica certificada de clase II TELSTARTM BIO IIA (Thermo Fischer Scientific, Villebon sur Yvette, Francia) y usando el equipo de protección personal apropiado, requerido para cumplir con los estándares de

bioseguridad dictados por la Organización Mundial de la Salud durante el año 2020 (**Wenzhong et al.**, 2020).

La elección de la siembra mediante el método directo a partir del hisopado nasofaríngeo a temperatura ambiente y el posterior sellado únicamente con la matriz HCCA, se basó en publicaciones previas en el área (**Nachtigall et al.,** 2020).

Preparación de la placa para MALDI-TOF MS. Al momento de generar los espectros, las muestras se descongelaron a temperatura ambiente y se agitaron suavemente en forma manual.
Sin otro enriquecimiento previo, se realizó la siembra por el método directo que es el que se utiliza en la rutina de los laboratorios clínicos por su simplicidad y según lo recomendado por el fabricante (**Wang et al.,** 2018; **MALDI Biotyper 3.1 User Manual**).
Brevemente, se colocó 1 ul de la muestra en cada pocillo de la placa de acero (acero rectificado de 96 pocillos, Bruker Daltonics), por triplicado; luego se dejó secar durante unos minutos a temperatura ambiente y se cubrió con 1 ul de matriz comercial HCCA (solución de ácido α-ciano-4-hidroxicinámico diluido en 500 µl de acetonitrilo, 250 µl de ácido trifluoroacético al 10% y 250 µl de agua de grado HPLC).
Al cabo de unos 15 minutos en la cabina de flujo laminar, la placa MALDI se transportó en un contenedor cerrado herméticamente y colocado dentro de una caja de telgopor cerrada hasta llegar al equipo; allí la placa se introdujo en el instrumento MicroFlex LT (Bruker Daltonics, Bremen, Alemania) y se generó la condición de alto vacío.

Generación de espectros proteicos. Los espectros de masa se adquirieron manualmente utilizando el software FlexControl v3.4 (Bruker Daltonics, Bremen, Alemania) en el modo *OFF*. Los datos se recopilaron en el rango de masas entre 2000 - 20000 Da en el modo lineal de ionización positiva. Cada espectro fue una suma de 240 disparos del láser en diferentes regiones del pocillo, recogidos en incrementos de a 40 a una frecuencia de 100 Hz. La adquisición de datos se llevó a cabo al 40% de la energía máxima de láser y los iones positivos fueron extraídos con voltaje de aceleración de 20 KV (**Figura 6**).
La plataforma fue calibrada previo a cada ensayo utilizando el estándar de prueba bacteriano *BTS* (Bruker Daltonics, Bremen, Alemania) que consiste en un extracto comercial de proteínas ribosomales de una cepa patrón de *Escherichia coli* DH5-α suplementado con dos proteínas adicionales (RNAasa y mioglobina), empleado para la detección de picos estándares en el rango de 3637 Da a 16952 Da dentro de un rango de tolerancia para la

posición de cada pico, según lo recomendado por el fabricante (**Flex control 3.4 user manual**, 2011).

Figura 6. Esquema general de la obtención de perfiles proteicos a partir de hisopados nasofaríngeos en MALDI-TOF.

Todos los espectros recopilados se guardaron en carpetas de archivos para su pre-procesamiento y análisis posterior. La información peptídica obtenida se evalúo con tres enfoques distintos:

ESTRATEGIA 1) Creación de una Base de Datos "*in house*" de espectros proteicos.

ESTRATEGIA 2) Detección manual y automatizada de potenciales biomarcadores.

ESTRATEGIA 3) Desarrollo de modelos predictivos de clasificación basados en *ML.*

7.1. ESTRATEGIA 1) Creación de una Base de Datos "*in house*" de espectros proteicos de referencia. Construcción de la biblioteca de MSPs.

Los espectros crudos se analizaron visualmente en el software Flex Analysis v3.4. Para construir la base de datos *"in house"* se seleccionaron los de mayor calidad, es decir, aquellos que contenían picos de alta intensidad y bajo ruido (**Tabla 1**). El procesamiento de espectros y análisis de picos característicos en Flex Analysis se describe en detalle más adelante.

Sobre los espectros seleccionados, se aplicó el método de pre-procesamiento estándar en el software Maldi Biotyper OC V3.1.66, que incluyó el ajuste de masa, el suavizado de los espectros con el método Savistsky-Golay, la resta de la línea de base y la normalización. Luego se crearon 20 perfiles de referencia con el método *MSP Creation* de la configuración predeterminada.

Se generaron perfiles a partir de: 9 muestras positivas-COVID-19, 8 muestras negativas-COVID-19 y 3 muestras positivas-otros virus respiratorios (FLU, SAR y VH).

Previo a la incorporación de cada MSP a la nueva base de datos denominada "***BE COVID-19***", se verificó que estos MSP cumplieran con el criterio de inclusión de tener al menos 40 picos conservados con una frecuencia del 100% para cada masa.

Esta decisión se fundamenta en la experiencia previa a partir de colonias bacterianas aisladas, donde el fabricante recomienda la presencia de 70 picos con una frecuencia superior al 75% para cada masa. Al tratarse de muestras clínicas complejas, el número de picos reproducibles suele ser menor, pero se estableció que esos 40 picos estén presentes en la totalidad de las muestras.

Tabla 1. Muestras empleadas para crear la Base de Datos ***"in house"*** **BE COVID-19**.

ID MSP	RESULTADO RT-PCR
8083	DETECTABLE
8103	DETECTABLE
8117	DETECTABLE
7562	DETECTABLE
7669	DETECTABLE
7834	DETECTABLE

ID MSP	RESULTADO RT-PCR
9334	DETECTABLE
9350	DETECTABLE
9597	DETECTABLE
116	NO DETECTABLE
131	NO DETECTABLE
143	NO DETECTABLE
19	NO DETECTABLE
6961	NO DETECTABLE
8040	NO DETECTABLE
1148	NO DETECTABLE
1153	NO DETECTABLE
984	DETECTABLE-VH
217	DETECTABLE- SAR
2974	DETECTABLE-FLU

A continuación, se construyó un dendrograma a partir de los MSPs, para evaluar el agrupamiento jerárquico (comúnmente conocido como *hierarchical clustering*) basado en las señales de masa e intensidad de los picos. Las relaciones se calcularon utilizando el coeficiente de similitud de Pearson. La distancia se consideró en unidades relativas, donde 0 indica similitud completa y 1000 indica completa disimilitud.

Es válido aclarar que en algunos trabajos científicos suele utilizarse el punto de corte del 10% en el nivel de distancia entre 0 (idéntico) y 1.000 (no relacionado), como criterio para la identificación confiable a nivel de especie, pero esto es aplicable únicamente al agrupamiento de aislamientos microbianos (**Cipolla et al.,** 2018).

Evaluación. La base de datos creada fue desafiada con 223 nuevas muestras caracterizadas por qRT-PCR (88 muestras COVID-19 positivo y 135 muestras COVID-19 negativo) y la identificación se realizó en el software MALDI Biotyper OC v3.1.66 para una clasificación denominada *offline* (del inglés, fuera de línea); es decir, se compararon los espectros de prueba generados manualmente y almacenados con anterioridad, con los espectros de la nueva Base de Datos de referencia.

La expresión del resultado acompañado de un valor de puntaje se obtiene de la cercanía con el MSP de referencia durante el proceso de emparejamiento de espectros como se detalla a continuación:

1) El algoritmo de coincidencia calcula tres valores por separado. Primero, el número de señales en el espectro de referencia que tienen una coincidencia cercana en el espectro desconocido; sin coincidencias devuelve un valor = 0 y la coincidencia completa devuelve un valor = 1.
2) Luego, el número de señales en el espectro desconocido que tienen una coincidencia cercana en el espectro de referencia; sin coincidencias devuelve un valor = 0 y una coincidencia completa devuelve un valor = 1.

3) Finalmente, se calcula la simetría de los pares de señales coincidentes. Si las intensidades de las señales en el espectro desconocido se corresponden con las intensidades del espectro de referencia y las señales de baja intensidad también se corresponden, esto da como resultado un alto valor de simetría y la llamada matriz de correlación produce un valor cercano a 1. Si los pares de señales coincidentes no muestran simetría en absoluto, esto da como resultado un valor cercano a 0.

Estos tres valores se multiplican juntos y el resultado se normaliza a 1000. El valor máximo de puntuación o *score* que se puede obtener en una identificación en MALDI-TOF MS es 3.00 (= log 1000).

7.2. ESTRATEGIA 2) Detección manual y automatizada de potenciales picos biomarcadores.

Se realizó la búsqueda de patrones biomarcadores sobre los datos de los mejores espectros obtenidos directamente de hisopados nasofaríngeo, de modo similar a lo descripto en ensayos previos a partir de otros fluidos corporales (**Pusch et al.,** 2003**; Zhang et al.,** 2015). Con el fin de detectar cambios en los niveles de proteínas que reflejasen las variaciones en el estadío de la enfermedad, se aplicaron algoritmos bioinformáticos que simplificaron la detección de picos de alta frecuencia (**Ressom et al**., 2005) aplicando métodos supervisados y no supervisados como se describe a continuación.

En el desarrollo de esta estrategia se emplearon dos softwares diferentes; Flex Analysis y ClinPro Tools (**Camoez et al.,** 2016; **Khot y Fisher,** 2013), para detectar picos característicos de muestras positivas de COVID-19, frente a picos propios de las muestras negativas de COVID-19, en el rango de masas de 2 a 20 kDa.

Para dicho análisis inicial, se utilizaron los 20 MSP generados durante la creación de la Base de Datos (Estrategia 1) ya que como se detalló previamente, estos perfiles cumplieron el criterio de inclusión predeterminado.

Análisis MANUAL mediante software Flex Analysis v3.4.

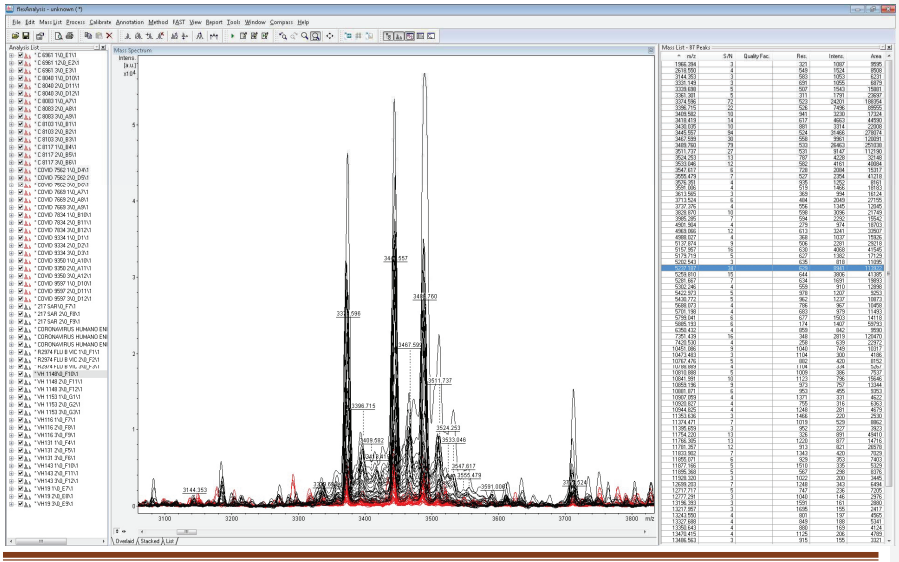
Los archivos de espectros de los MSPs se exportaron como archivos mzXML utilizando CompassXport CXP3.0.5. (Bruker Daltonics, Bremen, Alemania) y se verificaron los criterios de calidad del espectro para el aspecto global y la intensidad.
Dicho análisis incluyó: el ajuste de la masa, la suavización de espectros, resta de la línea base, normalización y selección de picos significativos en base a su área/intensidad. Los parámetros utilizados fueron todos los predeterminados para la normalización según la configuración estándar de Bruker.
El programa admite tres algoritmos diferentes de detección de masas: Centroide, SNAP y buscador de picos de suma. Cada buscador está optimizado para una tarea específica (**flexAnalysis 3.4 User Manual**, 2011). En este caso, se utilizó el centroide porque en el caso de espectros de proteínas beneficia el análisis al definir la masa con precisión, ya que utiliza la primera y la segunda derivada para detectar un pico. Para definir la posición de un pico en el centroide debe especificarse un nivel de altura específico; aquí se utilizó el nivel de corte por encima de la línea de base recomendado de alrededor del 80%. El resultado se almacenó en una lista de picos y el espectro de masas se analizó con las etiquetas de masas detectadas para cada pico. Esta lista de picos junto a otros parámetros (área, intensidad,

resolución, señal/ruido) se exportó a un archivo de Excel (Microsoft, Redmond, WA) para analizar los posibles biomarcadores. Los datos de los picos se binarizaron, es decir, si había un pico dentro del intervalo de tolerancia de 10 Da, se utilizó un valor de "1" para indicar su presencia; de lo contrario, se utilizó "0" para indicar una ausencia de pico dentro de ese intervalo.

Asimismo, se realizó el análisis visual de los espectros superpuestos en las diferentes opciones de vistas que ofrece el programa (stacked/gel view/overlaid), buscando diferencias significativas entre los dos grupos de estudio. Este tipo de análisis conlleva mucho tiempo y depende básicamente de la experiencia del operador, pero suele arrojar importantes resultados.

En la
Figura 7 se muestra, a modo de ejemplo, la imagen de la pantalla del software Flex Analysis donde se puede observar la superposición de espectros individuales, positivos para COVID-19 y negativos para COVID-19 en el tipo de vista *overlaid* y la lista de masas halladas por espectro, acompañadas de los parámetros de cada pico.

Figura 7. Análisis visual de espectros superpuestos en Flex Analysis v3.4. A la izquierda se observa la lista de espectros descargados y la selección de colores por el operador; en este caso se despliegan en rojo las muestras COVID-19 positivas y en negro las negativas.
A la derecha de la pantalla aparece la lista de masas de los picos presentes en cada uno de los espectros.



Análisis AUTOMATIZADO mediante software ClinPro Tools v 3.0.

Los mismos archivos de espectros de los 20 MSPs se importaron al software ClinPro Tools (Bruker Daltonics, Bremen, Alemania) para el reconocimiento de patrones diferenciales entre grupos. Para ello se aplicó el entrenamiento supervisado al ingresar los datos en diferentes clases, según el resultado conocido a partir de la técnica de RT-PCR.

Los pasos de pre-procesamiento de los datos se establecieron en base a la configuración predeterminada para todos los análisis según se indica en el manual del usuario (**ClinPro Tools User Manual Version 3.0,** 2011**; Camoez et al.,** 2016**; Zhang et al.,** 2015). Los espectros se suavizaron con 10 ciclos del algoritmo Savitzky / Golay con una anchura de 2 m/z. La resta inicial se realizó con el algoritmo Top Hat. La selección de picos se realizó sobre el espectro promedio de cada grupo, con un umbral de señal a ruido seteado en 5.

Los picos característicos de los dos grupos generados, clase 1 = muestras positivas COVID-19; clase 2 = muestras negativas COVID-19, se seleccionaron utilizando la función de "Tabla estadística de picos " en ClinPro Tools, seguida de la confirmación manual de que esos mismos picos eran distinguibles usando Flex Analysis (**Khot y Fisher**, 2013). Se obtuvo, además, la gráfica de distribución bidimensional (2D plot) de los dos mejores picos para la separación, calculados automáticamente por el software.
Los valores de *m/z* de los espectros promedio fueron extraídos y los “potenciales picos biomarcadores” se identificaron de acuerdo con su significancia estadística aplicando los diferentes test disponibles en ClinPro Tools (**Tabla Suplementaria S3**). Las pruebas estadísticas incluyeron la prueba t, el análisis de varianza (ANOVA), Wilcoxon o Kruskal - Prueba de Wallis (W / KW) y prueba de Anderson - Darling (AD) **(Stephens,** 1974**; Wang et al.,** 2018), cuyos fundamentos se detallan en el glosario.
Se estableció un valor p de 0.05 como punto de corte:
*Si p ≤0,05 en la prueba AD y ≤0.05 W / KW, el pico se considera “potencial pico biomarcador”.
*Si p >0.05 en AD, pero ≤0.05 en ANOVA o W / KW, el pico se debe seleccionar como “potencial pico biomarcador”.

El poder discriminatorio para cada biomarcador potencial se describió mediante el análisis del área bajo la curva ROC (característica operativa del receptor). Un valor de área bajo la curva (AUC) de 0 indica que el pico en cuestión no discrimina un perfil, mientras que un AUC de 1 indica que el pico considerado discrimina sensiblemente. Se determinó el AUC de cada pico y solo se seleccionaron picos con valores de AUC ≥ 0,80.

7.3. ESTRATEGIA 3) Diseño de modelos predictivos de clasificación rápida basados en herramientas de *Machine Learning*.

El objetivo de la generación de los modelos predictivos fue determinar una firma común entre los espectros de las clases que los componen de manera tal que luego se puedan clasificar otros espectros de prueba desconocidos.

Para la preparación de los datos del conjunto de entrenamiento, se utilizaron las funciones predeterminadas en el software ClinPro Tools, que involucró la resta de la línea de base (top hat 10% de ancho mínimo de la línea de base), la normalización de espectros (corriente de iones total), la recalibración (1,000 ppm de desplazamiento de pico máximo y 30% de coincidencia con los picos calibrantes, con exclusión de espectros nulos o fuera de rango), el cálculo del espectro promedio (resolución 800), el cálculo de la lista de picos promedio (umbral de señal/ruido seteado en 5), el cálculo de picos en los espectros individuales y la normalización de la lista de picos **(Clin Pro Tools user Manual versión 3.1,** 2011**; Stephens,** 1974**).**

Una vez procesados los datos, se evaluó la variabilidad del grupo de muestras en su conjunto, es decir, la posibilidad de formar dos clases bien separadas, pero cuya información sea homogénea. En conjuntos de datos con muchos grupos de variables, las mismas a menudo muestran un comportamiento similar y contienen información redundante. Para ello, se realizó el análisis de componentes principales (PCA) para evaluar la capacidad de distinción entre los perfiles de positivo / negativo para la infección, pero reduciendo la dimensionalidad del conjunto de datos y, al mismo tiempo, reteniendo la información.
En el análisis de PCA se reemplazan conjuntos de variables por una única variable, que se denomina componente principal 1, 2 o 3, las cuales suelen contener la mayor parte de la varianza.

Entrenamiento supervisado. Para la generación de cada modelo se ensayaron los tres tipos de algoritmos que existen en el software y cuyos fundamentos han sido previamente descriptos (GA, SNN, QC); la selección del número máximo de mejores picos se estableció como 100, y el número máximo de generaciones se estableció como 50. Los números de los vecinos más cercanos evaluados en el algoritmo GA-kNN fueron todos los disponibles (1, 3, 5 y 7) para cada clasificación binaria.

Establecimiento del conjunto de entrenamiento:

Se seleccionaron 432 espectros correspondientes a triplicados de:
55 muestras positivas-COVID-19,
57 muestras negativas-COVID-19,
24 muestras positivas-FLU,
8 muestras de otros virus respiratorios

Los cuales se emplearon para construir los modelos de clasificación que se detallan:

1- Modelo de dos clases denominado A
Clase 1 = muestras positivas COVID-19
Clase 2 = muestras negativas COVID-19

2- Modelo de tres clases denominado B
Clase 1 = muestras positivas para COVID-19
Clase 2 = muestras negativas para COVID-19
Clase 3 = muestras positivas para otros virus respiratorios

3- Modelo de dos clases denominado C
Clase 1 = muestras positivas para COVID-19
Clase 2 = muestras positivas para influenza

El diseño de este último modelo se decidió porque muchas muestras de FLU se clasificaron como positivas para COVID-19 cuando se utilizó el modelo de 3 clases, por lo que, para optimizar los resultados, se creó un modelo específico de COVID-19 versus FLU y se aplicó solo cuando una muestra fue positiva para COVID-19 u otro virus respiratorio en el modelo de 3 clases.

Validación cruzada y capacidad de reconocimiento.

A continuación, se llevó a cabo la validación cruzada (VC) para obtener una medición estadística imparcial del rendimiento. Los datos se dividieron en subconjuntos de forma aleatoria; cada subconjunto sirvió como conjunto de validación para el modelo entrenado por el resto de los subconjuntos iterativamente. La precisión de la clasificación se obtuvo de la media de las evaluaciones.

Se calculó además la capacidad de reconocimiento (CR), para evaluar el rendimiento teórico del algoritmo, es decir, la clasificación adecuada de un conjunto de datos.
Si la CR es demasiado baja (<80-90%), significa que el modelo no fue capaz de aprender a partir de las características de las muestras y esto ocurre cuando no es posible encontrar una relación entre los datos de esa marcación. Por otro lado, si la CR es alta, esto solo indicará que el modelo ha aprendido en base al conjunto de datos de entrenamiento, pero no quiere decir que la clasificación de nuevas muestras vaya a ser exitosa. Para ello se debe realizar una prueba de validación externa con un nuevo conjunto de datos.

Prueba de validación externa.

Se realizó una validación externa con un conjunto de prueba de 501 espectros correspondientes a 167 muestras (68 eran positivas para SARS-CoV-2, 89 negativas para SARS-CoV-2, 10 positivas para otros virus respiratorios), diferentes a las muestras utilizadas para crear el conjunto de entrenamiento; los mismos se clasificaron utilizando la función "clasificar" en ClinPro Tools.
Se presentó el espectro de cada muestra del grupo de validación, al modelo de clasificación seleccionado. Entonces el software arrojó un resultado que se comparó con la técnica de referencia, para evaluar la concordancia.
Esta clasificación se realizó de forma individual con los tres modelos calculados.

Análisis estadístico.

Se calcularon valores de precisión, sensibilidad, especificidad, VPP y VPN (**ClinPro Tools User Manual Version 3.0**, 2011; **Stephens,** 1974). La FDA recomendaba el uso de VPP y VPN debido a que no tenía probado un "estándar de oro" para la detección de SARS-CoV-2 hasta ese momento **(United States Food and Drug Administration Guidance on Statistical Methods for Evaluating In Vitro Diagnostic Tests. https://www.fda.gov/media/71147/).**

7.4. Evaluación del desempeño de las ESTRATEGIAS 2 y 3 sobre muestras frescas obtenidas de la rutina diaria de un laboratorio.

Comparación de los resultados con respecto a los obtenidos sobre muestras congeladas.

Debido a la complejidad de la situación en los centros de salud a principios del año 2020 y a la dificultad para conseguir muestras en las condiciones apropiadas; del total de espectros adquiridos en el Instituto Malbrán para los ensayos de validación, solo 30 (denominadas desafío 1) fueron adquiridos en las mismas condiciones que las utilizadas para crear los modelos predictivos.

Se recibieron otras 64 muestras adicionales provenientes del servicio de Virosis Respiratorias - ANLIS Malbrán, que no formaron parte del análisis de resultados de este proyecto de tesis, ya que fueron procesadas en fresco, es decir al momento que ingresaron al laboratorio sin el proceso de ultra-congelación (-80°C). Únicamente con el objetivo de aprovechar la disponibilidad de estas muestras, se realizó un ensayo adicional (desafíos 2 y 3) evaluando las estrategias 2 y 3, pero sobre muestras frescas.

7.5. Análisis de correlación entre los valores de CT y el resultado de la EM.

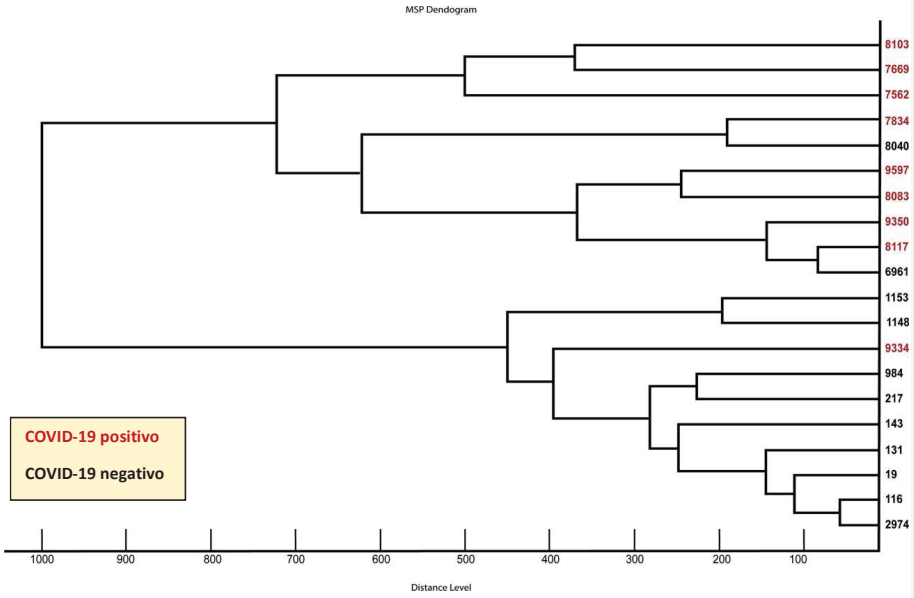
Finalmente, se compararon los valores de CT (*cycle treshold*) obtenidos de las muestras positivas a partir de la técnica de referencia, con respecto a los resultados de ML para evaluar una posible correlación entre ellos.
El CT es indicativo de presencia viral y su valor está asociado con la carga viral por ende con la cantidad de ARN presente en la muestra. Si el valor de CT es 35 o más en la prueba RT-PCR indica COVID-19 negativo. Si el valor de CT es inferior a 35 en la prueba RT-PCR, indica COVID-19 positivo.

8. RESULTADOS

8.1. ESTRATEGIA 1) Creación de una Base de Datos "*in house*" de espectros proteicos de referencia. Evaluación.

Basado en las similitudes de las señales de masa e intensidad de los picos se construyó el dendrograma a partir de los MSPs seleccionados que conformaron la base de datos.
En la **Figura 8** se observa la separación en dos clados mayoritarios correspondientes a las muestras positivas (rojo) y a las negativas (negro) para la enfermedad. En este análisis, los espectros fueron comparados entre sí de a pares, y el valor obtenido de dicha comparación permitió la agrupación jerárquica en ramas dentro de un árbol taxonómico según la cercanía existente entre ellos. La distancia de las ramas del árbol está relacionada con la similitud de los espectros y, por lo tanto, la similitud entre las muestras seleccionadas.
El análisis se realizó con la herramienta estadística integrada en el paquete del software.
Es válido recordar que aquí la distancia se muestra en unidades relativas, donde 0 indica similitud completa y 1000 indica completa disimilitud.
En la Figura se puede observar que los dos clados principales se separaron con el máximo nivel de distancia.

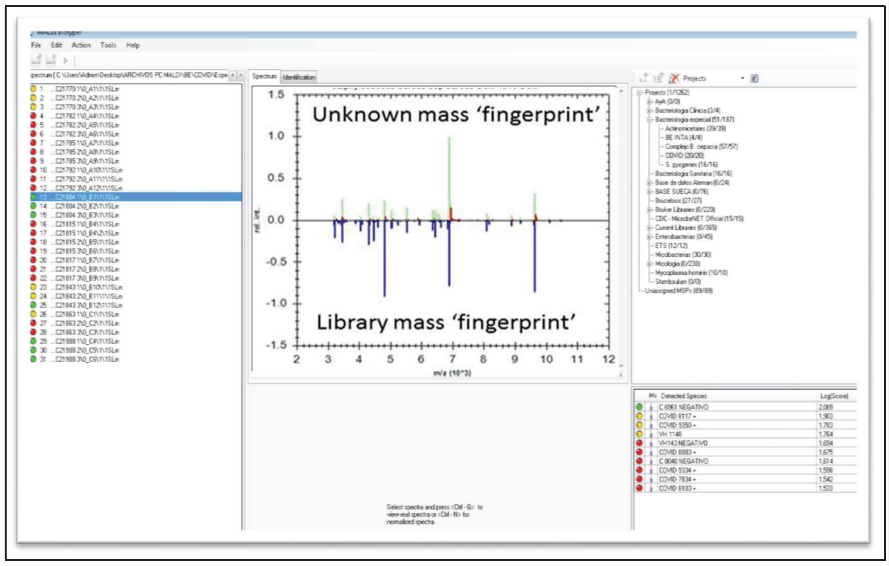
Figura 8. Dendrograma basado en los MSPs de las 20 muestras incorporadas en la nueva base de datos *"in house"*. El eje horizontal del dendrograma representa la distancia calculada en el agrupamiento, mostrada en unidades relativas, correspondiente a la similitud de los espectros proteómicos. El gráfico se creó empleando el software MALDI Biotyper OC v3.1.66 utilizando los parámetros estándares.



El espectro de masas de cada nueva muestra representó una "huella dactilar molecular" que se comparó con los MSPs de la base de datos complementaria.

En la **Figura 9** se muestra la representación gráfica de la identificación “fuera de línea” realizada sobre los espectros previamente almacenados a diferencia de la clasificación habitual en tiempo real.
En el extremo inferior derecho se observan los diez primeros valores de *score (top ten)* en orden decreciente de similitud, obtenidos para una muestra.

Figura 9. Representación gráfica de la coincidencia entre la muestra y la referencia obtenida en la ventana “identification view” del software MALDI Biotyper OC 3.1.66. La lista de picos del espectro de muestra desconocido se evidencia en la mitad superior del gráfico. El color de los picos refleja el grado de coincidencia con el MSP de referencia (verde = coincidencia completa, amarillo = coincidencia parcial, rojo = no coincidencia). La mitad inferior del gráfico muestra la lista de picos del MSP de referencia seleccionado en azul utilizando una escala de intensidad invertida.



Sólo el 22% de las muestras clasificadas (50/223; 16 COVID-19 positivas y 34 COVID-19 negativas) presentaron valores de puntaje ≥2.0, requerimiento necesario para una identificación confiable según recomendaciones del fabricante con el sistema de puntuación de Bruker **(Espinal et al., 2012)**. Los resultados se detallan en la **Tabla 2**.

El resto presentaron valores de *score* bajos, haciendo la identificación no confiable.

Tabla 2. Resultados de las muestras que arrojaron valor de *score* ≥2.0 del total de muestras utilizadas para evaluar la base de datos de referencia (N=50).

RESULTADO DE RT-PCR	ID muestra	RESULTADO EVALUACION DE BD	VALOR DE *SCORE*
SARS-COV-2 detectable	13468	COVID-19 Negativo	2.0
	13477	COVID-19 Negativo	2.185
	13478	COVID-19 Negativo	2.193
	21804	COVID-19 Negativo	2.170
	21843	COVID-19 Negativo	2.023
	21908	COVID-19 Positivo	2.107
	3.2.3/A10	COVID-19 Positivo	2.0
	5.2.1/B6	COVID-19 Positivo	2.0
	2152290	COVID-19 Positivo	2.1
	2152330	COVID-19 Negativo	2.0
	215051	COVID-19 Positivo	2.0
	2151105	COVID-19 Positivo	2.0
	2151593	COVID-19 Positivo	2.1
	2151816	COVID-19 Positivo	2.03
	2151941	COVID-19 Positivo	2.0
	2150151	COVID-19 Positivo	2.0
SARS-COV 2 no detectable	21902	COVID-19 Positivo	2.50
	21903	COVID-19 Positivo	2.038
	21904	COVID-19 Negativo	2.159
	21915	COVID-19 Positivo	2.120
	15342	COVID-19 Positivo	2.07

RESULTADO DE RT-PCR	ID muestra	RESULTADO EVALUACION DE BD	VALOR DE *SCORE*
	15343	COVID-19 Negativo	2.1
	15344	COVID-19 Positivo	2.76
	15353	COVID-19 Positivo	2.263
	15358	COVID-19 Positivo	2.084
	21920	COVID-19 Positivo	2.10
	M2150034	COVID-19 Positivo	2.0
	M2150063	COVID-19 Positivo	2.08
	M2150067	COVID-19 Positivo	2.1
	M2150071	COVID-19 Positivo	2.1
	M2150096	COVID-19 Positivo	2.0
	M215014	COVID-19 Positivo	2.1
	M2150120	COVID-19 Positivo	2.0
	M2150122	COVID-19 Negativo	2.0
	M2150128	COVID-19 Positivo	2.1
	8.1.3/c5	COVID-19 Positivo	2.0
	10.4.1/d4	COVID-19 Positivo	2.0
	2151242	COVID-19 Positivo	2.17
	2152065	COVID-19 Negativo	2.0
	2152068	COVID-19 Positivo	2.0
	2152069	COVID-19 Negativo	2.1
	2152317	COVID-19 Positivo	2.0
	2152321	COVID-19 Positivo	2.0
	2152343	COVID-19 Positivo	2.0
	2152346	COVID-19 Positivo	2.0
	2152348	COVID-19 Positivo	2.1
Influenza	R10	COVID-19 Negativo	2.182
	R19	COVID-19 Negativo	2.045
	R76	COVID-19 Negativo	2.351
	FLU86	COVID-19 Negativo	2.035

En la **Tabla 3** se presentan los parámetros analíticos del desempeño de la ESTRATEGIA 1-Creación de una Base de Datos *"in house*".

Tabla 3. Parámetros analíticos resultantes de la evaluación de desempeño de la ESTRATEGIA 1 (N=50).

PARÁMETROS EVALUADOS	(%)	95% IC (%)
Exactitud	38.0	24.65-52.83
Especificidad	26.5	12.88-44.36
Sensibilidad	62.5	35.43-84.80
VPP	28.6	20.65-38.07
VPN	60.0	39.19-52.83

Valores obtenidos con la herramienta MedCalc's Diagnostic test evaluation calculator.

8.2. ESTRATEGIA 2) Detección manual y automatizada de potenciales picos biomarcadores.

Análisis MANUAL mediante software Flex Analysis v3.4.

En base al análisis manual, se detectaron seis potenciales picos biomarcadores en los siguientes valores de *m/z*:

3372 Da	***3442 Da***	***3465 Da***	***3488 Da***	***6347 Da***	***10836 Da***

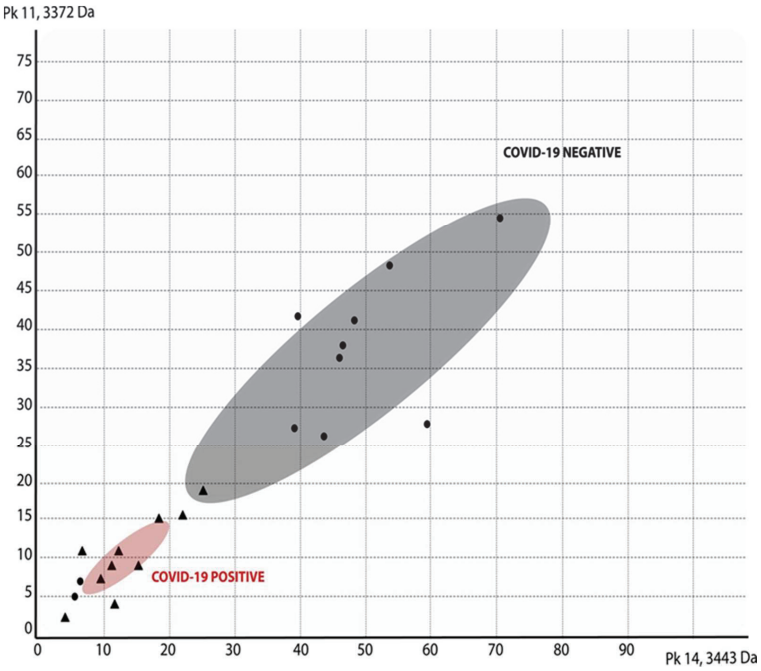
Análisis AUTOMATIZADO mediante software ClinPro Tools v 3.0.

Los mismos archivos de espectros de los MSPs, se importaron al software ClinPro Tools para el reconocimiento de patrones diferenciales entre grupos. Se ingresaron los datos en dos clases diferentes, según el resultado conocido de RT-PCR. Este tipo de análisis pertenece al método supervisado.

En base al cálculo estadístico de picos de cada clase (**Tabla suplementaria S3)**, se obtuvo: la distribución 2D de los dos mejores picos, las curvas ROC y la varianza, como se describe a continuación.

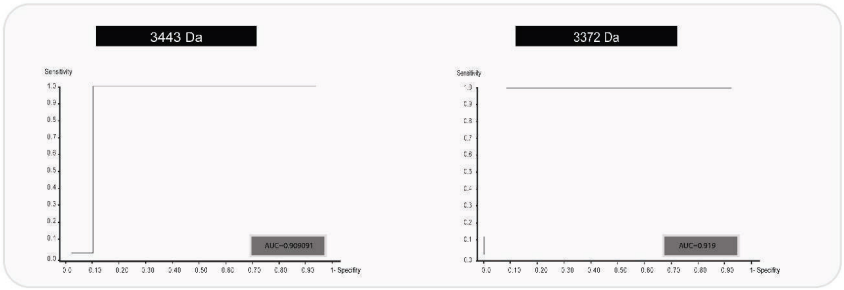
En la **Figura 10** se muestra la gráfica de distribución bidimensional (2D) de todos los espectros de cada clase en base a los dos mejores picos obtenidos para la separación. Los valores de *m/z* de estos se muestran en los ejes x e y.

Figura 10. Distribución de dos picos seleccionados en los espectros no excluidos de las dos clases durante la generación de un modelo de clasificación. Los datos se muestran en un plano bidimensional. Las elipses representan la desviación estándar del área / intensidad de los picos promedio. El eje X muestra los valores del área / intensidad del pico con respecto al pico más importante de acuerdo con el valor p, y el eje Y muestra los valores del área / intensidad del pico para el segundo pico más importante, respectivamente. Las medidas de los ejes se dan en unidades arbitrarias que se eligen automáticamente por el software para ajustarse de forma óptima en el plano. (ClinPro Tools v3.0).



El poder discriminatorio de cada pico biomarcador potencial se obtuvo mediante el análisis del área bajo la curva ROC (característica operativa del receptor). En la **Figura 11** se muestran las curvas ROC de los picos con la mayor capacidad de discriminación (AUC ≥ 0,80).
Estos mismos picos son los que aparecen en la **Figura 10** como los mejores picos para la separación de clases.

Figura 11. Curvas ROC de los picos con mayor poder discriminatorio en base al valor de área bajo la curva calculado. (ClinProTools v3.0).



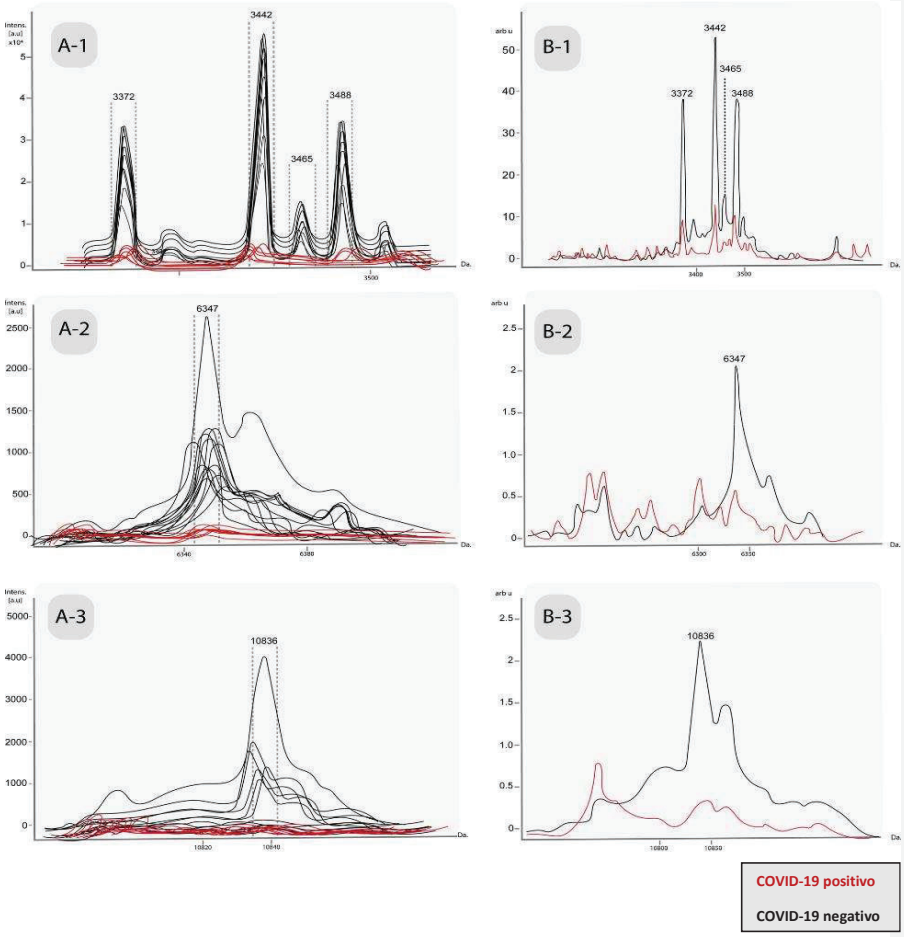
Los picos característicos de los dos grupos generados, clase 1 = muestras positivas COVID-19; clase 2 = muestras negativas COVID-19, se seleccionaron mediante la "Tabla estadística de picos ", seguida de la confirmación manual de que esos mismos picos eran distinguibles usando Flex Analysis.

El gráfico de los picos hallados en los espectros individuales de los MSPs, obtenidos mediante el análisis manual en Flex Analysis v3.4, se puede observar en la **Figura 12.A**.
Asimismo, en la **Figura 12.B**, se muestran los mismos picos en los espectros promedio de cada clase, obtenidos en forma automatizada con ClinProTools v3.0.

Figura 12. A Picos característicos en los espectros individuales de los MSPs que conforman la base de datos entre muestras COVID-19 positivo versus muestras COVID-19 negativo (Flex Analysis v3.4).

Figura 12. B Espectros promedio de los mismos picos.

A-1 / B-1: 3372 Da, 3442 Da, 3465 Da y 3488 Da. A-2 / B-2: 6347 Da. A-3 / B-3: 10836 Da (ClinProTools v3.0).



En base al perfil de biomarcadores hallados y en comparación con la técnica de referencia, se estableció un criterio de interpretación de las muestras:

Se consideró:
4, 5, 6 picos estaban presentes = muestra negativa.
0, 1, 2 picos estaban presentes = muestra positiva.
3 picos estuvieron presentes = Resultado No concluyente (NC).

La búsqueda de potenciales picos biomarcadores sobre los 20 MSPs que conformaron la BD *in house* y su interpretación según el criterio propuesto, se detallan en la **Tabla 4**.

Tabla 4. Análisis individual de cada MSP para la búsqueda de 6 potenciales picos biomarcadores y su interpretación (N = 20).

MSP ID	Potencial Biomarcador (Da)						INTERPRETACION	RESULTADO RT-PCR
	3372	3442	3465	3488	6347	10836		
8083	+	-	-	+	-	-	POSITIVO	DETECTABLE
8103	+	+	+	+	-	-	NEGATIVO	DETECTABLE
8117	-	+	-	+	-	-	POSITIVO	DETECTABLE
7562	+	+	+	+	-	-	NEGATIVO	DETECTABLE
7669	-	-	-	+	-	-	POSITIVO	DETECTABLE
7834	+	+	-	-	-	-	POSITIVO	DETECTABLE
9334	+	+	+	+	-	-	NEGATIVO	DETECTABLE
9350	-	-	-	-	-	-	POSITIVO	DETECTABLE
9597	+	+	-	-	-	-	POSITIVO	DETECTABLE
116	+	+	+	+	+	+	NEGATIVO	ND
131	+	+	+	+	+	+	NEGATIVO	ND
143	+	+	+	+	-	+	NEGATIVO	ND
19	+	+	+	+	-	-	NEGATIVO	ND
984	+	+	+	+	-	-	NEGATIVO	ND
6961	+	+	-	-	-	-	POSITIVO	ND
8040	+	+	+	+	-	-	NEGATIVO	ND
217	+	+	+	+	-	+	NEGATIVO	ND
974	+	+	-	+	+	+	NEGATIVO	ND
1148	+	+	+	+	-	-	NEGATIVO	ND
1153	+	+	+	+	-	-	NEGATIVO	ND

ND: NO DETECTABLE

+: pico presente / -: pico ausente.

En la **Tabla 5** se presentan los parámetros analíticos de desempeño de la ESTRATEGIA 2- Detección manual y automatizada de potenciales picos biomarcadores.

Tabla 5. Parámetros analíticos del desempeño de la ESTRATEGIA 2 sobre los 20 MSPs que conformaron la BD *in house*.

PARÁMETROS EVALUADOS	(%)	95% IC (%)
Sensibilidad	66.67	29.93- 92.51
Especificidad	90.91	58.72-99.77
VPP	85.71	46.67-97.63
VPN	76.92	56.50- 89.54
Exactitud	80.00	56.34- 94.27

Valores obtenidos con la herramienta MedCalc's Diagnostic test evaluation calculator

Asignación de identidad de péptido / proteína.

No fue posible identificar los biomarcadores potenciales mediante asignación bioinformática debido principalmente a la limitación que presentan estos equipos de EM en términos de resolución y precisión de masas.
Por lo tanto, ninguno de los picos encontrados en este trabajo de tesis pudo atribuirse molecularmente a proteínas específicas del virus o del huésped, pero sí resulta válido remarcar que todos los potenciales biomarcadores se detectaron de forma llamativa en la mayoría de las muestras negativas.

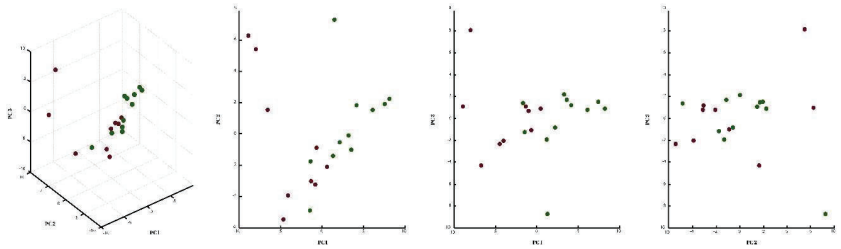
8.3. ESTRATEGIA 3) Diseño de modelos predictivos de clasificación rápida basados en herramientas de ML.

Los espectros a partir de este tipo de muestras clínicas son extremadamente complejos y los picos visualizados representaron el total del contenido ionizable recogido a partir de un hisopo nasal. Sin embargo, se logró la detección de patrones diferenciables por técnicas de ML que permitieron generar modelos apropiados para diferenciar los casos de COVID-19 de los no COVID-19.

Se realizó el análisis de componentes principales (PCA) para evaluar la varianza en el grupo de datos, pero reduciendo la dimensionalidad del conjunto y, al mismo tiempo, reteniendo la información.

En la **Figura 13**, se observa la distribución de los datos de cada clase en base al análisis de componentes principales (PCA) calculados automáticamente por la herramienta externa MATLAB integrada a ClinPro Tools, para extraer, mostrar y clasificar la varianza dentro del conjunto.

Figura 13. Resultados del PCA para las dos clases seleccionadas utilizando los datos de MSPs. En rojo se observan las muestras COVID-19 positivo y en verde las COVID-19 negativo (MATLAB-ClinPro Tools V3.0).



En el entrenamiento supervisado, para el diseño de los modelos finales de clasificación se optó por el algoritmo GA-kNN. Este algoritmo alcanzó los mejores resultados de validación cruzada (VC) y capacidad de reconocimiento (CR) entre todos los algoritmos ensayados, exhibiendo la mejor eficiencia en la clasificación de los pacientes de prueba.

Los resultados de los valores de CR y VC de los modelos se resumen en la **Tabla 6**.

Tabla 6. Indicadores de rendimiento de cada modelo diseñado.

MODELO	CR (%)	VC (%)
A	**100**	**93**
B	**100**	**87**
C	**100**	**93**

Los diez mejores picos fueron los que mostraron una diferencia significativa entre las clases durante la generación de los modelos y se detallan en la **Tabla 7** junto sus parámetros estadísticos. La tabla completa de picos obtenidos para cada modelo seleccionado en ClinPro Tools se puede encontrar en el material suplementario, **Tabla S4**.

El total de regiones de integración de los picos usados según los diferentes algoritmos para cada modelo se encuentran en el material suplementario, **Tabla S5**.

Tabla 7. Picos característicos (los diez mejores) obtenidos estadísticamente para cada modelo desarrollado (ClinPro Tools V3.0).

2 Clases modelo A				
Mass	DAve	PTTA	PWKW	PAD
3372,19	7,85	0,000171	0,0000138	< 0.000001
3443,14	7,79	0,00834	0,000042	< 0.000001
4966,6	4,79	0,000277	0,000093	< 0.000001
5236,08	3,88	0,00584	0,000764	< 0.000001
4078,24	3,49	0,000011	< 0.000001	0
3487,1	3,24	0,0737	0,0191	< 0.000001
4985,73	3,07	0,000033	0,0000101	< 0.000001
3359,22	2,66	0,000032	0,000015	< 0.000001
3393,78	2,54	0,00182	0,00418	< 0.000001
3475,85	2,48	0,000608	0,000148	< 0.000001

3 Clases modelo B				
Mass	DAve	PTTA	PWKW	PAD
3443,31	13,35	0,00149	0,00000568	< 0.000001
3372,29	11,52	0,000023	0,00000427	< 0.000001
3487,4	11,42	0,00969	0,0359	< 0.000001
5236,12	8,63	< 0.000001	< 0.000001	< 0.000001
4966,4	6,74	0,0000157	< 0.000001	< 0.000001
4985,38	4,12	0,00000296	< 0.000001	< 0.000001
5137,86	3,89	< 0.000001	< 0.000001	< 0.000001
3465,33	3,6	0,0183	0,00758	< 0.000001
5382,94	3,58	< 0.000001	0,00000425	< 0.000001
5157,12	3,52	< 0.000001	< 0.000001	< 0.000001

2 Clases modelo C				
Mass	DAve	PTTA	PWKW	PAD
3443,19	13,35	0,00332	0,000319	< 0.000001
3372,15	11,52	0,00108	0,00059	< 0.000001
4965,83	6,69	0,0000165	0,00000382	< 0.000001
5235,02	4,62	0,00000929	0,000128	< 0.000001
4985,26	4,2	0,00000651	0,00000106	< 0.000001
3465,33	3,61	0,0384	0,0414	< 0.000001
3394,17	3,3	0,00856	0,000829	< 0.000001
4939,73	3,06	< 0.000001	0,0000776	< 0.000001
5381,66	2,93	0,0000266	0,0117	< 0.000001
4077,47	2,5	0,00881	0,00794	< 0.000001

DAve: diferencia entre la máxima y la mínima intensidad del pico promedio de todas las clases.
PTTA: valor de P obtenido mediante la prueba t. donde 0: bueno y 1: malo
PWKW: valor de p obtenido mediante la prueba de Wilcoxon / Kruskal-Wallis. donde 0: bueno y 1: malo
PAD: valor de p obtenido mediante la prueba de Anderson-Darling. 0: distribución no normal, 1: distribución normal

Validación. Los resultados de desempeño de ML, sobre un conjunto de prueba de 167 nuevas muestras (501 espectros), se detallan en la **Tabla 8.**

Para ello, se presentó cada espectro del grupo de validación externa, al modelo de clasificación seleccionado y utilizando la función "clasificar". Entonces el software arrojó un resultado que se comparó con las técnicas de referencia actuales, para evaluar la concordancia entre los métodos.

Tabla 8. Parámetros analíticos resultantes de la evaluación de desempeño de la ESTRATEGIA 3 - Diseño de modelos predictivos de clasificación basados en *ML*.

Parámetros evaluados	(%)	95% IC (%)
Sensibilidad	50.72	38.41 - 62.98
Especificidad	67.35	57.13 - 76.48
VPP	52.24	43.10 - 61.23
VPN	66.00	59.56 - 71.90
Exactitud	60.48	52.63 - 67.95

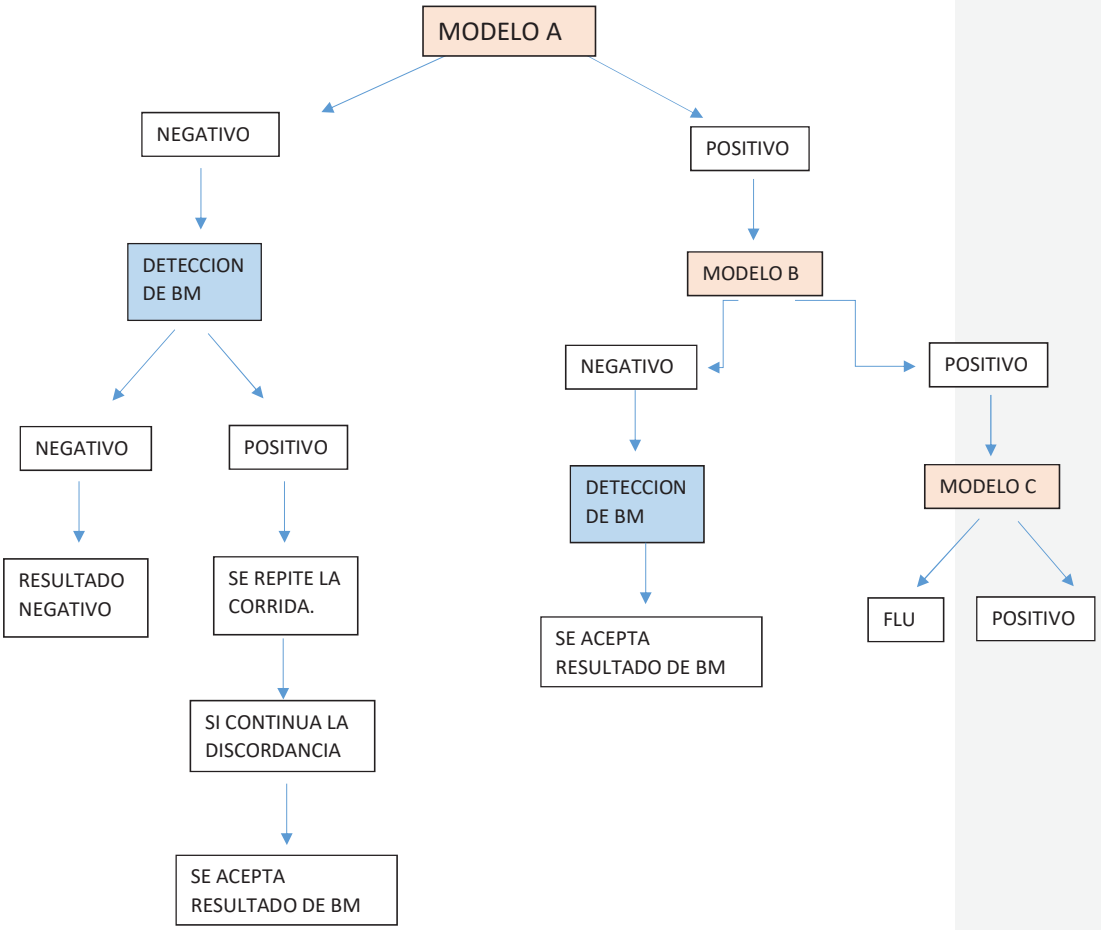
Valores obtenidos con la herramienta MedCalc's Diagnostic test evaluation calculator

ALGORITMO FINAL.

En base a la experiencia previa en el área y a los parámetros estadísticos calculados al comparar los resultados de las estrategias 2 y 3 de forma individual con respecto a la técnica de referencia, es que se decidió la aplicación de un algoritmo final basado en el uso conjunto de los tres modelos de ML diseñados y la detección de biomarcadores (estrategia 2+3), como se detalla debajo.

En los casos en que se produjeron discordancias entre los resultados de ML y detección de BM, la clasificación se priorizó sobre la base de la detección de BM, ya que mejoraron el rendimiento de la metodología implementada en todos los casos discordantes. Esta decisión está justificada en los parámetros estadísticos obtenidos.

ALGORITMO FINAL DE CLASIFICACION



Este enfoque basado en los algoritmos de aprendizaje automático con la combinación de biomarcadores potenciales se evaluó con las 167 muestras clínicas y los resultados del rendimiento final se resumen en la **Tabla 9**.

Los resultados obtenidos a partir del total de muestras procesadas, en comparación con la técnica de referencia, se puede encontrar en el material suplementario, **Tabla S6.**

Tabla 9. Parámetros analíticos resultantes de la evaluación de desempeño del aprendizaje automático combinado con biomarcadores potenciales (N=167).
ESTRATEGIA 2 y ESTRATEGIA 3.

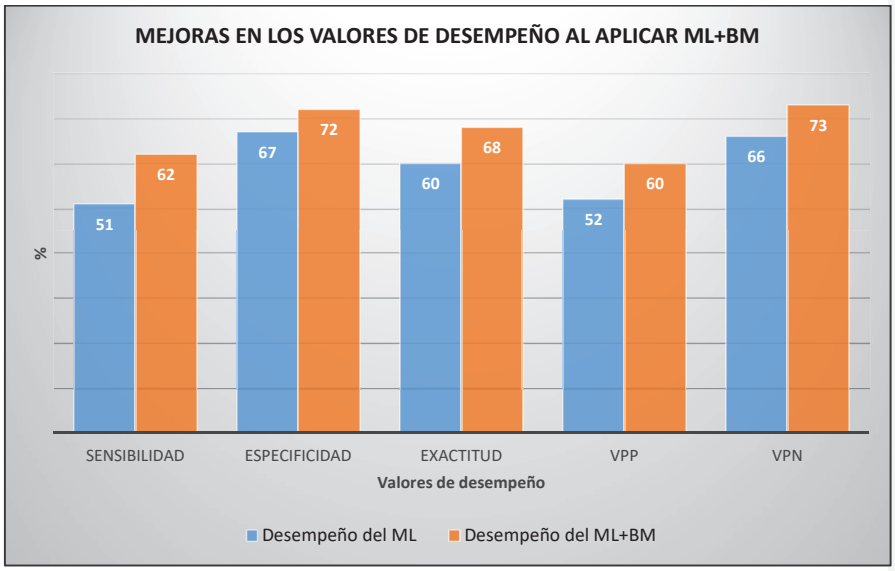
PARÁMETROS EVALUADOS	(%)	IC 95% (%)
EXACTITUD	67.66	60.00-74.69
ESPECIFICIDAD	71.72	61.78-80.31
SENSIBILIDAD	61.76	49.18-73.29
VALOR PREDICTIVO POSITIVO	60.00	51.01-68.37
VALOR PREDICTIVO NEGATIVO	73.20	66.33-79.10

Valores obtenidos con la herramienta MedCalc's Diagnostic test evaluation calculator

En la **Figura 14** se muestran las mejoras en los valores estadísticos de desempeño al aplicar ML + BM (columna naranja), en comparación al uso de ML únicamente (columna celeste) sobre el total de muestras del grupo de validación.

Es importante aclarar que este tipo de análisis no sería apropiado sobre la detección de BM individualmente, debido a que 117 muestras arrojaron resultado No Conclusivo (NC), por lo que no es posible calcular el desempeño general de los BM sobre el total de muestras, a pesar del excelente desempeño que presentaron cuando sí se detectaron los picos según el criterio de interpretación establecido.

Figura 14. Cambios en los valores del desempeño al utilizar la clasificación basada en el aprendizaje automático únicamente, en comparación con la combinación de ESTRATEGIAS 2 y 3 (Microsoft Excel 2016).



8.4. Evaluación del desempeño de las ESTRATEGIAS 2 y 3 sobre muestras frescas obtenidas de la rutina diaria de un laboratorio.

Comparación de los resultados con respecto a los obtenidos sobre muestras congeladas.

Para llevar a cabo este análisis adicional, se realizó el "desafío 1" con muestras que también habían sido crio preservadas (N=30) y para los "desafíos 2 y 3" (N=64), se obtuvieron espectros a partir de muestras frescas que no habían sido sometidas a un proceso de congelamiento-descongelamiento previo.

Los resultados se detallan en la **Tabla 10**; las celdas en verde corresponden a resultados concordantes con la RT-PCR, las celdas en amarillo corresponden a falsos positivos y las celdas en rojo a falsos negativos.

Tabla 10. Resultados del desempeño del algoritmo final sobre tres tandas de muestras adquiridas en distintas condiciones (congeladas vs frescas).

DESAFIO 1 (18-06-2020) **MUESTRAS PRESERVADAS A -80C**	RESULTADO MALDITOF	
RT-PCR	N	
SARS COV-2	10	POSITIVO
NO DETECTABLE	3	POSITIVO
NO DETECTABLE	7	NEGATIVO
FLU A	3	POSITIVO
FLU A	7	NEGATIVO

DESAFIO 2 (07/07/2020) MUESTRAS SEMBRADAS DURANTE LA RUTINA DIAGNÓSTICA (SIN CONGELAMIENTO PREVIO)		
RT-PCR	N	RESULTADO MALDITOF
NO DETECTABLE	10	POSITIVO
NO DETECTABLE	8	NEGATIVO
SARS COV-2	8	NEGATIVO
SARS COV-2	6	POSITIVO

DESAFIO 3 (09/07/2020) MUESTRAS SEMBRADAS DURANTE LA RUTINA DIAGNÓSTICA (SIN CONGELAMIENTO PREVIO)		
RT-PCR	N	RESULTADO MALDITOF
SARS COV-2	14	POSITIVO
NO DETECTABLE	8	POSITIVO
NO DETECTABLE	4	NEGATIVO
SARS COV-2	6	NEGATIVO

Evaluación del desempeño de las estrategias 2 y 3 en muestras frescas y en muestras congeladas.

El cálculo de los valores de desempeño del método propuesto sobre el total de muestras desafiadas (N=94) se muestra en la **Tabla 11**.

Se obtuvieron además los valores de desempeño del algoritmo en forma individual para cada grupo de muestras adquiridas en diferentes condiciones, y los resultados se muestran en la **Tabla 12**.

Tabla 11. Valores de desempeño del método propuesto para el total de 94 muestras adquiridas en distintas condiciones.

Parámetros del desempeño	%
SENSIBILIDAD	66
ESPECIFICIDAD	52
VPP	55
VPN	63
Exactitud	58

Tabla 12. Valores de desempeño del método propuesto para cada grupo de muestras (Largada) por separado.

Parámetros del desempeño (%)	Largada 1	Largada 2	Largada 3
SENSIBILIDAD	**100**	36	70
ESPECIFICIDAD	**70**	44	33
VPP	**63**	33	64
VPN	**100**	47	40
Exactitud	**80**	40	56

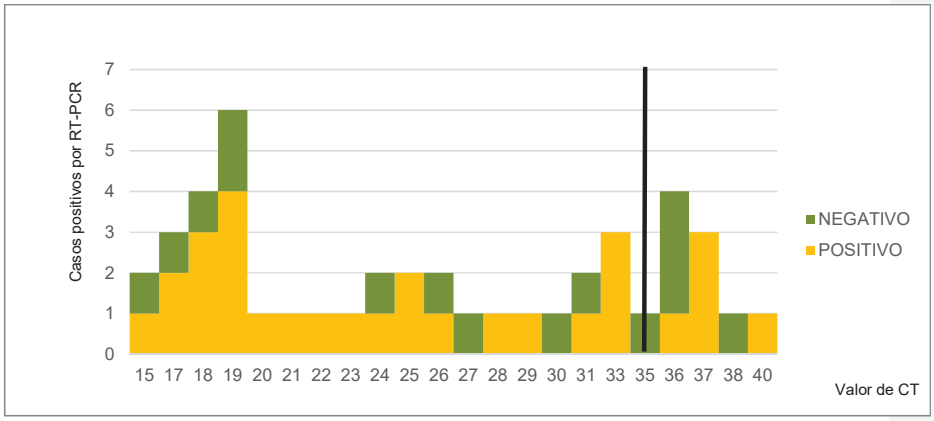
Se observó una marcada diferencia en el desempeño entre el desafío 1 y los desafíos 2 y 3. La única variable entre los mismos fue la característica de las muestras. El desarrollo de los modelos matemáticos fue llevado a cabo con espectros proteicos obtenidos a partir de muestras que habían sido preservadas en ultra-frío (-80C), la misma condición en la que se encontraban las muestras de la primera largada.

8.5. Análisis de correlación entre los valores de CT y el resultado de la EM.

Finalmente, se realizó un análisis de los resultados obtenidos con el desarrollo proteómico en comparación con los valores de CT surgidos de la qRT-PCR de 44 muestras positivas únicamente, de las que se pudo contar con ese dato (**Figura 15**).

No se observó una correlación directa entre los resultados del desarrollo proteómico y los valores de CT de los casos positivos por qRT-PCR. De las diez muestras con valor de CT ≥ a 35, el 50% dio negativo y el 50% positivo por la técnica de EM. De 34 muestras consideradas COVID-19 positivas mediante RT-PCR, el 71% (N=24), dieron también positivas empleando el nuevo desarrollo, sin embargo, no hay diferencias sustanciales en los resultados en base a los cambios en los valores de CT.

Figura 15. Histograma conteniendo los valores de CT de casos positivos confirmados mediante RT-PCR (eje X) en comparación con los resultados del desarrollo proteómico aplicado (verde=resultado negativo; amarillo=resultado positivo).



9. DISCUSIÓN

En la actualidad, la EM es una tecnología ampliamente utilizada debido a que presenta características analíticas únicas, entre las que se incluyen: fácil operación del equipo, simple preparación del analito, bajo consumo de muestra y de reactivos, alta sensibilidad, resultados objetivos y reproducibles y una alta capacidad de automatización. Estas ventajas llevaron a pensar que MALDI–TOF podría utilizarse como una herramienta para la detección de biomarcadores incluso en muestras biológicas diversas sin necesidad de una separación complicada. **Croxatto et al**., en el año 2012, describe la generación rápida y sencilla de perfiles peptídicos directamente a partir de muestras clínicas de pacientes como herramienta de diagnóstico microbiológico.

El análisis multivariado a través de técnicas de aprendizaje automatizado se suele aplicar en estos enfoques para determinar rápidamente diferencias en los patrones biomoleculares de muestras clínicas (**Fitzpatrick et al**., 2020).

Desde principios del año 2020, una nueva enfermedad por coronavirus azota a la población mundial (**COVID-19 Map - Johns Hopkins Coronavirus Resource Center**, 2020), es por ese entonces que la proteómica ha comenzado a discutirse como un potencial método de diagnóstico viral. Sin embargo, tal como describe **Grossegesse et al**., 2020, su uso de rutina aún está en discusión, básicamente porque la EM se enfrenta a dos grandes desafíos: el material viral no puede aislarse tan fácilmente como las bacterias y el ruido que representan las proteínas del huésped en la muestra se convierte en un importante interferente.

Al momento de comenzar este proyecto de tesis, era escaso el conocimiento acerca de la patología molecular y celular de la infección. Los casos iban de leves a graves con una alta proporción de la población que no presentaba síntomas, pero resultaba igualmente infecciosa (**Zhou**, 2020). Con el correr de los meses comenzaron a aparecer publicaciones o preimpresiones describiendo sus principales proteínas y detallando los posibles blancos biológicos del virus. Se conoce que los virus envueltos tienen la característica única de utilizar la membrana de la célula huésped como un revestimiento que cubre su genoma de ARN **(Walls et al**., 2017).

Paralelamente, se originó una respuesta extraordinaria de la comunidad científica a través de esfuerzos colectivos con el objetivo común de comprender la patogenia de la enfermedad, evaluar las estrategias de tratamiento y el desarrollo de vacunas a velocidades sin precedentes para minimizar su impacto en las personas y en la economía global (**Li,** 2016; **Wenzhong**, 2020). Si bien existían decenas de compuestos en evaluación, al momento de redactar este manuscrito, no se disponía de terapias antivirales específicas ni vacunas disponibles contra el nuevo coronavirus SARS-CoV-2; siendo la única metodología diagnóstica disponible en el laboratorio clínico, RT-PCR, que como se conoce, en los países

en desarrollo puede ser una técnica costosa y los reactivos son a menudo difíciles de conseguir, más aún en la actual situación de demanda.

En base a la posibilidad de disponer de un panel de muestras estudiadas por el LNR, Servicio de Virus Respiratorios, y al igual que lo planteado por un gran número de profesionales de la comunidad científica, se propone evaluar los conocimientos previos y la metodología al alcance, con la posibilidad de extender la aplicación de la EM como técnica de tamizaje en el diagnóstico de COVID-19 (**McDermott, et al**., 2020).

Se planteó, de modo similar a lo descripto por **Yi-Tzu Cho** et al., 2015, el desafió de evaluar la aplicación de la EM, en la identificación de casos de COVID-19, partiendo de la detección de biomarcadores específicos y de su aplicación en el diseño de modelos predictivos automatizados.

Iles et al. (2020) utiliza la EM para detectar glicoproteínas de la envoltura viral del SARS-CoV-2 como una técnica de tamizaje global con numerosas ventajas, entre ellas la sencillez de la toma de muestra (gárgaras), la velocidad del análisis y el menor costo de testeo. En nuestro caso, la mayor complejidad se dio al principio del desarrollo cuando se tuvo que superar la dificultad de generar un perfil propio de una muestra clínica con presencia viral, ya que, hasta el momento, la experiencia estaba centrada en protocolos conocidos y validados para la identificación bacteriana de rutina donde las señales medidas representan proteínas intracelulares, en su mayoría proteínas ribosómicas muy abundantes. El rango de masas varía de 2.000 a 20.000 Da y las proteínas reveladas no están glicosiladas, aunque pueden modificarse pos-traduccionalmente de otras formas, como la fosforilación. Sin embargo, para la detección de virus, este rango de masas es inadecuadamente bajo; las proteínas virales observadas en MALDI-TOF son mucho más grandes (10.000 - 300.000 Da) además de que suelen estar fuertemente glicosiladas y modificadas pos-traduccionalmente. Este rango de masas estaba fuera de los límites efectivos para el HCCA y no existía la posibilidad de evaluar otras matrices alternativas. La elección de su utilización se basó además en su desempeño conocido y en publicaciones en el área (**Nachtigall et al**., 2020).

El muestreo para la detección de COVID-19 en las pruebas de primera línea se realiza a través de la nasofaringe y / o hisopados orofaríngeos, seguido de la inactivación viral por calor o mediante el agregado de SDS, para la destrucción de proteínas virales, pero preservando el ácido nucleico. Por el contrario, para que la prueba sea efectiva en MALDI-TOF, la desactivación de la muestra clínica tiene que preservar la membrana viral y sus proteínas, pero destruyendo el ácido nucleico funcional, esto puede llevarse a cabo aplicando 15 minutos de radiación UV sobre la placa de acero sembrada o simplemente mediante el agregado de una matriz ácida y el secado a temperatura ambiente.

No se ensayaron muestras alternativas, debido a que las pruebas sobre gárgaras o saliva se publicaron en la literatura de forma posterior a este desarrollo, pero además debido a que se disponía únicamente de muestras de hisopado nasofaríngeo para llevar a cabo los ensayos. La solución salina del medio de transporte viral es un insumo ampliamente disponible y además era compatible con las técnicas espectrométricas aplicadas.

Originalmente se pensaba que, al restar las especies de alta abundancia en una muestra compleja, el proteoma de baja abundancia se haría visible y conduciría al descubrimiento de biomarcadores, sin embargo, tal como postuló **Righetti et al**. en el año 2013, esta suposición fracasó ya que las especies de baja abundancia permanecen aún más diluidas y difícilmente se pueden detectar. En este trabajo de tesis, tampoco se llevaron a cabo procedimientos de enriquecimiento de proteínas virales dentro de la muestra. La presencia de abundantes moléculas del huésped, como la albúmina, no solo puede enmascarar la detección de proteínas menos abundantes, sino que también suprime la ionización de otras proteínas de baja abundancia y afecta la reproducibilidad. Las técnicas de enriquecimiento y purificación consisten generalmente en separación por ultracentrifugación, precipitación selectiva por adición de sales o polietilenglicol, precipitación con acetona, mediante la adición de hielo, entre otras; metodologías poco prácticas para implementar en la rutina del laboratorio y más aún si nos encontramos frente al desafío que significa una enfermedad emergente de características desconocidas. Se propone entonces la siembra directa de la muestra en la placa de acero para ahorrar tiempo y dinero de modo similar a lo propuesto por **Dao et al**., 2021, donde se procesa para MALDI-TOF directamente el esputo de pacientes con sospecha de tuberculosis y en el trabajo de **Nachtigall et al**., 2020, que emplea las mismas muestras de hisopado nasofaríngeo que las empleadas en este trabajo, para su análisis proteómico. Diferente a esto es lo propuesto por **Lipi et al**., 2020, para estudios proteómicos en sueros donde sí resulta necesario el enriquecimiento de las proteínas de bajo peso molecular que suelen verse enmascaradas por la presencia en gran concentración de las de mayor peso molecular como la albúmina y la inmunoglobulinaG. Este tipo de obstáculos con el proteoma de bajo peso molecular, son bien conocidos en muestras como plasma y suero, pero no a partir de muestras de otra índole, como las utilizadas en este estudio.

Según **Fitzpatrick et al**., 2020 la IA ofrece un importante potencial para la prevención y el control de infecciones, la detección de brotes y la predicción de pacientes con alto riesgo de enfermar. El entrenamiento automatizado es considerado un subdominio de la IA, donde la computadora usa algoritmos para aprender de conjuntos de datos pasados para hacer predicciones sobre nuevos datos, en lugar de ejecutar un conjunto de reglas programadas. Uno de los principales desafíos que acarrea es el de lograr un conjunto de datos representativos de alta calidad para desarrollar modelos precisos para cada contexto en el

que se utilizan. Con el objetivo de aplicar la IA, se procesó un gran conjunto de muestras disponibles para obtener los datos proteómicos del conjunto. El análisis de los componentes principales, las curvas ROC y las técnicas de *Machine Learning* se utilizaron para construir e identificar los modelos predictivos y sus posibles combinaciones, que lograron la mejor distinción de los casos positivos COVID-19 de los negativos para la enfermedad.

La IA presenta muchas ventajas, incluida la velocidad de análisis de conjuntos de datos infinitamente grandes. Para desarrollos futuros, sin embargo, se requiere un cambio de cultura y comportamiento. La mayoría de los estudios hasta la fecha evalúan el desempeño retrospectivamente, por lo que existe la necesidad de una evaluación prospectiva en la vida real del entorno clínico para establecer una estrecha colaboración del personal de salud con los expertos en manejo de datos, que permita interpretar los resultados y garantizar la relevancia clínica. De lo contrario, los errores que se introducen durante el proceso de formación del aprendizaje automático pueden resultar en falsos negativos, clasificaciones erróneas o falta de aplicabilidad real debidas al sobreajuste de los datos que generaron el modelo. Tal como describe **Rhoads** (2020), aquí se utilizó un conjunto de entrenamiento y uno menor de validación al desarrollar y evaluar los algoritmos de la IA, un paso importante para implementar una aplicación de IA en el diagnóstico clínico.

A lo largo del desarrollo de la tesis, se obtuvieron resultados para cada estrategia evaluada: sólo el 22% de las muestras clasificadas con la base de datos complementaria pasaron el valor de *score* estipulado. El resto presentaron valores de puntaje bajos, haciendo la identificación no confiable. Por lo tanto, se decidió discontinuar la utilización de la base de datos complementaria denominada “BE-COVID-19”, debido a los bajos valores de desempeño obtenidos y a la frecuente ocurrencia de resultados variables en la mayoría de las muestras ensayadas. Esto puede deberse a que la identificación bacteriana habitual en MALDI-TOF, se basa en estrictas similitudes en la ubicación e intensidad de los picos entre la muestra incógnita y el espectro de referencia en la Base de Datos; siendo imposible la estandarización de estos criterios sobre espectros obtenidos a partir de muestras clínicas y sin un mayor procesamiento previo.

La detección manual y automatizada de picos biomarcadores en los softwares disponibles, arrojó 6 picos diferenciales a 3372 Da, 3442 Da, 3465 Da, 3488 Da, 6347 Da y 10836 Da, presentes llamativamente en las muestras negativas y ausentes en las positivas.

Los valores de desempeño observados en la detección de biomarcadores (estrategia 2) sobre los 20 MSPs fueron aceptables, sin embargo, al momento de decidir el algoritmo final de identificación se optó por la combinación del entrenamiento automatizado y la detección de los picos mencionados. Esta decisión diagnóstica está basada en la experiencia previa del investigador en diferentes enfoques (**Manfredi et al.**, 2021) y en la literatura; tal es el caso de lo propuesto por **Chan et al**., 2021, donde se aplica el metanálisis para investigar el

aprendizaje automático y los nuevos biomarcadores en el diagnóstico de la enfermad de Alzheimer.

Es válido aclarar, que el excelente desempeño de la estrategia 2 fue producto del análisis inicial únicamente sobre los 20 MSPs que consisten en los perfiles más limpios y reproducibles, con picos intensos y definidos. Pero luego, en la validación de esta estrategia con 167 nuevas muestras de hisopado nasofaríngeo, que se encontraban en distintas condiciones de almacenamiento hasta el momento de su procesamiento, resultó muy difícil independizarse de la variabilidad en los espectros, lo cual afectó directamente las señales en el peptidoma, arrojando solamente el 30% de las muestras (50/167) un resultado (positivo o negativo), es decir que complementó a la identificación de referencia.

Las 117 muestras restantes del desafío arrojaron resultado No Conclusivo en el análisis de picos biomarcadores.

Es por eso, que en este tipo de enfoques se deciden aplicar algoritmos basados en diferentes métodos de tamizaje proteómico; en este caso, se optó por la búsqueda de picos característicos y la clasificación rápida de espectros incógnita aplicando modelos matemáticos de ML (combinación de estrategias 2 y 3).

No fue posible la asignación bioinformática sobre los biomarcadores potenciales. Una de las limitaciones de la elaboración de perfiles por MALDI-TOF es justamente la falta de información sobre la identidad de las especies informativas. Hasta los espectrómetros de masas más precisos del tipo TOF-TOF tienen una potencia limitada en términos de resolución y precisión de masas incluso en un modo de reflector "encendido". Se sabe que la adquisición en modo lineal proporciona una mayor sensibilidad, pero el retorno es la pérdida de resolución y precisión de la masa. La idea de comparar la masa obtenida por MALDI-TOF operando en modo lineal para un pico dado, con la masa teórica de una proteína indexada en una base de datos, como puede ser uniprot (**https://www.uniprot.org/)** para su identificación, sería cometer un profundo error conceptual.

Ninguno de los picos hallados en este trabajo de tesis, puedo atribuirse a proteínas de la respuesta inflamatoria del huésped o a alguna de las proteínas propias del virus, ya que fueron detectadas curiosamente en la mayoría de las muestras negativas de COVID-19, de modo similar a lo descripto por **Nachtigall et al**., 2020.

Esto podría deberse a algún tipo de interacción huésped-virus que requiere de mayores estudios en un futuro, aunque no es incongruente pensar que estos picos podrían deberse a cambios en la respuesta inflamatoria del huésped y que además podrían estar asociados al microbioma humano que se ve seriamente afectado por la infección viral, tal como comienza a revelarse en estudios recientes **(Rhoades et al.,** 2021**).** Con respecto a la complejidad en la caracterización del peptidoma, **Gomila et al**. (2020) probaron la utilidad de la EM MALDI-TOF para clasificar y predecir la gravedad de pacientes COVID-19 en un

entorno clínico de forma urgente y para establecer un tratamiento eficaz. Mediante esta tecnología se pudieron identificar cinco proteínas que se regulaban significativamente en alza en el suero de los pacientes críticos; entre ellas, las proteínas amiloides séricas de fase aguda A1 y A2 inducidas por el virus, que produjeron los picos más intensos a 11.530 y 11.686 Da. Las mismas no fueron detectadas en este trabajo de tesis. Esto puede deberse a que, en la publicación, partieron de otro tipo de muestra clínica (sueros) y a que las proteínas halladas solo se encontraban en el suero de los pacientes más críticos (datos no disponibles).

Por su parte, en el trabajo de **Chivte et al.** (2021), publicación posterior a este proyecto de tesis, adoptaron la recolección de saliva estandarizada para detectar la infección por SARS-CoV-2, ya que las gotitas de saliva son consideradas un vehículo principal de transmisión viral y se informa que las glándulas salivales son un objetivo de infección, así como un reservorio del virus. Los autores han podido detectar Inmunoglobulinas humanas específicas de la enfermedad, pero lo han logrado aplicando un complejo procedimiento de extracción que incluyó filtraciones y precipitación con acetona y ditiotreitol. Utilizaron además controles de IgA, IgG, IgM humanas y empleando la siembra del tipo sándwich: matriz-muestra-matriz. Con este procedimiento se consume mucho más reactivo y optando por la lectura en el rango de 20.000 a 200.000 Da únicamente posible en plataformas de mayor poder de resolución y utilizando como matriz el ácido sinapínico en lugar de HCCA, reactivo no disponible en nuestro país en la actualidad. Es importante tener en cuenta que, aunque los picos encontrados dentro de estos rangos podrían usarse para separar razonablemente a los individuos positivos de COVID-19 de los negativos, la variación en las intensidades máximas fue evidente entre los espectros positivos de COVID-19; esto puede deberse al muestreo en diferentes tiempos en el curso de la infección y a la respuesta inmune variable de cada huésped.

De todas formas, todos estos estudios de asignación de identidades proteicas, posteriores al desarrollo de esta tesis, afirman que se requieren mayores análisis confirmatorios, como la secuenciación de proteínas del proteoma del paciente sano e infectado con SARS-CoV-2. Resulta extremadamente laboriosa la verificación de la identidad de los biomarcadores inmunitarios humanos y se prefiere, al igual que en este trabajo, trabajar fuertemente en el desarrollo de modelos de aprendizaje automático para lograr precisiones más altas e imparciales utilizando múltiples picos en su conjunto como criterio de diagnóstico.

Se resalta la importancia de trabajar con replicados técnicos que disminuyan el riesgo de cometer errores durante la clasificación y aumenten la reproducibilidad.

Cuando se aplicó el algoritmo diagnóstico final sobre el panel de muestras frescas, es decir procesadas inmediatamente en el momento de la apertura para realizar la RT-PCR, se observó una marcada diferencia en el desempeño entre ensayos donde la única variable

fue la característica de las muestras. El desarrollo de los modelos matemáticos fue realizado con espectros proteicos obtenidos a partir de muestras que habían sido preservadas en ultra-frío (-80ºC). Cuando se realizó el ensayo con nuevas muestras que también habían sido crio preservadas previamente (desafío 1), los resultados fueron mucho más reproducibles que cuando se obtuvieron espectros a partir de muestras que no habían sido sometidas a un proceso de congelamiento-descongelamiento previo (desafíos 2 y 3).

Dado que el proceso de congelamiento-descongelamiento de suspensiones que contienen células produce la ruptura de membranas y además la estrategia de la aplicación de MALDITOF se basa en la detección de un peptidoma característico, la hipótesis para explicar las diferencias observadas se focaliza en el efecto del crioproceso sobre los péptidos biomarcadores.

En conclusión, cuando se desarrollan modelos de Inteligencia Artificial, los resultados solo son reproducibles si se procesan las muestras en las mismas condiciones que las que se utilizaron para crear los algoritmos, ya que se ve afectada la calidad del espectro, por ende, la información varia afectando la reproducibilidad del desarrollo informático.

De considerarse la implementación de este desarrollo en un laboratorio clínico, las muestras recibidas deberían ser sometidas a un breve paso de congelamiento previo a -80°C o de no disponer de la condición de ultra frío, unas horas en el freezer, para asegurar la confiabilidad en los resultados obtenidos.

Cuando se analizaron los valores de CT obtenidos de la RT-PCR a partir de las muestras positivas, no se observó una correlación directa entre el valor de CT y los resultados del desarrollo proteómico. Los resultados positivos del desarrollo aparecieron con similar frecuencia entre muestras con alto valor de CT como en los pacientes que presentaron los valores más bajos de CT. Un pequeño estudio reciente llevado a cabo por **Shah et al**., 2021, encontró que no había una relación entre los valores de CT y la gravedad de la enfermedad, sino que los síntomas tienen una correlación más fuerte con los valores de CT; que indica básicamente la presencia del virus en la garganta, no en los pulmones.

Recientemente se ha cuestionado la capacidad de estos valores para reflejar la verdadera carga viral, expertos afirman que los valores de CT para una muestra varían entre diferentes kits y técnicas, también dependen del momento de la recolección de la muestra en relación con el inicio de los síntomas; las muestras recolectadas antes en la enfermedad tendrán valores de CT más bajos que las recolectados más tardíamente. Por lo tanto, el tiempo de muestreo desde el inicio de la enfermedad deberá estandarizarse al comparar los valores de CT entre la enfermedad leve y grave y en este trabajo de tesis no se contó con esa información.

Para calcular los parámetros estadísticos se debieron usar el VPP y el VPN como indicadores de sensibilidad y especificidad, recomendados por la FDA debido a la ausencia de un método "*gold standard"* de detección de SARS-CoV-2 validado hasta ese momento.

Según los resultados preliminares alcanzados a partir de este desarrollo (precisión = 67,66%, sensibilidad = 61,76%, especificidad = 71,72%, VPP =60.00%, VPN =73.20%), los métodos basados en EM junto con el análisis multivariado, demostraron que la proteómica es una herramienta interesante que merece ser explorada como un enfoque diagnóstico complementario de enfermedades infecciosas y no infecciosas, debido a su bajo costo y alto rendimiento (**Tran et al**., 2021, **Fitzpatrick et al**., 2020), incluso como describe **Wang et al**., en su trabajo, para la identificación precisa y rápida de los tipos de *Staphylococcus aureus* resistente a la meticilina (SARM) para el control de la infección; siempre que se logren mejorar los valores obtenidos en este primer enfoque.

Las limitaciones de este estudio incluyen el uso de muestras clínicas congeladas como prueba de concepto y que no se han podido identificar los picos hallados para lograr la caracterización de proteínas virales y factores de respuesta del huésped. Sin embargo, el objetivo de este estudio era determinar si MALDI-TOF-MS potenciado por el ML podía diferenciar entre los pacientes COVID-19 positivos por PCR frente a los que dieron negativo por la técnica de referencia. El estudio evaluó además la detección de otros coronavirus o enfermedades afines, como la influenza en la comunidad.

Recientemente comienzan a discutirse biomarcadores de proteínas para evaluar la progresión de la enfermedad COVID-19 y su posible asociación con la gravedad de los pacientes (**Grenga et al**., 2021) pero utilizando EM en tándem con instrumentos mucho más costosos y de alta resolución.

Desarrollar un sistema clínicamente eficiente y rentable basado en EM para la detección de pandemias virales futuras, requerirá de la comprensión de la química, la biología y la física de la interacción patógeno-huésped y de sus biomoléculas marcadoras únicas. Se requiere, además, el análisis de un gran número de muestras y la implementación de técnicas de enriquecimiento y purificación, para evaluar la aplicabilidad como técnica de tamizaje.

La adopción de la tecnología basada en MALDI-TOF ha tenido efectos dramáticos en la reducción de costos en la identificación microbiológica de infecciones y las estimaciones han sido de hasta un 80% de reducción en general (**Dhiman et al**., 2011)**.** Dado que este sistema generalmente depende de volúmenes extremadamente bajos de reactivos, el mayor costo es la compra del espectrómetro de masas ya que los costos por determinación son de menos de 1 USD por muestra. Todos los tubos de recogida de muestras e insumos asociados en el proceso rondan los 5 USD por muestra a diferencia de los costos de determinaciones moleculares que rondan los 15 USD por muestra.

Cuando se iniciaron los ensayos en abril del año 2020, aún no había antecedentes sobre este enfoque y los datos proteómicos se compartieron en repositorios de acceso abierto, poniéndolos a disposición de otros grupos de investigadores de todo el mundo (**https://zenodo.org/**). Además, los resultados preliminares de las investigaciones se

publicaron rápidamente en mayo de 2020, en el sitio de distribución en línea gratuito de preimpresiones biorXiv, siendo el primer artículo que mostró el potencial de la EM MALDI-TOF combinada con herramientas de inteligencia artificial para el diagnóstico de COVID-19. El artículo final publicado en octubre de 2020 en una revista revisada por pares (**Rocca et al**., 2020), es un artículo que sigue siendo citado por investigadores que han decidido explorar este trabajo original e innovador.

Este trabajo constituye las bases y alienta a los investigadores a explorar el potencial de la EM para evaluar la viabilidad de esta tecnología, ampliamente disponible en los laboratorios de microbiología clínica de nuestro país y de todo el mundo, como una herramienta rápida y económica de diagnóstico de enfermedades a partir de muestras clínicas **(Kallow et al.,** 2010**; He et al.,** 2010**; Cherkaoui et al.,** 2010).

10. CONCLUSIONES

- Se logró la generación de espectros de masas característicos, directamente a partir de hisopados nasofaríngeos utilizando un Espectrómetro de Masas del tipo MALDI-TOF en el rango de lectura de la rutina de un laboratorio clínico convencional, lo que lo convierte en una metodología accesible y más fácilmente transferible a los laboratorios de menor complejidad que cuenten con la metodología de EM.
- Los mejores valores de desempeño se alcanzaron al procesar las muestras en las mismas condiciones iniciales que las empleadas para generar los modelos de aprendizaje automático. Es por ese motivo que únicamente formaron parte del desarrollo y del análisis de los resultados de esta tesis, los hisopados nasofaríngeos que habían sido previamente conservados en ultra-frío.
- La Base de Datos *in house* presentó bajos valores de desempeño (S=62,5%, E=26,5%, VPP=28,6%, VPN=60%) durante su evaluación arrojando gran cantidad de No Identificaciones, por lo que su aplicación fue discontinuada.
- Se detectaron seis picos potenciales biomarcadores de negatividad, reproducibles en los dos softwares utilizados para el análisis, con alta sensibilidad y especificad de diagnóstico.
- No fue posible asignarles identidad a los péptidos hallados debido a la conocida limitación en la resolución del equipo. De todos modos, este tipo de caracterizaciones específicas no están validadas hoy en día mediante técnicas de EM.
- Se logró la detección de patrones diferenciables entre perfiles proteicos de pacientes sanos y enfermos de COVID-19 mediante técnicas de aprendizaje automático.
- En base a la capacidad de distinción evidenciada, se crearon modelos predictivos que funcionaron como clasificadores rápidos de nuevas muestras, lo cual implicaría un ahorro de tiempo y recursos de poder ser aplicado en el laboratorio convencional en el futuro cercano.
- Se observó que no basta el hallazgo de picos biomarcadores, sino que se requiere del análisis acoplado al entrenamiento automatizado para poder alcanzar una identificación más confiable.
- No es posible establecer una correlación directa entre los distintos valores de *CT*, la gravedad de la enfermedad y los resultados obtenidos por *Machine Learning*.
- Para mejorar el desempeño del desarrollo propuesto y extender su aplicación al diagnóstico de rutina, es necesario disponer de un N mucho mayor de espectros de muestras.

11. PERSPECTIVAS

La rápida publicación en línea de los resultados preliminares permitió la interacción con diferentes grupos proteómicos en todo el mundo. El artículo final publicado en octubre del año 2020 en el Journal of Virological methods, sigue siendo citado por investigadores que han decidido explorar este trabajo original e innovador. Algunas evidencias de la relevancia de este trabajo radican en la gran cantidad de contactos de todo el mundo que se han recibido, no sólo para consultar acerca de cómo replicar los ensayos, sino también proponiendo grupos de trabajo colaborativos a futuro (**Tran et al**., 2021; **Oishee et al**., 2021; **Chivte et al**., 2021, entre otros).

Toda la investigación, además se propuso en el marco de la Red Nacional que reúne a más de 30 laboratorios clínicos de proveedores de salud públicos y privados que han implementado MALDI-TOF en su diagnóstico de rutina. Esta perspectiva, animó a los participantes de la Red Nacional a iniciar varios proyectos de investigación para abordar este enfoque innovador aplicado al análisis de peptidomas de muestra y su asociación con diferentes enfermedades; por lo que se espera que, en un futuro próximo, MALDI–TOF no sea sólo una herramienta analítica en la investigación médica y clínica, sino que además se pueda utilizar como una herramienta de diagnóstico de rutina.

Entre las nuevas aplicaciones, se comenzó con el procesamiento de muestras para búsqueda de *Legionella* spp en el Laboratorio Bacteriología Especial-INEI ANLIS Malbrán, adquiriendo espectros para la búsqueda de biomarcadores distintivos de especie, serogrupo y capacidad toxigénica en base a lo reportado por **Kyritsi et al.** (2020) y para el análisis de la distribución de los perfiles obtenidos con el fin de crear modelos de clasificación rápida incluso a partir de muestras de orina directamente, modo similar a lo realizado en este trabajo de tesis. El objetivo es, una vez alcanzado un enorme conjunto de datos de entrenamiento, poder aplicar estos algoritmos como técnica de tamizaje de rutina para el diagnóstico de Legionellosis a partir de muestras clínicas y ambientales ahorrando tiempo y dinero sobre las técnicas de referencia actuales que incluyen desde la recolección y concentración de muestras de agua, la preparación de fórmulas específicas para medios de cultivo hasta la identificación definitiva (**Blanco et al**., 2021).

Por otra parte, se ha comenzado un trabajo colaborativo con el Laboratorio de Hantavirus-ANLIS Malbrán, adquiriendo en forma manual, los espectros de masas en sueros previamente caracterizados por métodos de referencia, para la identificación temprana de casos de síndrome pulmonar por hantavirus (SPH) mediante EM.

En Argentina, la enfermedad presenta una letalidad del 35%, no hay vacunas aprobadas ni terapias específicas disponibles y la baja prevalencia de la patología sumado a la ausencia de kits comerciales, retrasan notablemente el diagnóstico. Por ello, el objetivo es identificar un perfil exclusivo de expresión proteica en muestras de pacientes con SPH y evaluar la presencia de biomarcadores específicos, mediante la tecnología MALDI-TOF.

Este procedimiento podría realizarse con rapidez y precisión en los diferentes laboratorios del país, minimizando tiempos para la toma de decisiones epidemiológicas y clínicas, que redundarían en un mejor pronóstico del caso de SPH y seguimiento de sus contactos estrechos (resultados preliminares, **Torelli et al**, 2021).

Los resultados de las nuevas investigaciones serán transferidos al resto de los usuarios de la plataforma a lo largo y ancho del país, muchos de ellos ubicados en zonas endémicas y donde el diagnóstico se vuelve dificultoso.

12. BIBLIOGRAFIA

Comentado [i1]: Agregar las citas en el texo

- Antezack, A., Chaudet, H., Tissot-Dupont, H., Brouqui, P., & Monnet-Corti, V. (2020). Rapid diagnosis of periodontitis, a feasibility study using MALDI-TOF mass spectrometry. PloS one, 15(3), e0230334. https://doi.org/10.1371/journal.pone.0230334

- Bezstarosti, K., Lamers, M., Haagmans, M. L. & Demmers, J. A. (2020). Targeted Proteomics for the Detection of SARS-CoV-2 Proteins. BioRxiv. https://doi:10.1101/2020.04.23.057810.

- Blanco, S., Sanz, C., Gutiérrez, M. P., Simarro, M., López, I., Escribano, I., Eiros, J. M., Zarzosa, P., Orduña, A., López, J. C., & March, G. A. (2021). A new MALDI-TOF approach for the quick sequence type identification of Legionella pneumophila. *Journal of microbiological methods*, *188*, 106292. https://doi.org/10.1016/j.mimet.2021.106292.

- Brown, J. R., Bharucha, T., & Breuer, J. (2018). Encephalitis diagnosis using metagenomics: application of next generation sequencing for undiagnosed cases. The Journal of infection, 76(3), 225–240. https://doi.org/10.1016/j.jinf.2017.12.014

- Buckley, M. (2018). Paleoproteomics: An Introduction to the Analysis of Ancient Proteins by Soft Ionisation Mass Spectrometry. Paleogenomics, 31–52. Population Genomics. Springer, Cham. https://doi.org/10.1007/13836_2018_50.

- Camoez, M., Sierra, J. M., Dominguez, M. A., Ferrer-Navarro, M., Vila, J., & Roca, I. (2016). Automated categorization of methicillin-resistant Staphylococcus aureus clinical isolates into different clonal complexes by MALDI-TOF mass spectrometry. Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases, 22(2), 161.e1–161.e7. https://doi.org/10.1016/j.cmi.2015.10.009

- Cardozo, K. H. M., Lebkuchen, A., Okai, G. C., Schuch, R. A., Viana, L. G., Olive, A. N., Lazari, C. d. S., Fraga, A. M., Granato, C. & Carvalho, V. M. (2020). Fast and low-cost

detection of SARS-CoV-2 peptides by tandem mass spectrometry in clinical samples. Research Square. https://doi:10.21203/rs.3.rs-28883/v1.

- Chan, J.F., Yip, C.C., To, K.K., Tang, T.H., Wong, S.C., Leung, K.H., Fung, A.Y., Ng, A.C., Zou, Z., Tsoi, H.W., Choi, G.K., Tam, A.R., Cheng, V.C., Chan, K.H., Tsang, O.T. & Yuen K.Y. (2020). Improved molecular diagnosis of COVID-19 by the novel, highly sensitive and specific COVID-19-RdRp/hel real-time reverse Transcription-PCR assay validated in vitro and with clinical specimens. Journal of Clinical Microbiology, 58 (5), e00310–00320. https://doi.org/10.1128/JCM.00310-20.

- Chang, C. H., Lin, C. H. & Lane, H. Y. (2021). Machine Learning y nuevos biomarcadores para el diagnóstico de la enfermedad de Alzheimer. *Revista internacional de ciencias moleculares, 22*(5), 2761. https://doi.org/10.3390/ijms22052761

- Cherkaoui, A., Hibbs, J., Emonet, S., Tangomo, M., Girard, M., Francois, P., & Schrenzel, J. (2010). Comparison of two matrix-assisted laser desorption ionization-time of flight mass spectrometry methods with conventional phenotypic identification for routine identification of bacteria to the species level. Journal of clinical microbiology, 48(4), 1169–1175. https://doi.org/10.1128/JCM.01881-09

- Chin, A., Chu, J., Perera, M., Hui, K., Yen, H. L., Chan, M., Peiris, M., & Poon, L. (2020). Stability of SARS-CoV-2 in different environmental conditions. The Lancet. Microbe, 1(1), e10. https://doi.org/10.1016/S2666-5247(20)30003-3

- Chivte, P., LaCasse, Z., Seethi, V., Bharti, P., Bland, J., Kadkol, S. S., & Gaillard, E. R. (2021). MALDI-ToF protein profiling as a potential rapid diagnostic platform for COVID-19. Journal of mass spectrometry and advances in the clinical lab, 21, 31–41. https://doi.org/10.1016/j.jmsacl.2021.09.001

- Cios K.J., Swiniarski R.W., Pedrycz W. & Kurgan L.A. (2007). Unsupervised Learning: Association Rules. In: Data Mining. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-36795-8_10.

- Cipolla, L., Rocca, M.F., Armitano, R., Martinez, C., Almuzara, M., Faccone, D., Vay, C. & Prieto, M. (2018). Desarrollo y evaluación de una base de datos in house para la identificacion rapida de Burkholderia contaminans por EM MALDI-TOF. Revista

Argentina de Microbiología, 51 (3), 255-258. https://doi.org/10.1016/j.ram.2018.09.001.

- Claydon M.A., Davey S.N., Edwards-Jones V. & Gordon D.B. (1996). The rapid identification of intact microorganisms using mass spectrometry. Nature Biotechnology, 14(11):1584-6. https://doi:10.1038/nbt1196-1584

- ClinPro Tools User Manual Version 3.0. (2011). Bruker Daltonik GmbH, Bremen.

- Corman, V. M., Landt, O., Kaiser, M., Molenkamp, R., Meijer, A., Chu, D. K., Bleicker, T., Brünink, S., Schneider, J., Schmidt, M. L., Mulders, D. G., Haagmans, B. L., van der Veer, B., van den Brink, S., Wijsman, L., Goderski, G., Romette, J. L., Ellis, J., Zambon, M., Peiris, M. & Drosten, C. (2020). Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Euro surveillance: bulletin Europeen sur les maladies transmissibles=European comunicable disease bulletin, 25(3),2000045. https://doi.org/10.2807/1560-7917.ES.2020.25.3.2000045

- Cortez, M. F., Gomez, C., Ortiz, A., Pelossso, R., Pili, A., Rivero, J., Roldán, A. Sol., Sandoval, F. & San Juan, M. ESI-MALDI-TOF. http://ufq.unq.edu.ar/Docencia-Virtual/BQblog/MALDI-ESI-TOF.pdf

- COVID-19 Map - Johns Hopkins Coronavirus Resource Center, 2020. COVID-19 Map - Johns Hopkins Coronavirus Resource Center (Accessed 5 April 2020). https://coronavirus.jhu.edu/map.html.

- Croxatto, A., Prod'hom, G. & Greub, G. (2012), Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology. FEMS Microbiology Reviews, 36: 380-407. https://doi.org/10.1111/j.1574-6976.2011.00298.x

- Dao, T. L., Hoang, V. T., Ly, T., Lagier, J. C., Baron, S. A., Raoult, D., Parola, P., Courjon, J., Marty, P., Chaudet, H., & Gautret, P. (2021). Sputum proteomic analysis for distinguishing between pulmonary tuberculosis and non-tuberculosis using matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS): preliminary results. Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases, 27(11), 1694.e1–1694.e6. https://doi.org/10.1016/j.cmi.2021.02.031

- De Bel, A., Wybo, I., Vandoorslaer, K., Rosseel, P., Lauwers, S., & Piérard, D. (2011). Acceptance criteria for identification results of Gram-negative rods by mass spectrometry. Journal of medical microbiology, 60(Pt 5), 684–686. https://doi.org/10.1099/jmm.0.023184-0

- De Bruyne, K., Slabbinck, B., Waegeman, W., Vauterin, P., De Baets, B., & Vandamme, P. (2011). Bacterial species identification from MALDI-TOF mass spectra through data analysis and machine learning. Systematic and applied microbiology, 34(1), 20–29. https://doi.org/10.1016/j.syapm.2010.11.003

- Degand N., Carbonnelle E., Dauphin B., Beretti J.L., Le Bourgeois M., Sermet-Gaudelus I., Segonds C., Berche P., Nassif X. & Ferroni A. (2008). Matrix-assisted laser desorption ionization-time of flight mass spectrometry for identification of nonfermenting gram-negative bacilli isolated from cystic fibrosis patients. Journal of Clinical Microbiology, 46(10):3361-7. https://doi:10.1128/JCM.00569-08

- Demirev P.A., Ho Y.P., Ryzhov V. & Fenselau C. (1999). Microorganism Identification by Mass Spectrometry and Protein Database Searches. Analitical Chemistry, 71, 2732–2738. https://doi.org/10.1021/ac990165u

- Dhiman N, Hall L, Wohlfiel SL, Buckwalter SP, Wengenack NL. (2011). Performance and cost analysis of matrix-assisted laser desorption ionization-time of flight mass spectrometry for routine identification of yeast. Journal of Clinical Microbiology, 49(4):1614-1616. doi:10.1128/JCM.02381-10).

- Espinal, P., Seifert, H., Dijkshoorn, L., Vila, J., & Roca, I. (2012). Rapid and accurate identification of genomic species from the Acinetobacter baumannii (Ab) group by MALDI-TOF MS. Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases, 18(11), 1097–1103. https://doi.org/10.1111/j.1469-0691.2011.03696.x

- Espinosa, R.F., Rumi, V., Marchisio, M., Cejas, D., Radice, M., Vay, C., Barrios, R., Gutkind, G. & Di Conza, J. (2018). Fast and easy detection of CMY-2 in Escherichia coli by direct MALDI-TOF mass spectrometry. Journal of Microbiological Methods, 148:22-28. https://doi:10.1016/j.mimet.2018.04.001.

- Fenselau, C., Demirev, P.A. (2001). Characterization of intact microorganisms by MALDI mass spectrometry. Mass Spectrometry Reviews, 20:157–171.

- Feucherolles, M., Cauchie, H. M., & Penny, C. (2019). MALDI-TOF Mass Spectrometry and Specific Biomarkers: Potential New Key for Swift Identification of Antimicrobial Resistance in Foodborne Pathogens. Microorganisms, 7(12), 593. https://doi.org/10.3390/microorganisms7120593

- Fitzpatrick, F., Doherty, A., & Lacey, G. (2020). Using Artificial Intelligence in Infection Prevention. Current treatment options in infectious diseases, 1–10. Advance online publication. https://doi.org/10.1007/s40506-020-00216-7

- flexAnalysis 3.4 User Manual Revision 1 (November 2011). Bruker Daltonik GmbH, Bremen.

- Gomila, R.M., Martorell, G., Fraile-Ribot, P., Doménech-Sánchez, A., Oliver, A., García-Gasalla, M. & Albertí, S. (2020). Rapid classification and prediction of COVID-19 severity by MALDI-TOF mass spectrometry analysis of serum peptidome. medRxiv preprint doi: https://doi.org/10.1101/2020.10.30.20223057.

- Gonzalez de Buitrago, J. M. & Ferreira L. (2006). Proteómica clínica. Sociedad Española de Bioquimica clínica y patología molecular. Monografía. ISBN: 978-84-89975-25-5.

- Gouveia, D., Miotello, G., Gallais, F., Gaillard, J. C., Debroas, S., Bellanger, L., Lavigne, J. P., Sotto, A., Grenga, L., Pible, O., & Armengaud, J. (2020). Proteotyping SARS-CoV-2 Virus from Nasopharyngeal Swabs: A Proof-of-Concept Focused on a 3 Min Mass Spectrometry Window. Journal of proteome research, 19(11), 4407–4416. https://doi.org/10.1021/acs.jproteome.0c00535

- Granville V. (2017). Difference between Machine Learning, Data Science, AI, Deep Learning, and Statistics. https://www.datasciencecentral.com/profiles/blogs/difference-between-machine-learning-data-science-ai-deep-learning

- Grenga, L., Armengaud, J. (2020). Proteomics in the COVID-19 Battlefield: First SemesterCheckUp. *Proteomics*, 21,2000198. https://doi.org/10.1002/pmic.202000198

- Grossegesse, M., Hartkopf, F., Nitsche, A., Schaade, L., Doellinger, J. & Muth T. (2020). Perspective on Proteomics for Virus Detection in Clinical Samples. Journal of

Proteome Research, 19 (11), 4380-4388. https://:doi10.1021/acs.jproteome.0c00674.

- Guo, Y. R., Cao, Q. D., Hong, Z. S., Tan, Y. Y., Chen, S. D., Jin, H. J., Tan, K. S., Wang, D. Y., & Yan, Y. (2020). The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak - an update on the status. Military Medical Research, 7(1), 11. https://doi.org/10.1186/s40779-020-00240-0

- He, Y., Li, H., Lu, X., Stratton, C. W., & Tang, Y. W. (2010). Mass spectrometry biotyper system identifies enteric bacterial pathogens directly from colonies grown on selective stool culture media. Journal of clinical microbiology, 48(11), 3888–3892. https://doi.org/10.1128/JCM.01290-10

- Ihling, C., Tänzler, D., Hagemann, S., Kehlen, A., Hüttelmaier, S., Arlt, C. & Sinz, A. (2020). Mass Spectrometric Identification of SARS-CoV-2 Proteins from Gargle Solution Samples of COVID-19 Patients. Journal of proteome research, 19(11), 4389–4392. https://doi.org/10.1021/acs.jproteome.0c00280

- Iles, R.K., Zmuidinaite R., Iles J., Carnell G., Sampson A. & Heeney J. (2020). Development of a Clinical MALDI-ToF Mass Spectrometry Assay for SARS-CoV-2: Rational Design and Multi-Disciplinary Team Work. Diagnostics 10,746. https://doi.org/10.3390/diagnostics10100746

- Kallow, W., Erhard, M., Shah, H.N., Raptakis, E. & Welker, M. (2010). MALDI-TOF MS for microbial identification: years of experimental development to an established protocol. In Mass Spectrometry for Microbial Proteomics. (eds H.N. Shah and S.E. Gharbia). https://doi.org/10.1002/9780470665497.ch12

- Khot, P.D. & Fisher, M.A. (2013). Novel approach for differentiating Shigella species and Escherichia coli by matrix-assisted laser desorption ionization-time of flight mass spectrometry. Journal of Clinical Microbiology, 51 (11), 3711–3716. https://doi.org/10.1128/JCM.01526-13.

- Kyritsi, M. A., Kristo, I., & Hadjichristodoulou, C. (2020). Serotyping and detection of pathogenecity loci of environmental isolates of Legionella pneumophila using MALDI-TOF MS. International journal of hygiene and environmental health, 224, 113441. https://doi.org/10.1016/j.ijheh.2019.113441.

- Ledesma, M., Todero, M.F., Maceira L., Prieto, M., Vay, C., Galas, M., López, B., Yokobori, N. & Rearte, B. (2020). Plasma mass spectrometry fingerprints induced by bacterial endotoxins in murine models as markers of pathophysiological evolution in sepsis. Preprint bioRxiv 2020.12.29.424724. https://doi.org/10.1101/2020.12.29.424724

- Li, F. (2016). Structure, function, and evolution of coronavirus spike proteins. Annual Review of Virology, 3(1), 237–261. https://doi.org/10.1146/annurev-virology-110615-042301.

- Li, Z., Yi, Y., Luo, X., Xiong, N., Liu, Y., Li, S., Sun, R., Wang, Y., Hu, B., Chen, W., Zhang, Y., Wang, J., Huang, B., Lin, Y., Yang, J., Cai, W., Wang, X., Cheng, J., Chen, Z., Sun, K. & Ye, F. (2020). Development and clinical application of a rapid IgM-IgG combined antibody test for SARS-CoV-2 infection diagnosis. Journal of medical virology, 92(9), 1518–1524. https://doi.org/10.1002/jmv.25727

- Lipi D., Vedang M. & Ashok K. V. (2020). Comprehensive Analysis of Low Molecular Weight Serum Proteome Enrichment for Mass Spectrometric Studies. ACS Omega, 5 (44), 28877-28888. https://:doi10.1021/acsomega.0c04568

- McDermott, A. (2020). inner Workings: Molecular biologists offer “wartime service” in the effort to test for COVID-19 Proceedings of the National Academy of Sciences, 117 (18): 9656-9659. DOI: 10.1073/pnas.2006240117

- Maldonado, N., Robledo C. & Robledo J. (2018). La espectrometría de masas MALDI-TOF en el laboratorio de microbiología clínica. Infectio, 22(1): 35-45.

- Manfredi, E., Rocca, M.F., Barrios, R., Miliwebsky, E., Deza, N., Carbonari, C., Baschkier, A. & Chinen, I. (2019). Análisis proteómico comparativo para la detección de Ecoli O157 H7 empleando espectrometría de masas MALDI TOF. XV Congreso argentino de microbiología, Buenos Aires, Argentina.

- Marvin, L. F., Roberts, M. A., & Fay, L. B. (2003). Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry in clinical chemistry. Clinica chimica acta; international journal of clinical chemistry, 337(1-2), 11–21. https://doi.org/10.1016/j.cccn.2003.08.008

- Mellmann, A., Cloud, J., Maier, T., Keckevoet, U., Ramminger, I., Iwen, P., Dunn, J., Hall, G., Wilson, D., Lasala, P., Kostrzewa, M., & Harmsen, D. (2008). Evaluation of matrix-assisted laser desorption ionization-time-of-flight mass spectrometry in comparison to 16S rRNA gene sequencing for species identification of nonfermenting bacteria. Journal of clinical microbiology, 46(6), 1946–1954. https://doi.org/10.1128/JCM.00157-08

- Mischak, H., Apweiler, R., Banks, R. E., Conaway, M., Coon, J., Dominiczak, A., Ehrich, J. H., Fliser, D., Girolami, M., Hermjakob, H., Hochstrasser, D., Jankowski, J., Julian, B. A., Kolch, W., Massy, Z. A., Neususs, C., Novak, J., Peter, K., Rossing, K., Schanstra, J. & Yamamoto, T. (2007). Clinical proteomics: A need to define the field and to begin to set adequate standards. Proteomics. Clinical applications, 1(2), 148–156. https://doi.org/10.1002/prca.200600771

- Nachtigall, F. M., Pereira, A., Trofymchuk, O. S., & Santos, L. S. (2020). Detection of SARS-CoV-2 in nasal swabs using MALDI-MS. Nature biotechnology, 38(10), 1168–1173. https://doi.org/10.1038/s41587-020-0644-7

- Nikolaev, E. N., Indeykina, M. I., Brzhozovskiy, A. G., Bugrova, A. E., Kononikhin, A. S., Starodubtseva, N. L., Petrotchenko, E. V., Kovalev, G. I., Borchers, C. H., & Sukhikh, G. T. (2020). Mass-Spectrometric Detection of SARS-CoV-2 Virus in Scrapings of the Epithelium of the Nasopharynx of Infected Patients via Nucleocapsid N Protein. Journal of proteome research, 19(11), 4393–4397. https://doi.org/10.1021/acs.jproteome.0c00412

- Oishee, M. J., Ali, T., Jahan, N., Khandker, S. S., Haq, M. A., Khondoker, M. U., Sil, B. K., Lugova, H., Krishnapillai, A., Abubakar, A. R., Kumar, S., Haque, M., Jamiruddin, M. R., & Adnan, N. (2021). COVID-19 Pandemic: Review of Contemporary and Forthcoming Detection Tools. Infection and drug resistance, 14, 1049–1082. https://doi.org/10.2147/IDR.S289629

- Orsburn, B. C., Jenkins, C., Miller, S. M., Neely, B. A. & Bumpus, N. (2020). In silico approach toward the identification of unique peptides from viral protein infection: Application to COVID-19. bioRxiv. https://doi10.1101/2020.03.08.980383.

- Prodan, A., Brand, H., Imangaliyev, S., Tsivtsivadze, E., van der Weijden, F., de Jong, A., Paauw, A., Crielaard, W., Keijser, B., & Veerman, E. (2016). A Study of the

Variation in the Salivary Peptide Profiles of Young Healthy Adults Acquired Using MALDI-TOF MS. PloS one, 11(6), e0156707. https://doi.org/10.1371/journal.pone.0156707.

- Pusch, W., Flocco, M. T., Leung, S. M., Thiele, H., & Kostrzewa, M. (2003). Mass spectrometry-based clinical proteomics. Pharmacogenomics, 4(4), 463–476. https://doi.org/10.1517/phgs.4.4.463.22753.

- Ressom H.W., Varghese S., Orvisky E., Drake S.K., Hortin G., Abdel-Hamid, M., Lofredo, C.A. & Goldman R. (2005). Analysis of MALDI-TOF Serum Profiles for Biomarker Selection and Sample Classification. Comprehensive Cancer Center, Georgetown University, Washington, DC Clinical Chemistry Service, Department of Laboratory Medicine, Viral Hepatitis Research Laboratory, NHTMRI, Cairo, Egypt. hwr@georgetown.edu.

- Rhoades, N.S., Pinski, A.N., Monsibais, A.N., Jankeel, A., Doratt, B.M., Cinco, I.R., Ibraim, I. & Messaoudi, I. (2021). Acute SARS-CoV-2 infection is associated with an increased abundance of bacterial pathogens, including Pseudomonas aeruginosa in the nose. Cell Reports, 36 (9); 109637. https://doi.org/10.1016/j.celrep.2021.109637.

- Rhoads D. (2020). Computer vision and artificial intelligence are emerging diagnostic tools for the clinical microbiologist. Journal of Clinical Microbiology, 58:e00511-20. https://doi.org/10.1128/JCM.00511-20.

- Righetti P. (2013). Proteome. Brenner's Encyclopedia of Genetics (Second Edition), 504-507. https://doi.org/10.1016/B978-0-12-374984-0.01230-4.

- Rocca, M.F., Barrios, R., Zintgraff, J. & Prieto, M. (2018). MALDI-TOF como herramienta para la subtipificación de Streptococcus pyogenes. VII Congreso de la sociedad argentina de bacteriología, micología y parasitología clínicas (SADEBAC), Buenos Aires, Argentina.

- Rocca M.F., Barrios R., Zintgraff J., Martínez C., Irazu L. & Vay C. (2019). Utility of platforms Viteks MS and Microflex LT for the identification of complex clinical isolates that require molecular methods for their taxonomic classification. PLoS ONE 14(7): e0218077.

- Rocca M.F., Zintgraff J.C., Dattero M.E., Silva Santos L., Ledesma M., Vay C., Prieto M., Benedetti E., Avaro M., Russo M., Nachtigall F. M. & Baumeister E. (2020). A combined approach of MALDI-TOF mass spectrometry and multivariate analysis as a potential tool for the detection of SARS-CoV-2 virus in nasopharyngeal swabs. Journal of Virological Methods, 286,113991. https://doi.org/10.1016/j.jviromet.2020.113991.

- Rocca MF, Almuzara M, Barberis C, Vay C, Viñes P & Prieto M. (2020). Presentación del sitio web de la Red Nacional de Identificación Microbiológica por Espectrometría de Masas. Manual para la interpretación de resultados de MALDI-TOF MS. Revista Argentina de Microbiología, 52 (1), 83-84. https://doi:10.1016/j.ram.2019.03.001

- Shah, S., Singhal, T., Davar, N., & Thakkar, P. (2021). No correlation between Ct values and severity of disease or mortality in patients with COVID 19 disease. Indian Journal of Medical Microbiology, 39(1), 116–117. https://doi.org/10.1016/j.ijmmb.2020.10.021

- Shalev-Shwartz S. & Shai Ben D. (2014). Understanding Machine Learning: From Theory to Algorithms. book published by Cambridge University Press. http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning.

- Schubert, O. T., Röst, H. L., Collins, B. C., Rosenberger, G., & Aebersold, R. (2017). Quantitative proteomics: challenges and opportunities in basic and applied research. Nature protocols, 12(7), 1289–1294. https://doi.org/10.1038/nprot.2017.040

- Seng, P., Drancourt, M., Gouriet, F., La Scola, B., Fournier, P. E., Rolain, J. M., & Raoult, D. (2009). Ongoing revolution in bacteriology: routine identification of bacteria by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America, 49(4), 543–551. https://doi.org/10.1086/600885

- Stephens, M.A. (1974). EDF statistics for goodness of fit and some comparisons. Journal of the American Statistical Association, 69 (347), 730-737.

- Suthee Mangmee, Onrapak Reamtong, Thareerat Kalambaheti, Sittiruk Roytrakul & Piengchan Sonthayanon. (2020). MALDI-TOF mass spectrometry typing for

predominant serovars of non-typhoidal Salmonella in a Thai broiler industry. Food Control, 113, 107188. https://doi.org/10.1016/j.foodcont.2020.107188.

- Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y., Yoshida, T. and Matsuo, T. (1988), Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. Rapid Commun. Mass Spectrom., 2: 151-153. https://doi.org/10.1002/rcm.1290020802

- Torelli,C.A., Rocca, M.F., Coelho R., Zintgraff J., Kehl S., Prieto M., Alonso, D.A., Periolo N., Lopez B., Martinez V., Bellomo C. (2021). Identificación temprana de casos de síndrome pulmonar por hantavirus mediante espectrometría de masas MALDI-TOF. Congreso Argentino de Virología. Categoría: E-Poster. ID:267.

- Tran, N. K., Howard, T., Walsh, R., Pepper, J., Loegering, J., Phinney, B., Salemi, M. R., & Rashidi, H. H. (2021). Novel application of automated machine learning with MALDI-TOF-MS for rapid high-throughput screening of COVID-19: a proof of concept. Scientific reports, 11(1), 8219. https://doi.org/10.1038/s41598-021-87463-w

- van der Heide V. (2020). SARS-CoV-2 cross-reactivity in healthy donors. Nature reviews. Immunology, 20(7), 408. https://doi.org/10.1038/s41577-020-0362-x

- Walls, A.C., Tortorici, M.A., Snijder, J., Xiong, X., Bosch, B-J., Rey, F.A and Veesler, D. (2017). Tectonic conformational changes of a coronavirus spike glycoprotein promote membrane fusion. Proceedings of the National Academy of Sciences (PNAS) 114 (42): 11157-11162. https://doi.org/10.1073/pnas.1708727114

- Wang, H. Y., Lien, F., Liu, T. P., Chen, C. H., Chen, C. J., & Lu, J. J. (2018). Application of a MALDI-TOF analysis platform (ClinPro Tools) for rapid and preliminary report of MRSA sequence types in Taiwan. PeerJ, 6, e5784. https://doi.org/10.7717/peerj.5784

- Wayne, PA. 2017. Methods for the Identification of Cultured Microorganisms Using Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry. 1st ed. CLSI guideline M58. ISBN 1-56238-816-9.

- Westergren-Thorsson, G., Marko-Varga, G., Malmström, K. Larsen, J. (2006). PROTEOME. Encyclopedia of Respiratory Medicine, 527-532. https://doi.org/10.1016/B0-12-370879-6/00331-8.

- Wenzhong, L. & Hualan, L. (2020). COVID-19: attacks the 1-beta chain of hemoglobin and captures the porphyrin to inhibit human heme metabolism. ChemRxiv Preprint. https://doi.org/10.26434/chemrxiv.11938173.v8.

- Wilkins M.R., Sanchez J.C., Gooley A.A., Appel R.D., Humphery-Smith I., Hochstrasser D.F. & Williams K.L (1996). Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. Biotechnology and Genetic Engineering Reviews, 13:1, 19-50, https://doi:10.1080/02648725.1996.10647923

- World Health Organization, 2020. Rational Use of Personal Protective Equipment (PPE) for Coronavirus Disease (COVID-19): Interim Guidance, 19 March 2020. World Health Organization (Accessed 5 April 2020). https://apps.who.int/iris/handle/ 10665/331498.

- Yang, X., Yu, Y., Xu, J., Shu, H., Xia, J., Liu, H., Wu, Y., Zhang, L., Yu, Z., Fang, M., Yu, T.,Wang, Y., Pan, S., Zou, X., Yuan, S. & Shang, Y. (2020). Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a singlecentered, retrospective, observational study. Lancet Respiratory Medicine. Preprint. https://www.thelancet.com/lancet/article/S2213-2600(20)30079-5.

- Yi-Tzu Cho, Hung Su, Wen-Jeng Wu, Deng-Chyang Wu, Ming-Feng Houjj, Chao-Hung Kuo & Jentaie Shiea. (2015). Biomarker Characterization by MALDI–TOF/MS. Advances in Clinical Chemistry. http://dx.doi.org/10.1016/bs.acc.2015.01.001

- Zautner, A. E., Lugert, R., Masanta, W. O., Weig, M., Groß, U., & Bader, O. (2016). Subtyping of Campylobacter jejuni ssp. doylei Isolates Using Mass Spectrometry-based PhyloProteomics (MSPP). Journal of visualized experiments : JoVE, (116), 54165. https://doi.org/10.3791/54165

- Zhang, H., Cao, J., Li, L., Liu, Y., Zhao, H., Li, N., Li, B., Zhang, A., Huang, H., Chen, S., Dong, M., Yu, L., Zhang, J., & Chen, L. (2015). Identification of urine protein

biomarkers with the potential for early detection of lung cancer. Scientific reports, 5, 11805. https://doi.org/10.1038/srep11805

- Zhang, C., Shi, L., & Wang, F. S. (2020). Liver injury in COVID-19: management and challenges. The lancet. Gastroenterology & hepatology, 5(5), 428–430. https://doi.org/10.1016/S2468-1253(20)30057-1

- Zhou, M., Zhang, X., & Qu, J. (2020). Coronavirus disease 2019 (COVID-19): a clinical update. *Frontiers of medicine*, *14*(2), 126–135. https://doi.org/10.1007/s11684-020-0767-8

13. GLOSARIO DE MACHINE LEARNING

ALGORITMO DE CLASIFICACION: un algoritmo es un conjunto secuencial, definido y finito de reglas para obtener un determinado resultado en la realización de una actividad. En nuestro caso, son los algoritmos utilizados en la generación de un modelo de clasificación; pueden ser algoritmo genético (GA), red neuronal supervisada (SNN), clasificador rápido (QC).

ANALISIS DE CORRELACION: es un enfoque estadístico que analiza la correlación estocástica entre variables al azar (por ejemplo: picos y sus áreas) en un set de muestras. Hay dos tipos de información con la que se puede trabajar: los datos relativos a una sola variable se denominan datos univariados, en cambio, cuando el análisis de correlación se trata de estudiar la relación entre dos variables al mismo tiempo, los datos se denominan bivariados.

Si hay algún tipo de correlación entre dos variables, ambas se alteran juntas durante un período de tiempo. La correlación encontrada puede ser positiva o negativa, dependiendo de los valores numéricos medidos.

En los métodos estadísticos, el coeficiente de correlación "r" mide la fuerza, dirección y extensión de la relación entre dos variables, donde el valor de "r" siempre oscilará entre +1 y -1.

Dentro de los métodos de correlación de coeficientes más utilizados, se encuentra el método del diagrama de dispersión, que es un enfoque utilizado para encontrar la correlación entre dos variables, donde la relación se presenta en forma de diagrama para comprender cuán estrechamente se relacionan entre sí.

El diagrama tiene dos variables a lo largo de sus ejes `x' e `y', de las cuales una es independiente y la otra es la variable dependiente.

En los softwares empleados en EM, se aplica generalmente el coeficiente de correlación de Karl Pearson, que se define como un número entre -1 y 1. Un valor de 0 indica que no hay asociación entre las dos variables. Un valor mayor que 0 indica una asociación positiva. Es decir, a medida que aumenta el valor de una variable, también lo hace el valor de la otra. Un valor menor que 0 indica una asociación negativa; es decir, a medida que aumenta el valor de una variable, el valor de la otra disminuye.

La fórmula del coeficiente de correlación de Pearson es la siguiente:

$$r_{xy} = \frac{\sum z_x z_y}{N}$$

Donde:

"x" es igual a la variable número uno, "y" pertenece a la variable número dos, "zx" es la desviación estándar de la variable uno, "zy" es la desviación estándar de la variable dos y "N" es el número de datos.

Para mayor detalle, se recomienda entrar en el sitio https://www.questionpro.com/blog/es/coeficiente-de-correlacion-de-pearson/

ANDERSON DARLING TEST (AD): evalúa si la información de la muestra sigue una distribución específica o no. Si AD > 0.05, la población está distribuida normalmente y se pasa a t-test o anova test que deben ser menores a 0.05 para que se trate de un pico diferencial significativo.

ANÁLISIS DE COMPONENTES PRINCIPALES (PCA): Es una técnica matemática usada para graficar la varianza en un set de datos, independizándose de la dimensionalidad pero reteniendo toda la información. Esto se logra reemplazando grupos de variables por una variable nueva o componente principal; generalmente unos pocos PCA (1,2,3) contienen toda la varianza. En programas de Bruker, se grafica como scores donde cada punto es un espectro, o como *plots*, donde cada punto es un pico. El análisis de componentes principales, corresponde a las técnicas de agrupamiento no supervisado.

Los datos obtenidos por MALDI-TOF MS podrían orientar la predicción de la identidad de aislados desconocidos. Sin embargo, en el PCA, las cepas se diferencian en función de la presencia / ausencia de uno o más picos discriminantes sin presentar ninguna relación jerárquica entre ellos.

ARBOLES DE DECISIÓN (por su sigla en inglés, DT): son árboles de clasificación y regresión que crean una serie de reglas basadas en variables de entrada continuas y / o categóricas

para predecir un resultado. Son generalmente fáciles de interpretar, aunque pueden ser inestables incluso ante pequeños cambios de datos. Lo más relevante de estos algoritmos, es que son propensos al sobreajuste.

BIOMARCADOR DIAGNOSTICO: característica que se mide y evalúa objetivamente como indicador de procesos biológicos normales, patológicos o respuestas a una intervención terapéutica (National Institute of Health NIH, 2001). Su valor fundamental es su capacidad de diferenciar dos o más estados biológicos; cuando su concentración supera el umbral de detección, se considera predictor de enfermedad. Un biomarcador ideal debe ser fiable, es decir comportarse siempre de la misma forma y ser fácilmente medible. En su aplicación clínica no se pretende la detección de un único biomarcador, sino de un perfil característico de una determinada condición.

BOSQUES ALEATORIOS (*random forest*): es un tipo de aprendizaje conjunto, donde se combinan varios algoritmos de *machine learning* para crear un algoritmo más grande y de mejor rendimiento. Selecciona aleatoriamente 'k' puntos de datos del conjunto de entrenamiento, construye un árbol de decisión asociado con estos k puntos. Luego, se elige el número de árboles 'n' que se quieren construir y repetir. Para un nuevo punto de datos, se toman las predicciones de cada uno de los árboles de decisión 'n' y se asigna a la categoría de voto mayoritario.

CAPACIDAD DE RECONOCIMIENTO: describe el desempeño de un clasificador; en otras palabras, que tan bueno es un modelo para clasificar los datos que forman parte del mismo. Si la CR es demasiado baja (<80-90%), esto significa que el modelo no fue capaz de aprender a partir de las características de las muestras y esto ocurre cuando no es posible encontrar una relación entre los datos de esa marcación. Por otro lado, si la CR es alta, esto sólo indica que el modelo ha aprendido en base al conjunto de datos de entrenamiento, pero esto no quiere decir que la clasificación de nuevas muestras vaya a ser exitosa.

CLASE: grupo de espectros originados de muestras, como por ejemplo de un estadío de enfermedad, agrupados en base a sus características similares.

CLASIFICADORES BAYESIANOS (en inglés, Naïve Bayes): algoritmo que calcula la probabilidad asociada con cada clase en un conjunto de co-variables. El clasificador luego selecciona la clase con la mayor probabilidad como la "correcta". Su simplicidad contribuye a la popularidad de estos algoritmos; funciona relativamente bien en presencia de ruido, datos faltantes y características irrelevantes; requiere estimar menos parámetros y, por lo tanto, un menor conjunto de datos de entrenamiento, que otros algoritmos más complejos. La limitación más importante es que esta independencia a menudo varía en el mundo real por lo que los resultados pueden verse afectados y además la clase más probable puede pesar mucho en el funcionamiento del modelo, sesgando la predicción.

CONGLOMERADOS JERARQUICOS (*hierarchical clustering*): es un término genérico para una amplia variedad de procedimientos con un objetivo común: la formación de clases de sujetos o de variables similares pudiendo elegir entre una gran variedad de métodos de aglomeración y medidas de distancia. Por ejemplo, esta métrica puede ser la euclídea (distancias en un plano), aunque existen otras opciones más o menos robustas. El objetivo es identificar grupos de manera que la variabilidad intra-clase sea inferior a la variabilidad entre clases. Es una técnica aglomerativa, partiendo de muestras individuales que se van uniendo en conglomerados de acuerdo a su similitud (matriz de distancias entre los elementos de la muestra).

Entre las técnicas de *clustering* más utilizadas, se encuentra la llamada **k-medias o k-means** debido a que es rápida y sencilla. Se trata de un agrupamiento no jerárquico que requiere de cuatro etapas:

-seleccionar k puntos como centros

-calcular las distancias de cada elemento a los centros y asignar cada elemento al grupo cuyo centro se encuentre más cercano

-definir un criterio de optimidad y comprobar si reasignando alguno de los elementos mejora el criterio

-cuando no es posible mejorar el óptimo, termina el proceso

CURVAS ROC: del inglés, *Receiver Operating Characteristic*. Es un gráfico donde se evalúa la cualidad de discriminación de un pico; que da idea de la sensibilidad y especificidad de un test. Estas curvas solo pueden ser generadas en modelos de dos clases porque se debe decidir qué grupo es el verdadero. Los valores del área bajo la curva (AUC) de ese pico

variarán entre 0 y 1, cuanto más cercano a 1, mejor funciona el modelo y ese valor de discriminación para el pico. En epidemiología, suele recomendarse un valor de AUC > 0,8 en la curva ROC para que un pico pueda definirse como diferencial o probable biomarcador.

DENDROGRAMAS: es un tipo de representación gráfica o diagrama de datos en forma de árbol (gr. δένδρον déndron 'árbol') que organiza los datos en subcategorías que se van dividiendo en otras hasta llegar al nivel de detalle deseado (asemejándose a las ramas de un árbol que se van dividiendo sucesivamente). Este tipo de representación permite apreciar el agrupamiento jerárquico de las muestras, aunque no las relaciones de similitud o cercanía entre categorías. También se hace referencia al dendrograma como la ilustración de las agrupaciones derivadas de la aplicación de un algoritmo de agrupamiento jerárquico.

EMPAQUETAMIENTO (en inglés, *boostrap*): el algoritmo subyacente se ajusta a cada copia de los datos de entrenamiento originales y luego crea una predicción final basada en las salidas de los modelos resultantes, es decir que la predicción final para un resultado cuantitativo se obtiene promediando todas las predicciones individuales.

ESPECIFICIDAD: es la probabilidad de que un sujeto sano tenga un resultado negativo en la prueba. La especificidad es el porcentaje de verdaderos negativos o la probabilidad de que la prueba sea negativa si la enfermedad no está presente. Los falsos positivos son sujetos sanos diagnosticados como enfermos.

$$\text{Especificidad} = \frac{VN}{VN + FP}$$

Donde VN, serían los verdaderos negativos; y FP, los falsos positivos.

En diagnóstico clínico, cuando el valor de especificidad de una prueba supera el 80%, se considera buena.

K- VECINO MÁS CERCANO (por su sigla en inglés, *k-NN*): uno de los algoritmos de aprendizaje no supervisado más simple; divide las observaciones en un número preestablecido de grupos distintos (k), de modo que la variación dentro del grupo sea lo más pequeña posible. Es simple, fácil de interpretar y computacionalmente eficiente. La clase asignada se determina como la clase mayoritaria de los k puntos de datos de

entrenamiento más cercanos. La medida de similitud comúnmente utilizada es la distancia euclidiana. Sin embargo, una limitación importante es que el número de clases debe especificarse previamente; podrán ser 1, 3, 5 o 7, de acuerdo al número de muestras. Además, una ligera diferencia en k puede producir resultados muy diferentes por lo que suele ser conveniente ensayar un mismo modelo aplicando varios k.

MÁQUINA DE VECTORES DE SOPORTE (por su sigla en inglés, *SVM*): algoritmo del aprendizaje supervisado que encuentra el mejor margen de separación hiperplano entre las clases en una representación dimensional. SVM solo se usa como un algoritmo de selección de picos o características y la clasificación se realiza utilizando un algoritmo k-NN basado en esos picos seleccionados. Las SVM se utilizan tradicionalmente para clasificación binaria (modelos de dos clases) y generalmente demuestran un bajo error de clasificación.

MODELO DE CLASIFICACION: contiene las características de la preparación y de los clasificadores seleccionados, se guarda en formato XML para poder ser descargado y clasificar espectros nuevos desconocidos.

OVERFITTING: también conocido como sobreajuste; significa que un modelo clasifica mucho mejor las clases del modelo que el test set; en general es un indicador de que algunos parámetros durante la creación del modelo fueron fuertemente adaptados.

PRECISIÓN: se refiere a la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud. Cuanto menor es la dispersión mayor es la precisión. Se representa por la proporción entre el número de predicciones correctas (tanto positivas como negativas) y el total de predicciones. En forma práctica es el porcentaje de casos positivos detectados.

Se calcula como: VP/(VP+FP)

p-VALUE: es la probabilidad de que un efecto simplemente cambie; provee una medida de la fuerza de una asociación; entonces si un p-value es menor a 0.05 esos picos no están fuertemente asociados y serían útiles para la separación.

El p-value se calcula mediante test estadísticos diversos:

t-test y wilcoxon test (modelos de 2 clases).

Anova y kruskal wallis (modelos de más de dos clases).

Una prueba de hipótesis que se utiliza para comparar las medias de dos poblaciones se llama prueba t. Una técnica estadística que se utiliza para comparar las medias de más de dos poblaciones se conoce como Análisis de Varianza o ANOVA.

REGRESIÓN LOGÍSTICA: es un modelo de clasificación lineal que se utiliza para modelar variables dependientes binarias. Se utiliza para predecir la probabilidad (p) de que ocurra un evento. Si p> = 0.5, la salida es 1 si no 0.

SENSIBILIDAD: es la probabilidad de clasificar correctamente a un individuo enfermo, es decir, la probabilidad de que para un sujeto enfermo se obtenga en una prueba diagnóstica un resultado positivo. La sensibilidad es, por lo tanto, la capacidad de la prueba complementaria para detectar la enfermedad. La sensibilidad es el porcentaje de verdaderos positivos o la probabilidad de que la prueba sea positiva si la enfermedad está presente; los falsos negativos son sujetos enfermos diagnosticados como sanos.

$$Sensibilidad = \frac{VP}{VP + FN}$$

Donde VP es verdaderos positivos y FN falsos negativos.

En muchas de las técnicas realizadas, cuando este valor supera el 80%, se toma como bueno.

VALIDACION CRUZADA: en modelos de *ML*, es la validación automática que ocurre durante la creación de un modelo, donde una pequeña porción de todos los espectros es separada y queda afuera de la generación del modelo y del análisis de clases. Luego estos espectros son clasificados y se determina el número de predicciones correctas e incorrectas. Esto se repite muchas veces y el número de predicciones se acumula para cada clase.

VALIDACION EXTERNA: se utiliza un grupo separado de muestras de resultado conocido, es decir se conoce previamente a que clase pertenecen y que no ha sido utilizado en la generación del modelo. A partir de la comparación de los resultados predichos con los verdaderos positivos y negativos es que se calculan la sensibilidad (VP) y especificidad (VN).

MATERIAL SUPLEMENTARIO

14. MATERIAL SUPLEMENTARIO

CRONOGRAMA DE TRABAJO INICIAL PARA LA OBTENCIÓN DE LA INFORMACIÓN PEPTÍDICA EN EL INSTITUTO MALBRÁN, AÑO 2020.

21 Abril	Adquisición de espectros a partir muestras seleccionadas por el servicio de virosis respiratorias, para utilizar como *training set o set de entrenamiento.* Creación de perfiles proteicos o MSPs en base a los mejores espectros para el desarrollo de una Base de Datos de perfiles proteicos COVID-19 positivos y COVID-19 negativos. Análisis, preprocesamiento y búsqueda de potenciales Biomarcadores. *N=25 (SARS-COV 2, n=13; ND, n=4; FLU, n=4; RSV, n=1; Sar, n=1; CovH, n=2)*
28 Abril	Adquisición de espectros a partir muestras seleccionadas por el servicio de virosis respiratorias para utilizar como *training set.* Desafío de la base de datos de espectros proteicos desarrollada. Análisis, preprocesamiento y búsqueda de potenciales Biomarcadores. *N=32 (SARS-COV 2, n=18; ND, n=6; FLU, n=4; RSV, n=1; Sar, n=1; CovH, n=2)*
7 Mayo	Preparación de una publicación con resultados preliminares: "A Combined approach of MALDI-TOF Mass Spectrometry and multivariate analysis as a potential tool for the detection of SARS-CoV-2 virus in nasopharyngeal swabs." enviada preprint a **bioRxiv.**
8 Mayo	Publicación online del preprint en bioRxiv. https://www.biorxiv.org/content/10.1101/2020.05.07.082925v1 Enviada para publicación al **Journal of Virological Methods.**
2 Junio	Adquisición de espectros a partir muestras seleccionadas por el servicio de virosis respiratorias para utilizar como *training set.* *N=36 (SARS-COV 2, n=12; ND, n=12; FLU, n=12)*
10 junio	Revisión de publicación, reenvío de manuscrito modificado 21 de julio. A la espera de aceptación.
18 Junio	Ensayo Desafío 1 *N=30 (SARS-COV 2, n=10; ND, n=10; FLU, n=10)*
7 Julio	Ensayo Desafío 2 *N=32 (SARS-COV 2, n=14; ND, n=18)*
9 Julio	Ensayo Desafío 3 *N=32 (SARS-COV 2, n=20; ND, n=12)*
20 julio- 3 agosto	Análisis de resultados y elaboración de informes para las autoridades y LNR.

ND: no detectable; FLU: virus influenza; Sar: virus sarampión; CoV-H: coronavirus humano endémico.

Tabla S1. Total de muestras remitidas por el servicio de Virosis Respiratorias- ANLIS Malbrán y utilizadas para adquirir la información proteica (N=123).

1 LARGADA MALDI-TOF 21-04-2020 MUESTRAS CONGELADAS		
IDENTIFICACION NEUROVIROSIS	**N° de muestra**	**CT**
SARS-CoV-2	**C 8105**	**33**
	C 8082	**26**
	C 8083	**32**
	C 8097	**37**
	C 8103	**24**
	C 8117	**20**
	C 8188	**18**
	C 8163	**26**
	C 8128	**24**
	C 8139	**35**
	C 8057	**34**
	C 8157	**29**
	C 8172	**17**
SARS-CoV-2	**C 8035**	**ND**
	C 8037	**ND**
	C 8040	**ND**
	C 6961	**ND**
FLU A H1	**R 2742**	**28**

FLU A H3	**R 2970**	**30**
FLU B YAM	**R 2963**	**26**
FLU B VIC	**R 2974**	**27**
RSV	**RSV 1986**	**27**
Sar	**217/EFE/HNF/2019**	**28**
CoVH 229E	**VH 1148**	**25**
CoVH HKU1	**VH 1153**	**28**
2 LARGADA MALDI TOF 28-04-2020 **MUESTRAS CONGELADAS**		
IDENTIFICACION NEUROVIROSIS	**N° de muestra**	**CT**
SARS-CoV-2	**C 7693**	**24**
	C 9347	**25**
	C 7669	**26**
	C 9350	**36**
	C 7760	**25**
	C 9208	**25**
	C 8085	**21**
	C 7834	**16**
	C 7913	**29**
	C 7872	**18**
	C 9214	**29**
	C 7523	**22**
	C 9334	**34**
	C 7562	**27**

	C 7828	26
	C 9597	23
	C 7476	20
	C 8187	18
RINOVIRUS HUMANO	VH 19	NEG p/ VR*
	VH 27	NEG p/ VR*
	VH 59	NEG p/ VR*
	VH 131	NEG p/ VR*
	VH 116	NEG p/ VR*
	VH 143	NEG p/ VR*
INFLUENZA A H1	R 2658	30
INFLUENZAA H3	R 2872	27
INFLUENZA B YAMAGATA	R 2953	14
INFLUENZA B VICTORIA	R 2954	20
VIRUS SINCITIAL RESPIRATORIO	RSV 2003	30
SARAMPION	146/EFE/HNF/2020	25
CORONAVIRUS HUMANO OC43	VH 984	33
CORONAVIRUS HUMANO HKU1	VH 1010	29
3 LARGADA MALDITOF 02-06-2020 MUESTRAS CONGELADAS		
IDENTIFICACION NEUROVIROSIS	N° de muestra	CT
	C 13466	17
	C 13468	25
	C 13463	18

SARS-CoV-2	**C 13474**	**25**
	C 13476	**26**
	C 13477	**23**
	C 13478	**17**
	C 13488	**18**
	C 13496	**21**
	C 13530	**17**
	C 13548	**16**
	C 13551	**18**
FLU A 2019	**R 9**	**24**
	R 10	**22**
	R 11	**19**
	R 12	**18**
	R 13	**26**
	R 19	**32**
	R 20	**25**
	R 48	**26**
	R 60	**21**
	R 63	**29**
	R 75	**17**
	R 76	**25**
	C 15342	**ND**
	C 15343	**ND**
	C 15344	**ND**
	C 15345	**ND**

SARS-CoV-2	C 15348	ND
	C 15349	ND
	C 15351	ND
	C 15352	ND
	C 15353	ND
	C 15355	ND
	C 15356	ND
	C 15358	ND
4 LARGADA MALDITOF 18-06-2020 DESAFIO 1. MUESTRAS CONGELADAS		
IDENTIFICACION NEUROVIROSIS	NRO DE MUESTRA	CT
SARS COV-2	C21770	18
	C21782	17
	C21785	18
	C21792	19
	C21804	20
	C21815	17
	C21817	18
	C21843	19
	C21863	19
	C21908	19
	C21901	ND
	C21902	ND
	C21903	ND

SARS COV-2	**C21904**	**ND**
	C21905	**ND**
	C21914	**ND**
	C21915	**ND**
	C21916	**ND**
	C21919	**ND**
	C21920	**ND**
FLU A	**R81**	**23**
FLU A	**R82**	**26**
FLU A	**R83**	**30**
FLU A	**R84**	**20**
FLU A	**R85**	**23**
FLU A	**R86**	**23**
FLU A	**R87**	**23**
FLU A	**R88**	**26**
FLU A	**R89**	**26**
FLU A	**R90**	**23**
N MUESTRAS CONGELADAS	**123**	

QUINTA LARGADA MALDITOF **07-07-2020** **DESAFIO 2. MUESTRAS FRESCAS**		
IDENTIFICACION NEUROVIROSIS	**NRO DE MUESTRA**	**CT**

SARS-CoV-2	C24663	29
	C24664	25
	C24665	33
	C24654	36
	C24656	36
	C24657	40
	C24658	30
	C24659	36
	C24660	38
	C24670	15
	C24675	37
	C24676	17
	C24678	18
	C24681	27
SARS-CoV-2	C24679	ND
	C24680	ND
	C24661	ND
	C24662	ND
	C24666	ND
	C24667	ND
	C24668	ND
	C24669	ND
	C24671	ND
	C24672	ND
	C24673	ND

	C24674	ND
	C24677	ND
	C24650	ND
	C24651	ND
	C24652	ND
SEXTA LARGADA MALDI-TOF 09-07-2020 DESAFIO 3. MUESTRAS FRESCAS		
IDENTIFICACION NEUROVIROSIS	NRO DE MUESTRA	CT
SARS-CoV-2	C 25050	26
	C25051	37
	C25052	24
	C25053	36
	C25079	19
	C25056	37
	C25057	33
	C25076	25
	C25077	35
	C25061	28
	C25062	31
	C25063	15
	C25064	24
	C25066	21
	C25067	26

	C25068	**23**
	C25069	**22**
	C25070	**31**
	C25071	**19**
	C25073	**33**
SARS-CoV-2	**C25074**	**ND**
	C25075	**ND**
	C25072	**ND**
	C25058	**ND**
	C25065	**ND**
	C25078	**ND**
	C25080	**ND**
	C25081	**ND**
	C25054	**ND**
	C25055	**ND**
	C25059	**ND**
	C25060	**ND**
N MUESTRAS FRESCAS	**64**	
N TOTAL	**187 MUESTRAS**	

El resto de los espectros (N:188) utilizados en este trabajo de tesis fueron adquiridos en las mismas condiciones y remitidos por hospitales colaboradores de la red nacional de EM.

Tabla S2. Total de espectros remitidos por Laboratorios pertenencientes a la Red Nacional de Espectrometría de Masas-ReNaEM para completar la información proteica (N=188).

HOSPITAL NAVAL	
IDENTIFICACION DE REFERENCIA	**N° de muestra**
SARS-CoV-2 NO DETECTABLE	2149937
	2150023
	2150034
	2150046
	2150047
	2150049
	2150050
	2150063
	2150066
	2150067
	2150068
	2150069
	2150070
	2150071
	2150073
	2150075
	2150096
	2150097
	2150108
	2150114
	2150120
	2150122
	2150127
	2150128
	2150131
	2149937
	2150771
	2150772
	2150775
	2150776
	2150786
	2150788

	2150790
	2150793
	2150818
	2150830
	2150840
	2150860
	2150868
	2150869
	2151035
	2151083
	2151107
	2151112
	2151114
	2151240
	2151242
	2151244
	2151245
	2151250
	2152065
	2152068
	2152069
	2152070
	2152072
	2152075
	2152085
	2152304
	2152316
	2152317
	2152321
	2152337
	2152339
	2152343
	2152345
	2152346
	2152348
	2152352
	2152354
	2152355
	2152356

	2152357
	2152359
	2152385
	2152387
	2152388
	2152389
	2155115
	215077286
SARS-CoV-2 DETECTABLE	2152121
	2152128
	2152144
	2152290
	2152312
	2152319
	2152320
	2152330
	2152347
	2152386
	21477101
	2150151
	2150154
	2150158
	2150230
	2150265
	2150306
	2150319
	2150320
	2150730
	2150761
	2150767
	2150823
	2150828
	2150867
	2150987
	2151028
	2151032
	2151037
	2151038
	2151045

	2151105
	2151279
	2151498
	2151542
	2151544
	2151592
	2151593
	2151601
	2151719
	2151730
	2151754
	2151774
	2151784
	2151816
	2151818
	2151841
	2151899
	2151908
	2151935
	2151941
	2151989
	2152064
	2152066
	2152067
	2150151
	2150154
	2150158
	2150141
	2150126
	2150065
	2150131-1
	2150128-1
	2150127-1
	2150122-1
	2150120-1
	2150114-1

SARS-CoV-2 NO DETECTABLE (training set)	2150108-1
	2150097-1
	2150096-1
	2150075-1
	2150073-1
	2150071-1
	2150070-1
	2150069-1
	2150068-1
	2150067-1
	2150066-1
	2150063-1
	2150050-1
	2150049-1
	2150047-1
	2150046-1
	2150034-1
	2150023-1
HOSPITAL DE CLINICAS JOSE DE SAN MARTIN	
IDENTIFICACION DE REFERENCIA	**N° de muestra**
SARS-CoV-2 DETECTABLE	HC 12
	HC 11
	HC 5
	HC 4
	HC 3
	HC 2
	HC 1
	HC 13
	HC 14

INFLUENZA POSITIVO	HC 20
	HC 21
	HC 22
	HC 23
	HC 24
SARS-CoV-2 NO DETECTABLE	HC 6
	HC 7
	HC 8
	HC 9
	HC 10
	HC 15
	HC 16
	HC17
	HC 18
	HC 19
188	**N TOTAL**

Tabla S3. Picos característicos en los perfiles proteicos (20 MSPs) que componen la Base de Datos ***in house***, obtenidos mediante el software ClinProTools, con los respectivos valores de las pruebas estadísticas.

Mass	PTTA	PWKW	PAD
3099,28	0,128	0,0426	0,03580
3372,3	0,0133	0,0199	0,05450
3443,28	0,0133	0,0222	0,04480
3465,6	0,0133	0,0244	0,05790
3778,83	0,128	0,0107	0,00000
3805,86	0,156	0,0236	< 0.000001
4062,3	0,128	0,0459	0,00013
4242,57	0,128	0,0124	0,00019

4842	0,269	0,0244	0,00000
5529,72	0,129	0,0437	0,00000
5558,22	0,185	0,0459	< 0.000001
5944,52	0,0543	0,0199	0,00320
5968,94	0,128	0,0244	0,00016
5991,83	0,128	0,0437	0,03090
6026,52	0,705	0,0437	< 0.000001
6347,57	0,0146	0,0096	0,02110
6824,02	0,128	0,0096	< 0.000001
6845,48	0,128	0,0459	0,00006
6996,15	0,185	0,0459	0,00032
7013,01	0,128	0,0126	0,00034
7030,77	0,128	0,0236	0,00248
7111,92	0,0543	0,0011	0,00015
7134,23	0,0885	0,0151	0,00658
9436,43	0,135	0,0459	0,00115
10836,83	0,0885	0,0236	0,00019
12691,94	0,0284	0,0126	0,00671
13155,98	0,128	0,0426	< 0.000001
13273,88	0,12	0,0437	0,00003

PTTA: valor de p obtenido mediante la prueba t.

PWKW: valor de p obtenido mediante la prueba de Wilcoxon / Kruskal-Wallis.

PAD: valor de p obtenido mediante la prueba de Anderson-Darling.

Tabla S4. Picos característicos obtenidos en ClinProTools para cada modelo de clasificación desarrollado.

MODELO A DE 2 CLASES				
Mass	**DAve**	**PTTA**	**PWKW**	**PAD**
3037,3	0,55	0,669	0,00000704	0
3053,4	0,05	0,937	0,000401	0
3141,09	0,77	0,000982	0,0323	< 0.000001
3274,56	0,14	0,684	0,0224	< 0.000001
3298,16	0,54	0,0677	0,000164	< 0.000001
3318,59	1,24	0,00155	0,00267	< 0.000001
3328,68	1,94	0,00834	0,0000534	0
3338,31	1,76	0,000671	0,00000507	< 0.000001
3359,22	2,66	0,0000324	0,000015	< 0.000001
3372,19	7,85	0,000171	0,0000138	< 0.000001
3393,78	2,54	0,00182	0,00418	< 0.000001
3415,88	1,91	0,00279	0,00939	< 0.000001
3443,14	7,79	0,00834	0,000042	< 0.000001
3464,89	2,35	0,0203	0,00244	< 0.000001
3475,85	2,48	0,000608	0,000148	< 0.000001
3487,1	3,24	0,0737	0,0191	< 0.000001
3521,16	0,9	0,041	0,00239	< 0.000001
3792,31	0,51	0,041	0,00958	< 0.000001
3828,01	1,69	0,0000545	0,00532	< 0.000001
4063,7	1,42	< 0.000001	< 0.000001	< 0.000001
4078,24	3,49	0,0000111	< 0.000001	0
4121,73	0,76	0,00151	0,000157	< 0.000001
4215,93	0,6	0,0287	0,000572	< 0.000001
4228,01	0,31	0,0926	0,0241	< 0.000001
4551,55	1,46	0,000105	0,00313	< 0.000001
4940,53	2,27	0,0000326	0,000367	< 0.000001
4966,6	4,79	0,000277	0,000093	< 0.000001
4985,73	3,07	0,0000339	0,0000101	< 0.000001

MODELO A DE 2 CLASES				
Mass	**DAve**	**PTTA**	**PWKW**	**PAD**
5005,44	1,67	0,000104	0,000796	< 0.000001
5049	1,04	0,000688	0,00205	< 0.000001
5102,52	1,08	< 0.000001	0,000042	< 0.000001
5120,59	1,27	0,0000324	0,000646	< 0.000001
5137,74	2,06	0,0000183	0,000563	< 0.000001
5157,05	1,81	0,0000522	0,000222	< 0.000001
5177,39	1,24	< 0.000001	0,0000656	< 0.000001
5200,72	1,22	< 0.000001	0,0000273	< 0.000001
5222,16	2,33	9,23E-06	0,000029	< 0.000001
5236,08	3,88	0,00584	0,000764	< 0.000001
5257,56	1,65	0,00185	0,0187	< 0.000001
5281,88	1,43	0,000332	0,00537	< 0.000001
5382,8	0,67	0,471	0,0419	< 0.000001
5423,4	1,1	0,0000813	0,000194	< 0.000001
5821,18	0,13	0,0991	0,0168	< 0.000001
5886,24	0,57	0,00254	0,000607	< 0.000001
5928,77	0,49	< 0.000001	< 0.000001	< 0.000001
5947,23	0,53	3,81E-06	< 0.000001	< 0.000001
5970,51	0,46	0,00307	< 0.000001	< 0.000001
5993,74	0,38	4,81E-06	< 0.000001	< 0.000001
6178,26	0,58	1,05E-06	< 0.000001	< 0.000001
6192,52	0,64	< 0.000001	< 0.000001	< 0.000001
6349,58	0,31	0,000642	0,0108	< 0.000001
6734,9	0,1	0,0511	0,00732	< 0.000001
6826,67	0,86	0,00294	0,00111	0
6846,9	0,46	0,00155	0,000148	0
6889,14	0,23	0,0753	0,014	< 0.000001
7858,61	0,02	0,799	0,00311	0
8151,33	0,12	0,0677	< 0.000001	< 0.000001
8217,02	0,2	< 0.000001	< 0.000001	< 0.000001
8296,73	0,34	1,05E-06	< 0.000001	< 0.000001

MODELO A DE 2 CLASES				
Mass	**DAve**	**PTTA**	**PWKW**	**PAD**
8343,38	0,22	1,05E-06	< 0.000001	< 0.000001
8363,75	0,22	0,000142	0,00000507	0
8454,06	0,45	0,00254	0,00000464	< 0.000001
8526,85	0,31	< 0.000001	< 0.000001	< 0.000001
8566,46	0,61	1,05E-06	0,0000403	< 0.000001
8632,06	0,3	< 0.000001	< 0.000001	< 0.000001
8739,58	0,16	0,000663	0,00000507	< 0.000001
8801,09	0,09	0,00996	0,0056	< 0.000001
9065,22	0,03	0,455	0,00484	< 0.000001
9596,4	0,06	0,0914	0,00641	< 0.000001
9798,57	0,03	0,646	0,000886	0
9959,39	0,1	0,0304	0,00185	< 0.000001
10525,53	0,09	0,0027	0,0000834	< 0.000001
10654,22	0,13	0,000183	0,00000394	< 0.000001
10840,12	0,54	0,0232	0,00267	< 0.000001
11007,15	0,14	0,11	0,0269	< 0.000001
11327,7	0,16	0,000535	< 0.000001	< 0.000001
11368,08	0,1	0,249	0,000117	< 0.000001
11457,57	0,15	0,000277	< 0.000001	< 0.000001
11518,27	0,08	0,329	0,00173	< 0.000001
11730,23	0,76	0,0000366	< 0.000001	< 0.000001
11982,61	0,38	< 0.000001	< 0.000001	< 0.000001
12774,94	0,05	0,348	0,0385	< 0.000001
13460,83	0,1	0,00772	0,000404	< 0.000001
13782,09	0,21	1,98E-06	< 0.000001	< 0.000001

MODELO B DE 3 CLASES			
Mass	**PTTA**	**PWKW**	**PAD**
3037	0,12	< 0.000001	0
3053,15	0,909	0,000884	0
3274,44	0,768	0,0415	< 0.000001
3298,55	0,00000379	< 0.000001	< 0.000001
3318,38	0,0000046	0,00000382	< 0.000001
3329,32	0,0198	0,00033	0
3338,31	0,000249	< 0.000001	0
3358,9	0,0000601	0,00000464	< 0.000001
3372,29	0,000023	0,00000427	< 0.000001
3394,21	0,000688	0,000453	< 0.000001
3415,88	0,00572	0,0142	< 0.000001
3443,31	0,00149	0,00000568	< 0.000001
3465,33	0,0183	0,00758	< 0.000001
3476,25	0,002	0,000581	< 0.000001
3487,4	0,00969	0,0359	< 0.000001
3521,3	0,106	0,0023	< 0.000001
3779,51	0,000634	0,0469	< 0.000001
3792,29	0,0191	0,00343	< 0.000001
3805,63	0,00000264	0,00103	< 0.000001
3828,21	0,00000499	0,00000273	< 0.000001
4063,48	0,00000216	0,00000189	< 0.000001
4077,7	0,00000507	< 0.000001	0
4121,95	0,00359	0,000346	< 0.000001
4137,25	0,00792	0,0375	0
4159,67	0,000465	0,0000998	< 0.000001
4192,97	0,0976	0,00593	< 0.000001
4215,92	0,00565	0,0000143	< 0.000001
4228,2	0,219	0,026	< 0.000001
4244,24	< 0.000001	0,0168	0
4551,64	< 0.000001	< 0.000001	< 0.000001
4714,52	0,000572	0,00757	< 0.000001

MODELO B DE 3 CLASES			
Mass	**PTTA**	**PWKW**	**PAD**
4842,79	< 0.000001	0,00000524	< 0.000001
4900,18	0,0000216	0,000574	< 0.000001
4940,61	< 0.000001	0,00000786	< 0.000001
4966,4	0,0000157	< 0.000001	< 0.000001
4985,38	0,00000296	< 0.000001	< 0.000001
5005,74	0,0000207	0,0000593	< 0.000001
5027,01	< 0.000001	0,0000172	< 0.000001
5047,15	0,0000157	< 0.000001	< 0.000001
5102,48	< 0.000001	< 0.000001	< 0.000001
5120,27	< 0.000001	< 0.000001	< 0.000001
5137,86	< 0.000001	< 0.000001	< 0.000001
5157,12	< 0.000001	< 0.000001	< 0.000001
5177,37	< 0.000001	< 0.000001	< 0.000001
5200,9	< 0.000001	< 0.000001	< 0.000001
5236,12	< 0.000001	< 0.000001	< 0.000001
5257,66	< 0.000001	< 0.000001	< 0.000001
5282,52	< 0.000001	0,0000656	< 0.000001
5300,99	0,00000425	0,0349	< 0.000001
5382,94	< 0.000001	0,00000425	< 0.000001
5404,05	0,000918	0,0129	< 0.000001
5422,71	0,0000619	0,00132	< 0.000001
5595,25	< 0.000001	0,0151	< 0.000001
5618,66	0,00675	0,0185	0
5868,49	0,0635	0,0469	0
5945,31	< 0.000001	< 0.000001	< 0.000001
5970,65	0,00156	< 0.000001	< 0.000001
5996,93	0,0000874	< 0.000001	0
6027,04	0,0019	0,0408	0
6179,7	0,00000118	< 0.000001	< 0.000001
6192,6	< 0.000001	< 0.000001	< 0.000001
6349,12	0,00000752	0,0000438	0

MODELO B DE 3 CLASES			
Mass	**PTTA**	**PWKW**	**PAD**
6580,6	0,00353	0,0129	< 0.000001
6641,73	0,00678	0,00295	< 0.000001
6733,82	0,107	0,0408	< 0.000001
6826,65	0,00293	0,000282	0
6847,39	0,00378	0,000174	0
6889,92	0,111	0,0000707	< 0.000001
6933,06	0,379	0,0408	< 0.000001
6955,02	0,586	0,0000489	< 0.000001
6973,67	0,969	0,00201	< 0.000001
6998,6	0,882	0,00804	0
7858,43	0,94	0,00826	0
7882,47	0,186	0,00188	< 0.000001
7993,54	0,0716	0,00974	< 0.000001
8149,4	0,114	0,00000621	< 0.000001
8217,27	< 0.000001	< 0.000001	< 0.000001
8296,95	0,00000144	< 0.000001	< 0.000001
8362,83	0,000177	0,0000159	0
8454	0,00751	0,00000108	< 0.000001
8524,41	< 0.000001	< 0.000001	< 0.000001
8566,73	0,00000181	0,000116	< 0.000001
8609,06	< 0.000001	< 0.000001	< 0.000001
8631,32	< 0.000001	< 0.000001	< 0.000001
8736,43	0,00238	0,0000117	< 0.000001
8767,43	0,00541	0,000204	< 0.000001
8801,09	0,0217	0,00773	< 0.000001
8834,5	0,134	0,00395	< 0.000001
8964,46	0,00799	0,000426	< 0.000001
9028	0,0106	0,000795	< 0.000001
9066,02	0,157	0,00593	< 0.000001
9439,25	0,000115	0,000108	< 0.000001
9520,79	0,191	0,000291	0

MODELO B DE 3 CLASES			
Mass	**PTTA**	**PWKW**	**PAD**
9594,67	0,241	0,0303	< 0.000001
9797,6	0,768	0,0041	0
9959,09	0,00015	0,000158	< 0.000001
10095,92	< 0.000001	0,00148	0
10313,56	0,0153	0,00692	< 0.000001
10526,51	0,00553	0,000185	< 0.000001
10598,03	< 0.000001	< 0.000001	< 0.000001
10653,93	0,000783	0,0000242	< 0.000001
10839,07	0,00541	0,00563	< 0.000001
11007,03	0,163	0,0361	< 0.000001
11324,91	0,000272	< 0.000001	< 0.000001
11368,11	0,244	0,0000656	< 0.000001
11450,46	0,000655	< 0.000001	< 0.000001
11518,8	0,0737	0,0023	0
11731,47	0,000071	< 0.000001	< 0.000001
11983,45	< 0.000001	< 0.000001	< 0.000001
12322,66	< 0.000001	< 0.000001	< 0.000001
12694,69	0,0323	0,0142	< 0.000001
12774,25	0,0919	0,0128	0
13160,9	0,00836	0,0000593	0
13276,1	0,00919	0,000648	< 0.000001
13458,2	0,00378	0,000291	< 0.000001
13779,14	0,00000264	< 0.000001	< 0.000001
14971,99	0,00969	< 0.000001	0

MODELO C DE 2 CLASES			
Mass	**PTTA**	**PWKW**	**PAD**
3297,99	0,0105	0,00323	< 0.000001
3317,85	0,0194	0,0118	< 0.000001
3358,47	0,0425	0,000591	< 0.000001
3372,15	0,00108	0,00059	< 0.000001
3394,17	0,00856	0,000829	< 0.000001
3443,19	0,00332	0,000319	< 0.000001
3465,33	0,0384	0,0414	< 0.000001
3509,33	0,0671	0,91	< 0.000001
3612,3	0,00135	0,00218	< 0.000001
3792,09	0,0107	0,00276	< 0.000001
3827,59	0,0945	0,00229	< 0.000001
4063,14	0,062	0,00224	< 0.000001
4077,47	0,00881	0,00794	< 0.000001
4159,53	0,0199	0,0334	< 0.000001
4191,54	0,0319	0,0115	< 0.000001
4205,05	0,96	0,000687	0
4213,98	0,00143	6,2505	< 0.000001
4227,82	0,783	0,0374	< 0.000001
4243,08	0,000313	0,00793	< 0.000001
4550,62	< 0.000001	6,89E-05	< 0.000001
4713,67	0,00101	0,00313	< 0.000001
4841,81	2,58E-06	8,45E-06	< 0.000001
4898,87	0,000188	0,000919	< 0.000001
4939,73	< 0.000001	7,7605	< 0.000001
4965,83	1,65E-05	3,8206	< 0.000001
4985,26	6,51E-06	1,0606	< 0.000001
5005,63	0,000019	0,000382	< 0.000001
5026,8	6,51E-06	6,2505	< 0.000001
5045,54	0,0543	0,000382	< 0.000001
5099,3	0,00443	0,0104	< 0.000001
5114,91	< 0.000001	3,8206	< 0.000001

MODELO C DE 2 CLASES			
Mass	**PTTA**	**PWKW**	**PAD**
5136,14	< 0.000001	< 0.000001	< 0.000001
5156,31	< 0.000001	< 0.000001	< 0.000001
5177,16	0,00102	1,4705	< 0.000001
5235,02	9,29E-06	0,000128	< 0.000001
5256,78	0,000153	0,000919	< 0.000001
5381,66	2,66E-05	0,0117	< 0.000001
5513,02	0,236	0,91	< 0.000001
5577,12	0,0129	0,0447	< 0.000001
5867,67	0,0386	0,0266	< 0.000001
5942,18	< 0.000001	0,000208	< 0.000001
5969,47	0,00155	0,00731	< 0.000001
6025,34	0,0191	0,0414	< 0.000001
6179,29	0,256	0,0126	< 0.000001
6348,6	0,000975	7,7605	< 0.000001
6579,71	0,00762	0,00794	< 0.000001
6641,18	0,00819	0,00182	< 0.000001
6825,7	0,00155	0,00177	< 0.000001
6846,55	0,0325	0,00955	< 0.000001
6888,9	0,182	8,4506	< 0.000001
6928,44	0,291	0,0042	< 0.000001
6953,46	0,859	0,000113	< 0.000001
6996,41	0,883	0,00255	< 0.000001
8215,93	0,0108	0,00218	1,37E-05
8296,6	0,0744	0,0414	< 0.000001
8453,28	0,0734	0,00014	< 0.000001
8522,47	0,062	0,0103	< 0.000001
8609,38	0,431	0,91	< 0.000001
8798,77	0,465	0,91	9,65E-06
8833,89	0,934	0,91	4,05E-05
9437,42	0,0106	0,0181	9,65E-06
9956,84	8,34E-06	7,7605	< 0.000001

MODELO C DE 2 CLASES			
Mass	**PTTA**	**PWKW**	**PAD**
10093,81	9,29E-06	0,00116	< 0.000001
10175,09	0,00342	0,0394	< 0.000001
10309,89	0,0123	0,00294	5,5E-06
10446,67	0,0235	0,0464	0
11324,28	0,783	0,91	< 0.000001
11367,07	0,25	0,00218	< 0.000001
11517,39	0,0376	0,0179	< 0.000001
11983,22	0,00443	0,0303	< 0.000001
12694,14	0,0191	0,0266	< 0.000001
13160,17	0,00819	0,00116	0
13275,6	0,00665	0,00515	< 0.000001

DAve: diferencia entre la máxima y la mínima intensidad promedio de picos de todas las clases.

PTTA: valor de p obtenido mediante la prueba t.

PWKW: valor de p obtenido mediante la prueba de Wilcoxon / Kruskal-Wallis.

PAD: valor de p obtenido mediante la prueba de Anderson-Darling.

Tabla S5. Regiones de integración empleadas para la clasificación de acuerdo a los diferentes algoritmos desarrollados.

MODELO A	MODELO B	MODELO C
3141,09	3274,44	3297,99
3153,2	3298,55	3317,85
3274,56	3318,38	3329,53
3298,16	3329,32	3337,77
3318,59	3338,31	3358,47
3328,68	3358,9	3415,86
3338,31	3394,21	3443,19
3359,22	3415,88	3475,1
3372,19	3443,31	3509,33
3393,78	3465,33	3520,17
3415,88	3476,25	3612,3
3443,14	3487,4	3710,42
3464,89	3509,27	3792,09
3475,85	3521,3	3827,59
3487,1	3752,73	3981,9
3508,64	3792,29	4063,14
3521,16	3828,21	4191,54
3545,88	4159,67	4213,98
3710,89	4215,92	4227,82
3753,3	4357,11	4243,08
3779,64	4551,64	4419,13
3792,31	4572,95	4550,62
3805,73	4842,79	4713,67
3828,01	4900,18	4841,81
3983,02	4940,61	4898,87
4063,7	4985,38	4939,73
4121,73	5005,74	4985,26

MODELO A	MODELO B	MODELO C
4137,23	5027,01	5005,63
4215,93	5047,15	5026,8
4357,05	5102,48	5099,3
4395,66	5236,12	5114,91
4551,55	5257,66	5177,16
4572,95	5282,52	5256,78
4714,43	5300,99	5279,69
4966,6	5382,94	5300,77
4985,73	5404,05	5362,53
5005,44	5422,71	5403,26
5026,6	5595,25	5421,88
5049	5695,34	5493,96
5102,52	5799,14	5513,02
5120,59	5945,31	5529,62
5137,74	5996,93	5577,12
5200,72	6179,7	5686,29
5222,16	6192,6	5867,67
5281,88	6308,19	5942,18
5301,44	6349,12	5999,92
5403,51	6387,64	6025,34
5423,4	6580,6	6179,29
5696,12	6641,73	6192,48
5799,18	6733,82	6306,15
5821,18	6889,92	6348,6
5868,4	7348,4	6387,16
5886,24	7452,34	6579,71
5928,77	7567,1	6641,18
5947,23	7882,47	6825,7
5970,51	7935,93	6846,55
6178,26	7993,54	6888,9

MODELO A	MODELO B	MODELO C
6192,52	8149,4	6928,44
6308,15	8217,27	7347,45
6349,58	8296,95	7449,95
6580,52	8362,83	7672,31
6642,38	8454	7858,69
6734,9	8524,41	7935,75
6931,67	8566,73	7993,58
6997,5	8609,06	8215,93
7348,85	8631,32	8361,11
7486,74	8736,43	8453,28
7570,04	8767,43	8522,47
7611,17	8801,09	8609,38
7630,92	8964,46	8630,83
7858,61	9028	8735,05
7933,72	9066,02	8768,49
8217,02	9439,25	8798,77
8296,73	9520,79	9066,37
8343,38	9594,67	9437,42
8363,75	9959,09	9592,93
8454,06	10313,56	9794,42
8526,85	10403,9	9956,84
8566,46	10446,36	10093,81
8632,06	10526,51	10175,09
8739,58	10598,03	10309,89
8801,09	10653,93	10402,61
9065,22	10839,07	10446,67
9596,4	11007,03	10526,81
9959,39	11324,91	10597,93
10096,13	11368,11	10650,39
10177,17	11450,46	10762,49

MODELO A	MODELO B	MODELO C
10313,98	11518,8	10838,35
10403,65	11731,47	10920,15
10525,53	11983,45	11324,28
10654,22	12322,66	11367,07
10808,06	12694,69	11448,18
11007,15	12774,25	11517,39
11327,7	13160,9	11983,22
11368,08	13276,1	12694,14
11457,57	13458,2	12773,58
11982,61	13779,14	13160,17
12774,94	14971,99	13275,6
13460,83		13457,24
13782,09		14690,06

Tabla S6. Resultados obtenidos de las muestras clínicas (N= 167) empleando Machine Learning y detección de Biomarcadores, comparados con la técnica de referencia actual (qRT-PCR).

ID	RESULTADOS DEL ML	BIOMARCADORES INTERPRETACION	ML + BIOMARCADORES	RESULTADO DE LA RT - PCR
21770	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO
21782	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 POSITIVO
21785	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO
21792	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 POSITIVO
21804	COVID-19 NEGATIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO
21815	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 POSITIVO
21817	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO
21843	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 POSITIVO
21863	COVID-19 NEGATIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO
21901	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 NEGATIVO
21902	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
21903	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
21904	COVID-19 POSITIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO
21905	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
21908	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO
21914	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO
21915	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
21916	COVID-19 NEGATIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 NEGATIVO
21919	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 NEGATIVO
21920	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2149937	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150023	COVID-19 NEGATIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 NEGATIVO
2150034	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150046	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO

ID	RESULTADOS DEL ML	BIOMARCADORES INTERPRETACION	ML + BIOMARCADORES	RESULTADO DE LA RT - PCR
2150047	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150049	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150050	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150063	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150066	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150067	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150068	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150069	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150070	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150071	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150073	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150075	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150096	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150097	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150108	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150114	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150120	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150122	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150127	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150128	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150131	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2149937	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150771	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 NEGATIVO
2150772	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150775	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150776	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 NEGATIVO
2150786	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150788	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO

ID	RESULTADOS DEL ML	BIOMARCADORES INTERPRETACION	ML + BIOMARCADORES	RESULTADO DE LA RT - PCR
2150790	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 NEGATIVO
2150793	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 NEGATIVO
2150818	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 NEGATIVO
2150830	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150840	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2150860	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 NEGATIVO
2150868	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 NEGATIVO
2150869	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2151035	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2151083	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 NEGATIVO
2151107	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2151112	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2151114	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2151240	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2151242	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 NEGATIVO
2151244	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 NEGATIVO
2151245	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2151250	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2152065	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 NEGATIVO
2152068	COVID-19 POSITIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2152069	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2152070	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 NEGATIVO
2152072	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 NEGATIVO
2152075	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 NEGATIVO
2152085	COVID-19 POSITIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2152304	COVID-19 POSITIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2152316	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 NEGATIVO
2152317	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO

ID	RESULTADOS DEL ML	BIOMARCADORES INTERPRETACION	ML + BIOMARCADORES	RESULTADO DE LA RT - PCR
2152321	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 NEGATIVO
2152337	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 NEGATIVO
2152339	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2152343	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2152345	COVID-19 NEGATIVO	COVID-19 POSITIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2152346	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2152348	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 NEGATIVO
2152352	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2152354	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 NEGATIVO
2152355	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2152356	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 NEGATIVO
2152357	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2152359	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 NEGATIVO
2152385	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2152387	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2152388	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2152389	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2155115	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
215077286	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
2152121	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO
2152128	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 POSITIVO
2152144	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 POSITIVO
2152290	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 POSITIVO
2152312	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 POSITIVO
2152319	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 POSITIVO
2152320	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 POSITIVO
2152330	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO
2152347	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 POSITIVO

ID	RESULTADOS DEL ML	BIOMARCADORES INTERPRETACION	ML + BIOMARCADORES	RESULTADO DE LA RT - PCR
2152386	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 POSITIVO
21477101	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 POSITIVO
2150151	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 POSITIVO
2150154	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 POSITIVO
2150158	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 POSITIVO
2150230	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 POSITIVO
2150265	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 POSITIVO
2150306	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 POSITIVO
2150319	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO
2150320	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 POSITIVO
2150730	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO
2150761	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 POSITIVO
2150767	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 POSITIVO
2150823	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 POSITIVO
2150828	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 POSITIVO
2150867	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO
2150987	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 POSITIVO
2151028	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO
2151032	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 POSITIVO
2151037	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO
2151038	COVID-19 NEGATIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO
2151045	COVID-19 NEGATIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO
2151105	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO
2151279	COVID-19 NEGATIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO
2151498	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO
2151542	COVID-19 NEGATIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO
2151544	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO
2151592	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO

ID	RESULTADOS DEL ML	BIOMARCADORES INTERPRETACION	ML + BIOMARCADORES	RESULTADO DE LA RT - PCR
2151593	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO
2151601	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO
2151719	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 POSITIVO
2151730	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 POSITIVO
2151754	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 POSITIVO
2151774	COVID-19 NEGATIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO
2151784	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 POSITIVO
2151816	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 POSITIVO
2151818	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO
2151841	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 POSITIVO
2151899	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 POSITIVO
2151908	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 POSITIVO
2151935	COVID-19 POSITIVO	NC	COVID-19 NEGATIVO	COVID-19 POSITIVO
2151941	COVID-19 NEGATIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO
2151989	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 POSITIVO
2152064	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 POSITIVO
2152066	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 POSITIVO
2152067	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO
2150151	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 POSITIVO
2150154	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 POSITIVO
2150158	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 POSITIVO
R81	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 NEGATIVO
R82	COVID-19 POSITIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO
R83	COVID-19 POSITIVO	NC	COVID-19 POSITIVO	COVID-19 NEGATIVO
R84	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO
R85	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO
R86	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO
R87	COVID-19 POSITIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO

ID	RESULTADOS DEL ML	BIOMARCADORES INTERPRETACION	ML + BIOMARCADORES	RESULTADO DE LA RT - PCR
R88	COVID-19 NEGATIVO	NC	COVID-19 NEGATIVO	COVID-19 NEGATIVO
R89	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 POSITIVO	COVID-19 NEGATIVO
R90	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO	COVID-19 NEGATIVO

NC: No Conclusivo.