Universidad CEU San Pablo

CEINDO – CEU Escuela Internacional de Doctorado

PROGRAMA en Ciencia y Tecnología de la Salud

Design, validation and implementation of a software tool for metabolites annotation and identification

A Doctoral Thesis Presented in
Partial Fulfilment of the Requirements for
the Degree of Doctor at Universidad San Pablo CEU by
Alberto Gil de la Fuente

Supervised by
Coral Barbas Arribas
Abraham Otero Quintana

Universidad CEU San Pablo
Escuela Politécnica Superior
CEMBIO
Madrid 2019

Este trabajo ha sido realizado en el Centro de Metabolómica y Bioanálisis (CEMBIO), Departamento de Química y Bioquímica, Área de Química Analítica de la Facultad de Farmacia de la Universidad San Pablo CEU (Madrid), y en el laboratorio de Ingeniería Biomédica, Departamento de Ingeniería Biomédica, Área de Ingeniería Biomédica de la Escuela Politécnica Superior de la Universidad CEU-San Pablo, bajo la dirección de la Dra. Coral Barbas Arribas y del Dr. Abraham Otero Quintana. Este trabajo es un compendio de trabajos publicados en revistas indexadas en la primera mitad del índice del Journal Citation of Reports (JCR). Estos trabajos se listan a continuación:

1) **Gil-de-la-Fuente, A.**, Armitage, E.G., Otero, A., Barbas, C. and Godzien, J.; *Differentiating signals to make biological sense - a guide through databases for MS-based non-targeted metabolomics*; **Electrophoresis**, 2017, 38(18), 2242-2256
   - Impact factor: 2.744, Q2 (Analytical chemistry; Biochemical research methods)

2) **Gil-de-la-Fuente, A.**, Godzien, J., Fernández López, M., Rupérez, F.J., Barbas, C. and Otero, A.; *Knowledge-based metabolite annotation tool: CEU Mass Mediator*; **Journal of Pharmaceutical and Biomedical Analysis**, 2018, 154, 138-149
   - Impact factor 3.255, Q1 (Analytical chemistry; Pharmacology and pharmacy)

3) **Gil-de-la-Fuente, A.**, Traldi, F., Siroka, J., Kretowski, A., Ciborowski, M., Otero, A., Barbas, C. and Godzien, J.; *Characterization and annotation of oxidized glycerophosphocholines for non-targeted metabolomics with LC-QTOF-MS data;* **Analytica Chimica Acta**, 2018, 1037(11), 358-368
   - Impact factor 5.123, Q1 (Analytical chemistry)

4) **Gil-de-la-Fuente, A.**, Godzien, J., Saugar, S., Garcia-Carmona, R., Badran, H., Wishart, D.S., Barbas, C. and Otero, A.; *CEU Mass Mediator 3.0: A Metabolite Annotation Tool*; **Journal of Proteome Research**, 2019, 18(2), 797-802
   - Impact factor 3.950, Q1 (Biochemical research methods)

La Dra. Coral Barbas Arribas, y el Dr. Abraham Otero Quintana, directores de este trabajo, expresan su conformidad para la presentación del mismo por considerar que reúne los requisitos necesarios y constituye una aportación original al tema tratado.

Fdo. Dra. Coral Barbas Arribas Fdo. Dr. Abraham Otero Quintana

“Doubt is the beginning of wisdom”

Aristotle

TABLE OF CONTENTS

TABLE OF CONTENTS

List of abbreviations

API	Application Programming Interface
CCS	Collision Cross Section
CE	Capillary Electrophoresis
CEMBIO	Centre for Metabolomics and Bioanalysis
CER	Ceramide
CFM-ID	Competitive Fragmentation Modelling for Metabolites Identification
CID	Collision induced dissociation
CMM	CEU Mass Mediator
CS	Composite Spectrum
CSI	Compound Structure Identification
DG	Diacylglycerols
DFA	Discriminant Function Analysis
EI	Electron Impact
ESI	Electrospray Ionization
FA	Fatty Acid
GC	Gas Chromatography
HILIC	Hydrophilic Interaction Liquid Chromatography
HMDB	Human Metabolome Database
InChI	International Chemical Identifier
InChI Key	Hashed International Chemical Identifier
IP	Ionization Product
IS	Internal Standard
IT	Ion Trap
J2EE	Java 2 Platforms, Enterprise Edition
JS	JavaScript
JSON	JavaScript Object Notation
LC	Liquid Chromatography
LipidMaps	LIPID Metabolites and Pathways Strategy
LPC	Lysophosphocholine
MG	Monoradylglycerol
MINE	Metabolic *In silico* Network Expansion Databases
MS	Mass Spectrometry

MSI	Metabolomics Standard Initiative
MS^n	Multi-Stage Mass Spectrometry
MT	Migration Time
m/z	Mass to charge
NIMS	Nanostructure Imaging Mass Spectrometry
NMR	Nuclear Magnetic Resonance
oxPCs	oxidized glycerophosphocholines
PA	Glycerophosphates
PC	Phosphocholines/Glycerophosphocholines
PCA	Principal Component Analysis
PE	Phosphoetanolamine
PG	Phosphoglycerol
PI	Phosphoinositol
PIF	Precursor Ion Fingerprinting
ppm	Parts per million
PS	Phosphoserine
RefMet	Reference set of Metabolites
REST	Representational State Transfer
RI	Retention Index
RMT	Relative Migration Time
RP	Reversed-Phase
RT	Retention Time
SM	phosphosphingolipid
ST	Sterol (cholesterol ester)
TG	Triacylglycerol
TOF	Time Of Flight
Q	Quadrupole
QQQ	Triple Quadrupole

ACKNOWLEDGEMENTS

The projects leading to the results presented in this thesis have been carried out in the University CEU-San Pablo, concretely in the Centre for Metabolomics and Bioanalysis (CEMBIO), and in the Biomedical Engineering Laboratory. I thank the University CEU-San Pablo for giving me the opportunity to work on this exciting venture.

I would like to convey my gratitude to my supervisors, Dr. Coral Barbas Arribas and Dr. Abraham Otero Quintana, for giving me the chance to work on this project. I am especially obliged for their continuous support and patience. Their expectatives for excellent work has motivated me day after day to finish my journey as a PhD student.

Many thanks to Dr. Joanna Godzien, who welcomed me with a big smile since the beginning of the project. She taught me how hard you need to work to enhance your performance, to accept the new challenges and to meet the expectations. She was an authentic leader and she supported me on each stage of this project.

It has been an honour to work in the CEMBIO group with such a diverse group of scientists and students from different backgrounds and cultures. I would like to express my sincere gratitude to all the people who have directly or indirectly helped and encouraged me during my PhD.

I would like to thank to the Biomedical Engineering Group. They integrated me in their group even though I was not physically with them all the time and they led me patiently to the right track when I felt lost. Special thanks to Sergio Saugar, Rodrigo García Carmona, Mariano Fernández García and Andrés Esteban Fernández.

I was able to collaborate with different scientists who contributed to the development of the project. I enjoyed a period abroad in the University of Alberta, where I worked in the Bioinformatics group of the Wishart Lab. I learnt there how a large bioinformatics group is organized, how to join a different team for a short period of time, and how to be useful to them in return to all the things that they provided me with. I am particularly thankful to the head of the group, Dr. David. S. Wishart. He welcomed me very kindly, he allowed all the researchers to participate in different kind of projects and I felt very close with everyone. Particularly Dr. Danuta Chamot, Dr. Yannick Djoumbou-Feunang and Hasan Badran made this period quite fun and enjoyable.

In this stage I would like to sincerely thank Tammy and Darrell Pidner who helped me a lot to settle down in Alberta, making me feel like home.

During this trip, I have had the great pleasure to share my experiences with a lot of people aiming to share their knowledge and finding solutions. Thanks to all my friends that they have contributed with their support and patience. I would like to mention a special collaborator, MSc. Yaoxiang Li, a researcher from Georgetown university that is actively collaborating in the development of new features in the CMM tool with a proactive attitude and a critical vision.

Finally, I am very grateful to my family for being supportive and positive. They taught me the importance of the responsibilities and duties in our decisions.

They all have significantly contributed to the presented thesis.

Sincerely,

Alberto

RESUMEN

La metabolómica es una subárea de la biología de sistemas que tiene como objetivo el estudio de las moléculas de pequeño tamaño (normalmente < 1,000 Da) producidas por los procesos metabólicos que concurren en una célula. Desde finales del siglo anterior la metabolómica con un enfoque no dirigido se ha empleado con éxito en diferentes aplicaciones, como el descubrimiento de biomarcadores, el descubrimiento de dianas terapeúticas, la medicina personalizada o simplemente conocer los mecanismos biológicos del organismo estudiado. Al tratarse de estudios no dirigidos, la investigación trata de obtener tanta información como sea posible para cubrir el mayor número de metabolitos presentes, siendo esta fase clave en el éxito de la investigación.

El número de metabolitos extraídos y posteriormente identificados con cierto nivel de confianza puede definirse como cobertura de metabolitos o “metabolite coverage”. Esta fase de identificación de metabolitos es actualmente el principal cuello de botella en los análisis metabolómicos, puesto que la información obtenida analíticamente requiere de una extensa cantidad de trabajo y conocimiento para permitir obtener identificaciones con éxito. Las fases de separación y detección proporcionan valiosa información que puede ser utilizada de forma automática por herramientas software. Por otra parte, existen actualmente un gran número de fuentes de datos de metabolómica que proporcionan una correlación entre la señal analítica y la identificación del compuesto.

El objetivo de esta tesis es la creación de una herramienta software que permita la consulta simultánea a las bases de datos metabolómicas más relevantes existentes para ofrecerles a los investigadores la posibilidad de obtener datos de ellas a partir de una única consulta. Esta consulta simultánea va a permitir el acceso a más información tanto en profundidad, puesto que los investigadores podrán acceder a la información complementaria sobre metabolitos contenidos en distintas bases de datos, como en amplitud, pues hay un gran número de compuestos que se encuentran en una única base de datos. Los investigadores que no consultan dicha base de datos están reduciendo el conjunto de metabolitos sobre los que están realizando el proceso de identificación, y aumentando el potencial número de metabolitos sin identificar en sus experimentos.

Además, la herramienta debe explotar la información analítica y no analítica para facilitar la anotación e identificación de metabolitos, ampliar la cobertura de metabolitos en los estudios y reducir el número de identificaciones erróneas que pueden conducir a interpretaciones biológicas erróneas.

La herramienta creada se denomina CEU Mass Mediator (CMM). Actualmente contiene 332,665 compuestos experimentales provenientes de las fuentes de datos HMDB (Human Metabolome Database), KEGG, LipidMaps (LIPID Metabolites and Pathways Strategy), Metlin y una librería propia creada en el Centro de Excelencia de Metabolómica y Bioanálisis (CEMBIO) y 681,198 compuestos generados mediante aproximaciones computacionales (in-silico) provenientes de HMDB y MINE (Metabolic In silico Network Expansion Databases). CMM permite la consulta simultánea a estas bases de datos desde una misma interfaz y en una única consulta a partir de las *m/z* y, opcionalmente, el tiempo de retención (RT) y la agrupación de picos obtenidos mediante el agrupamiento de señales provenientes de un mismo metabolito primario como son los isótopos, los aductos, las moléculas con múltiple carga, los multímeros o los fragmentos, llamado en esta tesis Composite Spectrum (CS).

El sistema experto utilizando información de MS^1

CMM permite aplicar diferentes filtros a los investigadores para mejorar el proceso de filtrado y la eficiencia de las reglas. CMM permite restringir la búsqueda en función de los elementos presentes, con tres posibilidades: CHNOPS, CHNOPS+Cl y todos los elementos, permitiendo incluir o excluir en las búsquedas compuestos que contengan deuterio. CMM permite también restringir la búsqueda a lípidos, e incluir o excluir péptidos en la búsqueda, así como los potenciales aductos formados en modo positivo o negativo.

CMM utiliza esta información por su sistema experto (CMM-ES) basado en 122 reglas para puntuar estas anotaciones basándose en la probabilidad de los tipos de compuestos de formar un determinado aducto (puntuación χ_1), la presencia o ausencia de aductos esperados para un determinado tipo de compuestos y la relación entre estos aductos entre diferentes señales (puntuación χ_2) y el orden de elución según la hidrofobicidad en lípidos pertenecientes a una misma clase, ya que tienen la misma estructura y sólo diferen en la longitud de la cadena de los ácidos grasos y su nivel de saturación, medido en el número de dobles enlaces (puntuación χ_3). Estas tres puntuaciones están integradas en una puntuación general que se calcula según la siguiente media geómetrica:

$$\chi = exp\left(\frac{\sum_{i=1}^{3} \omega_i \cdot ln\chi_i}{\sum_{i=1}^{3} \omega_i}\right)$$

donde ω_i es el peso de cada puntuación, ω_1 = 1, ω_2 = 1 y ω_3 ϵ [0, 2]. $\omega_3$depende del del número de reglas aplicadas para el orden de elución, ya que el número es variable y cuanto mayor número de anotaciones provenientes de otras señales, mayor es la evidencia que proporcionan. CMM tiene en cuenta el modificador utilizado en la fase móvil para la formación de aductos, puesto que la presencia del NH_3 va a modificar los potenciales aductos formados.

Un ejemplo de las reglas de ionización sería la probabilidad de los monoglicéridos (MG) de formar determinado tipo de aductos. Los MG son difícilmente detectados en modo de ionización negativa, y en modo de ionización positiva el aducto $[M+H]^+$ es comúnmente formado, al igual que el $[M+NH_4]^+$ si se utiliza amonio como modificador en la fase móvil. El aducto $[M+Na]^+$ se puede formar, pero siempre con una intensidad menor que el $[M+H]^+$. Las reglas de orden de elución se aplicarían en el caso de dos señales (S_1,S_2) con un RT(RT_{S1}, RT_{S2}) y RT_{S1}>RT_{S2} y dos anotaciones putativas (AP_{S1}, AP_{S2}). Si AP_{S1} se corresponde con un MG(20:0) y AP_{S2} con un MG(22:0), si el análisis se ha realizado mediante fase reversa (RP por sus siglas en inglés, Reversed-Phase), hay una evidencia negativa en estas dos anotaciones, puesto que el compuesto MG(22:0) debería eluir más tarde que el compuesto MG(20:0), y estas anotaciones tendrán una puntuación baja. Sin embargo, si RT_{S1}<RT_{S2}, entonces la evidencia acerca de las dos señales S_1 y S_2 perteneciendo a los metabolitos MG(20:0) y MG(22:0) respectivamente es positiva y la puntuación de ambas anotaciones será incrementada. CMM permite incluir en la búsqueda señales no significativas tras hacer el estudio estadístico entre dos o más grupos de estudio. Estas señales que no tienen una significancia estadística pueden no ser útiles como biomarcadores, pero aportan evidencia para la anotación e identificación de señales de las significativas, que son potencialmente biomarcadores y el objetivo principal del estudio. CMM utiliza la evidencia y no muestra las anotaciones de ellas al usuario.

Una herramienta semi-automática para la identificación de oxPCs

Por otro lado, CMM ofrece un servicio para la identificación de glicerofosfocolinas oxidadas (oxPCs). Los oxPCs están siendo estudiados recientemente como biomarcadores relevantes en los mecanismos de la salud y de la enfermedad. Debido a ello, la identificación de los mismos en los experimentos metabolómicos resulta de especial interés, y la aparición de herramientas que permitan anotarlas y estudiar su función biológica es un gran avance. CMM utiliza la información analítica de experimentos realizados mediante cromatografía líquida, ionización por electrospray y detección por espectrometría de masas (LC-ESI-MS). Integra conocimiento de la fragmentación producida por los oxPCs e incluye compuestos derivados de los lípidos oxidados no presentes en otras bases de datos. Basándose en este conocimiento analítico, CMM compara el espectro experimental introducido por el usuario con el presente en la base de datos y obtenido en el análisis de los estándares y el usuario puede de esta forma identificar si el espectro corresponde a un oxPC.

Una herramienta que utiliza información no analítica

CMM permite la agrupación de compuestos para la posterior interpretación biológica. Una vez el usuario ha filtrado, anotado e identificado su lista de señales, esta puede ser introducida al servicio de análisis de pathways para agruparlas en función de los pathways donde están presentes. CMM ordena estos pathways en base al número de compuestos de cada pathway presentes en el experimento y la relevancia de estos compuestos dentro del pathway. La relevancia de los compuestos identificados dentro del pathway, medida como el número de pathways en los que un compuesto está presente.

Una herramienta de búsqueda con información de MS/MS

CMM ofrece también un servicio de búsqueda con información proveniente de MS/MS para soportar la identificación de metabolitos con un nivel de confianza mayor. Esta búsqueda esta basada en la similitud del espectro experimental y la librería de espectros experimentales e *in-silico* obtenida de la base de datos HMDB.

Además, CMM ofrece una funcionalidad única, como es una interfaz para calcular la calidad de un espectro MS/MS para su posterior identificación. Las condiciones experimentales son clave a la hora de obtener un espectro claro e interpretable que habilite una identificación con un mayor nivel de confianza. Un espectro con gran cantidad de ruido y una intensidad baja no va a permitir distinguir los picos provenientes de un metabolito con la contaminación presente en el espectrómetro, lo que puede llevar a identificaciones erróneas y, en consecuencia, a interpretaciones biológicas equivocadas. CMM puntúa los espectros experimentales en función de la intensidad de la señal en MS^1 y MS/MS, al nivel de ruido presente, al número de escaneos realizados para el análisis de MS/MS y/o el número de muestras utilizadas para obtener el espectro de MS/MS (la correspondencia de señales en diferentes análisis reduce el efecto de las contaminaciones), la presencia de más de un metabolito en la celda de colisión para el análisis de MS/MS y el *crosstalk*, un fenómeno que se produce cuando en la celda de colisión o en el espectrómetro de masas aún hay iones provenientes del anterior análisis.

Una API REST para el acceso a la herramienta

Todos los servicios de CMM se ofrecen tanto a usuarios sin conocimiento informático a través de una página web, como a desarrolladores que quieran utilizar los servicios a través de una interfaz de programación de aplicaciones (API) de transferencia de estado representacional (REST). Esta segunda opción es muy útil para integrar CMM en otras herramientas, para facilitar el uso dentro de *workflows* o para el acceso a través de otras interfaces. Actualmente CMM está integrada en la base de datos metabolómica con mayor número de citas: HMDB. Los usuarios de HMDB pueden realizar consultas a CMM desde su interfaz y explotar las funcionalidades de CMM previamente explicadas, reduciendo la curva de aprendizaje necesaria para utilizar una nueva herramienta. Este servicio está disponible en http://www.hmdb.ca/spectra/ms_cmm/search.

CMM también está accesible a través de un paquete de R disponible en el CRAN (Comprehensive **R** Archive Network). Los usuarios que estén habituados a trabajar con R también pueden utilizar todas las funcionalidades desde sus programas de R. Este paquete está accesible en https://rdrr.io/github/lzyacht/cmmr/.

CMM es una aplicación J2EE (Java 2 Platforms, Enterprise Edition) de código abierto cuyo código está disponible en https://github.com/albertogilf/ceuMassMediator y actualmente está desplegada en un servidor Apache TomEE 7.0.2 y cuya base de datos está alojado en un servidor MySQL Server 5.7.24. La aplicación puede ser accedida desde cualquier navegador en la dirección http://ceumass.eps.uspceu.es/ o a través de los diferentes servicios en su REST API http://ceumass.eps.uspceu.es/mediator/api/v3. CMM actualiza la información de los compuestos de las bases de datos integradas aproximadamente cada 6 meses.

Organización del documento

En el primer capítulo se ha realizado una revisión de los recursos y fuentes de datos disponibles para la identificación de metabolitos utilizando electrospray como fuente de ionización. La información contenida en estos recursos es en muchas ocasiones complementaria y el nivel de solapamiento de metabolitos presentes en las bases de datos se puede calificar como bajo, por lo que los investigadores deben consultar diferentes recursos para ampliar la cobertura de metabolitos en los estudios metabolómicos. El segundo capítulo presenta la primera versión de CMM. En él se desarrolla una aproximación heurística para la anotación de metabolitos a partir de información proveniente de MS^1 y del tiempo de retención obtenido en la separación previa, ya sea mediante cromatografía líquida o electroforesis capilar. El capítulo tercero supone un paso adelante en la estrategia, ya que integra conocimiento propio del CEMBIO, no solo obtenido de fuentes de datos externas. En él se describe la obtención de conocimiento analítico acerca de glicerofosfocolinas oxidadas y la creación de un método semi automático para su detección e identificación utilizando el tiempo de retención y la información de MS^1 y MS^2. El capítulo cuarto describe las actualizaciones llevadas a cabo en CMM. Se han incorporado nuevos servicios progresivamente para dar soporte a la identificación de metabolitos tales como un medidor de calidad del espectro para información proveniente de MS^2, la incorporación de información referente a ontología y taxonomía, y el soporte de identificación a partir de información proveniente de MS^2. Todos los servicios presentes en CMM y desarrollados durante esta tesis están disponibles a través de una API REST para facilitar el acceso automático y la comunicación con otras herramientas.

Los metabolitos son los productos y los responsables de la situación final de un sistema biológico. La correcta y completa identificación de los metabolitos va a resultar en una mayor información para la interpretación biológica. En consecuencia, se remarca la necesidad de combinar información analítica y no analítica para obtener un nivel de confianza mayor en la identificación de metabolitos, así como la utilidad de proporcionar herramientas de software a los investigadores para facilitarles el éxito en sus experimentos.

ABSTRACT

Metabolomics is a subarea of the systems biology devoted to the study of the small size molecules (usually < 1,000 Da) produced by the metabolic processes happening in a cell. Since the end of the previous century untargeted metabolomics has been successfully applied to different domains such as biomarker discovery, therapeutical targets discovery, personalized medicine or providing knowledge about organisms and mechanisms of health and disease. Untargeted metabolomics, by nature, aims to obtain as much information as possible to maximize the number of detected and identified metabolites, being the metabolite identification vital in the final success of the studies.

The number of extracted metabolites and subsequently identified with certain confidence level can be defined as “metabolite coverage”. The identification is the main bottleneck in metabolomic studies since the analytical information acquired requires a high amount of work and knowledge to be successfully exploited. On the one hand, separation and detection provide a valuable information that can be exploited in an automatic way by software tools. On the other hand, currently there are a large number of metabolomic data sources containing information about the metabolites they store. Both information coming from the analyses and the data sources can be used to provide a higher confidence level in the metabolite identification.

The final goal of this thesis is the design, validation and implementation of a software tool that allows the simultaneous query over different metabolomic databases to offer the researchers the possibility of retrieving data from them in a single step. This simultaneous query will allow the access to more data both in depth, since they will be able to access the complementary information stored in distinct databases about metabolites contained in more than one database, and width, since there are a high number of compounds only present in a single database, with the consequent risk for the researchers of skipping metabolites during the

annotation and identification process, thus potentially increasing the number of unknows in the experiment.

Furthermore, the tool should exploit the analytical and non-analytical information to aid during the metabolite annotation and identification, therefore increasing the metabolite coverage in the metabolomic studies and reducing the number of misidentifications that lead to potential wrong biological interpretations.

The first chapter reviews the available resources and data sources for the metabolite identification using Electrospray as ionization technique. The information contained in those resources is often complementary and the metabolite overlap is low. Therefore, the researchers should query different resources to boost the metabolite coverage in their studies. The second chapter introduces the first version of the software tool performed in this thesis: CEU Mass Mediator (CMM). The tool develops a heuristic approach for metabolite annotation from information coming from MS^1 and the RT or MT obtained in the chromatographic or electrophoretic separation. The third chapter presents the acquisition of analytical knowledge from oxidized glycerophosphocholines and the creation of a semi-automated approach for their detection and identification using the RT and information obtained in MS^1 and MS^2 analysis. The fourth chapter describes the updates performed in CMM. New services have been gradually incorporated such as a spectral quality assessment, the incorporation of ontology and taxonomy information, and the support of MS^2 searches. All the services present in CMM are available through a REST API to facilitate the automatic access and the communication with other software tools.

The metabolites are the end products and the responsible of the biological systems status. The correctness and completeness of metabolite identification result in a higher amount of information for the subsequent biological interpretation. Consequently, we remark the necessity of combining analytical and non-analytical information to obtain and provide a higher confidence level in the metabolite identification, as well as the utility

of the software tools in helping researchers to successfully conduct their experiments.

INTRODUCTION

1.1 Introduction to Metabolomics

Systems biology is an integrative discipline that requires the contribution of different fields such as chemistry, biology, computer science, physics or mathematics to unravel the insights of the complex living organisms by integrating quantitative assessments with mathematical models.[1] The state of an organism is a dynamic and constantly evolving phenomenon resulting from multi-interactions between internal and external factors.[2] The internal factors are defined as the levels of an organism function including genes, transcripts, proteins and metabolites. Changes in this multivariate homeostasis can lead to disorders or diseases. These multi-interactions between different factors result in a phenotype response that should be investigated holistically, considering the relationships between different molecules.[3] Although systems biology pursues a holistic approach, it starts by reducing the organism into sub-components in order to understand their structure and functions and then, the behaviour and interactions between components can be studied. To achieve a general and deep understanding of the full biological system, all its sub-components should be studied. In the -omics field, these sub-components can be summarized in genomics, transcriptomics, proteomics and metabolomics,[4] and the integration of all these fields yields a full picture about the biological system that is known as multi-omics. The multi-omics approach provides a more holistic molecular perspective compared to the traditional approaches.[5]

Metabolomics is the last -omic science in the -omics cascade. It studies the intermediate and end-products of the metabolism, allowing scientists to observe and track subtle changes in the organism.[6,7] Thus, it is considered as the omic that best reflects the phenotype response,[8-10] generating a high volume of information about the organism and being currently one of the fastest growing research areas.

Metabolomics started to be treated as an independent area in the 90's. In that moment, Jeremy Nicholson et al. defined the field as "*a* *measurement of the dynamic multiparametric metabolic response of living system to pathophysiological stimuli or genetic modification*".[11] Olivier Fiehn defined metabolomics in 2002 as "*a comprehensive and quantitative analysis of all metabolites in a system*".[8] Metabolomics has been used in different applications such as discovery of biomarkers[12-15], providing knowledge about mechanisms of disease[16-18], discovery of therapeutical targets[19-21] or personalized medicine.[22-24]

There are two approaches to integrate the -omics sciences. On the one hand, a top-down data reduction strategy based on the genes and transcripts to predict the phenotypic changes, which is achieved using targeted proteomic and metabolomic analyses. On the other hand, a bottom up data reduction strategy, using targeted or untargeted metabolomics as starting point to guide the other -omics sciences (see Figure 1).[5]

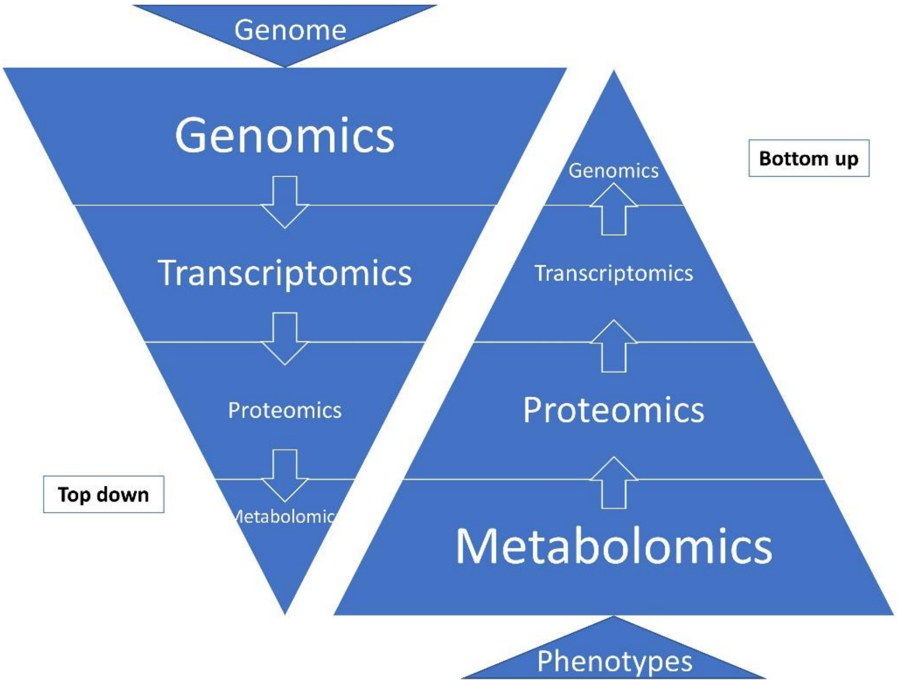


Figure 1 Data reduction integration approaches in -omics sciences: top down and bottom up.

Targeted analyses differ from untargeted analyses in the existence of a prior hypothesis which enables researchers to create a list of metabolites of interest that need to be quantified. Therefore, the number of metabolites to be measured is higher in untargeted approaches than in targeted ones. Although the low coverage of the metabolome compared to the genome, transcriptome and proteome may limit and difficult the biological interpretation of the final results, untargeted analyses (bottom up approaches) enable the discovery of unexpected changes, and often in earlier states, than proteomics, transcriptomics or genomics analyses.

Untargeted approaches are especially interesting in cases where targeted approaches are not successful, i.e., if the experimental hypothesis turned out not to be right. These situations are not uncommon since living organisms have a high complexity that hinders the control of all the variables involved. The metabolites present in an organism include the endogenous molecules; the xenobiome, consisting of compounds derived from sources outside the organism;[25,26] the nutribiome, resulting from food-derived xenobiotics,[27,28] and the gut microbiome, formed by the molecules produced by bacteria living in the organism.[29,30] To maximize the information obtained from the sample in untargeted approaches, the experiment should be performed optimizing the metabolite coverage. This term can be defined as (1) the number of metabolites present in a sample, (2) the number of metabolites separated and detected by analytical methods, or (3) the number of identified metabolites. In this dissertation the metabolite coverage will refer to the number of metabolites identified with a confidence level 3 or higher (see Table 1).

The impressive success of genomics and proteomics is not easy to repeat in the field of metabolomics because of the problematic nature of metabolites themselves: the enormous physiochemical diversity, the broad ranges of concentrations, and the large and yet undetermined size of the metabolome. Combined, these issues constitute the source of many challenges along the metabolomic workflow, being particularly important the undetermined size of the metabolome, which refers to the complete set of metabolites present in an organism.

Confidence level	*Description*	*Matching requirement*
Level 0	Unequivocal 3D structure, including full stereochemistry.	Determination of 3D structure following natural product guidelines.
Level 1	Confident 2D structure, using reference standard or full 2D structure elucidation.	At least two orthogonal characteristics, such as MS/MS fragmentation pattern, RT or CCS.
Level 2	Probable structure using literature data and/or fragmentation spectra and/or knowledge over the RT.	At least two orthogonal characteristics matching and evidences for excluding the rest of candidates.
Level 3	Possible structure, isomers or class.	More than one candidate, only one characteristic matched is required for supporting the proposed candidate.
Level 4	Unknown.	Detectable feature in a sample.

RT: retention time, CSS: collision cross section

Table 1 Updated confidence levels proposed by the Metabolomics Society (2017).

1.2 The Metabolomic workflow

The metabolomic workflow starts with a biological question that requires a cascade of sequential stages to be answered, including experimental design, sample preparation, data acquisition, data preprocessing and statistical analysis, identification and biological interpretation, and a final hypothesis generation. Biological samples are treated to extract their “crude” metabolite content, removing interferences and nonrelevant elements such as genes, proteins, or salts. The measurement of the extracted metabolites can be performed by two different analytical techniques: Nuclear Magentic Resonance (NMR) Mass Spectrometry (MS) approaches. NMR provide information about the spectra to be subsequentally interpreted. The MS approaches are carried out via direct analysis (shotgun metabolomics)[31] or by separation prior to the analysis, which uses either liquid, gas or ion chromatographic (LC, GC or IC respectively) or capillar electrophoretic (CE) principles. Detection is performed using MS methods that provide information about chemical shift or mass to charge (*m/z*) ratio, respectively (see Figure 2 illustrating the MS metabolomic workflow). From a historical point of view, the use of NMR has a long tradition; however, due to its superb sensitivity and high resolution,[32] MS has more recently dominated the metabolomic field, especially when the amount of material for the experiment and the funding is limited.[33] The data obtained is then preprocessed and analyzed to ideally identify the compounds present in the sample and proceed with the biological interpretation.

1.2.1 Data acquisition

The success of metabolomics depends on the metabolite coverage. A vital stage for broadening the metabolite coverage is the data acquisition, which depends on the capabilities of the equipment used. The more powerful separation techniques combined with more sophisticated analyzers and more sensitive detectors increase the quality and the quantity of the data obtained.[34,35] Considering the size and the diversity of the metabolome, the samples should be analyzed using different experimental techniques to increase the metabolite coverage, since a single experimental technique cannot separate and detect all the metabolites present in the sample.[36] Multi-platform analyses increase the metabolite coverage because the separation techniques and the different solvents are focused on separating different types of molecules. For example, the ionic and polar compounds are well separated using CE or IC, the volatile compounds can be well detected using GC, while LC provides separation for the broadest range of metabolites depending on the mobile phases, modifiers, columns and parameters used.

For the data acquisition, MS requires the ionization of molecules prior to their measurement using ionization techniques such as electron ionization (EI) or electrospray (ESI), among others. Depending on the type of mass analyzer used, spectrometric measurements can provide either nominal (quadrupole (Q), triple quadrupole (QQQ), ion trap (IT)) or accurate (time of flight (TOF), QTOF, OrbiTrap) monoisotopic mass. However, in untargeted studies, only high accuracy mass spectrometers are used since

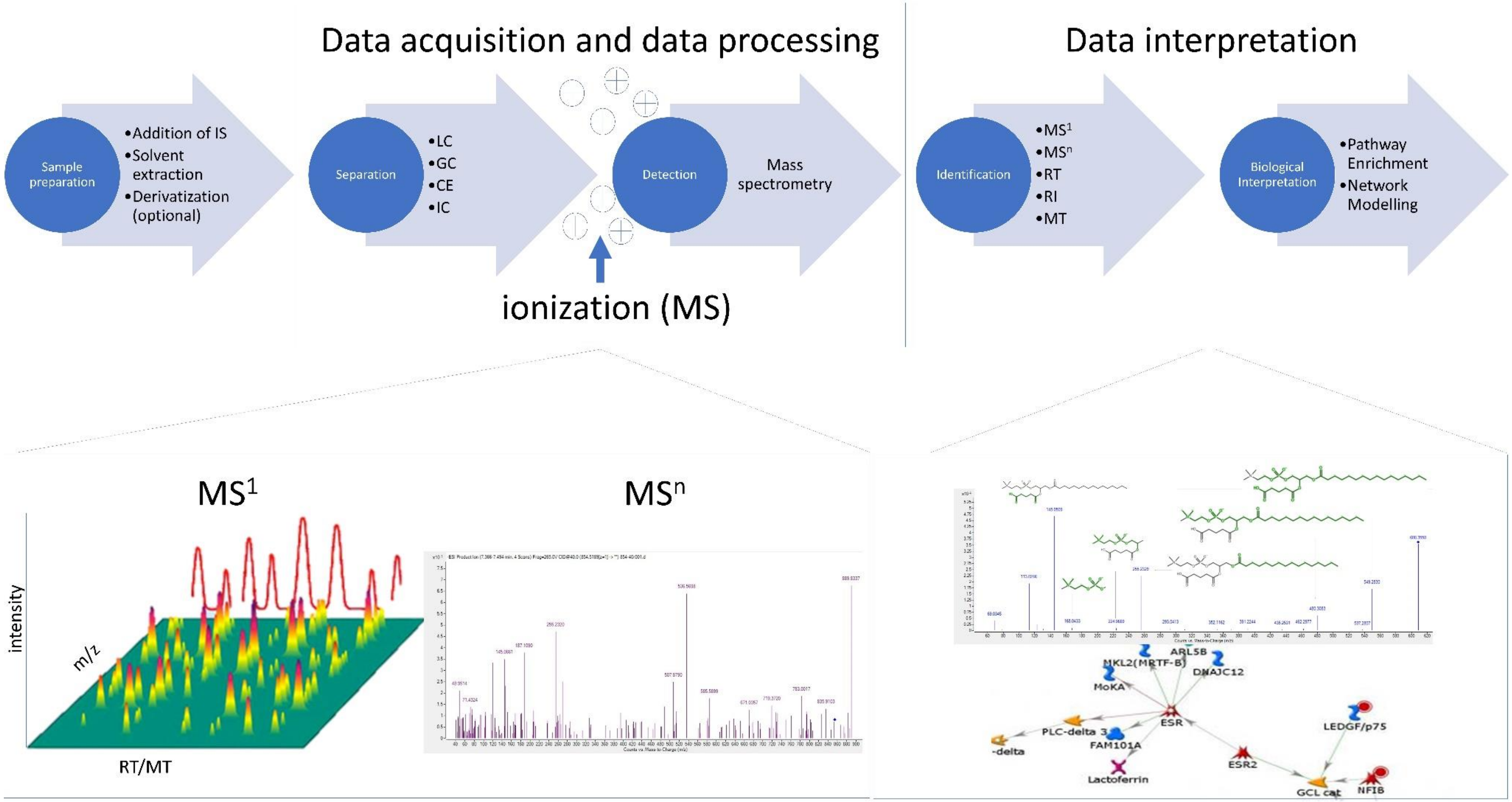


Figure 2 A general overview of the metabolomic workflow using MS.

the nominal mass hinders considerably the identification of the metabolites corresponding to the signals acquired.

For the data acquisition, MS requires the ionization of molecules prior to their measurement using ionization techniques such as electron ionization (EI) or electrospray (ESI), among others. Depending on the type of mass analyzer used, spectrometric measurements can provide either nominal (quadrupole (Q), triple quadrupole (QQQ), ion trap (IT)) or accurate (time of flight (TOF), QTOF, OrbiTrap) monoisotopic mass. However, in untargeted studies, only high accuracy mass spectrometers are used since the nominal mass hinders considerably the identification of the metabolites corresponding to the signals acquired.

1.2.2 Data preprocessing and statistical analysis

The information obtained must be converted from spectra and chromatograms/electropherograms to a three-dimensional matrix consisting of mass (*m/z* or monoisotopic mass), chromatographic time (retention time -RT- for LC and GC, migration time -MT- for CE), and intensity or abundance. Each peak in the three-dimensional matrix is called feature. This matrix is often subjected to statistical analysis to compare the metabolite content between different conditions, e.g., control and case. This leads to the selection of compounds causing the observed phenotypic changes. These initially anonymous signals are then assigned to actual metabolites in order to allocate them to the corresponding metabolic pathways.

There are a number of software tools for the data preprocessing[37,38] using different algorithms but they all share a similar purpose.[2,39-44] All of them translate the raw data into features corresponding to the previously explained three-dimensional matrix. Some of them process the co-eluting signals to group them based on different transformations suffered by the primal metabolite (adducts, charges, neutral loses), adding a fourth dimension to the matrix.

The data preprocessing includes the filtration considering the analytical and biological aspects of the data. It aims to reduce the matrix complexity by removing unreliable or previously known non-related signals.[45,46] Furthermore, it is recommended to perform a data normalization consisting in the application of several operations to reduce the analytical or biological variation arising from drifts in the sample analyses. These drifts can be successfully controlled using internal standards (IS).

Ideally, during this process the monoisotopic mass is calculated based on the ionization products (IPs) such as isotopes (C^{12}, C^{13}, H^1, H^2, etc.), adducts ($[M+H]^+$, $[M+NH_4]^+$, $[M-H]^-$, $[M+Cl]^-$, etc.), multimers ($[2M+H]^+$, $[2M+Na]^+$, $[2M-H]^-$, etc.), multiple charge adducts ($[M+2H]^{2+}$, $[M+H+Na]^{2+}$, $[M-2H]^{2-}$, etc.) and/or neutral loss fragments ($[M+H-H_2O]^+$, $[M-H-H_2O]^-$, etc.). These signals can be grouped into a single pseudospectrum to calculate the monoisotopic mass. The relationship between different IPs can be established based on the peak shape criteria and the correlation analysis.[47] However, this grouping carried out by the software tools sometimes fails, or it is not possible to detect due to analytical conditions.[48] This data preprocessing is very important since it will reduce the chance of obtaining false positive annotations. The impact of data preprocessing and ion annotation in the identification process can be very large, since it has been reported that up to 90% of high-quality signals detected using LC/MS might correspond to contaminants, artefacts, or IPs.[49]

When the metabolomic study is devoted to the comparison between two different groups, e.g. case and control, a statistical analysis to reveal significant changes between these groups is performed. Some examples of statistical test performed are univariate (Student’s *t*-test, Mann Whitney *U*-test or ANOVA) or multivariate analysis (Principal Component Analysis -PCA-, Discriminant Function Analysis -DFA-). The statistical analysis permits researchers to devote their efforts to the analysis of the metabolites with an intensity significantly different between the groups, since they are prone to being biomarkers or indicatives of a difference to be studied for understanding these significant differences.

1.2.3 Metabolite identification and confidence levels

Once the metabolomic study has unveiled a list of features, they must be identified to provide a biological meaning. Metabolite identification plays a vital role within the metabolomic workflow. This stage aims to find an identifier (structure, database id, name) for each feature since the features by themselves cannot be analyzed under a biological point of view. Only if researchers truly identify them, the biological meaning can be elucidated. Therefore, the final outcome of the metabolomic experiments strongly depends on the identification process.[48]

The metabolomic community agrees that the identification is essential to convert analytical data into meaningful biological knowledge.[50] However, metabolite identification is one of the most challenging stages in metabolomics, being often the main bottleneck in the entire workflow.[51] A low identification rate hinders biological interpretation of the experiment due to many pieces of the puzzle being missing. A significant misidentification rate could lead to inconsistent analysis of the results and even to wrong biological interpretations.[52]

Regardless the technique employed to separate our samples prior to the mass spectrometer (GC, LC or CE), there are several classifications of confidence levels for the identification of metabolites.[33] The most popular is the one proposed by the Metabolomics Standards Initiative (MSI), a consortium formed in 2005 to provide the metabolomic community with a set of standards and protocols to improve the quality of the metabolomic studies.[53] It includes five different confidence levels (see Table 1) for the identification of each compound derived from metabolomic experiments.[54,55]

A number of tools exist to annotate the metabolites with different confidence levels.[48,55] Commonly, the first step in metabolite identification is the assignment of a unique or a set of putative metabolite candidates to the *m/z* values obtained by MS^1 (see Table 1, confidence level 3). The confidence level 3 will be referred to as **annotation** from now on in this dissertation. Then, further analysis using Multi-Stage Mass Spectrometry (MS^n) or exploiting information from other (orthogonal) sources, such as collision cross section (CCS) ion mobility, RT from GC or LC analysis or MT for CE data, is carried out.[56] This additional information enables researchers to apply their chemistry knowledge to support or refute the identifications, and to possibly achieve confidence level 2 (see Table 1 and Figure 3 A). Confidence level 2 or higher will be referred to as **identification** from now on in this dissertation. The highest confident levels (level 0 and 1) require the analysis of an authentic standard to compare its properties obtained under identical analytical conditions to those of the identified feature. Achieving these confidence levels is hindered by the availability of authentic standards and the funding.[2]

Different and/or complementary confidence levels than those of the Metabolomics Standard Initiative have been proposed. Schrimpe-Rutledge et al. proposed a framework (see Figure 3 B) that split the confidence levels based on the knowledge about molecular formula (level 4), tentative structure based on MS^1 database match (level 3), putative identification using fragmentation patterns and orthogonal information such as RT or CCS (level 2) and the validated identification using authentic standards (level 1).[57] Although nowadays the spectrometers are extremely accurate, for the majority of the features acquired there is no possibility yet to calculate a unique molecular formula based on the "*seven golden rules*",[58] Lewis or Senior chemical rules, and hydrogen/carbon ratio or elemental ratio probabilities.[59] Therefore the level 4 is not achieved for all the features obtained in a metabolomic experiment.

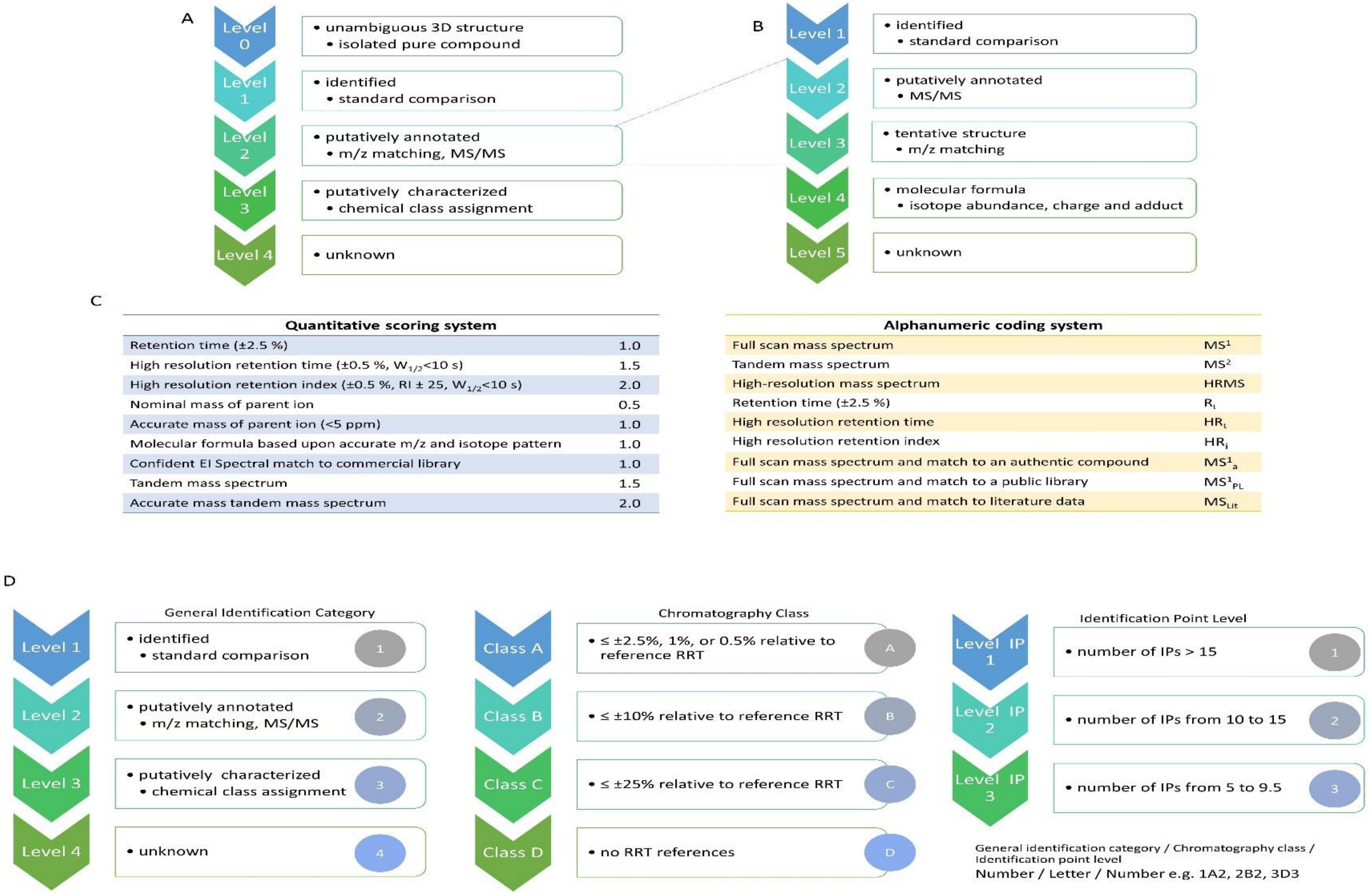
Sumner et al. proposed a system with a quantitative scoring and an alphanumeric coding system to gauge the confidence of our peers (see Figure 3 C). This solution expands the reported confidence levels by including more detailed information about how the researchers can achieve a particular level of confidence depending on the analytical technique, the mass accuracy, the resolution, the elution time or the MS^n analysis.[60]

The most recent proposal includes a confidence scale and an ID score.[61] The confidence level assigned to a metabolite identification includes a number, a letter and another number. The first number corresponds to one of the four main categories: identified using authentic standards (level 1), putatively annotated using m/z matching and MS/MS fragmentation (level 2), putatively characterized by its IPs to a chemical class assignment (level 3) and unknowns (level 4). The letter indicates the chromatographic characteristics based on relative retention time (RRT), and the last number reflects the number of IPs encountered between the feature of interest and the authentic standard or information from the database (see Figure 3 D).

1.3 Current challenges in metabolite identification

There are major differences for the metabolite identification workflow depending on the analytical technique employed. GC/MS is usually equipped with EI as ionization source. EI provokes a high and reproducible fragmentation of the molecules, and the Kovats retention indices (RIs) are easily calculated once the retention times have been obtained.[62] Furthermore, GC/MS often measures derivatized forms (analytes) instead of primary metabolites. Therefore, there are well and known established methods for the GC/MS metabolite identification.[63-65] GC/MS databases contain information about monoisotopic mass, fragmentation patterns and Kovats RI. Some of the most popular GC/MS databases are NIST, Wiley, the Fiehn Library, Mass bank or the MassBank of North America (MoNA).

However, CE and LC separation techniques are less reproducible, which yields a substantively different workflow for metabolite identification. The assignment of *m/z* values to a set of annotations is performed by querying accessible databases, which ideally makes this process very accurate and efficient (confidence level 3).[66]



Quantitative scoring system	
Retention time (±2.5 %)	1.0
High resolution retention time (±0.5 %, $W_{1/2}$<10 s)	1.5
High resolution retention index (±0.5 %, RI ± 25, $W_{1/2}$<10 s)	2.0
Nominal mass of parent ion	0.5
Accurate mass of parent ion (<5 ppm)	1.0
Molecular formula based upon accurate m/z and isotope pattern	1.0
Confident EI Spectral match to commercial library	1.0
Tandem mass spectrum	1.5
Accurate mass tandem mass spectrum	2.0

Alphanumeric coding system	
Full scan mass spectrum	MS^1
Tandem mass spectrum	MS^2
High-resolution mass spectrum	HRMS
Retention time (±2.5 %)	R_t
High resolution retention time	HR_t
High resolution retention index	HR_i
Full scan mass spectrum and match to an authentic compound	MS^1_a
Full scan mass spectrum and match to a public library	MS^1_{PL}
Full scan mass spectrum and match to literature data	MS_{Lit}

D
General Identification Category
Level 1
• identified
• standard comparison
1
Level 2
• putatively annotated
• m/z matching, MS/MS
2
Level 3
• putatively characterized
• chemical class assignment
3
Level 4
• unknown
4
Chromatography Class
Class A
• ≤ ±2.5%, 1%, or 0.5% relative to reference RRT
A
Class B
• ≤ ±10% relative to reference RRT
B
Class C
• ≤ ±25% relative to reference RRT
C
Class D
• no RRT references
D
Identification Point Level
Level IP 1
• number of IPs > 15
1
Level IP 2
• number of IPs from 10 to 15
2
Level IP 3
• number of IPs from 5 to 9.5
3
General identification category / Chromatography class / Identification point level
Number / Letter / Number e.g. 1A2, 2B2, 3D3

Figure 3 Confidence level systems in metabolite identification.

Currently, there are a considerable number of databases either exclusively devoted to metabolomics or easily applicable to metabolomic data. However, the vast amount of data repositories and the low overlap requires also manual querying and integration of the results from different sources.[67] Data sources can be specific for certain types of compounds; this is the case for the Human Metabolome Database (HMDB),[68] which covers the human metabolome, the (LIPID Metabolites and Pathways Strategy, (LipidMaps)[69] and LipidBank[70], which contain only lipids, the Universal Natural Product Database (UNPD) devoted to primary and secondary plant metabolites,[71] or the Milk Composition Database (MCDB) made up by compounds present in milk. Other sources, such as Metlin,[72] KEGG,[73] MassBank[74] or mzCloud, contain all kind of compounds. To overcome the lack of experimentally detected compounds forming the metabolome, some databases have incorporated *in-silico* generated compounds, like MyCompoundId,[75] HMDB or the Metabolic In Silico Network Expansion Databases (MINEs),[76] using general biotransformations of previously detected compounds.[77] The current databases differ in size, search modes, available adducts to search or mass searching tolerance. That has caused the emergence of a number of software tools offering a common interface to query multiple databases, with some of them providing additional processing features unavailable in the original databases.[78,79]

The long list of software tools with different and complementary purposes illustrates the importance of Workflows to integrate them, allowing the researchers to use a common interface and consequently, saving time and incrementing the visibility of the tools integrated there. They also allow researchers to save and share their data, increasing the reproducibility and the collaboration between different laboratories. Some examples of integrative workflows are Taverna,[80] KNIME,[81] Workflow4Metabolomics[82] or GNPS.[83] The decoupling between the software tools and the platforms where they were developed eases their integration.

1.3.1 Use of fragmentation obtained by MS^n

Despite all the available information, further work is usually needed to accept or reject the annotations retrieved from the databases. The chromatographic data and fragmentation pattern are the two characteristics most commonly used to possibly achieve confidence level 2. Mass spectrometers allow the researches to isolate the compounds within a specified elution time and m/z range in a collision cell for a further analysis, usually applying a voltage to acquire the fragmentation pattern. Then, a comparison between the fragmentation pattern obtained with a reference spectrum can be calculated, either from a database or to the one obtained from an authentic standard (if available), as well as structural elucidation to provide meaningful to each IP. The most common method is to compare them against MS/MS libraries, that provide different methods for the fragmentation matching: e.g. “peak counting”, that counts the number of matching peaks, or the dot product, that processes a two-dimensional comparison based on the *m/z* and intensities. This method is widely applied, with an arithmetic or geometric mean to calculate the final matching score.

This strategy can be applied a number of times coupling the output from the first collision cell to another collision cell where a new voltage can be applied to a specific IP or to all IPs produced in the first collision cell (MS^n analysis, where n > 2, also called tandem MS), obtaining a fragmentation tree that might be useful to distinguish between compounds with similar structures and fragmentation patterns. In any case, the co-elution of compounds when isolating them in the collision cell hampers the identification, since the spectrometer will acquire the fragmentation of different compounds and it is difficult to distinguish which ones correspond to each precursor ion. To overcome the lack of available standards, and therefore experimentally spectra of compounds in the existing databases, there are a number of tools that predict the fragmentation pattern based on the compound structure. There are substantial differences between the approaches that the tools use to predict the spectra: heuristic approaches,[84-86] machine learning,[87,88] quantum chemistry[89,90] or combinatorial approaches[91-94]. Despite the added value that *in-silico* prediction tools for molecules fragmentation provide, they are not as precise as the experimental ones. Most of them have a high recall but low precision i.e., they produce more IPs than those actually observed experimentally.

A second problem that can arise in the future regarding *in-silico* spectra is the generation of millions of highly similar structures, which could lead to the generation of millions of highly similar spectra, hindering the unequivocal identification of compounds through the fragmentation pattern.[54] Although the MS/MS and MS^n provides structural information and often it is enough to reach confidence level 2 for the putative annotations, sometimes there is not enough evidence to determine a unique structure and therefore achieve identification. If there is not enough evidence for the identification of an unequivocal compound, the presence or absence of particular chemical groups provides valuable insight into the membership of a molecule within a specific chemical class, providing a higher confidence level than those features only annotated through *m/z* match, but not fully achieving the level 2. Then, research has to focus on orthogonal characteristics such as ion mobility, CCS or RT.[55]

1.3.2 Use of analytical information

Chromatography and CE offer additional information about the metabolite structure through the RT and the MT, respectively. They can be used as orthogonal filters during the metabolite identification. However, the number of databases containing RT and/or MT information is certainly low. The high number of different separation columns and possible combinations of solvent buffers and chromatographic conditions to separate the metabolites makes the RT highly variable. Therefore, the RT value has a very low reproducibility between different laboratories. Furthermore, minor changes in analytical conditions can alter it. There are models to predict RT, but they are restricted to very specific conditions.[95-98] The lack of large and diverse training sets difficults the generation of robust retention prediction models for metabolite identification since the success of the RT modelling depends on it. As consequence, the RT from LC can be used to reject false positive identifications rather than to confirm the true positive ones when there are not authentic standards available to reach confidence level 1 in the identification.

In GC, retention is a function of the boiling point of the molecule and its interactions with the column film. RT can be predicted using Quantitative Structure Retention Relationships (QSRR) considering the overall structural properties or additive retention contribution of individual chemical substructures. In practice, instead of absolute values of RT, they are converted into system-independent constants using Kovats RI for isothermal conditions, linear RI for ramped temperatures,[99] and Lee RI.[100] RI changes with the column and temperature program, but they can be easily converted using known software tools like iMatch2.[101] The databases for GC/MS (NIST, Wiley, the Fiehn Library, MoNA or MINE) contain the Kovats RI and the researchers can perform the identification using orthogonal information based on *m/z* and Kovats RI.

CE uses electrophoretic principles to separate molecules, therefore MT is used instead of RT. The MT represents the time that a molecule spends in the migration from the sample inlet to the detector. The reproducibility in CE is lower than in GC and LC, but knowing the exact analytical conditions, there are software tools to predict the MT of a particular molecule from the structure of the cations[102] or to calculate the effective mobility of the molecules knowing the MT and the analytical conditions.[103] Specific databases can be created and used for the metabolite identification based on the collected MT or the relative MT (RMT) regarding a background electrolyte, usually methionine sulfone or paracetamol. The number of software tools for metabolite identification in CE is still low due to the low number of databases containing experimental information obtained through CE and the lower number of users applying this technique.

1.3.3 Use of non-analytical information

The non-analytical information can be also used to support or refute the putative annotations. Applying the same principle as for the use of chromatographic and electrophoretic information, it is easier to discard some identifications rather than to confirm them between different isomers with similar pathway information. For example, if the matrix obtained was acquired from a human sample, there is a high evidence that plant metabolites are not likely to be a true positive annotation. Analogously, knowing the connectivity and dependency between metabolites present in a pathway can be used as an additional source of positive or negative evidence to support or reject a putative annotation. Some software tools use already this information from pathways.[104-106]

1.4 Research objectives

The final objective of this research is the creation of a software tool to help the metabolomic community to overcome the main challenges previously explained. It is important to highlight that each analytical technique faces different challenges regarding the identification of compounds. Metabolite annotation using GC-EI-MS is relatively well established, but the metabolite coverage of this technique is the lowest compared to LC-ESI-MS and CE-ESI-MS. CE-ESI-MS has a higher metabolite coverage than GC-EI-MS, but the identification process may be difficult due to the low number of databases containing experimental information obtained by this technique and the MT shifts. Also, in CE-ESI-MS there is a limited capability of using tandem MS to obtain fragmentation patterns. LC-ESI-MS has the highest coverage due to the high number of experimental setups available. It is the technique where metabolite identification is most challenging. The high number of experimental conditions makes the RT very variable, and there are not general prediction methods yet. Hence, most of the times the annotation starts with the single information of the *m/z* obtained through MS analysis and identification is performed with a high amount of manual work applying analytical and biological knowledge. Given this variability among analytical techniques, the tool must be versatile to accomplish the needs of the different experimental techniques used.

Although metabolite identification will never have perfect precision and recall, it is important that systematic solutions arise to help researchers during this process, without undermining the researchers' experience during the final metabolite identification. The more information is exploited, the more precise the identification will be, and a higher confidence level will be reached. Moreover, the greater number of software tools are available to support exploiting this information, the higher metabolite coverage will be achieved, and a better standardization of the annotation process will be possible, while the misidentifications will be reduced. This thesis was born from the hypothesis that the metabolomic applications can be improved in terms of exploiting analytical and non-analytical information, such as chromatography, IPs patterns and biological knowledge, and providing a proper explanation to the putative annotations. The chromatographic information is directly related to the polarity of the compounds; therefore, the RT and MT provide vital information for identification. The structure of the molecules determines the possible IPs formed in the ionization source, while the nature of the organism studied offers insights about the compounds there analyzed. The main goal of the tool developed in this thesis, CMM, is to exploit as much as possible the analytical and nonanalytical information available to support the researcher in the metabolite identification process.

References

[1] Breitling, R., Frontiers in Physiology, 2010, 1:9

[2] Brown, M.; Wedge, D.C.; Goodacre, R.; Kell, D.B.; Baker, P.N.; Kenny, L.C.; Mamas, M.A.; Neyses, L. and Dunn, W.B., Bioinformatics, 2011, 27(8), 1108-1112

[3] Kell, D.B. and Oliver, S.G., Bioessays, 2004, 26. 99-105

[4] Dettmer, K. and Hammock, B.D., Environ. Health Perspect, 2004, 112(7), A396-A397

[5] Pinu, F.R.; Beale, D.J.; Paten, A.M.; Kouremenos, K.; Swarup, S.; Schirra, H.J. and Wishart, D.S., Metabolites, 2019, 4, 76

[6] Lewis, G.D.; Asnani, A. and Gerszten, R.E.J, Am. Coll. Cardiol., 2008, 52. 117-123

[7] Griffin, J.L.; Atherton, H. and Shockcor, J. and Atzori, L., Nat. Rev. Cardiol., 2011, 11, 630-644

[8] Fiehn, O. Plant Mol. Biol., 2002, 48(1-2), 155-171

[9] Nielsen, J. and Oliver, S., Trends Biotechnol., 2005, 23, 544-546

[10] Gieger, C.; Geistlinger, L.; Altmaier, E.; Hrabé de Angelis, M.; Kronenberg, F.; Meitinger, T.; Mewes, H.W.; Wichmann, H.E.; Weinberger, K.M.; Adamski, J.; Illig, T. and Suhre, K. PLoS Genet 2008, 4(11), 1-12

[11] Nicholson, J.K.; Lindon, J.C. and Holmes, E., Xenobiotica, 1999, 29 (11), 1181-1189

[12] Kell, D.B., Expert Rev. Mol. Diagn, 2007, 4, 329-333

[13] Kenny, L.C.; Broadhurst, D.I.; Dunn, W.; Brown, M.; North, R.A.; McCowan, L.; Roberts, C.; Cooper, G.J.S.; Kell, D.B. and Baker, P.N., Hypertension, 2010, 56(4), 741-749

[14] Johnson, C.H.; Ivanisevic, J. and Siuzdak, G., Nat. Rev. Mol. Cell Biol., 2016, 17(7), 451-460

[15] Liu, W.; Liu, Y.; Yang, Y.; Ou, W.; Chen, X.; Huang, B.; Wang, H. and Liu, M., J. Nutr. Health Aging, 2018, 22(10), 1189-1197

[16] Schauer, N. and Fernie, A.R., Trends Plant Sci., 2006, 11(10), 508-516

[17] Zhao, S.; Wang, M.; He, D.; Liu, L.; Shu, Y.; Li, H.; Liu, Y.; Liu, Z.; Song, Z. and Lu, A., Phytomedicine, 2018, 50, 61-72

[18] Zhang, M.; Liu, Y.; Liu, B.; Li, N.; Dong, X.; Hong, Z.; Chai, Y. and Liu, M., Metabolomics, 2019, 15(2):13

[19] Kumar, B.; Prakash, A.; Ruhela, R.K. and Medhi, B., Pharmacological Reports, 2014, 66(6), 956-963

[20] Sun, H.; Zhang, A.; Liu,S.; Qiu, S.; Li,X.; Zhang, T.; Liu, L. and Wang, X., J. Chromatogr. B Analyt. Technol. Biomed Life Sci., 2018, 1102, 143-151

[21] Douillet, D.C.; Pinson, B.; Ceschin, J.; Hürlimann, H.C.; Saint-Marc, C.; Laporte, D.; Daignan-Fornier, B.; Claverol, S.; Bonneu, M. and Konrad, M., J.Biol.Chem., 2019, 294(3), 805-815

[22] Van-der-Greef, J.; Hankemeier, T. and McBurney, R.N., PHARMACOGENOMICS, 2006, 7(7), 1087-1094

[23] Li, B.; He, X.; Jia, W. and Li, H., Molecules, 2017, 22(7), 1173

[24] Jacob, M.; Lopata, A.L.; Dasouki, M. and Abdel-Rahman, A.M., Mass Spectrom. Rev., 2019, 38(3), 221-238

[25] Johnson, C.H.; Patterson, A.D.; Idle, J.R. and Gonzalez, F.J., Annu Rev Pharmacol Toxicol 2012, 52, 37-56

[26] Kummen, M. and Hov, J.R., Liver Int., 2019, https://doi.org/10.1111/liv.14153

[27] Barnes, S., J. Am. Diet Assoc. 2008, 108(11), 1888-95

[28] Kesavan, P.; Banerjee, A.; Banerjee, A.; Murugesan, R.; Marotta, F. and Pathak, S., 2018, Chapter 17, 221-235

[29] Yatsunenko, T.; Rey, F.E.; Manary, M.J.; Trehan, I.; Dominguez-Bello, M.; Contreras, M.; Magris, M.; Hidalgo, G.; Baldassano, R.N.; Anokhin, A.P.; Heath, A.C.; Warner, B.; Reeder, J.; Kuczynski, J.; Caporaso, J.G.; Lozupone, C.A.; Lauber, C.; Clemente, J.C.; Knights, D. and Knight, R., Nature, 2012, 486 (7402), 222-227

[30] Goodrich, J.K.; Waters, J.L.; Poole, A.C.; Sutter, J.L.; Koren, O.; Blekhman, R.; Beaumont, M.; Van Treuren, W.; Knight, R.; Bell, J.T.; Spector, T.D.; Clark, A.G. and Ley, R.E., Cell, 2014, 159 (4), 789-799

[31] Sun, G.; Yang, K.; Zhao, Z.; Guan, S.; Han, X. and Gross, R. W., Anal. Chem., 2007, 79, 17, 6629-6640

[32] Dettmer, K.; Aronov, P.A. and Hammock, B.D., Mass Spectrom. Rev. 2007, 26(1), 51-78

[33] Godzien, J.; Gil-de-la-Fuente, A.; Otero, A. and Barbas, C., 2018, Chapter 15, 415-441

[34] Glauser, G.; Veyrat, N.; Rochat, B.; Wolfender, J.L. and Turlings, T.C., J. Chromatogr. A, 2013, 1292, 151-159

[35] Gowda, G.A. and Djukovic, D., Methods Mol. Biol. 2014, 1198. 3-12

[36] Zhang, A.; Sun, H.; Wang, P.; Han, Y. and Wang, X., Analyst, 2012, 137(2), 293-300

[37] Katajamaa, M. and Orešič, M., J Chromatogr A., 2007, 1158(1), 318-328

[38] Sugimoto, M.; Kawakami, M.; Robert, M.; Soga, T. and Tomita, M., Curr. Bioinform., 2012, 7(1), 96-108

[39] Hendriks, M.W.B.; van-Eeuwijk, F.A.; Jellema, R.H.; Westerhuis, J.A.; Reijmers, T.H.; Hoefsloot, H.C.J. and Smilde, A.K., Trends Anal. Chem., 2011, 30(10), 1685-1698

[40] Scheltema, R.A.; Jankevics, A.; Jansen, R.C.; Swertz, M.A. and Breitling, R., Anal.Chem., 2011, 83(7), 2786-2793

[41] Kuhl, C.; Tautenhahn, R.; Bottcher, C.; Larson, T.R. and Neumann, S., Anal.Chem., 2012, 84(1), 283-289

[42] Clasquin, M.F.; Melamud, E. and Rabinowitz, J.D., Curr. Protoc. Bioinformatics., 2012, Chapter 14(14), Unit14.11

[43] Uppal, K.; Soltow, Q.A.; Strobel, F.H.; Stephen-Pittard, W.; Gernert, K.M.; Yu, T. and Jones, D.P., BMC Bioinformatics, 2013, 14(1), 1-12

[44] Chong, J.; Soufan, O.; Li, C.; Caraus, I.; Li, S.; Bourque, G.; Wishart, D.S. and Xia, J., Nucleic Acids Res., 2018, 46, W486-W494

[45] Dunn, W.B.; Wilson, I.D.; Nicholls, A.W. and Broadhurst, D., Bioanalysis, 2012, 4(18), 2249-2264

[46] Godzien, J.; Ciborowski, M.; Angulo, S.; Ruperez, F.J.; Martinez, M.P.; Senorans, F.J.; Ciluentes, A.; Ibanez, E. and Barbas, C., J. Proteome Res., 2011, 10(2), 837-844

[47] Zhang, W.; Chang, J.; Lei, Z.; Huhman, D.; Sumner, L.W. and Zhao, P.X., Anal. Chem., 2014, 86 (13) 6245–6253

[48] Gil-de-la-Fuente, A.; Grace Armitage, E.; Otero, A.; Barbas, C. and Godzien, J., Electrophoresis, 2017, 38 (18), 2242-2256

[49] Mahieu, N.G. and Patti, G.J., Anal.Chem., 2017, 89(19), 10397-10407

[50] Creek, D.J.; Dunn, W.B.; Fiehn, O.; Griffin, J.L.; Hall, R.D.; Lei, Z.; Sumner, L.W.; Mistrik, R.; Neumann, S.; Schymanski, E.L.; Trengove, R.; Wolfender, J.L., Metabolomics, 2014, 10(3), 350-353

[51] Uppal, K.; Walker, D.I.; Liu, K.; Li, S.; Go, Y.M. and Jones, D.P., Chem. Res. Toxicol., 2016, 29(12), 1956-1975

[52] Bharti, S.K. and Raja, R. Curr. Metabolomics, 2014, 2(3), 163-173

[53] Fiehn, O.; Robertson, D.;Griffin, J.; van-der-Werf, M.; Nikolau, B.; Morrison, N.; Sumner, L.W.; Goodacre, R.; Hardy, N.W.; Taylor, C.; Fostel, J.; Kristal, B.; Kaddurah-Daouk, R.; Mendes, P.; van-Ommen, B.; Lindon, J.C. and Sansone, S.A., Metabolomics, 2007, 3(3), 175-178

[54] Blaženovic, I.; Kind, T.; Ji, J. and Fiehn,O., Metabolites, 2018, 8(2), 1989-2218

[55] Chaleckis, R.; Meister, I.; Zhang, P. and Wheelock, C.E., Curr. Opin. Biotechnol., 2019, 55, 44-50

[56] Feng-Xiao, J.; Zhou, B. and Ressom, H.W., TRAC-trend anal. chem., 2012, 32, 1-14

[57] Schrimpe-Rutledge, A.; Codreanu, S.G.; Sherrod, S.D. and McLean, J.A., J.Am.Soc.Mass Spectrom., 2016, 27(12), 1897-1905

[58] Kind, T. and Fiehn, O., BMC Bioinformatics, 2007, 8, 105-120

[59] Rogers, S.; Scheltema, R.A.; Girolami, M. and Breitling, R., Bioinformatics 25 (4) (2009) 512–518

[60] Sumner, L.W.; Lei, Z.; Nikolau, B.J.; Saito, K.; Roessner, U. and Trengove, R., Metabolomics, 2014, 10(6), 1047-1049

[61] Rochat, B., J.Am.Soc.Mass Spectrom., 2017, 28(4), 709-723

[62] Stein, S.E.; Babushok, V.I.; Brown, R.L. and Linstrom, P.J., J. Chem. Inf. Model., 2007, 47(3), 975-980

[63] Jiang, Y.; Zhao, L.; Yuan, M. and Fu, A., J. Food Biochem., 2017, 41(3)

[64] Schauer, N.; Steinhauser, D.; Strelkov, S.; Schomburg, D.; Allison, G.; Moritz, T.; Lundgren, K.; Roessner-Tunali, U.; Forbes, M.G. and Willmitzer, L., FEBS Lett., 2005, 579(6), 1332-1337

[65] Wei, X.; Koo, I.; Kim, S. and Zhang, X., Analyst, 2014, 139(10), 2507-2514

[66] Bingol, K.; Bruschweiler-Li, L.; Li, D.; Zhang, B.; Xie, M. and Brüschweiler, R., Bioanalysis, 2016, 8(6), 557-573

[67] Gil-De-La-Fuente, A.; Godzien, J.; Saugar, S.; Garcia-Carmona, R.; Badran, H.; Wishart, D.S.; Barbas, C. and Otero, A., J. Proteome Res., 2019, 18(2), 797-802

[68] Wishart, D.S.; Djoumbou-Feunang, Y.; Marcu, A.; Guo, A.C.; Liang, K.; Vazquez-Fresno, R.; Johnson, D.; Li, C.; Karu, N.; Lo, E.; Assempour, N.; Berjanskii, M.; Singhal, S.; Arndt, D.; Liang, Y.; Badran, H.; Grant, J.; Serra-Cayuela, A.; Mandal, R.; Pon, A.; Knox, C.; Wilson, M.; Sajed, T.; Sayeeda, Z.; Liu, Y.; Neveu, V.; Scalbert, A. and Manach, C., Nucleic Acids Res., 2018, 46(D1), D608-D617

[69] Sud, M.; Fahy, E.; Cotter, D.; Brown, A.; Dennis, E.A.; Glass, C.K.; Merrill, A.H. Jr.; Murphy, R.C.; Raetz, C.R.H.; Russell, D.W. and Subramaniam, S., Nucleic Acids Res, 2007, 35(1), D527-D532

[70] Yasugi, E. and Watanabe, K., Tanpakushitsu Kakusan Koso, 2002, 47(7), 837-841

[71] Banerjee,P.; Erehman,J.; Gohlke, B.O.; Wilhelm, T.; Preissner, R. and Dunkel, M., Nucleic Acids Res., 2015, 43(D1), D935-D939

[72] Smith, C.A.; O'Maille, G.; Want, E.J.; Qin, C.; Trauger, S.A.; Brandon, T.R.; Custodio, D.E.; Abagyan, R. and Siuzdak, G., Ther. Drug Monit., 2005, 27(6), 747-751

[73] Kanehisa, M., Methods Mol.Biol., 2016, 1374, 55-70

[74] Horai, H.; Arita, M.; Nihei, Y.; Ikeda, T.; Ojima, Y.; Kakazu, Y.; Soga, T.; Nishioka, T.; Suwa, K.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M.Y.; Saito, K.; Kanaya, S.; Tanaka, K.; Takahashi, H.; Tanaka, S.; Aoshima, K.; Oda, Y.; Nakanishi, H.; Ikeda, K.; Taguchi, R.; Akimoto, N.; Maoka, T.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K. and Matsuura, F., Journal of Mass Spectrometry, 2010, 45, 703-714

[75] Li, L.; Zuniga, A.; Stanislaus, A.E.; Wu, Y.; Huan, T.; Zheng, J.; Li, R.; Zhou, J.; Shi, Y.; Wishart, D.S. and Lin, G., Anal.Chem., 2013, 85(6), 3401-3408

[76] Jeffryes, J.G.; Broadbelt, L.J.; Tyo, K.E.J.; Colastani, R.L.; Henry, C.S.; Elbadawi-Sidhu, M.; Kind, T.; Fiehn, O.; Niehaus, T.D. and Hanson, A.D., J. Cheminformatics, 2015, 7:44

[77] Djoumbou-Feunang, Y.; Fiamoncini, J.; Gil-de-la-Fuente, A.; Greiner, R.; Manach, C. and Wishart, D.S., J. Cheminformatics, 2019, 11:2, 1-25

[78] Domingo-Almenara, X.; Montenegro-Burke, J.; Benton, H.P. and Siuzdak, G., Anal. Chem., 2018, 90(1), 480-489

[79] Nash, W.J. and Dunn, W.B., Trends Analyt. Chem., 2018, In press, DOI: 10.1016/j.trac.2018.11.022

[80] Wolstencroft, K.; Haines, R.; Fellows, D.; Williams, A.; Withers, D.; Owen, S.; Soiland-Reyes, S.; Dunlop, I.; Nenadic, A.; Fisher, P.; Bhagat, J.; Belhajjame, K.; Bacall, F.; Hardisty, A.; Nieva-de-la-Hidalga, A.; Vargas, M.P.B.; Sufi, S. and Goble, C., Nucleic Acids Res., 2013, 41, W557-W561

[81] Beisken, S.; Meinl, T.; Wiswedel, B.; Figueiredo, L.F.; Berthold, M. and Steinbeck, C., BMC Bioinformatics, 2013, 14, 257

[82] Pétéra, M.; Le-Corguille, G.; Landi, M.; Monsoor, M.; Tremblay-Franco, M.; Duperier, C.; Martin, J.F.; Jacob, D.; Guitton, Y.; Lefebvre, M.; Pujos-Guillot, E.; Giacomoni, F.; Thévenot, E. and Caron, C., Bioinformatics, 2015, 31(9), 1495-1495

[83] Wang, M.;Carver, J.J.;Pevzner, P.;Mohimani, H.;Bandeira, N.;Phelan, V.V.;Sanchez, L.M.;Garg, N.;Watrous, J.;Luzzatto-Knaan, T.;Porto, C.;Bouslimani, A.;Melnik, A.V.;Meehan, M.J.;Pace, L.A.;Gonzalez, D.J.;Koyama, N.;Dorrestein, K.;Duggan, B.M.;Almaliti, J.;Gerwick, W.H.;Moore, B.S.;Dorrestein, P.C.;Peng, Y.;Nguyen, D.D.;Kapono, C.A.;Hsu, C.C.;Floros, D.J.;Zeng, Y.;Liu, W.T.;Crüsemann, M.;Boudreau, P.D.;Duncan, K.R.;Kleigrewe, K.;Gerwick, L.;Larson, C.B.;O'Neill, E.C.;Briand, E.;Glukhov, E.;Kharbush, J.J.;Mascuch, S.J.;Jensen, P.R.;Esquenazi, E.;Pociute, E.;Houson, H.;Vuong, L.;Macherla, V.;Sandoval-Calderón, M.;Sohlenkamp, C.;Kersten, R.D.;Quinn, R.A.;Gavilan, R.G.;Sedio, B.E.;Boya, C.A.P.;Torres-Mendoza, D.;Gutiérrez, M.;Northen, T.;Jenkins, S.;Dutton, R.J.;Parrot, D.;Tomasi, S.;Carlson, E.E.;Aigle, B.;Michelsen, C.F.;Jelsbak, L.;Maansson, M.;Klitgaard, A.;Nielsen, K.F.;Edlund, A.;McLean, J.;Shi, W.;Piel, J.;Helfrich, E.J.N.;Ryffel, F.;Vorholt, J.A.;Murphy, B.T.;Elfeki, M.;Liaw, C.C.;Yang, Y.L.;Humpf, H.U.;Keyzers, R.A.;Sims, A.C.;Baric, R.;Johnson, A.R.;Sidebottom, A.M.;Silva, D.B.;Marques, L.M.;Demarque, D.P.;Silva, R.R.;Rodríguez, A.M.C.;Lopes, N.P.;Granatosky, E.A.;Kurita, K.L.;Linington, R.G.;Charusanti, P.;Palsson, B.O.;McPhail, K.L.;Vining, O.B.;Traxler, M.F.;Engene, N.;Hoffman, T.;Müller, R.;Agarwal, V.;Williams, P.G.;Dai, J.;Neupane, R.;Gurr, J.;Lamsa, A.;Pogliano, K.;Zhang, C.;Allard, P.M.;Wolfender, J.L.;Phapale, P.;Alexandrov, T.;Nothias, L.F.;Litaudon, M.;Kyle, J.E.;Metz, T.O.;Waters, K.M.;Peryea, T.;Nguyen, D.T.;VanLeer, D.;Shinn, P.;Jadhav, A.;Liu, X.;Zhang, L. and Knight, R., Nat. Biotechnol., 2016, 34(8), 828-837

[84] Meringer, M. and Schymanski, E.L., Metabolites, 2013, 3(2), 440-462

[85] Kind, T.; Liu, K.H.; Lee, D.Y.; Defelice, B.; Meissen, J.K.; Fiehn, O., Nat. Methods, 2013, 10(8), 755-758

[86] Tsugawa, H.; Kind, T.; Nakabayashi, R.; Yukihira, D.; Tanaka, W.; Cajka, T.; Saito, K.; Fiehn, O. and Arita, M., Anal.Chem., 2016, 88(16), 7946-7958

[87] Heinonen, M.; Rantanen, A.; Mielikainen, T.; Kokkonen, J.; Kiuru, J.; Ketola, R. A. and Rousu, J., Rapid Commun. Mass Spectrom., 2008, 22(19), 3043-3052

[88] Kangas, L.J.; Metz, T.O.; Isaac, G.; Schrom, B.T.; Ginovska-Pangovska, B.; Wang, L.; Tan, L.; Lewis, R. R. and Miller, J.H., Bioinformatics, 2012, 28(13), 1705-1713

[89] Janesko, B.G.; Li, L. and Mensing, R., Anal. Chim. Acta, 2017, 995, 52-64

[90] Asgeirsson, V.; Bauer, C.A. and Grimme, S., Chem. Sci., 2017, 8(7), 4879-4895

[91] Heinonen, M.; Shen, H.; Zamboni, N. and Rousu, J., Bioinformatics, 2012, 28(18), 2333-2341

[92] Dührkop, K.; Shen, H.; Meusel, M.; Rousu, J. and Böcker, S., PNAS, 2015, 112, 12580-12585

[93] Ruttkies, C.; Wolf, S.; Neumann, S.; Schymanski, E.L. and Hollender, J., J. Cheminformatics, 2016, 8(1), 1, 3

[94] Djoumbou-Feunang, Y.; Pon, A.; Karu, N.; Zheng, J.; Li, C.; Arndt, D.; Gautam, M.; Allen, F. and Wishart, D. S., Metabolites, 2019, 9(4), 72

[95] Stanstrup, J.; Neumann, S. and Vrhovsek, U., Anal. Chem., 2015, 87(18), 9421-9428

[96] Navarro-Reig, M.; Ortiz-Villanueva, E.; Tauler, R. and Jaumot, J., Metabolites, 2017, 7(4), 54

[97] Hall, L.M.; Hill, D.W.; Bugden, K.; Cawley, S.; Hall, L.H.; Chen, M. and Grant, D.F., J. Chem. Inf. Model., 2018, 58(3), 591-604

[98] Codesido, S.; Giuseppe, M.R.; Lehmann, F.; González-Ruiz, V.; García, A.; Xenarios, I.; Liechti, R.; Bridge, A.; Boccard, J. and Rudaz, S., Metabolites, 2019, 9(5), 85

[99] Boswell, P.G.; Carr, P.W.; Cohen, J.D. and Hegeman, A.D., J. Chromatogr A., 2012, 1263, 179-188

[100] Lee, M.L.; Vassilaros, D.L. and White, C. M., Anal. Chem., 1979, 51(6), 768-773

[101] Koo, I.; Shi, X.; Kim, S. and Zhang, X., J Chromatogr A., 2014, 1337, 202-210

[102] Sugimoto, M.; Hirayama, A.; Robert, M.; Abe, S.; Soga, T. and Tomita, M., Electrophoresis, 2010, 14, 2311-2318

[103] González-Ruiz, V.; Gagnebin, Y.; Drouin, N.; Rudaz, S.; Schappler, J. and Codesido, S., Electrophoresis, 2018, 39(9), 1222-1232

[104] Weber, R.J.M. and Viant, M.R., Chemometrics Intellig.Lab.Syst., 2010, 104(1), 75-82

[105] Alden, N.; Krishnan, S.; Porokhin, V.; Raju, R.; McElearney, K.; Gilbert, A. and Lee, K., Anal.Chem., 2017, 89(24), 13097-13104

[106] Uppal, K.; Walker, D.I. and Jones, D.P., Anal. Chem., 2017, 89(2), 1063-1067

GOALS, SCOPE AND OUTLINE

Untargeted Metabolomics is a useful and powerful tool to approach the biological systems. The success of untargeted metabolomics is closely connected to metabolite identification. Reducing the false positives and increasing the true positives and true negatives provides a more accurate and complete picture for the subsequent biological interpretation. A larger metabolite coverage helps establishing relationships, while the misidentifications lead to wrong conclusions.

Metabolite identification is probably the most significant and persistent challenge in untargeted metabolomics. Regardless the confidence level required, this task is generally very slow and tedious. The first step in this task is the metabolite annotation, the assignation of putative structures to the already processed data matrix containing the features. Publicly available databases such as MassBank, KomicMarket, HMDB, Metlin, KEGG, LipidMaps or MINE contain metabolites that enable researchers to retrieve the putative candidates for the features. However, they only cover a fraction of the full metabolome, since a large portion of it is yet to be discovered. Moreover, the databases have a low overlap among them since they are devoted to different purposes and/or target metabolites.

The main goal of this dissertation is the creation of a software tool to support researchers in metabolite annotation and identification: CEU Mass Mediator (CMM). This tool aims to exploit as much analytical and non-analytical information as possible, both coming from the CEMBIO and from external sources. In particular, it aims to develop the next features:

- A single interface to query simultaneously distinct databases, with the automatic unification of compounds coming from them.

- A MS^1 annotation expert-system using information coming from RT, from experimental IPs formed by lipid class and relationships between signals coeluting in LC-ESI-MS experiments. The RT information has not been used before for such purpose by other tools. Additionally, it will provide optional filters to restrict the putative annotations returned for a given query.

- A semi-automated service for the annotation and identification of oxidized glycerophosphocholines (oxPCs).
- A MS/MS search service comparing the experimental fragmentation with the ones available in the data sources and scoring the putative annotations returned as result of a query.
- A spectral quality controller to guide the researchers about how good is the MS/MS data obtained in an experiment with the purpose of metabolite annotation.
- A Representational State Transfer (REST) Application Programming Interface (API) to access the previously described services to facilitate integration the communication with other tools.

Although CMM intends to complement the tools already available for the metabolite annotation and identification, it also aims to be a self-contained tool to provide a better user experience. Therefore, it will provide services already available in other tools when this duplicity makes sense for the overall functionality of the tool; i.e., to avoid the need for the users of using a different tool and interface for each step in the metabolite identification process. The final target user is both the analytical chemist with low knowledge about computer science that shall use the tool through its web interface, and the developer that is interested in integrating CMM in their tools or workflows through its REST API. According to these assumptions, the tool will be:

- Open-source, to increase the potential audience and to involve people from the metabolomic community.

- Accessible from a simple web interface, to avoid the need of computer science/programming knowledge to work with it.
- Accessible from a REST API, to provide automatic mechanisms to communicate with scripts, other tools or being integrated in metabolomic workflows.

To fulfill these requirements, the state of the art should be deeply studied to provide innovative and efficient solutions in the metabolite annotation and identification. **Chapter one** reviews the ESI-MS-based databases for untargeted metabolomics. A separation between GC/MS databases and general-purpose databases should be done, since GC/MS databases usually contain information about Kovats RI. This data helps significantly during the identification process because it provides an orthogonal filter. In LC/MS and CE/MS, the RT and MT respectively are far less reproducible, therefore their use as an orthogonal filter is not trivial.

To increase the metabolite coverage when using LC or CE, the researchers usually need to access, retrieve, merge and filter the results from different databases manually, investing a large amount of time in this process (see level 3 in Table 1). Once they have merged the putative annotations retrieved from the databases, it is time to discard or confirm them applying analytical, biological and chemical knowledge. One strategy consists in the application of analytical and nonanalytical knowledge about compounds. For example, if the experiment has been run using a Reversed-Phase (RP) column, the non-polar analytes are retained more than polar ones by the separation column. Therefore, a putative annotation of a feature with a high RT pointing to a polar compound can be discarded because the polar compounds are the ones eluting first. Meanwhile, biological knowledge can also be applied to discard compounds. For example, if a human sample is being analyzed and a putative annotation corresponds to a plant metabolite, there is a strong evidence pointing towards discarding the putative annotation.

Chapter two describes the software tool developed in this thesis: CMM. This first revision of the tool (CMM 2.0) provided researchers with a metabolite annotation tool using information coming from MS^1 analyses (confidence level 3). It integrates and unifies experimental compounds from HMDB, KEGG, Lipidmaps and Metlin and in-silico predicted compounds from MINE. In addition, CMM 2.0 scores the putative annotations which matched the query parameters based on ionization, adduct relation and RT rules developed in Drools, a business rules management system for Java.

Lipidomics is a newly emerged discipline within the -omics sciences that studies cellular lipids on a large scale. Within lipidomics, the biological role of oxidized glycerophosphocholines (oxPCs) is a current topic of research contributing to the understanding of health and disease. The identification of oxidized lipids is one of the challenges in the metabolite identification due to the low presence of them in the general databases. However, a systematic approach to identifying the oxPCs can be extracted from experimental knowledge. The chromatographic characteristics and the spectral information from MS^2 provide very useful information for the identification of oxPCs.

Chapter three outlines a proposal for the identification of oxPCs in untargeted metabolomics. It requires the insertion of the oxPCs into the CMM database to subsequently create a systematic approach for the recognition, annotation and identification of oxPCs. This approach uses information from the fragmentation obtained by MS^2 analyses for both long chain and short chain oxidations; the hydrophilicity of the new oxPCs; and the experimental knowledge about adduct formation. It incorporates a list of IPs and neutral losses of PC(16:0/20:4) known as PAPC. The presence of these compounds potentially increases the true putative annotations and decreases the false putative annotations.

The confidence level during the metabolite annotation can be raised comparing the experimental fragmentation spectra of the features with the spectra available in the databases. A current challenge in Metabolomics is to improve the quality of the MS^2 spectra obtained in the data acquirement and the data preprocessing. Nevertheless, the time and the funding to perform metabolomic experiments are limited, and the low availability of authentic standards often hinders the metabolite identification process in untargeted approaches. Consequently, a systematic method to evaluate the spectral quality obtained could permit researchers to focus on the most promising features with identification purposes, since a high-quality spectrum is paramount to increase true positive identifications and a low-quality spectrum can lead to false positive identifications.

A spectral quality controller, among other new functionalities, is described in **chapter four**, a major update in CMM: CMM 3.0. Besides that, the integrated data sources had grown qualitative and quantitatively since the release of CMM 2.0 (see **chapter two**), therefore an update of the information coming from them was performed. Due to the availability of the fragmentation spectra in the publicly available databases, a new service for the metabolite identification consisting in a MS/MS search was added; this service allow researchers to possible achieve a confidence level 2 in the putative annotations previously obtained by MS^1. It also provides a service to identify the oxPCs based on the approach presented in **chapter three**. In parallel, CMM 3.0 provides the users with a RESTful API that encapsulates its services, allowing the integration within automated workflows or within other software tools without the necessity of using the CMM web interface. This chapter presents the integration of CMM 3.0 into HMDB.

Finally, in the last chapter of this thesis, a summary of all the contributions made is presented, conclusions are drawn and possible lines of future work to extend the functionality of CMM are presented.

CHAPTER 1: DIFFERENTIATING SIGNALS TO MAKE BIOLOGICAL SENSE - A GUIDE THROUGH DATABASES FOR MS-BASED NON-TARGETED METABOLOMICS

Alberto Gil de la Fuente[1,2]
Emily Grace Armitage[3,4]
Abraham Otero[2]
Coral Barbas[1]
Joanna Godzien[1]

[1] Centre for Metabolomics and Bioanalysis (CEMBIO), Facultad de Farmacia, Universidad CEU San Pablo, Campus Montepríncipe, Boadilla del Monte, Madrid, Spain
[2] Department of Information Technology, Universidad CEU San Pablo, Campus Montepríncipe, Boadilla del Monte, Madrid, Spain
[3] Wellcome Centre for Molecular Parasitology, Institute of Infection, Immunity and Inflammation, College of Medical Veterinary and Life Sciences, University of Glasgow, Glasgow, UK
[4] Glasgow Polyomics, Wolfson Wohl Cancer Research Centre, College of Medical Veterinary and Life Sciences, University of Glasgow, Glasgow, UK

Received February 14, 2017
Revised March 17, 2017
Accepted March 17, 2017

Review

Differentiating signals to make biological sense – A guide through databases for MS-based non-targeted metabolomics

Metabolite identification is one of the most challenging steps in metabolomics studies and reflects one of the greatest bottlenecks in the entire workflow. The success of this step determines the success of the entire research, therefore the quality at which annotations are given requires special attention. A variety of tools and resources are available to aid metabolite identification or annotation, offering different and often complementary functionalities. In preparation for this article, almost 50 databases were reviewed, from which 17 were selected for discussion, chosen for their online ESI-MS functionality. The general characteristics and functions of each database is discussed in turn, considering the advantages and limitations of each along with recommendations for optimal use of each tool, as derived from experiences encountered at the Centre for Metabolomics and Bioanalysis (CEMBIO) in Madrid. These databases were evaluated considering their utility in non-targeted metabolomics, including aspects such as identifier assignment, structural assignment and interpretation of results.

Keywords:
Database / Mediator / Metabolite identification / Metabolomics / Structural elucidation
DOI 10.1002/elps.201700070

Additional supporting information may be found in the online version of this article at the publisher's web-site

1 Introduction

The importance of metabolomics and its utility is still increasing, both in terms of the range of applications and their frequency. Among the different applications, non-targeted metabolomics plays a vital role, revealing new and unexpected findings that can lead to further research in a particular direction [1–4]. However, the success of this approach highly depends on the possibility to understand and interpret the information hidden within a complex metabolomics dataset. Most metabolomics studies are based on ESI-MS [5–7], usually with a preceding separation step such as LC, tending to measure the ratio of mass to charge (*m/z*) and abundance of each ion that originate from chromatographically separated molecules. After data pre-processing and statistical analysis, a list of discriminating signals between sample groups can be obtained [8]. However, to understand the nature of this separation and its cause, masses must be annotated with metabolite identifications, which can be mapped onto biochemical pathways to understand their origins. Metabolite identification is influenced by a range of factors, which should be taken into consideration from the initial experimental design to the interpretation of results (Fig. 1).

To annotate measured masses with metabolite identifiers (IDs), a data source is needed for comparison. One solution would be to use an in-house library based on the authentic standards analysed under particular conditions. In this way, at least two independent and orthogonal characteristics (e.g. mass and retention time) could be used for comparison, providing first, the highest level of identification confidence according to MSI (Metabolomics Standards Initiative) guidelines [9]. This method is rather restrictive though, since only commercially available metabolites can be introduced to the library and used for annotation. New strategies

Correspondence: Dr. Joanna Godzien, Centre for Metabolomics and Bioanalysis (CEMBIO), Facultad de Farmacia, Universidad CEU San Pablo, Campus Monteprincipe, Boadilla del Monte, 28668 Madrid, Spain
E-mail: joannabarbara.godzien@ceu.es

Abbreviations: API, Application Programming Interfaces; **CeuMM**, CEU mass mediator; **CSI:FingerID**, Compound Structure Identification:FingerID; **HMDB**, Human Metabolome Database; **ID**, identifier; **KomicMarket**, Kazusa Omics Data Market; **LipidMap**, LIPID Metabolites and Pathways Strategy; **MINE**, Metabolic In Silico Network Expansion Databases; **MSI**, Metabolomics Standards Initiative; **MS^n**, multi-stage MS; **NIMS**, nanostructure imaging MS; **PEP search**, peptides search; **PIF**, Precursor Ion Fingerprinting; **ppm**, part per million; **Workbench**, UCSD metabolomics workbench

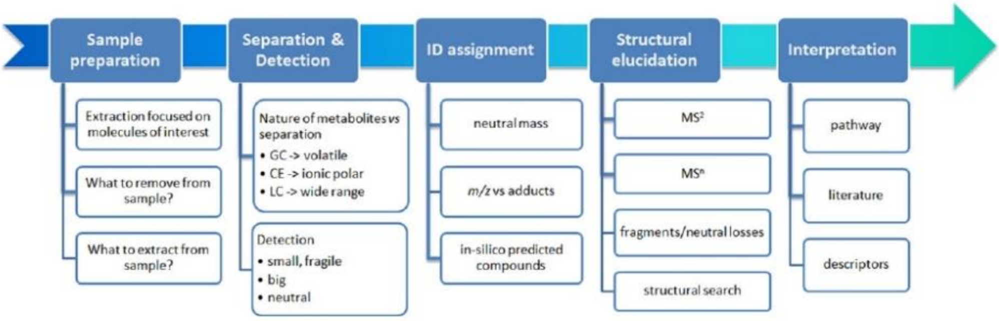


Figure 1. Different aspects of metabolite identification in the metabolomics workflow.

utilising online accessible databases that contain a large array of information have emerged to mitigate this shortfall [10–15]. Cross comparison of experimental data to databases can be performed using only one characteristic (mass) (second level of confidence for MSI), which highlights a limitation compared to using in-house libraries. Nevertheless, the amount of information provided is huge, covering different subclasses and including not only endogenous metabolites, but also substances originating from the microbiome, diet, plants, or supplementation. Therefore, the coverage of annotations across the data is much more promising. Furthermore, *in silico* predicted compounds are now available, considering biological modifications of known metabolites that may occur under particular conditions [16]. This somehow responds to the clear need to open metabolomics research to consider new or previously unidentified metabolites. Moreover, databases are continuously growing due to the contribution of many researchers.

In 2011, Fiehn et al. divided databases into two categories, making a clear distinction between pathway- and compound-centric databases [17]. In this review, only compound-centric databases are examined, omitting databases such as KEGG (www.genome.jp/kegg), Reactome (www.reactome.org) and Wikipathways (wikipathways.org). Additionally, only online, open-access databases are included, omitting commercial resources. Finally, only ESI-MS dedicated resources allowing exact mass searching are assessed. Following these restrictions, 17 data sources were selected for review from a total of 47 considered. For a comprehensive list of those rejected, refer to Supporting Information Table 1. Data sources covered in this article are as follows: BioCyc Database Collection (BioCyc) (biocyc.org), Ceu mass mediator (CeuMM) (ceumass.eps.uspceu.es), Compound Structure Identification:FingerID (CSI:FingerID) (www.csi-fingerid.org), Human Metabolome Database (HMDB) (www.hmdb.ca), Kazusa Omics Data Market (KomicMarket) (webs2.kazusa.or.jp/komicmarket/index.php), LipidBank (lipidbank.jp), LIPID Metabolites and Pathways Strategy (LipidMaps) (www.lipidmaps.org), MAGMa (www.emetabolomics.org/magma), MassBank (www.massbank.jp), MassTRIX (masstrix3.helmholtz-muenchen.de/masstrix3/), MetFrag (msbi.ipb-halle.de/MetFragBeta), METLIN (metlin.scripps.edu), Metabolic In Silico Network Expansion Databases (MINE) (minedatabase.mcs.anl.gov), MycompoundID (www.mycompoundid.org), MzCloud (www.mzcloud.org), MZedDB (maltese.dbs.aber.ac.uk:8888/hrmet/search/addsearch0.php) and UCSD metabolomics workbench (Workbench) (www.metabolomicsworkbench.org). The number of compounds contained in each is depicted in Fig. 2. Figure 3 illustrates the number of citations of each data source in google scholar, while information on the initial release data and latest updates for each are given in Supporting Information Table 2. All information given on each database is true as of 15 January 2017. It is important to highlight that this review was constructed based not only on literature research, but also on usage and revision of databases at the Centre for Metabolomics and Bioanalysis (CEMBIO), Madrid.

Of the data sources reviewed, BioCyc, HMDB, KomicMarket, LipidBank, LipidMaps, MassBank, METLIN, MzCloud and Workbench are considered databases sensu stricto. All the other online tools reviewed are mediators that use the information provided by databases: CeuMM, CSI:FingerID, MAGMa, MassTRIX, MetFrag, MINE and MZedDB. Detailed information on the sources used by each database and mediator is stated in Supporting Information Table 3. Both types of online tool are very important for the metabolomics society and both require continued improvement. Different databases focus on different types of molecules, therefore it is recommended to use a combination of resources for optimal coverage. In this way, mediators are advantageous since they perform searches across different sources through a single interface. However, not all mediators offer multi-source usage. For example MAGMa, MetFrag and MINE permit the use of only one source at once. MassTRIX on the other hand searches KEGG, HMDB and LipidMaps together or separately (as defined by the user) and CeuMM permits the search between all combinations of HMDB, KEGG, LipidMaps, Metlin and MINE as required. Within this review, the general characteristics of each of the data sources are detailed, followed by a discussion of functionality to compare and contrast the advantages and limitations of each for different aspects.

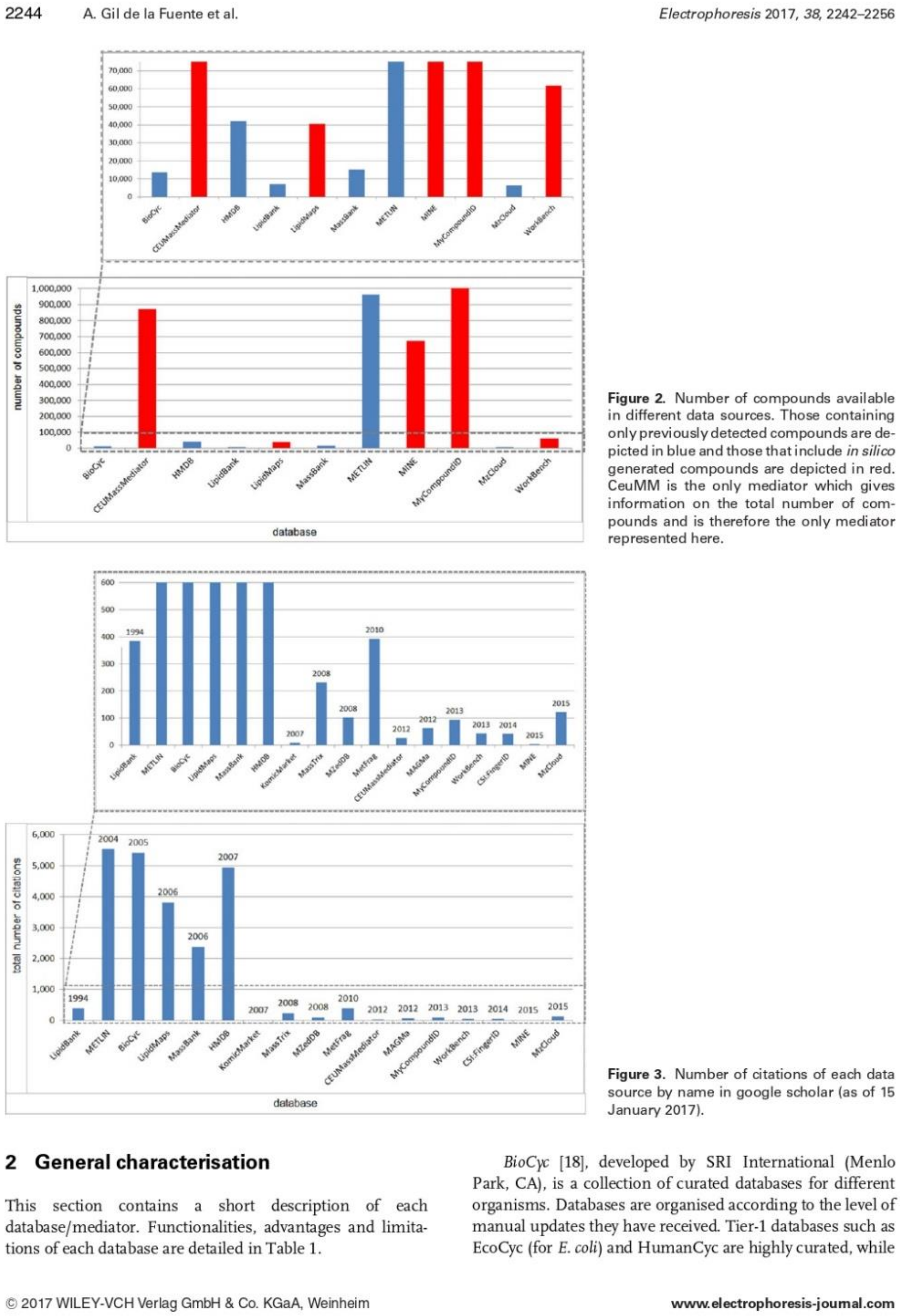


Figure 2. Number of compounds available in different data sources. Those containing only previously detected compounds are depicted in blue and those that include *in silico* generated compounds are depicted in red. CeuMM is the only mediator which gives information on the total number of compounds and is therefore the only mediator represented here.

Figure 3. Number of citations of each data source by name in google scholar (as of 15 January 2017).

2 General characterisation

This section contains a short description of each database/mediator. Functionalities, advantages and limitations of each database are detailed in Table 1.

BioCyc [18], developed by SRI International (Menlo Park, CA), is a collection of curated databases for different organisms. Databases are organised according to the level of manual updates they have received. Tier-1 databases such as EcoCyc (for *E. coli*) and HumanCyc are highly curated, while

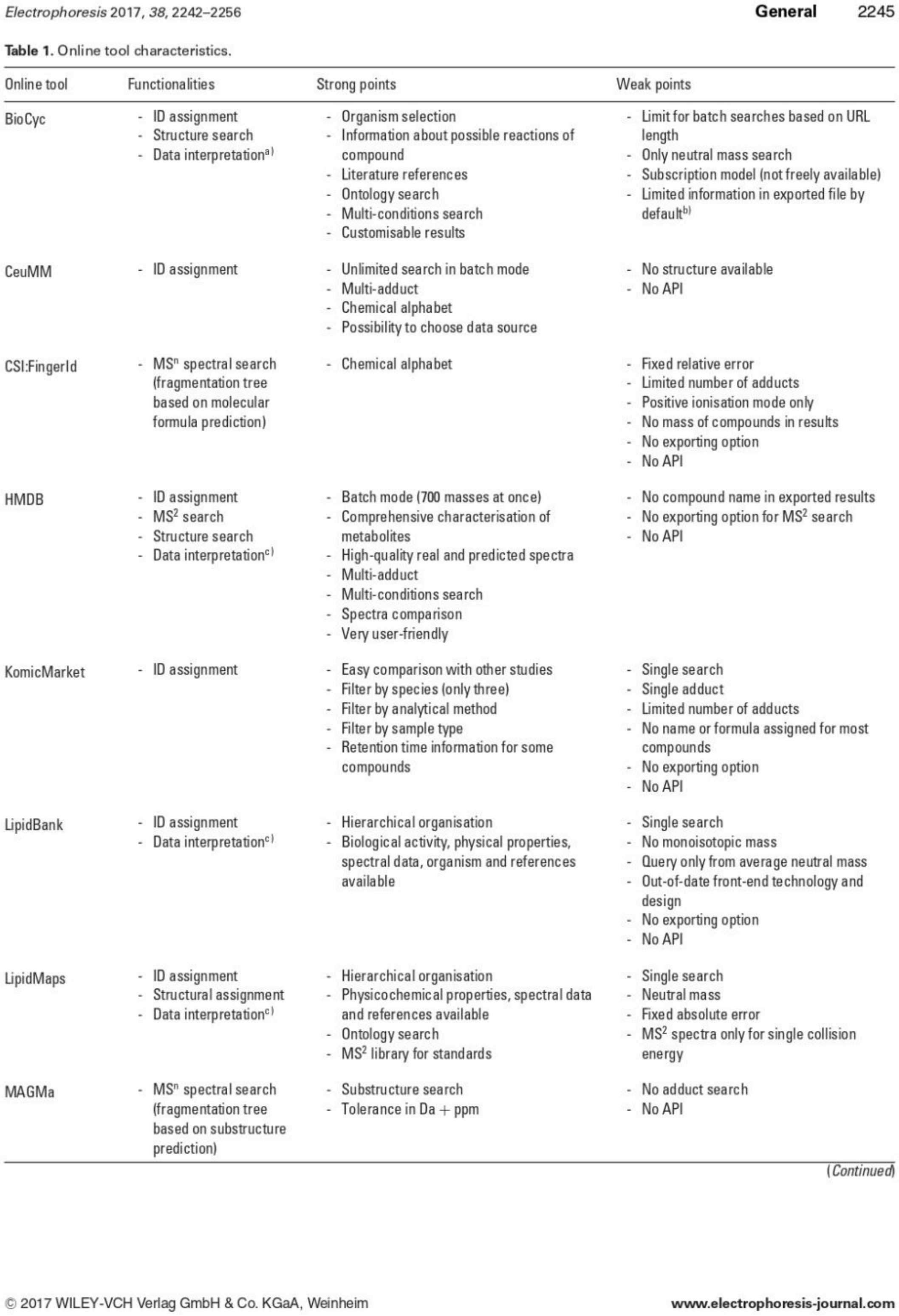


Table 1. Online tool characteristics.

Online tool	Functionalities	Strong points	Weak points
BioCyc	- ID assignment - Structure search - Data interpretation[a)]	- Organism selection - Information about possible reactions of compound - Literature references - Ontology search - Multi-conditions search - Customisable results	- Limit for batch searches based on URL length - Only neutral mass search - Subscription model (not freely available) - Limited information in exported file by default[b)]
CeuMM	- ID assignment	- Unlimited search in batch mode - Multi-adduct - Chemical alphabet - Possibility to choose data source	- No structure available - No API
CSI:FingerId	- MS^n spectral search (fragmentation tree based on molecular formula prediction)	- Chemical alphabet	- Fixed relative error - Limited number of adducts - Positive ionisation mode only - No mass of compounds in results - No exporting option - No API
HMDB	- ID assignment - MS^2 search - Structure search - Data interpretation[c)]	- Batch mode (700 masses at once) - Comprehensive characterisation of metabolites - High-quality real and predicted spectra - Multi-adduct - Multi-conditions search - Spectra comparison - Very user-friendly	- No compound name in exported results - No exporting option for MS^2 search - No API
KomicMarket	- ID assignment	- Easy comparison with other studies - Filter by species (only three) - Filter by analytical method - Filter by sample type - Retention time information for some compounds	- Single search - Single adduct - Limited number of adducts - No name or formula assigned for most compounds - No exporting option - No API
LipidBank	- ID assignment - Data interpretation[c)]	- Hierarchical organisation - Biological activity, physical properties, spectral data, organism and references available	- Single search - No monoisotopic mass - Query only from average neutral mass - Out-of-date front-end technology and design - No exporting option - No API
LipidMaps	- ID assignment - Structural assignment - Data interpretation[c)]	- Hierarchical organisation - Physicochemical properties, spectral data and references available - Ontology search - MS^2 library for standards	- Single search - Neutral mass - Fixed absolute error - MS^2 spectra only for single collision energy
MAGMa	- MS^n spectral search (fragmentation tree based on substructure prediction)	- Substructure search - Tolerance in Da + ppm	- No adduct search - No API

(*Continued*)

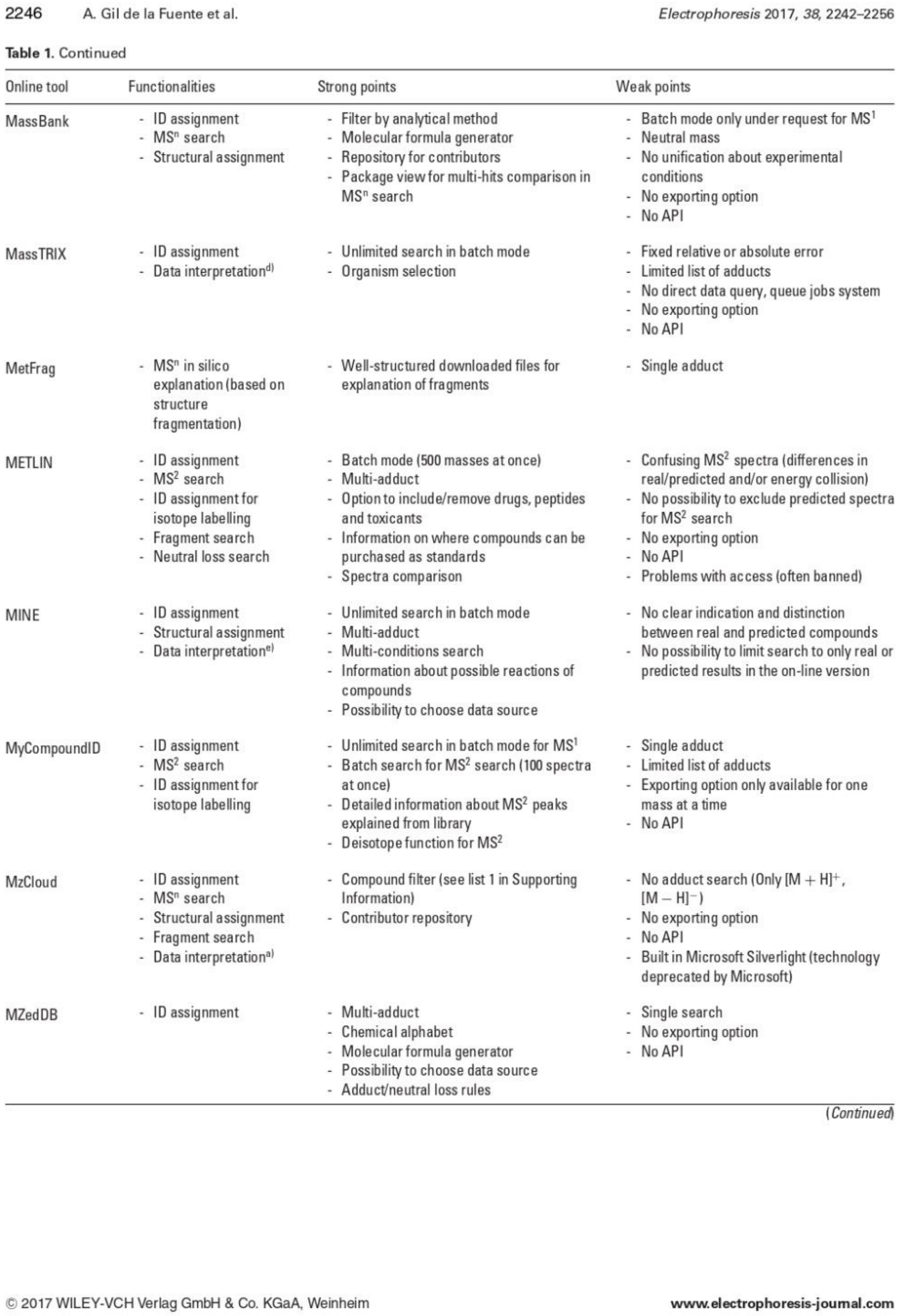


Table 1. Continued

Online tool	Functionalities	Strong points	Weak points
MassBank	- ID assignment - MS^n search - Structural assignment	- Filter by analytical method - Molecular formula generator - Repository for contributors - Package view for multi-hits comparison in MS^n search	- Batch mode only under request for MS^1 - Neutral mass - No unification about experimental conditions - No exporting option - No API
MassTRIX	- ID assignment - Data interpretation[d)]	- Unlimited search in batch mode - Organism selection	- Fixed relative or absolute error - Limited list of adducts - No direct data query, queue jobs system - No exporting option - No API
MetFrag	- MS^n in silico explanation (based on structure fragmentation)	- Well-structured downloaded files for explanation of fragments	- Single adduct
METLIN	- ID assignment - MS^2 search - ID assignment for isotope labelling - Fragment search - Neutral loss search	- Batch mode (500 masses at once) - Multi-adduct - Option to include/remove drugs, peptides and toxicants - Information on where compounds can be purchased as standards - Spectra comparison	- Confusing MS^2 spectra (differences in real/predicted and/or energy collision) - No possibility to exclude predicted spectra for MS^2 search - No exporting option - No API - Problems with access (often banned)
MINE	- ID assignment - Structural assignment - Data interpretation[e)]	- Unlimited search in batch mode - Multi-adduct - Multi-conditions search - Information about possible reactions of compounds - Possibility to choose data source	- No clear indication and distinction between real and predicted compounds - No possibility to limit search to only real or predicted results in the on-line version
MyCompoundID	- ID assignment - MS^2 search - ID assignment for isotope labelling	- Unlimited search in batch mode for MS^1 - Batch search for MS^2 search (100 spectra at once) - Detailed information about MS^2 peaks explained from library - Deisotope function for MS^2	- Single adduct - Limited list of adducts - Exporting option only available for one mass at a time - No API
MzCloud	- ID assignment - MS^n search - Structural assignment - Fragment search - Data interpretation[a)]	- Compound filter (see list 1 in Supporting Information) - Contributor repository	- No adduct search (Only $[M + H]^+$, $[M - H]^-$) - No exporting option - No API - Built in Microsoft Silverlight (technology deprecated by Microsoft)
MZedDB	- ID assignment	- Multi-adduct - Chemical alphabet - Molecular formula generator - Possibility to choose data source - Adduct/neutral loss rules	- Single search - No exporting option - No API

(*Continued*)

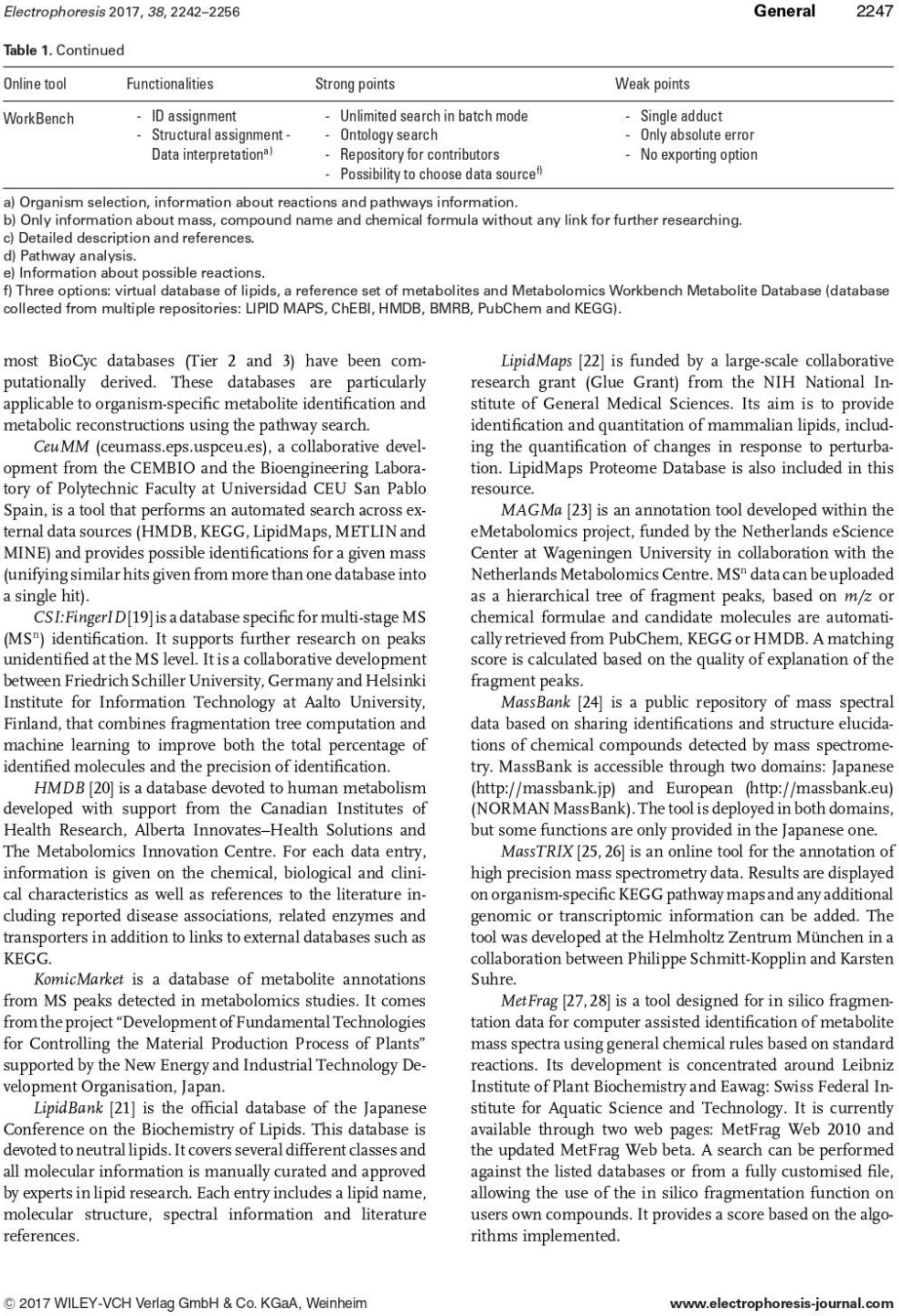


Table 1. Continued

Online tool	Functionalities	Strong points	Weak points
WorkBench	- ID assignment - Structural assignment - Data interpretation[a)]	- Unlimited search in batch mode - Ontology search - Repository for contributors - Possibility to choose data source[f)]	- Single adduct - Only absolute error - No exporting option

a) Organism selection, information about reactions and pathways information.
b) Only information about mass, compound name and chemical formula without any link for further researching.
c) Detailed description and references.
d) Pathway analysis.
e) Information about possible reactions.
f) Three options: virtual database of lipids, a reference set of metabolites and Metabolomics Workbench Metabolite Database (database collected from multiple repositories: LIPID MAPS, ChEBI, HMDB, BMRB, PubChem and KEGG).

most BioCyc databases (Tier 2 and 3) have been computationally derived. These databases are particularly applicable to organism-specific metabolite identification and metabolic reconstructions using the pathway search.

CeuMM (ceumass.eps.uspceu.es), a collaborative development from the CEMBIO and the Bioengineering Laboratory of Polytechnic Faculty at Universidad CEU San Pablo Spain, is a tool that performs an automated search across external data sources (HMDB, KEGG, LipidMaps, METLIN and MINE) and provides possible identifications for a given mass (unifying similar hits given from more than one database into a single hit).

CSI:FingerID[19] is a database specific for multi-stage MS (MS^n) identification. It supports further research on peaks unidentified at the MS level. It is a collaborative development between Friedrich Schiller University, Germany and Helsinki Institute for Information Technology at Aalto University, Finland, that combines fragmentation tree computation and machine learning to improve both the total percentage of identified molecules and the precision of identification.

HMDB [20] is a database devoted to human metabolism developed with support from the Canadian Institutes of Health Research, Alberta Innovates–Health Solutions and The Metabolomics Innovation Centre. For each data entry, information is given on the chemical, biological and clinical characteristics as well as references to the literature including reported disease associations, related enzymes and transporters in addition to links to external databases such as KEGG.

KomicMarket is a database of metabolite annotations from MS peaks detected in metabolomics studies. It comes from the project "Development of Fundamental Technologies for Controlling the Material Production Process of Plants" supported by the New Energy and Industrial Technology Development Organisation, Japan.

LipidBank [21] is the official database of the Japanese Conference on the Biochemistry of Lipids. This database is devoted to neutral lipids. It covers several different classes and all molecular information is manually curated and approved by experts in lipid research. Each entry includes a lipid name, molecular structure, spectral information and literature references.

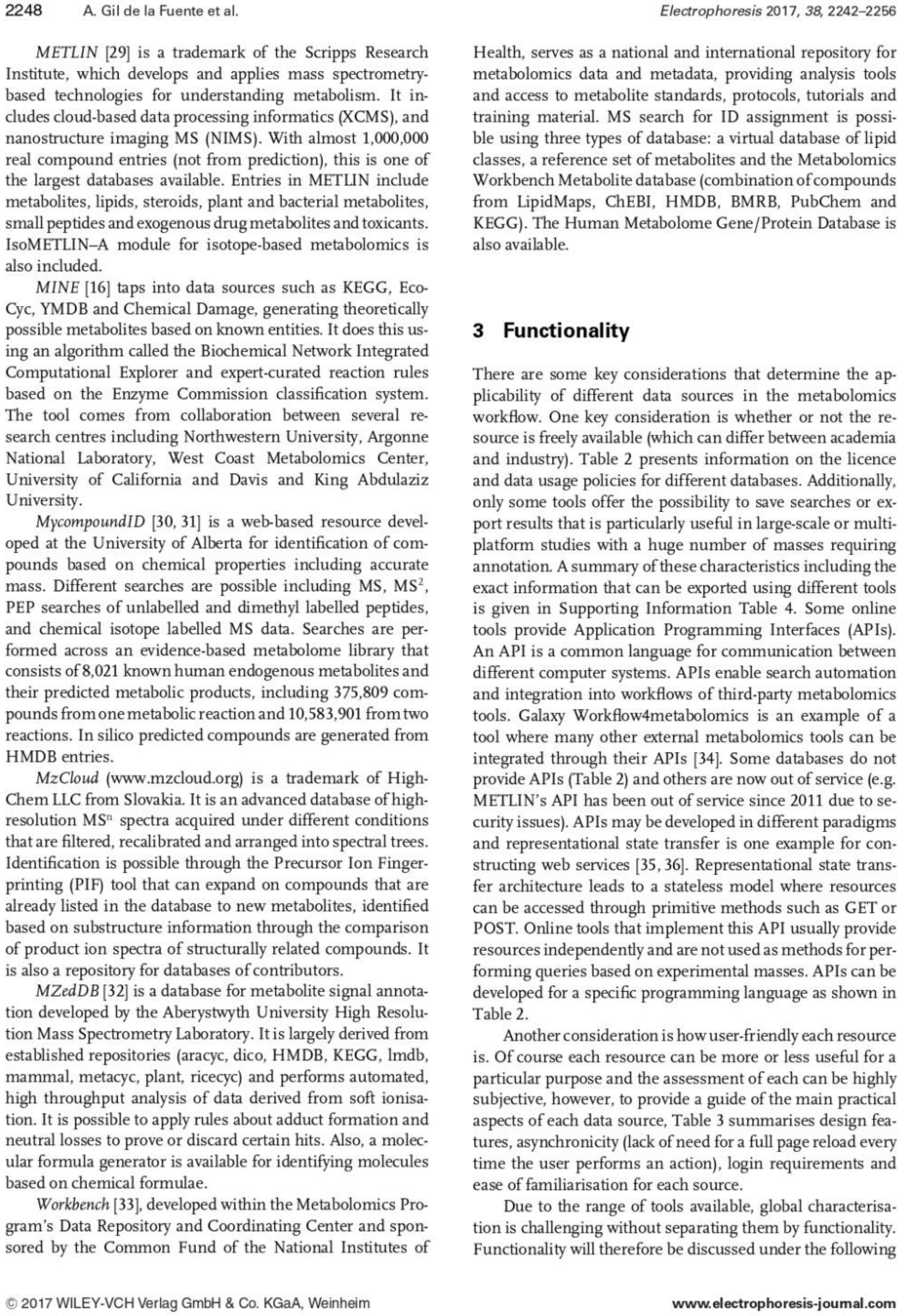
LipidMaps [22] is funded by a large-scale collaborative research grant (Glue Grant) from the NIH National Institute of General Medical Sciences. Its aim is to provide identification and quantitation of mammalian lipids, including the quantification of changes in response to perturbation. LipidMaps Proteome Database is also included in this resource.

MAGMa [23] is an annotation tool developed within the eMetabolomics project, funded by the Netherlands eScience Center at Wageningen University in collaboration with the Netherlands Metabolomics Centre. MS^n data can be uploaded as a hierarchical tree of fragment peaks, based on m/z or chemical formulae and candidate molecules are automatically retrieved from PubChem, KEGG or HMDB. A matching score is calculated based on the quality of explanation of the fragment peaks.

MassBank [24] is a public repository of mass spectral data based on sharing identifications and structure elucidations of chemical compounds detected by mass spectrometry. MassBank is accessible through two domains: Japanese (http://massbank.jp) and European (http://massbank.eu) (NORMAN MassBank). The tool is deployed in both domains, but some functions are only provided in the Japanese one.

MassTRIX [25, 26] is an online tool for the annotation of high precision mass spectrometry data. Results are displayed on organism-specific KEGG pathway maps and any additional genomic or transcriptomic information can be added. The tool was developed at the Helmholtz Zentrum München in a collaboration between Philippe Schmitt-Kopplin and Karsten Suhre.

MetFrag [27, 28] is a tool designed for in silico fragmentation data for computer assisted identification of metabolite mass spectra using general chemical rules based on standard reactions. Its development is concentrated around Leibniz Institute of Plant Biochemistry and Eawag: Swiss Federal Institute for Aquatic Science and Technology. It is currently available through two web pages: MetFrag Web 2010 and the updated MetFrag Web beta. A search can be performed against the listed databases or from a fully customised file, allowing the use of the in silico fragmentation function on users own compounds. It provides a score based on the algorithms implemented.



METLIN [29] is a trademark of the Scripps Research Institute, which develops and applies mass spectrometry-based technologies for understanding metabolism. It includes cloud-based data processing informatics (XCMS), and nanostructure imaging MS (NIMS). With almost 1,000,000 real compound entries (not from prediction), this is one of the largest databases available. Entries in METLIN include metabolites, lipids, steroids, plant and bacterial metabolites, small peptides and exogenous drug metabolites and toxicants. IsoMETLIN–A module for isotope-based metabolomics is also included.

MINE [16] taps into data sources such as KEGG, EcoCyc, YMDB and Chemical Damage, generating theoretically possible metabolites based on known entities. It does this using an algorithm called the Biochemical Network Integrated Computational Explorer and expert-curated reaction rules based on the Enzyme Commission classification system. The tool comes from collaboration between several research centres including Northwestern University, Argonne National Laboratory, West Coast Metabolomics Center, University of California and Davis and King Abdulaziz University.

MycompoundID [30, 31] is a web-based resource developed at the University of Alberta for identification of compounds based on chemical properties including accurate mass. Different searches are possible including MS, MS^2, PEP searches of unlabelled and dimethyl labelled peptides, and chemical isotope labelled MS data. Searches are performed across an evidence-based metabolome library that consists of 8,021 known human endogenous metabolites and their predicted metabolic products, including 375,809 compounds from one metabolic reaction and 10,583,901 from two reactions. In silico predicted compounds are generated from HMDB entries.

MzCloud (www.mzcloud.org) is a trademark of HighChem LLC from Slovakia. It is an advanced database of high-resolution MS^n spectra acquired under different conditions that are filtered, recalibrated and arranged into spectral trees. Identification is possible through the Precursor Ion Fingerprinting (PIF) tool that can expand on compounds that are already listed in the database to new metabolites, identified based on substructure information through the comparison of product ion spectra of structurally related compounds. It is also a repository for databases of contributors.

MZedDB [32] is a database for metabolite signal annotation developed by the Aberystwyth University High Resolution Mass Spectrometry Laboratory. It is largely derived from established repositories (aracyc, dico, HMDB, KEGG, lmdb, mammal, metacyc, plant, ricecyc) and performs automated, high throughput analysis of data derived from soft ionisation. It is possible to apply rules about adduct formation and neutral losses to prove or discard certain hits. Also, a molecular formula generator is available for identifying molecules based on chemical formulae.

Workbench [33], developed within the Metabolomics Program's Data Repository and Coordinating Center and sponsored by the Common Fund of the National Institutes of Health, serves as a national and international repository for metabolomics data and metadata, providing analysis tools and access to metabolite standards, protocols, tutorials and training material. MS search for ID assignment is possible using three types of database: a virtual database of lipid classes, a reference set of metabolites and the Metabolomics Workbench Metabolite database (combination of compounds from LipidMaps, ChEBI, HMDB, BMRB, PubChem and KEGG). The Human Metabolome Gene/Protein Database is also available.

3 Functionality

There are some key considerations that determine the applicability of different data sources in the metabolomics workflow. One key consideration is whether or not the resource is freely available (which can differ between academia and industry). Table 2 presents information on the licence and data usage policies for different databases. Additionally, only some tools offer the possibility to save searches or export results that is particularly useful in large-scale or multiplatform studies with a huge number of masses requiring annotation. A summary of these characteristics including the exact information that can be exported using different tools is given in Supporting Information Table 4. Some online tools provide Application Programming Interfaces (APIs). An API is a common language for communication between different computer systems. APIs enable search automation and integration into workflows of third-party metabolomics tools. Galaxy Workflow4metabolomics is an example of a tool where many other external metabolomics tools can be integrated through their APIs [34]. Some databases do not provide APIs (Table 2) and others are now out of service (e.g. METLIN's API has been out of service since 2011 due to security issues). APIs may be developed in different paradigms and representational state transfer is one example for constructing web services [35, 36]. Representational state transfer architecture leads to a stateless model where resources can be accessed through primitive methods such as GET or POST. Online tools that implement this API usually provide resources independently and are not used as methods for performing queries based on experimental masses. APIs can be developed for a specific programming language as shown in Table 2.

Another consideration is how user-friendly each resource is. Of course each resource can be more or less useful for a particular purpose and the assessment of each can be highly subjective, however, to provide a guide of the main practical aspects of each data source, Table 3 summarises design features, asynchronicity (lack of need for a full page reload every time the user performs an action), login requirements and ease of familiarisation for each source.

Due to the range of tools available, global characterisation is challenging without separating them by functionality. Functionality will therefore be discussed under the following

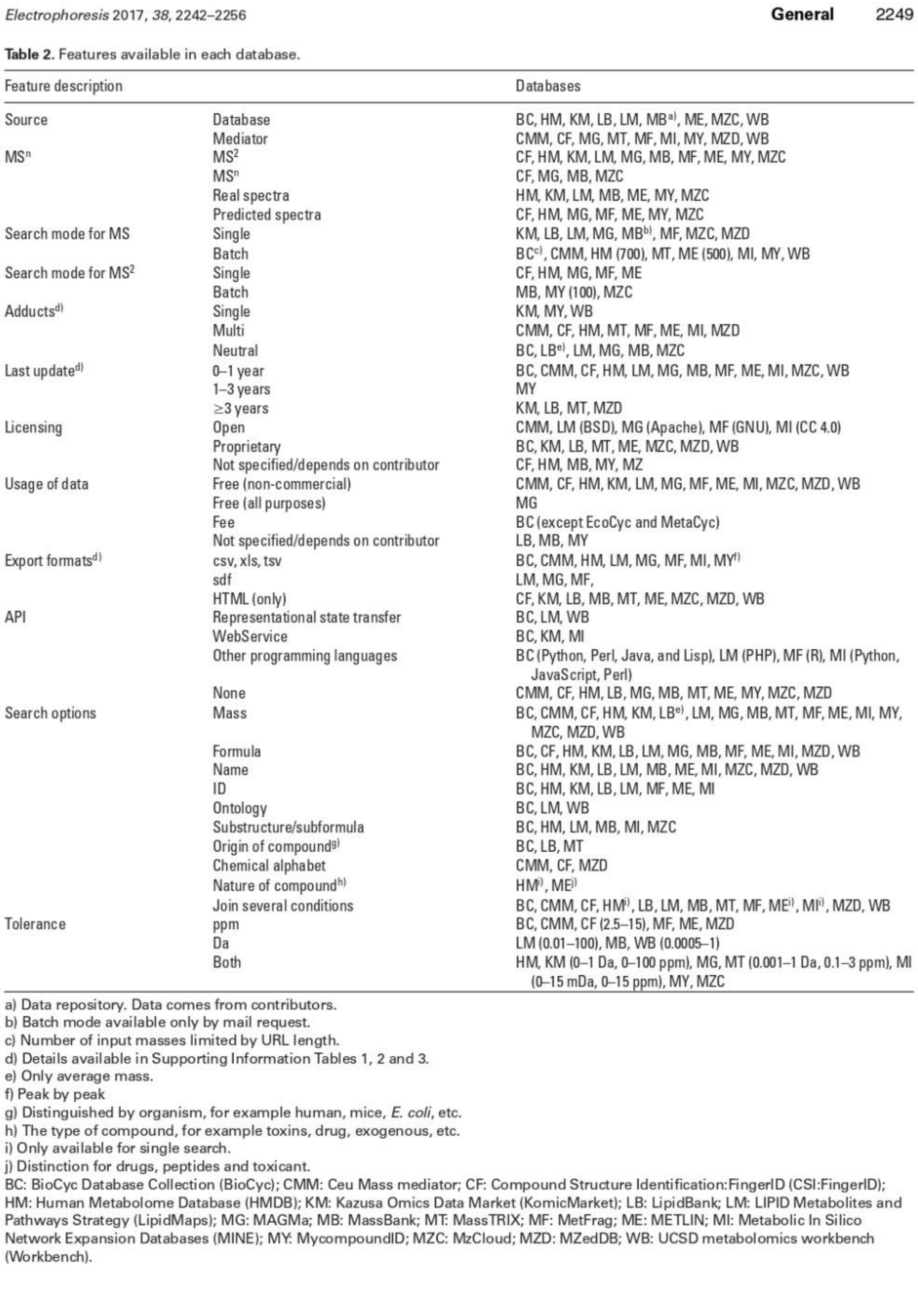


Table 2. Features available in each database.

Feature description		Databases
Source	Database	BC, HM, KM, LB, LM, MB[a], ME, MZC, WB
	Mediator	CMM, CF, MG, MT, MF, MI, MY, MZD, WB
MS^n	MS^2	CF, HM, KM, LM, MG, MB, MF, ME, MY, MZC
	MS^n	CF, MG, MB, MZC
	Real spectra	HM, KM, LM, MB, ME, MY, MZC
	Predicted spectra	CF, HM, MG, MF, ME, MY, MZC
Search mode for MS	Single	KM, LB, LM, MG, MB[b], MF, MZC, MZD
	Batch	BC[c], CMM, HM (700), MT, ME (500), MI, MY, WB
Search mode for MS^2	Single	CF, HM, MG, MF, ME
	Batch	MB, MY (100), MZC
Adducts[d]	Single	KM, MY, WB
	Multi	CMM, CF, HM, MT, MF, ME, MI, MZD
	Neutral	BC, LB[e], LM, MG, MB, MZC
Last update[d]	0–1 year	BC, CMM, CF, HM, LM, MG, MB, MF, ME, MI, MZC, WB
	1–3 years	MY
	≥3 years	KM, LB, MT, MZD
Licensing	Open	CMM, LM (BSD), MG (Apache), MF (GNU), MI (CC 4.0)
	Proprietary	BC, KM, LB, MT, ME, MZC, MZD, WB
	Not specified/depends on contributor	CF, HM, MB, MY, MZ
Usage of data	Free (non-commercial)	CMM, CF, HM, KM, LM, MG, MF, ME, MI, MZC, MZD, WB
	Free (all purposes)	MG
	Fee	BC (except EcoCyc and MetaCyc)
	Not specified/depends on contributor	LB, MB, MY
Export formats[d]	csv, xls, tsv	BC, CMM, HM, LM, MG, MF, MI, MY[f]
	sdf	LM, MG, MF,
	HTML (only)	CF, KM, LB, MB, MT, ME, MZC, MZD, WB
API	Representational state transfer	BC, LM, WB
	WebService	BC, KM, MI
	Other programming languages	BC (Python, Perl, Java, and Lisp), LM (PHP), MF (R), MI (Python, JavaScript, Perl)
	None	CMM, CF, HM, LB, MG, MB, MT, ME, MY, MZC, MZD
Search options	Mass	BC, CMM, CF, HM, KM, LB[e], LM, MG, MB, MT, MF, ME, MI, MY, MZC, MZD, WB
	Formula	BC, CF, HM, KM, LB, LM, MG, MB, MF, ME, MI, MZD, WB
	Name	BC, HM, KM, LB, LM, MB, ME, MI, MZC, MZD, WB
	ID	BC, HM, KM, LB, LM, MF, ME, MI
	Ontology	BC, LM, WB
	Substructure/subformula	BC, HM, LM, MB, MI, MZC
	Origin of compound[g]	BC, LB, MT
	Chemical alphabet	CMM, CF, MZD
	Nature of compound[h]	HM[i], ME[j]
	Join several conditions	BC, CMM, CF, HM[i], LB, LM, MB, MT, MF, ME[i], MI[i], MZD, WB
Tolerance	ppm	BC, CMM, CF (2.5–15), MF, ME, MZD
	Da	LM (0.01–100), MB, WB (0.0005–1)
	Both	HM, KM (0–1 Da, 0–100 ppm), MG, MT (0.001–1 Da, 0.1–3 ppm), MI (0–15 mDa, 0–15 ppm), MY, MZC

a) Data repository. Data comes from contributors.
b) Batch mode available only by mail request.
c) Number of input masses limited by URL length.
d) Details available in Supporting Information Tables 1, 2 and 3.
e) Only average mass.
f) Peak by peak
g) Distinguished by organism, for example human, mice, *E. coli*, etc.
h) The type of compound, for example toxins, drug, exogenous, etc.
i) Only available for single search.
j) Distinction for drugs, peptides and toxicant.
BC: BioCyc Database Collection (BioCyc); CMM: Ceu Mass mediator; CF: Compound Structure Identification:FingerID (CSI:FingerID); HM: Human Metabolome Database (HMDB); KM: Kazusa Omics Data Market (KomicMarket); LB: LipidBank; LM: LIPID Metabolites and Pathways Strategy (LipidMaps); MG: MAGMa; MB: MassBank; MT: MassTRIX; MF: MetFrag; ME: METLIN; MI: Metabolic In Silico Network Expansion Databases (MINE); MY: MycompoundID; MZC: MzCloud; MZD: MZedDB; WB: UCSD metabolomics workbench (Workbench).

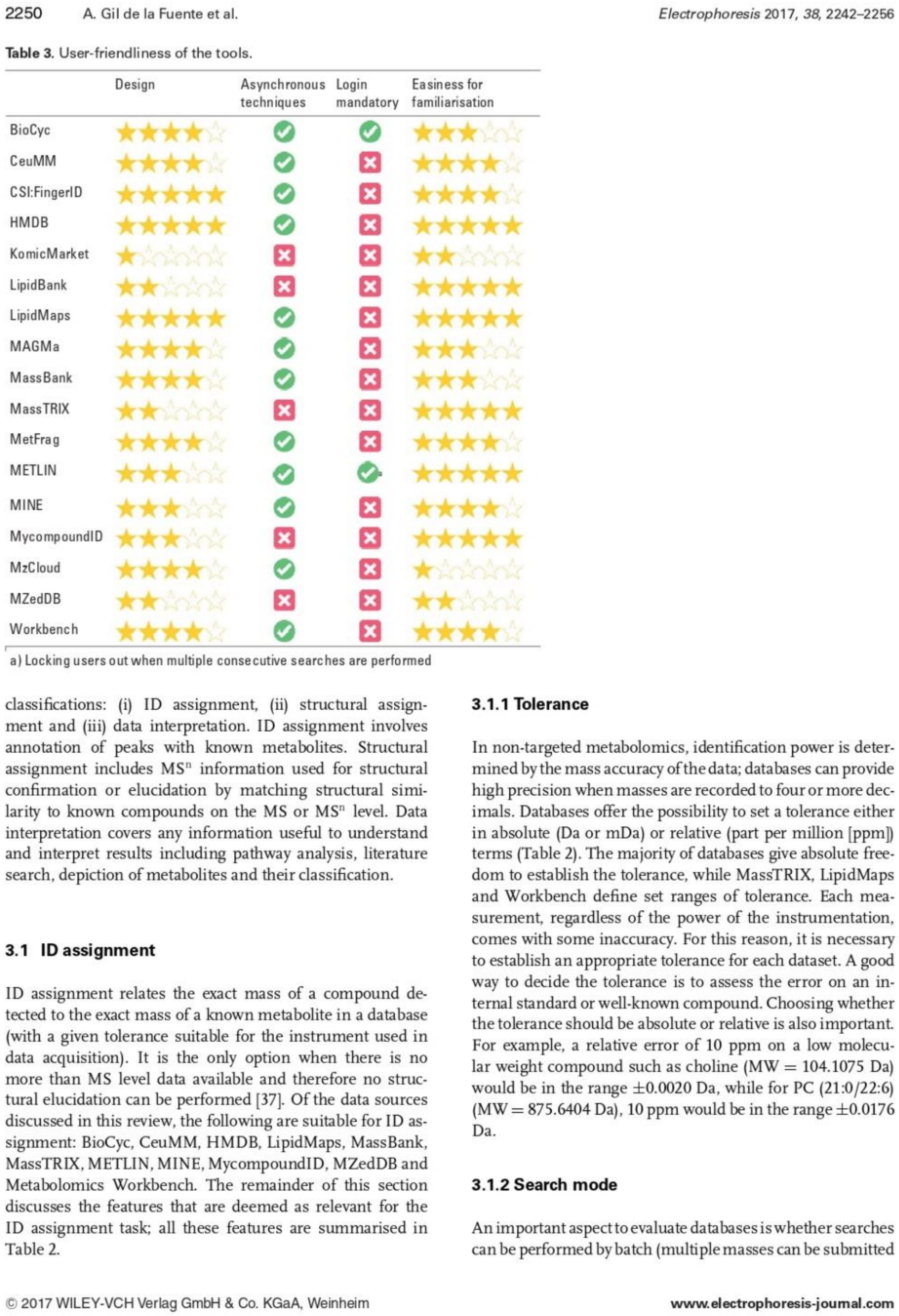


Table 3. User-friendliness of the tools.

	Design	Asynchronous techniques	Login mandatory	Easiness for familiarisation
BioCyc	★★★★☆	✓	✓	★★★☆☆
CeuMM	★★★★☆	✓	☒	★★★★☆
CSI:FingerID	★★★★★	✓	☒	★★★★☆
HMDB	★★★★★	✓	☒	★★★★★
KomicMarket	★☆☆☆☆	☒	☒	★★☆☆☆
LipidBank	★★☆☆☆	☒	☒	★★★★★
LipidMaps	★★★★★	✓	☒	★★★★★
MAGMa	★★★★☆	✓	☒	★★★☆☆
MassBank	★★★★☆	✓	☒	★★★☆☆
MassTRIX	★★☆☆☆	☒	☒	★★★★★
MetFrag	★★★★☆	✓	☒	★★★★☆
METLIN	★★★☆☆	✓	✓[a]	★★★★★
MINE	★★★☆☆	✓	☒	★★★★☆
MycompoundID	★★★☆☆	☒	☒	★★★★★
MzCloud	★★★★☆	✓	☒	★☆☆☆☆
MZedDB	★★☆☆☆	☒	☒	★★☆☆☆
Workbench	★★★★☆	✓	☒	★★★★☆

a) Locking users out when multiple consecutive searches are performed

classifications: (i) ID assignment, (ii) structural assignment and (iii) data interpretation. ID assignment involves annotation of peaks with known metabolites. Structural assignment includes MS^n information used for structural confirmation or elucidation by matching structural similarity to known compounds on the MS or MS^n level. Data interpretation covers any information useful to understand and interpret results including pathway analysis, literature search, depiction of metabolites and their classification.

3.1 ID assignment

ID assignment relates the exact mass of a compound detected to the exact mass of a known metabolite in a database (with a given tolerance suitable for the instrument used in data acquisition). It is the only option when there is no more than MS level data available and therefore no structural elucidation can be performed [37]. Of the data sources discussed in this review, the following are suitable for ID assignment: BioCyc, CeuMM, HMDB, LipidMaps, MassBank, MassTRIX, METLIN, MINE, MycompoundID, MZedDB and Metabolomics Workbench. The remainder of this section discusses the features that are deemed as relevant for the ID assignment task; all these features are summarised in Table 2.

3.1.1 Tolerance

In non-targeted metabolomics, identification power is determined by the mass accuracy of the data; databases can provide high precision when masses are recorded to four or more decimals. Databases offer the possibility to set a tolerance either in absolute (Da or mDa) or relative (part per million [ppm]) terms (Table 2). The majority of databases give absolute freedom to establish the tolerance, while MassTRIX, LipidMaps and Workbench define set ranges of tolerance. Each measurement, regardless of the power of the instrumentation, comes with some inaccuracy. For this reason, it is necessary to establish an appropriate tolerance for each dataset. A good way to decide the tolerance is to assess the error on an internal standard or well-known compound. Choosing whether the tolerance should be absolute or relative is also important. For example, a relative error of 10 ppm on a low molecular weight compound such as choline (MW = 104.1075 Da) would be in the range ±0.0020 Da, while for PC (21:0/22:6) (MW = 875.6404 Da), 10 ppm would be in the range ±0.0176 Da.

3.1.2 Search mode

An important aspect to evaluate databases is whether searches can be performed by batch (multiple masses can be submitted simultaneously) or only single searches are permitted. Manually querying hits mass by mass can be tedious and repetitive if not impractical.

3.1.3 Adducts

During the process of ionisation using ESI, adducts that alter the detected mass of the metabolite can be formed [38]. Working in positive mode, the most common adduct formations are as follows: $[M + H]^+$, $[M + Na]^+$, $[M + NH_4]^+$ and $[M + H - H_2O]^+$ and in negative mode: $[M - H]^-$, $[M + HCOO]^-$, $[M + Cl]^-$ and $[M - H - H_2O]^-$ [39]. A great deal of time can be saved with the option of searching multiple adducts and multimers simultaneously [32]. This is of particular importance for datasets obtained using high sensitivity equipment, where different adducts are detected, even those with very low abundance. This plays an even more relevant role when multi-signals originating from a single molecule are not combined into single values during data reprocessing. On inspection of the data sources, three types of search can be distinguished: neutral mass search only, *m/z* search for a single adduct and *m/z* for multi-adducts. Information on the search mode for each database is presented in Table 2 and a detailed list of possible adducts is given in Supporting Information Table 5. Lipids are best identified by their *m/z* and applying knowledge about possible ionisation and adduct formations in order to select adequate hits. By ordering these possible hits by retention time, different adducts corresponding to the same molecule can easily be identified. Moreover, this method allows the identification of mis-assignments considering the chemical properties and elution order. It is important though, when selecting possible adducts for ID assignment, only to allow those expected to minimise the risk of mis-assignment. Small molecules and acids should be also searched considering possible in-source fragmentation with the most common neutral loss of water [40, 41]. Some databases, for example MZedDB, offer the option to select multi-adducts following a list of defined rules regarding adduct formation [39] (putative ionisation product tab). These rules were established considering aspects such as the number of particular elements or chemical groups in a molecule (-OH, -COOH, $-NH_2$, etc.), the number of electrons or charges and information on non-covalently bound products and solvents.

Although there are no online tools that can combine metabolic features split by multi-adducts, some tools (e.g. METLIN) do offer the option to calculate the mass of different adducts, multimers and charges for any given compound. Similar options are also offered in LipidBank, LipidMaps and Metabolomics Workbench where *m/z* value is given for single adduct. In MZedDB, even when there is no compound listed for an exact mass in the database, the generated chemical formula can be used to predict *m/z* values for different adducts (adduct manipulation tab).

The possibility for batch searching and searching considering multi-adducts are of vital importance when considering the usefulness of a resource. Figure 4 depicts these functions for the different data sources considered in this review.

3.1.4 Exporting options

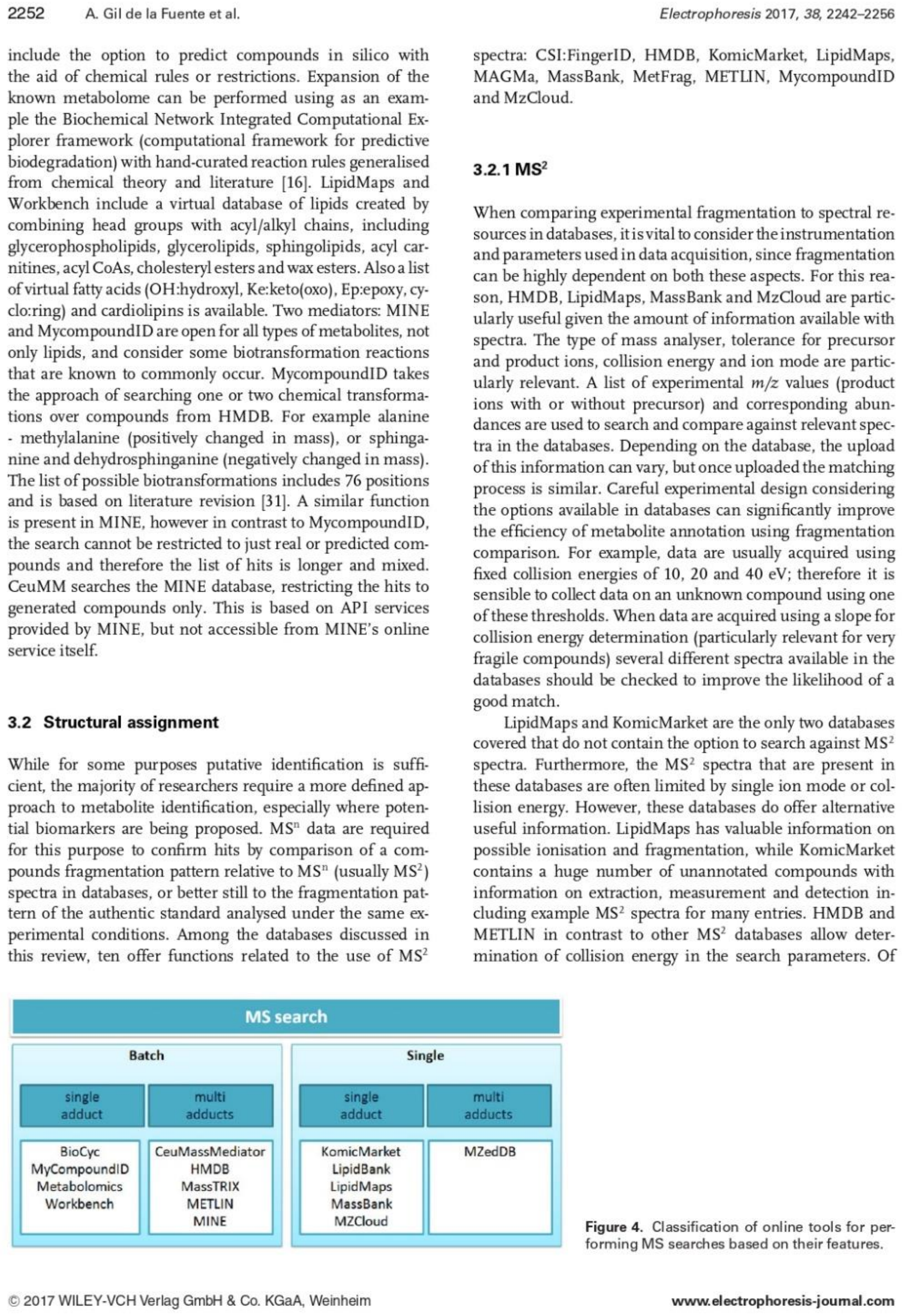
The purpose of ID assignment can be to provide a quick putative hit for detected masses, or to generate a longer list of options that can be later used in ID confirmation by MS^n analysis. Regardless of the purpose, the list of hits should be easily exportable. Most of the databases offer the possibility to save search results in a chosen format, for example csv, xls or sdf. KomicMarket, LipidBank, MassBank, MassTRIX, METLIN, MzCloud, MZedDB and Workbench do not offer automatic data download options for MS searches, thus results must be manually copied from the webpage. Workbench offers the option to save results but only one compound at a time which can render it ineffective for larger datasets.

3.1.5 Filters

The number of hits for any given search mass can be quite high. Careful filtration of this list to reduce the number of plausible hits is required. This filtration is generally performed manually, however CeuMM, CSI:FingerID and MZedDB offer functions to aid this process by restricting hits based on chemical alphabet (a list of elements selected based on expectation in given samples) or by restricting or including halogens and metals in the hits based on expectation. LipidMaps, BioCyc and Workbench offer the alternative option of allowing selection of expected compound classes (e.g. lipids, carnitines, amino acids), and excluding all other hits in order to filter the number of matches. LipidMaps, by definition, searches only lipids and related compounds, however it is possible to restrict the search to a particular class, category, or chemical composition in the ontology section (e.g. considering number of carbons, double bonds, rings or particular functional groups). MzCloud offers a useful list of filter categories (see list 1 in Supporting Information) to aid both MS and MS^n searches. One relevant possibility is to exclude some compounds from it, an option also present in METLIN. Since most databases were constructed considering utility in human studies [15], the option to restrict certain types of compound can be particularly useful when using databases for different (model) organisms with a more controlled metabolome [42]. Such options are possible in BioCyc, LipidBank and MassTRIX, where the former two use different data sources based on the restrictions and the latter highlights more plausible hits by organism selection in the output.

3.1.6 In silico compounds

Since ID assignment is restricted to available database entries, many experimental masses can be left unannotated after a search. As a solution to this, some databases now



include the option to predict compounds in silico with the aid of chemical rules or restrictions. Expansion of the known metabolome can be performed using as an example the Biochemical Network Integrated Computational Explorer framework (computational framework for predictive biodegradation) with hand-curated reaction rules generalised from chemical theory and literature [16]. LipidMaps and Workbench include a virtual database of lipids created by combining head groups with acyl/alkyl chains, including glycerophospholipids, glycerolipids, sphingolipids, acyl carnitines, acyl CoAs, cholesteryl esters and wax esters. Also a list of virtual fatty acids (OH:hydroxyl, Ke:keto(oxo), Ep:epoxy, cyclo:ring) and cardiolipins is available. Two mediators: MINE and MycompoundID are open for all types of metabolites, not only lipids, and consider some biotransformation reactions that are known to commonly occur. MycompoundID takes the approach of searching one or two chemical transformations over compounds from HMDB. For example alanine - methylalanine (positively changed in mass), or sphinganine and dehydrosphinganine (negatively changed in mass). The list of possible biotransformations includes 76 positions and is based on literature revision [31]. A similar function is present in MINE, however in contrast to MycompoundID, the search cannot be restricted to just real or predicted compounds and therefore the list of hits is longer and mixed. CeuMM searches the MINE database, restricting the hits to generated compounds only. This is based on API services provided by MINE, but not accessible from MINE's online service itself.

3.2 Structural assignment

While for some purposes putative identification is sufficient, the majority of researchers require a more defined approach to metabolite identification, especially where potential biomarkers are being proposed. MS^n data are required for this purpose to confirm hits by comparison of a compounds fragmentation pattern relative to MS^n (usually MS^2) spectra in databases, or better still to the fragmentation pattern of the authentic standard analysed under the same experimental conditions. Among the databases discussed in this review, ten offer functions related to the use of MS^2 spectra: CSI:FingerID, HMDB, KomicMarket, LipidMaps, MAGMa, MassBank, MetFrag, METLIN, MycompoundID and MzCloud.

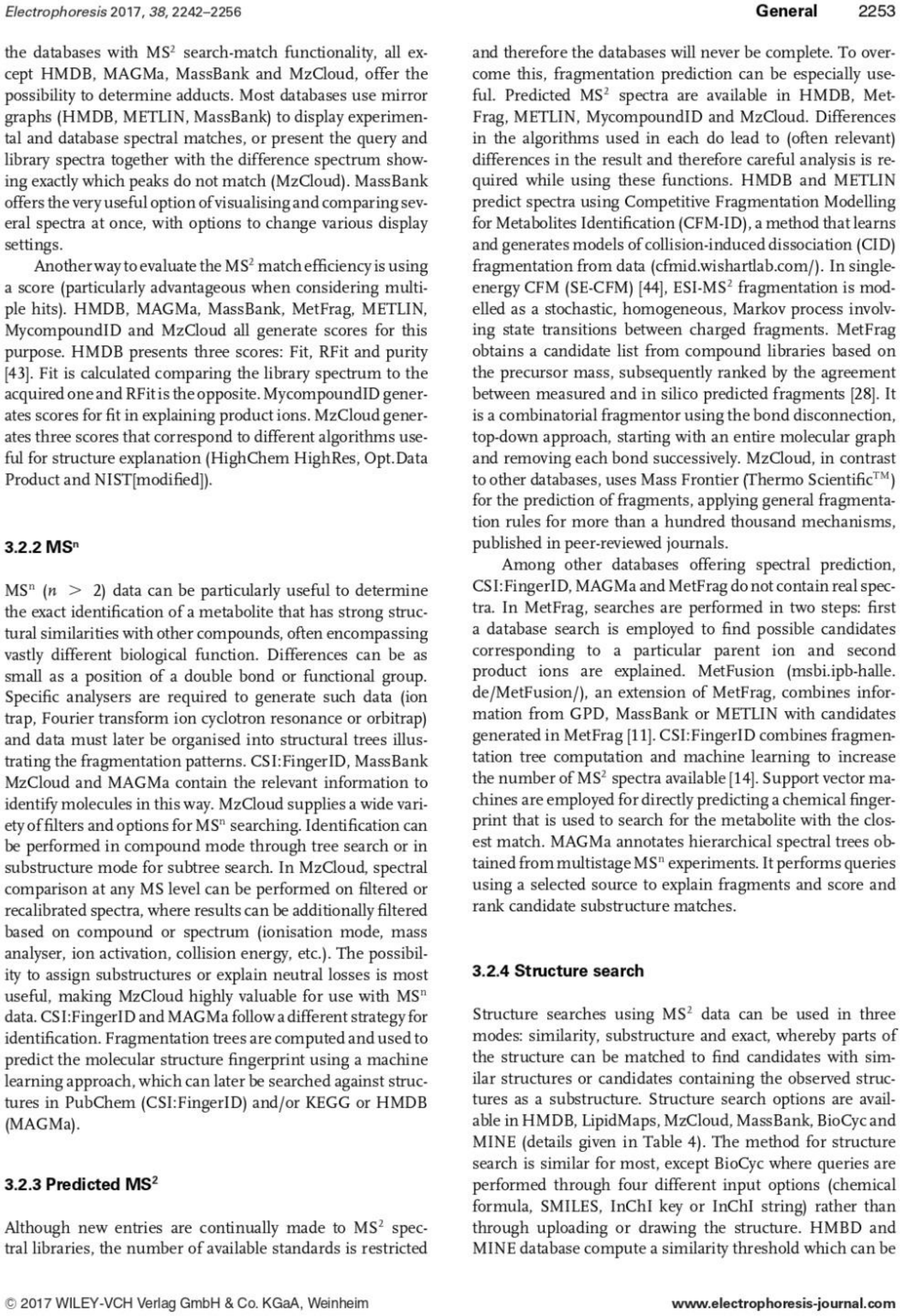
3.2.1 MS^2

When comparing experimental fragmentation to spectral resources in databases, it is vital to consider the instrumentation and parameters used in data acquisition, since fragmentation can be highly dependent on both these aspects. For this reason, HMDB, LipidMaps, MassBank and MzCloud are particularly useful given the amount of information available with spectra. The type of mass analyser, tolerance for precursor and product ions, collision energy and ion mode are particularly relevant. A list of experimental *m/z* values (product ions with or without precursor) and corresponding abundances are used to search and compare against relevant spectra in the databases. Depending on the database, the upload of this information can vary, but once uploaded the matching process is similar. Careful experimental design considering the options available in databases can significantly improve the efficiency of metabolite annotation using fragmentation comparison. For example, data are usually acquired using fixed collision energies of 10, 20 and 40 eV; therefore it is sensible to collect data on an unknown compound using one of these thresholds. When data are acquired using a slope for collision energy determination (particularly relevant for very fragile compounds) several different spectra available in the databases should be checked to improve the likelihood of a good match.

LipidMaps and KomicMarket are the only two databases covered that do not contain the option to search against MS^2 spectra. Furthermore, the MS^2 spectra that are present in these databases are often limited by single ion mode or collision energy. However, these databases do offer alternative useful information. LipidMaps has valuable information on possible ionisation and fragmentation, while KomicMarket contains a huge number of unannotated compounds with information on extraction, measurement and detection including example MS^2 spectra for many entries. HMDB and METLIN in contrast to other MS^2 databases allow determination of collision energy in the search parameters. Of

MS search			
Batch		Single	
single adduct	multi adducts	single adduct	multi adducts
BioCyc MyCompoundID Metabolomics Workbench	CeuMassMediator HMDB MassTRIX METLIN MINE	KomicMarket LipidBank LipidMaps MassBank MZCloud	MZedDB

Figure 4. Classification of online tools for performing MS searches based on their features.



the databases with MS^2 search-match functionality, all except HMDB, MAGMa, MassBank and MzCloud, offer the possibility to determine adducts. Most databases use mirror graphs (HMDB, METLIN, MassBank) to display experimental and database spectral matches, or present the query and library spectra together with the difference spectrum showing exactly which peaks do not match (MzCloud). MassBank offers the very useful option of visualising and comparing several spectra at once, with options to change various display settings.

Another way to evaluate the MS^2 match efficiency is using a score (particularly advantageous when considering multiple hits). HMDB, MAGMa, MassBank, MetFrag, METLIN, MycompoundID and MzCloud all generate scores for this purpose. HMDB presents three scores: Fit, RFit and purity [43]. Fit is calculated comparing the library spectrum to the acquired one and RFit is the opposite. MycompoundID generates scores for fit in explaining product ions. MzCloud generates three scores that correspond to different algorithms useful for structure explanation (HighChem HighRes, Opt.Data Product and NIST[modified]).

3.2.2 MS^n

MS^n ($n > 2$) data can be particularly useful to determine the exact identification of a metabolite that has strong structural similarities with other compounds, often encompassing vastly different biological function. Differences can be as small as a position of a double bond or functional group. Specific analysers are required to generate such data (ion trap, Fourier transform ion cyclotron resonance or orbitrap) and data must later be organised into structural trees illustrating the fragmentation patterns. CSI:FingerID, MassBank MzCloud and MAGMa contain the relevant information to identify molecules in this way. MzCloud supplies a wide variety of filters and options for MS^n searching. Identification can be performed in compound mode through tree search or in substructure mode for subtree search. In MzCloud, spectral comparison at any MS level can be performed on filtered or recalibrated spectra, where results can be additionally filtered based on compound or spectrum (ionisation mode, mass analyser, ion activation, collision energy, etc.). The possibility to assign substructures or explain neutral losses is most useful, making MzCloud highly valuable for use with MS^n data. CSI:FingerID and MAGMa follow a different strategy for identification. Fragmentation trees are computed and used to predict the molecular structure fingerprint using a machine learning approach, which can later be searched against structures in PubChem (CSI:FingerID) and/or KEGG or HMDB (MAGMa).

3.2.3 Predicted MS^2

Although new entries are continually made to MS^2 spectral libraries, the number of available standards is restricted and therefore the databases will never be complete. To overcome this, fragmentation prediction can be especially useful. Predicted MS^2 spectra are available in HMDB, MetFrag, METLIN, MycompoundID and MzCloud. Differences in the algorithms used in each do lead to (often relevant) differences in the result and therefore careful analysis is required while using these functions. HMDB and METLIN predict spectra using Competitive Fragmentation Modelling for Metabolites Identification (CFM-ID), a method that learns and generates models of collision-induced dissociation (CID) fragmentation from data (cfmid.wishartlab.com/). In single-energy CFM (SE-CFM) [44], ESI-MS^2 fragmentation is modelled as a stochastic, homogeneous, Markov process involving state transitions between charged fragments. MetFrag obtains a candidate list from compound libraries based on the precursor mass, subsequently ranked by the agreement between measured and in silico predicted fragments [28]. It is a combinatorial fragmentor using the bond disconnection, top-down approach, starting with an entire molecular graph and removing each bond successively. MzCloud, in contrast to other databases, uses Mass Frontier (Thermo Scientific^{TM}) for the prediction of fragments, applying general fragmentation rules for more than a hundred thousand mechanisms, published in peer-reviewed journals.

Among other databases offering spectral prediction, CSI:FingerID, MAGMa and MetFrag do not contain real spectra. In MetFrag, searches are performed in two steps: first a database search is employed to find possible candidates corresponding to a particular parent ion and second product ions are explained. MetFusion (msbi.ipb-halle.de/MetFusion/), an extension of MetFrag, combines information from GPD, MassBank or METLIN with candidates generated in MetFrag [11]. CSI:FingerID combines fragmentation tree computation and machine learning to increase the number of MS^2 spectra available [14]. Support vector machines are employed for directly predicting a chemical fingerprint that is used to search for the metabolite with the closest match. MAGMa annotates hierarchical spectral trees obtained from multistage MS^n experiments. It performs queries using a selected source to explain fragments and score and rank candidate substructure matches.

3.2.4 Structure search

Structure searches using MS^2 data can be used in three modes: similarity, substructure and exact, whereby parts of the structure can be matched to find candidates with similar structures or candidates containing the observed structures as a substructure. Structure search options are available in HMDB, LipidMaps, MzCloud, MassBank, BioCyc and MINE (details given in Table 4). The method for structure search is similar for most, except BioCyc where queries are performed through four different input options (chemical formula, SMILES, InChI key or InChI string) rather than through uploading or drawing the structure. HMBD and MINE database compute a similarity threshold which can be

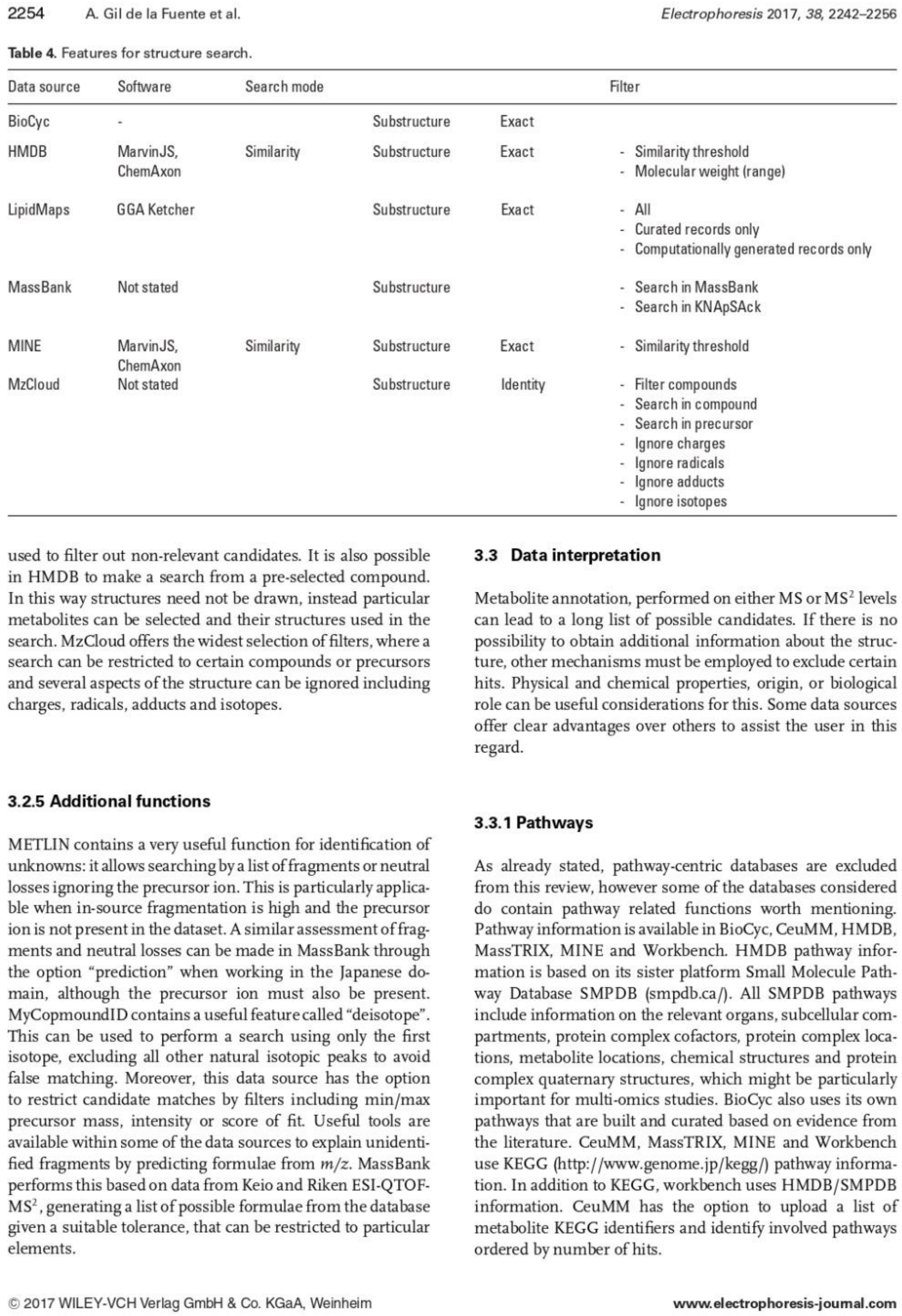


Table 4. Features for structure search.

Data source	Software	Search mode			Filter
BioCyc	-		Substructure	Exact	
HMDB	MarvinJS, ChemAxon	Similarity	Substructure	Exact	- Similarity threshold - Molecular weight (range)
LipidMaps	GGA Ketcher		Substructure	Exact	- All - Curated records only - Computationally generated records only
MassBank	Not stated		Substructure		- Search in MassBank - Search in KNApSAck
MINE	MarvinJS, ChemAxon	Similarity	Substructure	Exact	- Similarity threshold
MzCloud	Not stated		Substructure	Identity	- Filter compounds - Search in compound - Search in precursor - Ignore charges - Ignore radicals - Ignore adducts - Ignore isotopes

used to filter out non-relevant candidates. It is also possible in HMDB to make a search from a pre-selected compound. In this way structures need not be drawn, instead particular metabolites can be selected and their structures used in the search. MzCloud offers the widest selection of filters, where a search can be restricted to certain compounds or precursors and several aspects of the structure can be ignored including charges, radicals, adducts and isotopes.

3.2.5 Additional functions

METLIN contains a very useful function for identification of unknowns: it allows searching by a list of fragments or neutral losses ignoring the precursor ion. This is particularly applicable when in-source fragmentation is high and the precursor ion is not present in the dataset. A similar assessment of fragments and neutral losses can be made in MassBank through the option "prediction" when working in the Japanese domain, although the precursor ion must also be present. MyCopmoundID contains a useful feature called "deisotope". This can be used to perform a search using only the first isotope, excluding all other natural isotopic peaks to avoid false matching. Moreover, this data source has the option to restrict candidate matches by filters including min/max precursor mass, intensity or score of fit. Useful tools are available within some of the data sources to explain unidentified fragments by predicting formulae from m/z. MassBank performs this based on data from Keio and Riken ESI-QTOF-MS^2, generating a list of possible formulae from the database given a suitable tolerance, that can be restricted to particular elements.

3.3 Data interpretation

Metabolite annotation, performed on either MS or MS^2 levels can lead to a long list of possible candidates. If there is no possibility to obtain additional information about the structure, other mechanisms must be employed to exclude certain hits. Physical and chemical properties, origin, or biological role can be useful considerations for this. Some data sources offer clear advantages over others to assist the user in this regard.

3.3.1 Pathways

As already stated, pathway-centric databases are excluded from this review, however some of the databases considered do contain pathway related functions worth mentioning. Pathway information is available in BioCyc, CeuMM, HMDB, MassTRIX, MINE and Workbench. HMDB pathway information is based on its sister platform Small Molecule Pathway Database SMPDB (smpdb.ca/). All SMPDB pathways include information on the relevant organs, subcellular compartments, protein complex cofactors, protein complex locations, metabolite locations, chemical structures and protein complex quaternary structures, which might be particularly important for multi-omics studies. BioCyc also uses its own pathways that are built and curated based on evidence from the literature. CeuMM, MassTRIX, MINE and Workbench use KEGG (http://www.genome.jp/kegg/) pathway information. In addition to KEGG, workbench uses HMDB/SMPDB information. CeuMM has the option to upload a list of metabolite KEGG identifiers and identify involved pathways ordered by number of hits.

3.3.2 Description and classification

HMDB provides a great deal of information about each metabolite entry. This information is stored in a "metabocard" that details the taxonomy, ontology, physical, chemical and biological properties, spectra, expected physiological concentrations, literature references and appropriate links. LipidBank also contains very useful information for data interpretation, including genetic, bioactivity and metabolic data in addition to literature references, and Workbench provides literature references too. Supporting Information Table 6 details the information provided in each data source.

Classification approaches can be used to help filter or interpret hits given in databases using a forest or tree approach, for which taxonomy and ontology can be useful [45]. These data are available in BioCyc and HMDB for all metabolites and in LipidMaps and Workbench for lipids only, calling on LipidMaps whose nomenclature is the recognised standard for lipid classification. BioCyc includes additional useful information including metabolic reactions in which metabolites are involved or information on their presence or abundance in culture medium, for example. This is particularly useful when considering the plausibility of a metabolite as a statistically significant feature of a study and can also be useful in the experimental design stage to choose certain experimental conditions if there are particular metabolites of interest that may be affected by that. Similarly, MINE provides information about enzymes and products of reactions in which metabolites are involved.

Workbench contains information about previous projects and research where particular metabolites were already found. The highly detailed data include a further description of project, samples used, conditions applied and treatments and analytical conditions employed. Even measured abundances for particular masses across all the samples are stated.

4 Conclusions

Data analysis is a critical, but often an under-considered aspect of metabolomics research. In general, close to 50% of features detected in a non-targeted metabolomics study are unidentified compounds, leading to an important loss of information. Moreover, if features are mis-identified, data are wrongly interpreted and false conclusions are drawn onto which new experiments can be proposed. It is therefore vital to get this step right and be aware of the advantages and limitations of the tools at our disposal. As discussed, there is a range of different open access resources, with different characteristics that have been critically reviewed here. On-line tools will benefit from the input of a broad spectrum of scientists interested in metabolomics. However, the community as a whole should contribute to establish rules about data collected using different extraction protocols and analytical methods.

This work was supported by grants from the Spanish Ministerio de Economía y Competitividad (grant CTQ2014-55279-R) and USPCEU PCON10/2016. AGF acknowledges Fundación Universitaria San Pablo CEU for his PhD fellowship.

The authors have declared no conflict of interest.

5 References

[1] Dunn, W. B., Lin, W., Broadhurst, D., Begley, P., Brown, M., Zelena, E., Vaughan, A. A., Halsall, A., Harding, N., Knowles, J. D., Francis-McIntyre, S., Tseng, A., Ellis, D. I., O'Hagan, S., Aarons, G., Benjamin, B., Chew-Graham, S., Moseley, C., Potter, P., Winder, C. L., Potts, C., Thornton, P., McWhirter, C., Zubair, M., Pan, M., Burns, A., Cruickshank, J. K., Jayson, G. C., Purandare, N., Wu, F. C. W., Finn, J. D., Haselden, J. N., Nicholls, A. W., Wilson, I. D., Goodacre, R., Kell, D. B., *Metabolomics* 2015, *11*, 9–26.

[2] Peng, B., Li, H., Peng, X.-X., *Protein Cell* 2015, *6*, 628–637.

[3] Zhang, A., Sun, H., Yan, G., Wang, P., Wang, X., *Biomed. Res. Int.* 2015, *2015*, 1–6.

[4] Johnson, C. H., Ivanisevic, J., Siuzdak, G., *Nat. Rev. Mol. Cell Biol.* 2016, *17*, 451–459.

[5] Dunn, W. B., Ellis, D. I., *Trends Anal. Chem.* 2005, *24*, 285–294.

[6] Cao, M., Fraser, K., Rasmussen, S., *Metabolites* 2013, *3*, 1036–1050.

[7] Werner, E., Heilier, J.-F., Ducruix, C., Ezan, E., Junot, C., Tabet, J.-C., *J. Chromatogr. B* 2008, *871*, 143–163.

[8] Godzien, J., Ciborowski, M., Angulo, S., Barbas, C., *Electrophoresis* 2013, *34*, 2812–2826.

[9] Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., Fan, T. W., Fiehn, O., Goodacre, R., Griffin, J. L., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A. N., Lindon, J. C., Marriott, P., Nicholls, A. W., Reily, M. D., Thaden, J. J., Viant, M. R., *Metabolomics* 2007, *3*, 211–221.

[10] Bingol, K., Bruschweiler-Li, L., Li, D., Zhang, B., Xie, M., Brüschweiler, R., *Bioanalysis* 2016, *8*, 557–573.

[11] Gerlich, M., Neumann, S., *J. Mass Spectrom.* 2013, *48*, 291–298.

[12] Hufsky, F., Rempt, M., Rasche, F., Pohnert, G., Böcker, S., *Anal. Chim. Acta* 2012, *739*, 67–76.

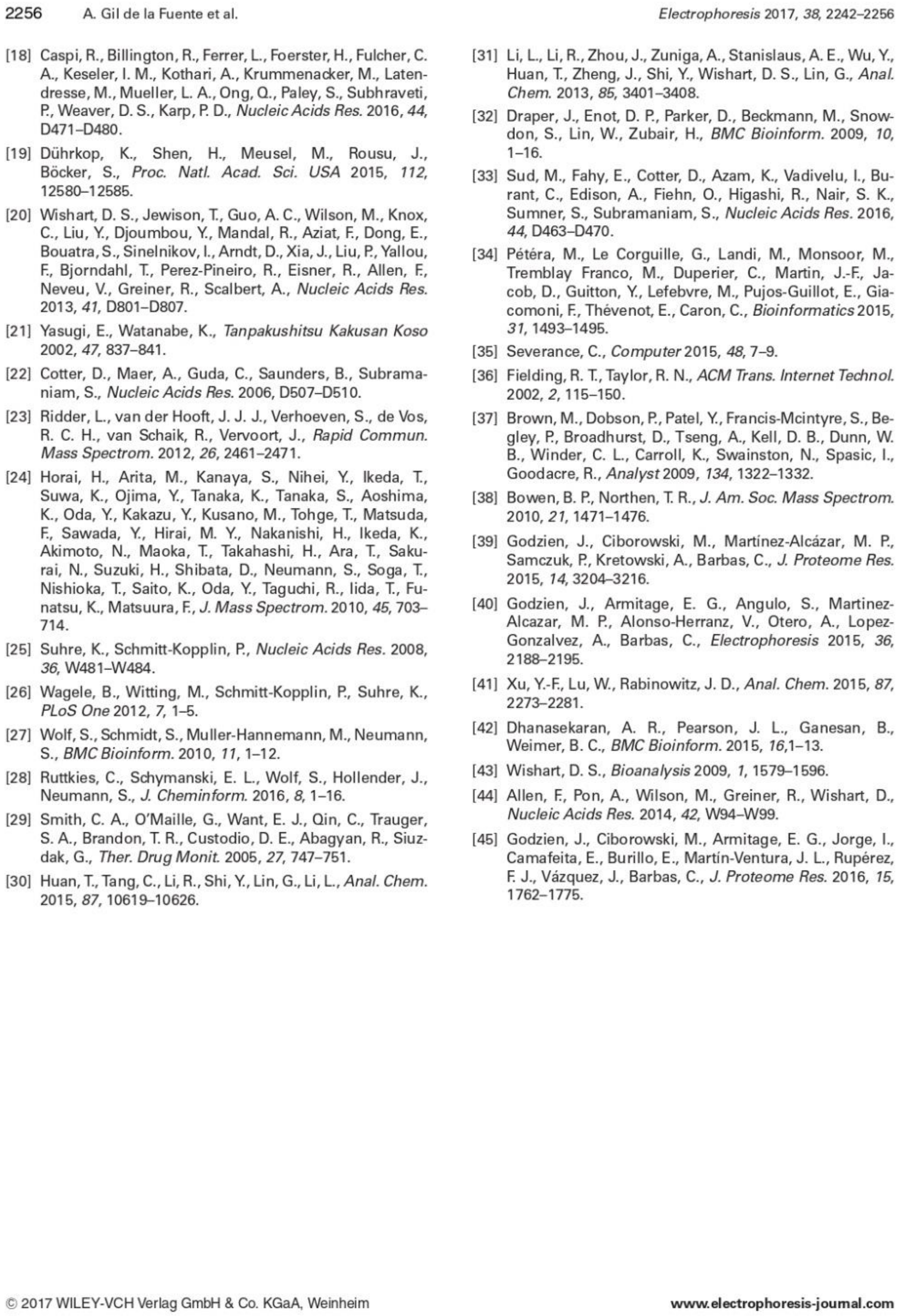
[13] Rojas-Cherto, M., Peironcely, J. E., Kasper, P. T., Van Der Hooft, J. J. J., De Vos, R. C. H., Vreeken, R., Hankemeier, T., Reijmers, T., *Anal. Chem.* 2012, *84*, 5524–5534.

[14] Shen, H. B., Duhrkop, K., Bocker, S., Rousu, J., *Bioinformatics* 2014, *30*, 157–164.

[15] Vinaixa, M., Schymanski, E., Neumann, S., Navarro, M., Salek, R., Yanes, O., *Trends Anal. Chem.* 2016, *78*, 23–35.

[16] Jeffryes, J. G., Colastani, R. L., Elbadawi-Sidhu, M., Kind, T., Niehaus, T. D., Broadbelt, L. J., Hanson, A. D., Fiehn, O., Tyo, K. E. J., Henry, C. S., *J. Cheminform.* 2015, *7*, 1–8.

[17] Fiehn, O., Barupal, D. K., Kind, T., *J. Biol. Chem.* 2011, *286*, 23637–23643.



[18] Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C. A., Keseler, I. M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D. S., Karp, P. D., *Nucleic Acids Res.* 2016, *44*, D471–D480.

[19] Dührkop, K., Shen, H., Meusel, M., Rousu, J., Böcker, S., *Proc. Natl. Acad. Sci. USA* 2015, *112*, 12580–12585.

[20] Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., Bouatra, S., Sinelnikov, I., Arndt, D., Xia, J., Liu, P., Yallou, F., Bjorndahl, T., Perez-Pineiro, R., Eisner, R., Allen, F., Neveu, V., Greiner, R., Scalbert, A., *Nucleic Acids Res.* 2013, *41*, D801–D807.

[21] Yasugi, E., Watanabe, K., *Tanpakushitsu Kakusan Koso* 2002, *47*, 837–841.

[22] Cotter, D., Maer, A., Guda, C., Saunders, B., Subramaniam, S., *Nucleic Acids Res.* 2006, D507–D510.

[23] Ridder, L., van der Hooft, J. J. J., Verhoeven, S., de Vos, R. C. H., van Schaik, R., Vervoort, J., *Rapid Commun. Mass Spectrom.* 2012, *26*, 2461–2471.

[24] Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., Oda, Y., Kakazu, Y., Kusano, M., Tohge, T., Matsuda, F., Sawada, Y., Hirai, M. Y., Nakanishi, H., Ikeda, K., Akimoto, N., Maoka, T., Takahashi, H., Ara, T., Sakurai, N., Suzuki, H., Shibata, D., Neumann, S., Soga, T., Nishioka, T., Saito, K., Oda, Y., Taguchi, R., Iida, T., Funatsu, K., Matsuura, F., *J. Mass Spectrom.* 2010, *45*, 703–714.

[25] Suhre, K., Schmitt-Kopplin, P., *Nucleic Acids Res.* 2008, *36*, W481–W484.

[26] Wagele, B., Witting, M., Schmitt-Kopplin, P., Suhre, K., *PLoS One* 2012, *7*, 1–5.

[27] Wolf, S., Schmidt, S., Muller-Hannemann, M., Neumann, S., *BMC Bioinform.* 2010, *11*, 1–12.

[28] Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J., Neumann, S., *J. Cheminform.* 2016, *8*, 1–16.

[29] Smith, C. A., O'Maille, G., Want, E. J., Qin, C., Trauger, S. A., Brandon, T. R., Custodio, D. E., Abagyan, R., Siuzdak, G., *Ther. Drug Monit.* 2005, *27*, 747–751.

[30] Huan, T., Tang, C., Li, R., Shi, Y., Lin, G., Li, L., *Anal. Chem.* 2015, *87*, 10619–10626.

[31] Li, L., Li, R., Zhou, J., Zuniga, A., Stanislaus, A. E., Wu, Y., Huan, T., Zheng, J., Shi, Y., Wishart, D. S., Lin, G., *Anal. Chem.* 2013, *85*, 3401–3408.

[32] Draper, J., Enot, D. P., Parker, D., Beckmann, M., Snowdon, S., Lin, W., Zubair, H., *BMC Bioinform.* 2009, *10*, 1–16.

[33] Sud, M., Fahy, E., Cotter, D., Azam, K., Vadivelu, I., Burant, C., Edison, A., Fiehn, O., Higashi, R., Nair, S. K., Sumner, S., Subramaniam, S., *Nucleic Acids Res.* 2016, *44*, D463–D470.

[34] Pétéra, M., Le Corguille, G., Landi, M., Monsoor, M., Tremblay Franco, M., Duperier, C., Martin, J.-F., Jacob, D., Guitton, Y., Lefebvre, M., Pujos-Guillot, E., Giacomoni, F., Thévenot, E., Caron, C., *Bioinformatics* 2015, *31*, 1493–1495.

[35] Severance, C., *Computer* 2015, *48*, 7–9.

[36] Fielding, R. T., Taylor, R. N., *ACM Trans. Internet Technol.* 2002, *2*, 115–150.

[37] Brown, M., Dobson, P., Patel, Y., Francis-Mcintyre, S., Begley, P., Broadhurst, D., Tseng, A., Kell, D. B., Dunn, W. B., Winder, C. L., Carroll, K., Swainston, N., Spasic, I., Goodacre, R., *Analyst* 2009, *134*, 1322–1332.

[38] Bowen, B. P., Northen, T. R., *J. Am. Soc. Mass Spectrom.* 2010, *21*, 1471–1476.

[39] Godzien, J., Ciborowski, M., Martínez-Alcázar, M. P., Samczuk, P., Kretowski, A., Barbas, C., *J. Proteome Res.* 2015, *14*, 3204–3216.

[40] Godzien, J., Armitage, E. G., Angulo, S., Martinez-Alcazar, M. P., Alonso-Herranz, V., Otero, A., Lopez-Gonzalvez, A., Barbas, C., *Electrophoresis* 2015, *36*, 2188–2195.

[41] Xu, Y.-F., Lu, W., Rabinowitz, J. D., *Anal. Chem.* 2015, *87*, 2273–2281.

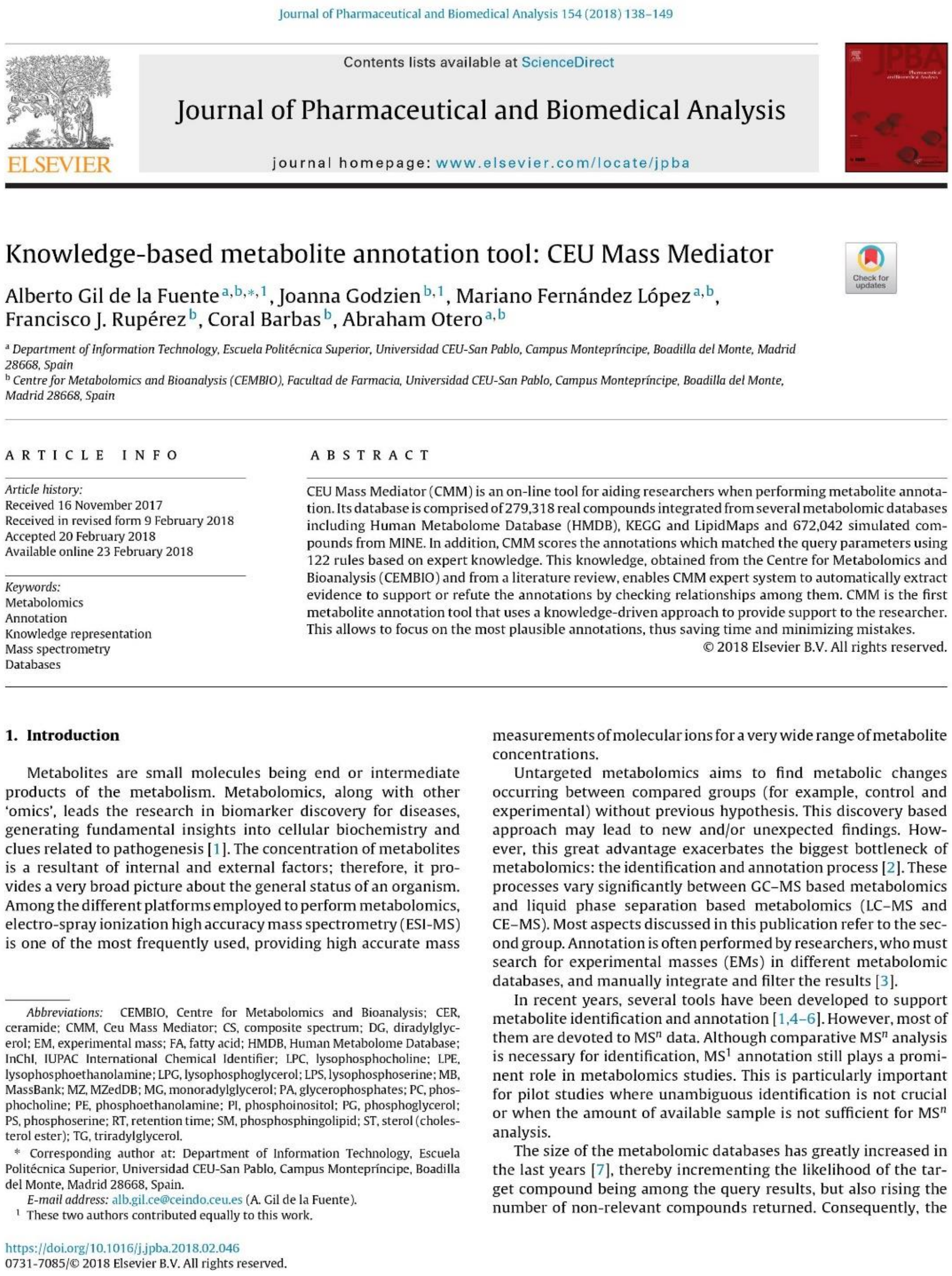
[42] Dhanasekaran, A. R., Pearson, J. L., Ganesan, B., Weimer, B. C., *BMC Bioinform.* 2015, *16*,1–13.

[43] Wishart, D. S., *Bioanalysis* 2009, *1*, 1579–1596.

[44] Allen, F., Pon, A., Wilson, M., Greiner, R., Wishart, D., *Nucleic Acids Res.* 2014, *42*, W94–W99.

[45] Godzien, J., Ciborowski, M., Armitage, E. G., Jorge, I., Camafeita, E., Burillo, E., Martín-Ventura, J. L., Rupérez, F. J., Vázquez, J., Barbas, C., *J. Proteome Res.* 2016, *15*, 1762–1775.

CHAPTER 2: KNOWLEDGE-BASED METABOLITE ANNOTATION TOOL: CEU MASS MEDIATOR



Journal of Pharmaceutical and Biomedical Analysis 154 (2018) 138–149

Contents lists available at ScienceDirect

Journal of Pharmaceutical and Biomedical Analysis

journal homepage: www.elsevier.com/locate/jpba

Knowledge-based metabolite annotation tool: CEU Mass Mediator

Alberto Gil de la Fuente[a,b,*,1], Joanna Godzien[b,1], Mariano Fernández López[a,b], Francisco J. Rupérez[b], Coral Barbas[b], Abraham Otero[a,b]

[a] *Department of Information Technology, Escuela Politécnica Superior, Universidad CEU-San Pablo, Campus Montepríncipe, Boadilla del Monte, Madrid 28668, Spain*

[b] *Centre for Metabolomics and Bioanalysis (CEMBIO), Facultad de Farmacia, Universidad CEU-San Pablo, Campus Montepríncipe, Boadilla del Monte, Madrid 28668, Spain*

ARTICLE INFO

Article history:
Received 16 November 2017
Received in revised form 9 February 2018
Accepted 20 February 2018
Available online 23 February 2018

Keywords:
Metabolomics
Annotation
Knowledge representation
Mass spectrometry
Databases

ABSTRACT

CEU Mass Mediator (CMM) is an on-line tool for aiding researchers when performing metabolite annotation. Its database is comprised of 279,318 real compounds integrated from several metabolomic databases including Human Metabolome Database (HMDB), KEGG and LipidMaps and 672,042 simulated compounds from MINE. In addition, CMM scores the annotations which matched the query parameters using 122 rules based on expert knowledge. This knowledge, obtained from the Centre for Metabolomics and Bioanalysis (CEMBIO) and from a literature review, enables CMM expert system to automatically extract evidence to support or refute the annotations by checking relationships among them. CMM is the first metabolite annotation tool that uses a knowledge-driven approach to provide support to the researcher. This allows to focus on the most plausible annotations, thus saving time and minimizing mistakes.

1. Introduction

Metabolites are small molecules being end or intermediate products of the metabolism. Metabolomics, along with other 'omics', leads the research in biomarker discovery for diseases, generating fundamental insights into cellular biochemistry and clues related to pathogenesis [1]. The concentration of metabolites is a resultant of internal and external factors; therefore, it provides a very broad picture about the general status of an organism. Among the different platforms employed to perform metabolomics, electro-spray ionization high accuracy mass spectrometry (ESI-MS) is one of the most frequently used, providing high accurate mass measurements of molecular ions for a very wide range of metabolite concentrations.

Untargeted metabolomics aims to find metabolic changes occurring between compared groups (for example, control and experimental) without previous hypothesis. This discovery based approach may lead to new and/or unexpected findings. However, this great advantage exacerbates the biggest bottleneck of metabolomics: the identification and annotation process [2]. These processes vary significantly between GC–MS based metabolomics and liquid phase separation based metabolomics (LC–MS and CE–MS). Most aspects discussed in this publication refer to the second group. Annotation is often performed by researchers, who must search for experimental masses (EMs) in different metabolomic databases, and manually integrate and filter the results [3].

In recent years, several tools have been developed to support metabolite identification and annotation [1,4–6]. However, most of them are devoted to MS^n data. Although comparative MS^n analysis is necessary for identification, MS^1 annotation still plays a prominent role in metabolomics studies. This is particularly important for pilot studies where unambiguous identification is not crucial or when the amount of available sample is not sufficient for MS^n analysis.

The size of the metabolomic databases has greatly increased in the last years [7], thereby incrementing the likelihood of the target compound being among the query results, but also rising the number of non-relevant compounds returned. Consequently, the

Abbreviations: CEMBIO, Centre for Metabolomics and Bioanalysis; CER, ceramide; CMM, Ceu Mass Mediator; CS, composite spectrum; DG, diradylglycerol; EM, experimental mass; FA, fatty acid; HMDB, Human Metabolome Database; InChI, IUPAC International Chemical Identifier; LPC, lysophosphocholine; LPE, lysophosphoethanolamine; LPG, lysophosphoglycerol; LPS, lysophosphoserine; MB, MassBank; MZ, MZedDB; MG, monoradylglycerol; PA, glycerophosphates; PC, phosphocholine; PE, phosphoethanolamine; PI, phosphoinositol; PG, phosphoglycerol; PS, phosphoserine; RT, retention time; SM, phosphosphingolipid; ST, sterol (cholesterol ester); TG, triradylglycerol.

[*] Corresponding author at: Department of Information Technology, Escuela Politécnica Superior, Universidad CEU-San Pablo, Campus Montepríncipe, Boadilla del Monte, Madrid 28668, Spain.

E-mail address: alb.gil.ce@ceindo.ceu.es (A. Gil de la Fuente).

[1] These two authors contributed equally to this work.

https://doi.org/10.1016/j.jpba.2018.02.046
0731-7085/© 2018 Elsevier B.V. All rights reserved.

already tedious task of filtering the query results is even more time-consuming. This is especially true when the analysis involves lipids because there are many isomeric and isobaric compounds.

The filtering task can be guided by the researchers' knowledge about the chemical properties of the compounds and the experimental set-up. Such filtering is performed manually and it is highly affected by the level of experience of the researcher, making it time consuming and prone to errors. Therefore, there is interest in the standardization and automation of this step. In this task, there are different types of information which can be used to support a proper annotation: information related to the analytical aspects of the measurement, to the data re-processing and to the databases themselves. This includes both separation and detection since most metabolomics studies are performed on hyphenated set-ups [8,9]. We shall present the main ones here.

1.1. Chromatographic order of elution

Retention time (RT) reflects the time that a particular molecule spends inside a column being retained by the stationary phase. This time depends on the mechanisms of retention, column geometry and temperature, instrument dwell volume, mobile phase, modifier and gradient. In the case of reversed phase chromatography, the most common, polar molecules have low interaction with the non-polar bed, therefore they elute very early. On the other hand, non-polar molecules will be retained for longer, eluting later. Although the flexibility of LC and the possibility to modify many experimental parameters (mobile phase, gradient, modifiers, flow, temperature, type of column, its length and its diameter) make this technique very powerful, they also prevent obtaining a reproducible RT. This is probably why, to the best of our knowledge, there is only one proposal for metabolite annotation where the dynamic RT prediction based on the chromatographic linear solvent strength for RP-LC data used to support steroid identification [10]. The behavior of molecules inside a chromatographic column under some particular chemical conditions is well defined, especially for compounds belonging to the same class [8]. In consequence, although absolute RT is very difficult to predict (even though it is not impossible [11–13]), prediction of the relative order of elution is feasible for certain compounds belonging to the same chemical class which are analyzed under the same analytical conditions. This can be a valuable aid in the analytical process.

1.2. Ionization and adduct formation

The majority of the molecules are ionized by simple protonation $[M+H]^+$ in positive ionization mode or deprotonation $[M-H]^-$ in negative mode. Some compounds, due to their structure, cannot form such adducts and can be ionized only by formation of other adducts [14]. For example, phosphoinositols (PIs) cannot be ionized by protonation $[M+H]^+$, therefore they are not detectable in positive mode, unless gaining sodium $[M+Na]^+$ or potassium $[M+K]^+$. Phosphocholines (PCs) are never ionized by deprotonation ($[M-H]^-$). Consequently, to ionize them, a formate or acetate adduct is needed. Although the majority of molecules are ionizable in both polarity modes, some of them only form positive ions and others only form negative ions.

A list of possible, impossible and/or preferred ions for distinct compound types can be established, as well as relationships between the expected intensities of the different ions. For example, PCs can be ionized in positive mode by protonation $[M+H]^+$, but they can also be ionized with sodium $[M+Na]^+$ and potassium $[M+K]^+$. However, the main signal is $[M+H]^+$ and all others have a lower intensity. Hence, a putative annotation for $[PC+Na]^+$ can be right only if the signal corresponding to $[PC+H]^+$ is also present, otherwise it is a misassignment. This knowledge about ionization and adduct formation can be used to support the annotation process. A different approach to represent this knowledge was already implemented in MZedDB (MZ) [15]. Adduct formation rules based on the structure of the molecule are applied to reject some putative annotations from ChemSpider and PubChem. Unfortunately, this very useful functionality is highly limited by the tool supporting a single mass in the search and by the lack of updates in the tool (the last update was in 2009).

1.3. Relative intensity of the composite spectrum signals

The same molecule can be ionized in several ways, leading to the formation of various signals: different adducts, dimers, multiple charges, etc. [16]. During data reprocessing, acquired chromatographic and spectrometric data is combined to represent each measured compound in three dimensions: mass, RT and intensity. Theoretically, all co-eluting signals corresponding to the same molecule should be clustered by the reprocessing software into a single set called feature. The composite spectrum (CS) is made up of all the signals that arise from the same feature. An example of a CS is given in supplementary file S1.

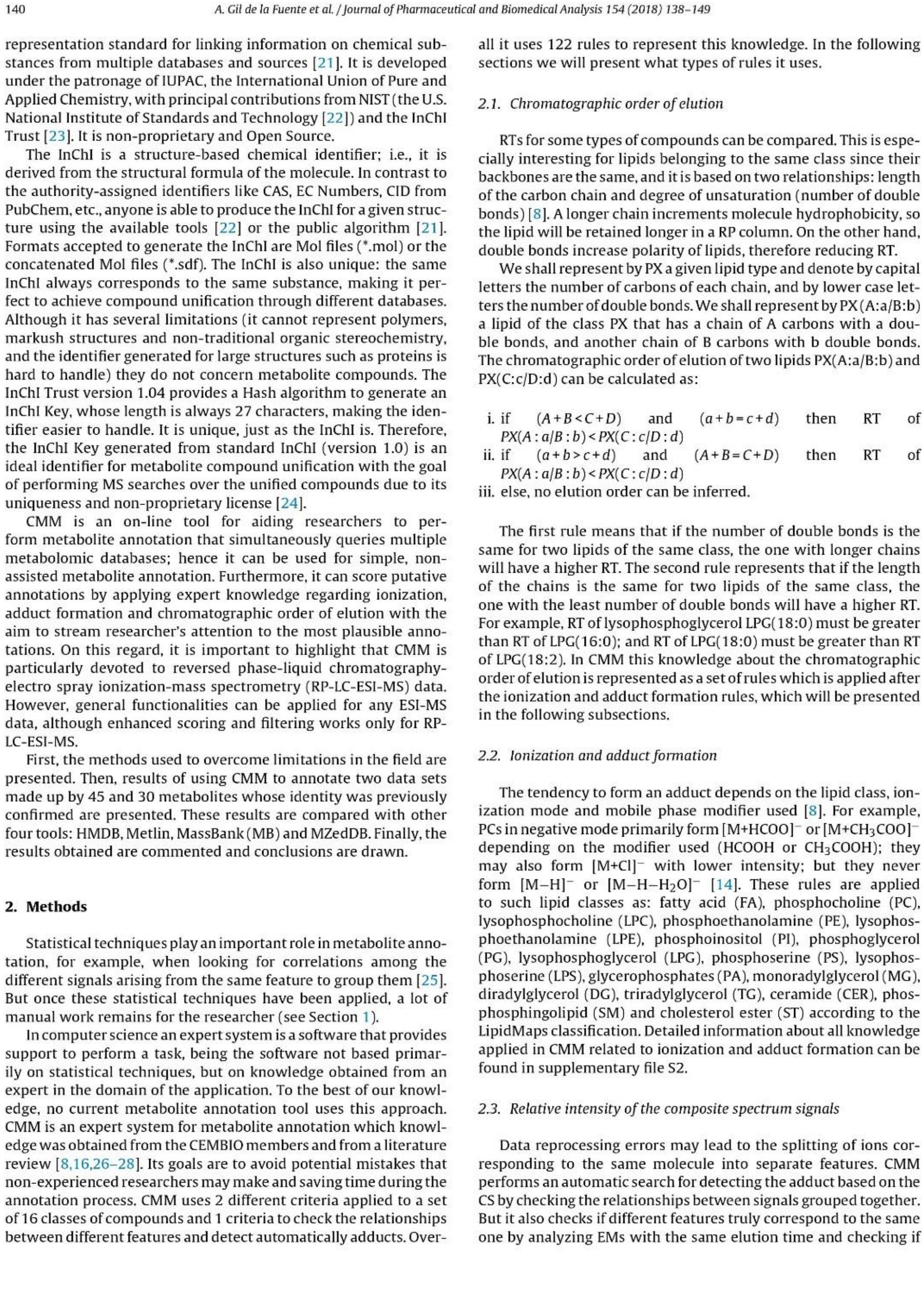
Generally, many multi-signals are correctly clustered together to produce a single CS but others are split in separate features. For this reason twofold data checking should be performed: researchers have to look into the features with the same elution time to check if they correspond to signals arising from the same feature.

1.4. Retrieval of multiple putative annotations from the databases

To improve the possibility of obtaining a match, frequently the features are searched over several databases. Consequently, putative annotations obtained from different databases have to be merged. Merging is challenging because often the same compound is named in a slightly different way in different databases, thus it is difficult to perform unification without the adequate experience and it consumes a lot of time. Moreover, the features and knowledge present in each database differs: some of them represent biological roles (e.g. KEGG) while others contain characteristics about the chemical structure and properties (e.g. LipidMaps) [17]. These databases often are complementary and the metabolomic community needs to work in increasing the interoperability between them.

Furthermore, each database reports results in diverse manners and this aspect has to be considered when merging annotations. Due to that, the use of mediators to perform searches across different libraries/databases is gaining momentum. Some of them offer the possibility of querying a single database at the same time (MINE [18]). Others present to the user all the retrieved annotations, including duplicates for the same compounds present in different databases (Masstrix [19]). Ideally, mediators should present annotations unified to the user, such as Metabolomics Workbench [20]. However, the total number of compounds available in this tool is relatively low [7] and it is not specified exactly how the compounds were unified.

Unification of the information present across databases is a good solution for the aforementioned issues [17]. Two main approaches can be followed: unify compounds based on their structure or unify them based on expert knowledge. The first approach requires the standardization of the representation of the structures, based on which unification can be performed, to avoid the compound mapping based on expert knowledge. Although neither approach is perfect, the first method is less prone to errors and it may be automated. The IUPAC International Chemical Identifier (InChI) is one of the structures which can be used for performing the unification of compounds. The InChI is the worldwide chemical structure



representation standard for linking information on chemical substances from multiple databases and sources [21]. It is developed under the patronage of IUPAC, the International Union of Pure and Applied Chemistry, with principal contributions from NIST (the U.S. National Institute of Standards and Technology [22]) and the InChI Trust [23]. It is non-proprietary and Open Source.

The InChI is a structure-based chemical identifier; i.e., it is derived from the structural formula of the molecule. In contrast to the authority-assigned identifiers like CAS, EC Numbers, CID from PubChem, etc., anyone is able to produce the InChI for a given structure using the available tools [22] or the public algorithm [21]. Formats accepted to generate the InChI are Mol files (*.mol) or the concatenated Mol files (*.sdf). The InChI is also unique: the same InChI always corresponds to the same substance, making it perfect to achieve compound unification through different databases. Although it has several limitations (it cannot represent polymers, markush structures and non-traditional organic stereochemistry, and the identifier generated for large structures such as proteins is hard to handle) they do not concern metabolite compounds. The InChI Trust version 1.04 provides a Hash algorithm to generate an InChI Key, whose length is always 27 characters, making the identifier easier to handle. It is unique, just as the InChI is. Therefore, the InChI Key generated from standard InChI (version 1.0) is an ideal identifier for metabolite compound unification with the goal of performing MS searches over the unified compounds due to its uniqueness and non-proprietary license [24].

CMM is an on-line tool for aiding researchers to perform metabolite annotation that simultaneously queries multiple metabolomic databases; hence it can be used for simple, non-assisted metabolite annotation. Furthermore, it can score putative annotations by applying expert knowledge regarding ionization, adduct formation and chromatographic order of elution with the aim to stream researcher's attention to the most plausible annotations. On this regard, it is important to highlight that CMM is particularly devoted to reversed phase-liquid chromatography-electro spray ionization-mass spectrometry (RP-LC-ESI-MS) data. However, general functionalities can be applied for any ESI-MS data, although enhanced scoring and filtering works only for RP-LC-ESI-MS.

First, the methods used to overcome limitations in the field are presented. Then, results of using CMM to annotate two data sets made up by 45 and 30 metabolites whose identity was previously confirmed are presented. These results are compared with other four tools: HMDB, Metlin, MassBank (MB) and MZedDB. Finally, the results obtained are commented and conclusions are drawn.

2. Methods

Statistical techniques play an important role in metabolite annotation, for example, when looking for correlations among the different signals arising from the same feature to group them [25]. But once these statistical techniques have been applied, a lot of manual work remains for the researcher (see Section 1).

In computer science an expert system is a software that provides support to perform a task, being the software not based primarily on statistical techniques, but on knowledge obtained from an expert in the domain of the application. To the best of our knowledge, no current metabolite annotation tool uses this approach. CMM is an expert system for metabolite annotation which knowledge was obtained from the CEMBIO members and from a literature review [8,16,26–28]. Its goals are to avoid potential mistakes that non-experienced researchers may make and saving time during the annotation process. CMM uses 2 different criteria applied to a set of 16 classes of compounds and 1 criteria to check the relationships between different features and detect automatically adducts. Overall it uses 122 rules to represent this knowledge. In the following sections we will present what types of rules it uses.

2.1. Chromatographic order of elution

RTs for some types of compounds can be compared. This is especially interesting for lipids belonging to the same class since their backbones are the same, and it is based on two relationships: length of the carbon chain and degree of unsaturation (number of double bonds) [8]. A longer chain increments molecule hydrophobicity, so the lipid will be retained longer in a RP column. On the other hand, double bonds increase polarity of lipids, therefore reducing RT.

We shall represent by PX a given lipid type and denote by capital letters the number of carbons of each chain, and by lower case letters the number of double bonds. We shall represent by PX (A:a/B:b) a lipid of the class PX that has a chain of A carbons with a double bonds, and another chain of B carbons with b double bonds. The chromatographic order of elution of two lipids PX(A:a/B:b) and PX(C:c/D:d) can be calculated as:

i. if $(A+B<C+D)$ and $(a+b=c+d)$ then RT of $PX(A:a/B:b)<PX(C:c/D:d)$

ii. if $(a+b>c+d)$ and $(A+B=C+D)$ then RT of $PX(A:a/B:b)<PX(C:c/D:d)$

iii. else, no elution order can be inferred.

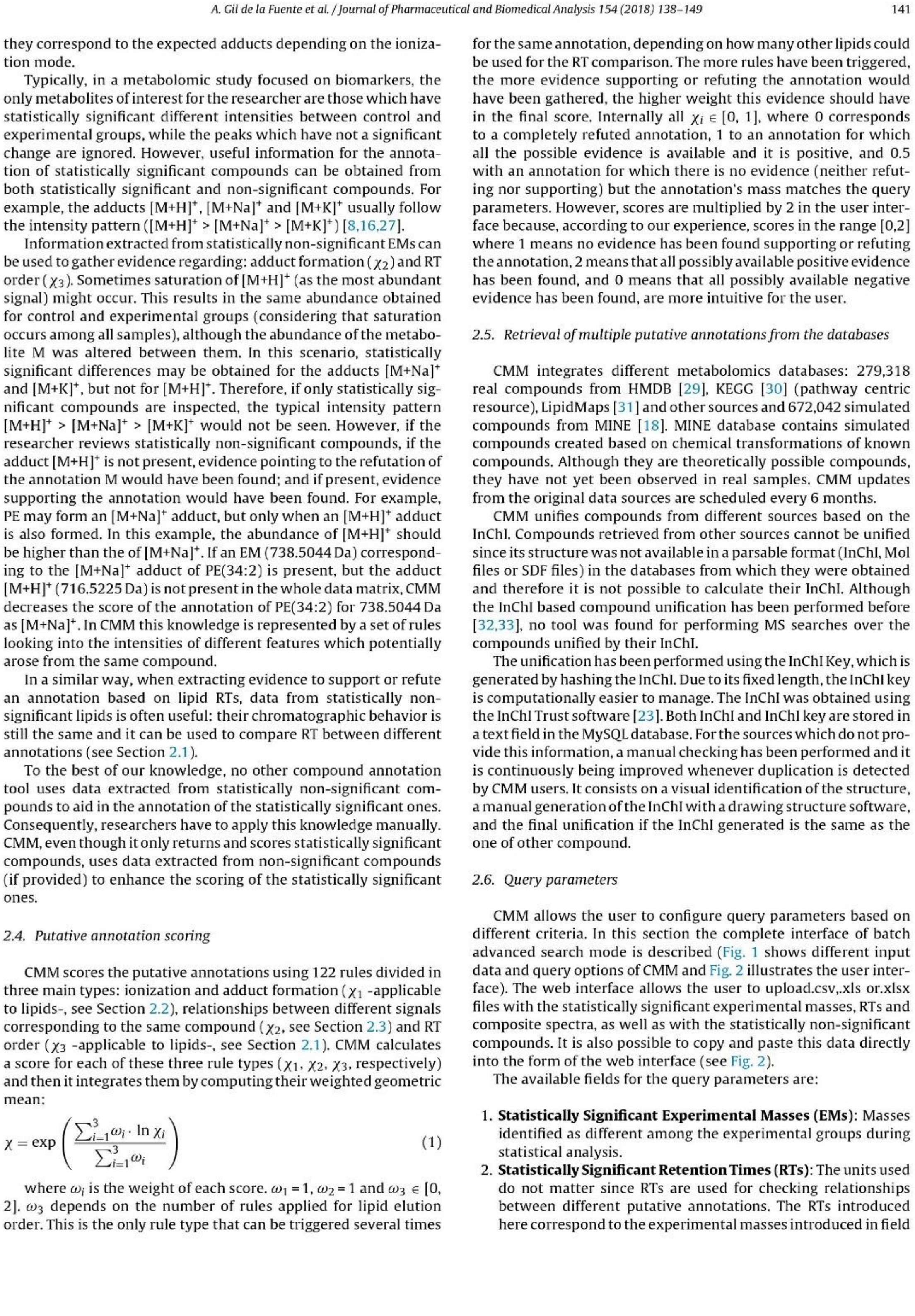
The first rule means that if the number of double bonds is the same for two lipids of the same class, the one with longer chains will have a higher RT. The second rule represents that if the length of the chains is the same for two lipids of the same class, the one with the least number of double bonds will have a higher RT. For example, RT of lysophosphoglycerol LPG(18:0) must be greater than RT of LPG(16:0); and RT of LPG(18:0) must be greater than RT of LPG(18:2). In CMM this knowledge about the chromatographic order of elution is represented as a set of rules which is applied after the ionization and adduct formation rules, which will be presented in the following subsections.

2.2. Ionization and adduct formation

The tendency to form an adduct depends on the lipid class, ionization mode and mobile phase modifier used [8]. For example, PCs in negative mode primarily form $[M+HCOO]^-$ or $[M+CH_3COO]^-$ depending on the modifier used (HCOOH or CH_3COOH); they may also form $[M+Cl]^-$ with lower intensity; but they never form $[M-H]^-$ or $[M-H-H_2O]^-$ [14]. These rules are applied to such lipid classes as: fatty acid (FA), phosphocholine (PC), lysophosphocholine (LPC), phosphoethanolamine (PE), lysophosphoethanolamine (LPE), phosphoinositol (PI), phosphoglycerol (PG), lysophosphoglycerol (LPG), phosphoserine (PS), lysophosphoserine (LPS), glycerophosphates (PA), monoradylglycerol (MG), diradylglycerol (DG), triradylglycerol (TG), ceramide (CER), phosphosphingolipid (SM) and cholesterol ester (ST) according to the LipidMaps classification. Detailed information about all knowledge applied in CMM related to ionization and adduct formation can be found in supplementary file S2.

2.3. Relative intensity of the composite spectrum signals

Data reprocessing errors may lead to the splitting of ions corresponding to the same molecule into separate features. CMM performs an automatic search for detecting the adduct based on the CS by checking the relationships between signals grouped together. But it also checks if different features truly correspond to the same one by analyzing EMs with the same elution time and checking if



they correspond to the expected adducts depending on the ionization mode.

Typically, in a metabolomic study focused on biomarkers, the only metabolites of interest for the researcher are those which have statistically significant different intensities between control and experimental groups, while the peaks which have not a significant change are ignored. However, useful information for the annotation of statistically significant compounds can be obtained from both statistically significant and non-significant compounds. For example, the adducts $[M+H]^+$, $[M+Na]^+$ and $[M+K]^+$ usually follow the intensity pattern ($[M+H]^+ > [M+Na]^+ > [M+K]^+$) [8,16,27].

Information extracted from statistically non-significant EMs can be used to gather evidence regarding: adduct formation (χ_2) and RT order (χ_3). Sometimes saturation of $[M+H]^+$ (as the most abundant signal) might occur. This results in the same abundance obtained for control and experimental groups (considering that saturation occurs among all samples), although the abundance of the metabolite M was altered between them. In this scenario, statistically significant differences may be obtained for the adducts $[M+Na]^+$ and $[M+K]^+$, but not for $[M+H]^+$. Therefore, if only statistically significant compounds are inspected, the typical intensity pattern $[M+H]^+ > [M+Na]^+ > [M+K]^+$ would not be seen. However, if the researcher reviews statistically non-significant compounds, if the adduct $[M+H]^+$ is not present, evidence pointing to the refutation of the annotation M would have been found; and if present, evidence supporting the annotation would have been found. For example, PE may form an $[M+Na]^+$ adduct, but only when an $[M+H]^+$ adduct is also formed. In this example, the abundance of $[M+H]^+$ should be higher than the of $[M+Na]^+$. If an EM (738.5044 Da) corresponding to the $[M+Na]^+$ adduct of PE(34:2) is present, but the adduct $[M+H]^+$ (716.5225 Da) is not present in the whole data matrix, CMM decreases the score of the annotation of PE(34:2) for 738.5044 Da as $[M+Na]^+$. In CMM this knowledge is represented by a set of rules looking into the intensities of different features which potentially arose from the same compound.

In a similar way, when extracting evidence to support or refute an annotation based on lipid RTs, data from statistically non-significant lipids is often useful: their chromatographic behavior is still the same and it can be used to compare RT between different annotations (see Section 2.1).

To the best of our knowledge, no other compound annotation tool uses data extracted from statistically non-significant compounds to aid in the annotation of the statistically significant ones. Consequently, researchers have to apply this knowledge manually. CMM, even though it only returns and scores statistically significant compounds, uses data extracted from non-significant compounds (if provided) to enhance the scoring of the statistically significant ones.

2.4. *Putative annotation scoring*

CMM scores the putative annotations using 122 rules divided in three main types: ionization and adduct formation (χ_1 -applicable to lipids-, see Section 2.2), relationships between different signals corresponding to the same compound (χ_2, see Section 2.3) and RT order (χ_3 -applicable to lipids-, see Section 2.1). CMM calculates a score for each of these three rule types (χ_1, χ_2, χ_3, respectively) and then it integrates them by computing their weighted geometric mean:

$$\chi = \exp\left(\frac{\sum_{i=1}^{3}\omega_i \cdot \ln \chi_i}{\sum_{i=1}^{3}\omega_i}\right) \qquad (1)$$

where ω_i is the weight of each score. $\omega_1 = 1$, $\omega_2 = 1$ and $\omega_3 \in [0, 2]$. ω_3 depends on the number of rules applied for lipid elution order. This is the only rule type that can be triggered several times for the same annotation, depending on how many other lipids could be used for the RT comparison. The more rules have been triggered, the more evidence supporting or refuting the annotation would have been gathered, the higher weight this evidence should have in the final score. Internally all $\chi_i \in [0, 1]$, where 0 corresponds to a completely refuted annotation, 1 to an annotation for which all the possible evidence is available and it is positive, and 0.5 with an annotation for which there is no evidence (neither refuting nor supporting) but the annotation's mass matches the query parameters. However, scores are multiplied by 2 in the user interface because, according to our experience, scores in the range [0,2] where 1 means no evidence has been found supporting or refuting the annotation, 2 means that all possibly available positive evidence has been found, and 0 means that all possibly available negative evidence has been found, are more intuitive for the user.

2.5. *Retrieval of multiple putative annotations from the databases*

CMM integrates different metabolomics databases: 279,318 real compounds from HMDB [29], KEGG [30] (pathway centric resource), LipidMaps [31] and other sources and 672,042 simulated compounds from MINE [18]. MINE database contains simulated compounds created based on chemical transformations of known compounds. Although they are theoretically possible compounds, they have not yet been observed in real samples. CMM updates from the original data sources are scheduled every 6 months.

CMM unifies compounds from different sources based on the InChI. Compounds retrieved from other sources cannot be unified since its structure was not available in a parsable format (InChI, Mol files or SDF files) in the databases from which they were obtained and therefore it is not possible to calculate their InChI. Although the InChI based compound unification has been performed before [32,33], no tool was found for performing MS searches over the compounds unified by their InChI.

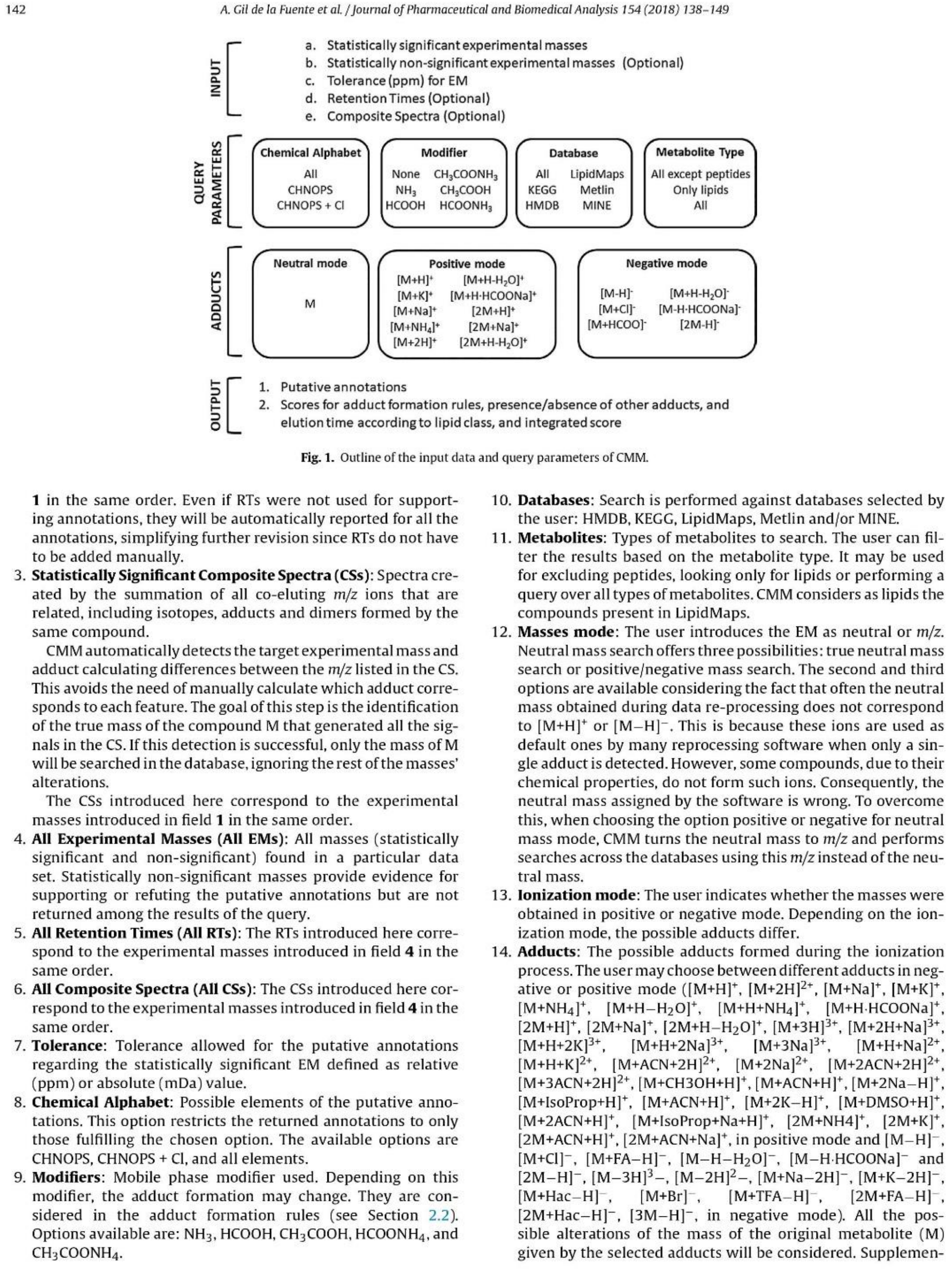
The unification has been performed using the InChI Key, which is generated by hashing the InChI. Due to its fixed length, the InChI key is computationally easier to manage. The InChI was obtained using the InChI Trust software [23]. Both InChI and InChI key are stored in a text field in the MySQL database. For the sources which do not provide this information, a manual checking has been performed and it is continuously being improved whenever duplication is detected by CMM users. It consists on a visual identification of the structure, a manual generation of the InChI with a drawing structure software, and the final unification if the InChI generated is the same as the one of other compound.

2.6. *Query parameters*

CMM allows the user to configure query parameters based on different criteria. In this section the complete interface of batch advanced search mode is described (Fig. 1 shows different input data and query options of CMM and Fig. 2 illustrates the user interface). The web interface allows the user to upload.csv,.xls or.xlsx files with the statistically significant experimental masses, RTs and composite spectra, as well as with the statistically non-significant compounds. It is also possible to copy and paste this data directly into the form of the web interface (see Fig. 2).

The available fields for the query parameters are:

1. **Statistically Significant Experimental Masses (EMs)**: Masses identified as different among the experimental groups during statistical analysis.
2. **Statistically Significant Retention Times (RTs)**: The units used do not matter since RTs are used for checking relationships between different putative annotations. The RTs introduced here correspond to the experimental masses introduced in field



INPUT
a. Statistically significant experimental masses
b. Statistically non-significant experimental masses (Optional)
c. Tolerance (ppm) for EM
d. Retention Times (Optional)
e. Composite Spectra (Optional)
QUERY PARAMETERS
Chemical Alphabet
All
CHNOPS
CHNOPS + Cl
Modifier
None
NH3
HCOOH
CH3COONH3
CH3COOH
HCOONH3
Database
All
KEGG
HMDB
LipidMaps
Metlin
MINE
Metabolite Type
All except peptides
Only lipids
All
ADDUCTS
Neutral mode
M
Positive mode
[M+H]+
[M+K]+
[M+Na]+
[M+NH4]+
[M+2H]+
[M+H-H2O]+
[M+H·HCOONa]+
[2M+H]+
[2M+Na]+
[2M+H-H2O]+
Negative mode
[M-H]-
[M+Cl]-
[M+HCOO]-
[M+H-H2O]-
[M-H·HCOONa]-
[2M-H]-
OUTPUT
1. Putative annotations
2. Scores for adduct formation rules, presence/absence of other adducts, and elution time according to lipid class, and integrated score

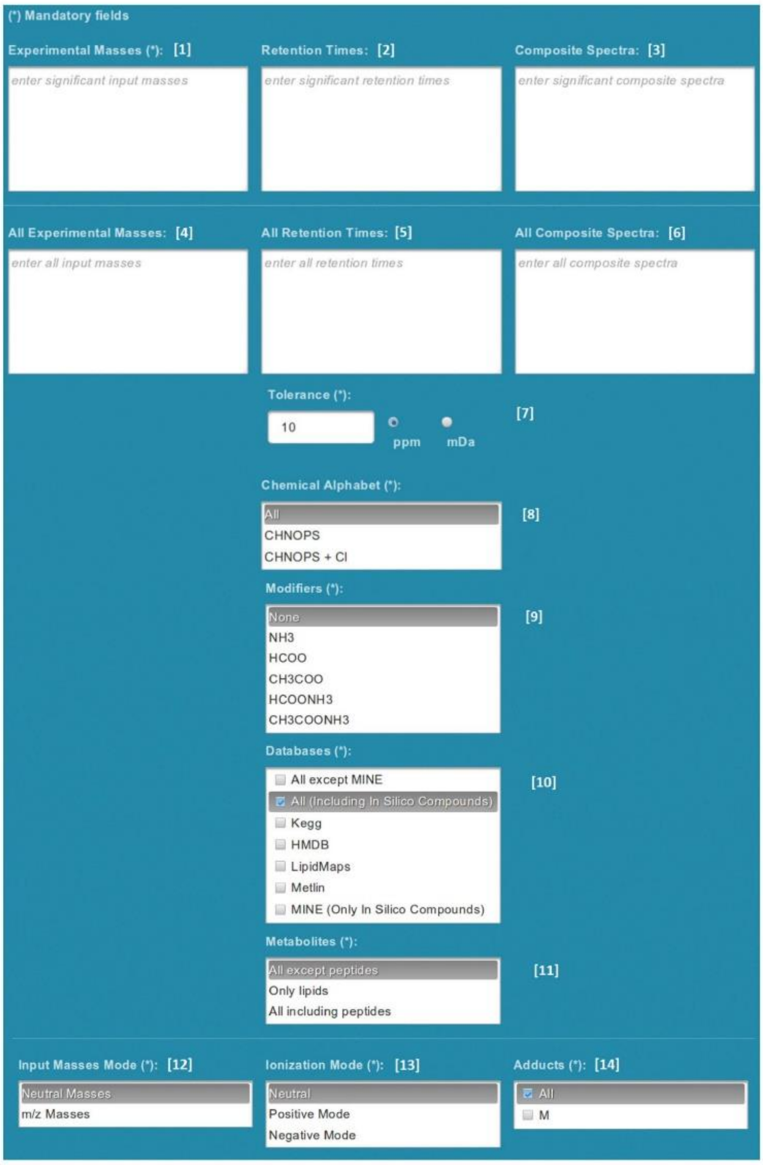
Fig. 1. Outline of the input data and query parameters of CMM.

1 in the same order. Even if RTs were not used for supporting annotations, they will be automatically reported for all the annotations, simplifying further revision since RTs do not have to be added manually.

3. **Statistically Significant Composite Spectra (CSs)**: Spectra created by the summation of all co-eluting *m/z* ions that are related, including isotopes, adducts and dimers formed by the same compound.

   CMM automatically detects the target experimental mass and adduct calculating differences between the *m/z* listed in the CS. This avoids the need of manually calculate which adduct corresponds to each feature. The goal of this step is the identification of the true mass of the compound M that generated all the signals in the CS. If this detection is successful, only the mass of M will be searched in the database, ignoring the rest of the masses' alterations.

   The CSs introduced here correspond to the experimental masses introduced in field **1** in the same order.
4. **All Experimental Masses (All EMs)**: All masses (statistically significant and non-significant) found in a particular data set. Statistically non-significant masses provide evidence for supporting or refuting the putative annotations but are not returned among the results of the query.
5. **All Retention Times (All RTs)**: The RTs introduced here correspond to the experimental masses introduced in field **4** in the same order.
6. **All Composite Spectra (All CSs)**: The CSs introduced here correspond to the experimental masses introduced in field **4** in the same order.
7. **Tolerance**: Tolerance allowed for the putative annotations regarding the statistically significant EM defined as relative (ppm) or absolute (mDa) value.
8. **Chemical Alphabet**: Possible elements of the putative annotations. This option restricts the returned annotations to only those fulfilling the chosen option. The available options are CHNOPS, CHNOPS + Cl, and all elements.
9. **Modifiers**: Mobile phase modifier used. Depending on this modifier, the adduct formation may change. They are considered in the adduct formation rules (see Section 2.2). Options available are: NH_3, HCOOH, CH_3COOH, $HCOONH_4$, and CH_3COONH_4.
10. **Databases**: Search is performed against databases selected by the user: HMDB, KEGG, LipidMaps, Metlin and/or MINE.
11. **Metabolites**: Types of metabolites to search. The user can filter the results based on the metabolite type. It may be used for excluding peptides, looking only for lipids or performing a query over all types of metabolites. CMM considers as lipids the compounds present in LipidMaps.
12. **Masses mode**: The user introduces the EM as neutral or *m/z*. Neutral mass search offers three possibilities: true neutral mass search or positive/negative mass search. The second and third options are available considering the fact that often the neutral mass obtained during data re-processing does not correspond to $[M+H]^+$ or $[M-H]^-$. This is because these ions are used as default ones by many reprocessing software when only a single adduct is detected. However, some compounds, due to their chemical properties, do not form such ions. Consequently, the neutral mass assigned by the software is wrong. To overcome this, when choosing the option positive or negative for neutral mass mode, CMM turns the neutral mass to *m/z* and performs searches across the databases using this *m/z* instead of the neutral mass.
13. **Ionization mode**: The user indicates whether the masses were obtained in positive or negative mode. Depending on the ionization mode, the possible adducts differ.
14. **Adducts**: The possible adducts formed during the ionization process. The user may choose between different adducts in negative or positive mode ($[M+H]^+$, $[M+2H]^{2+}$, $[M+Na]^+$, $[M+K]^+$, $[M+NH_4]^+$, $[M+H-H_2O]^+$, $[M+H+NH_4]^+$, $[M+H{\cdot}HCOONa]^+$, $[2M+H]^+$, $[2M+Na]^+$, $[2M+H-H_2O]^+$, $[M+3H]^{3+}$, $[M+2H+Na]^{3+}$, $[M+H+2K]^{3+}$, $[M+H+2Na]^{3+}$, $[M+3Na]^{3+}$, $[M+H+Na]^{2+}$, $[M+H+K]^{2+}$, $[M+ACN+2H]^{2+}$, $[M+2Na]^{2+}$, $[M+2ACN+2H]^{2+}$, $[M+3ACN+2H]^{2+}$, $[M+CH3OH+H]^+$, $[M+ACN+H]^+$, $[M+2Na-H]^+$, $[M+IsoProp+H]^+$, $[M+ACN+H]^+$, $[M+2K-H]^+$, $[M+DMSO+H]^+$, $[M+2ACN+H]^+$, $[M+IsoProp+Na+H]^+$, $[2M+NH4]^+$, $[2M+K]^+$, $[2M+ACN+H]^+$, $[2M+ACN+Na]^+$, in positive mode and $[M-H]^-$, $[M+Cl]^-$, $[M+FA-H]^-$, $[M-H-H_2O]^-$, $[M-H{\cdot}HCOONa]^-$ and $[2M-H]^-$, $[M-3H]^3-$, $[M-2H]^2-$, $[M+Na-2H]^-$, $[M+K-2H]^-$, $[M+Hac-H]^-$, $[M+Br]^-$, $[M+TFA-H]^-$, $[2M+FA-H]^-$, $[2M+Hac-H]^-$, $[3M-H]^-$, in negative mode). All the possible alterations of the mass of the original metabolite (M) given by the selected adducts will be considered. Supplemen-



(*) Mandatory fields

Experimental Masses (*): [1]
enter significant input masses

Retention Times: [2]
enter significant retention times

Composite Spectra: [3]
enter significant composite spectra

All Experimental Masses: [4]
enter all input masses

All Retention Times: [5]
enter all retention times

All Composite Spectra: [6]
enter all composite spectra

Tolerance (*): [7]
10 ○ ppm ● mDa

Chemical Alphabet (*): [8]
All
CHNOPS
CHNOPS + Cl

Modifiers (*): [9]
None
NH3
HCOO
CH3COO
HCOONH3
CH3COONH3

Databases (*): [10]
☐ All except MINE
☑ All (Including In Silico Compounds)
☐ Kegg
☐ HMDB
☐ LipidMaps
☐ Metlin
☐ MINE (Only In Silico Compounds)

Metabolites (*): [11]
All except peptides
Only lipids
All including peptides

Input Masses Mode (*): [12]
Neutral Masses
m/z Masses

Ionization Mode (*): [13]
Neutral
Positive Mode
Negative Mode

Adducts (*): [14]
☑ All
☐ M

Fig. 2. Batch advanced search interface.

tary file S3 shows the full list of adducts and the calculations performed by CMM.

2.7. Search modes

CMM provides researchers with different types of search, depending on what information they want to use for performing the metabolite annotation and depending on whether they want to look for multiple compounds or just one. The following search modes are available:

1. **Simple Search**: it allows the user to query a single mass within an established tolerance with basic choices of databases, metabolite types, masses mode, ionization mode and adducts.
2. **Batch Search**: it allows the user to query a batch of masses within an established tolerance with basic choices of databases, metabolite types, masses mode, ionization mode and adducts.
3. **Advanced Search**: it allows the user to query a single mass within an established tolerance with the next parameters: RT, CS, chemical alphabet, modifiers, databases, metabolite types, masses mode, ionization mode and adducts. As the user only

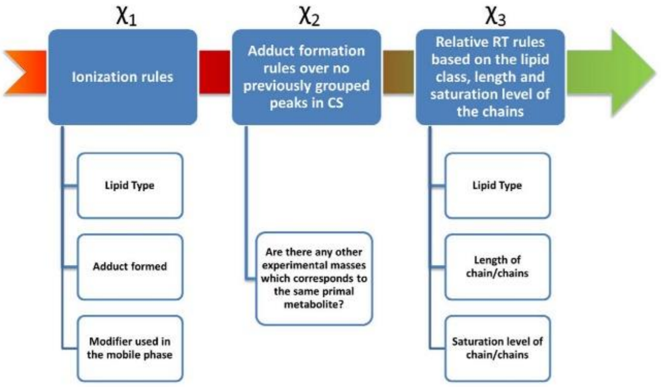

χ_1
χ_2
χ_3
Ionization rules
Adduct formation rules over no previously grouped peaks in CS
Relative RT rules based on the lipid class, length and saturation level of the chains
Lipid Type
Adduct formed
Modifier used in the mobile phase
Are there any other experimental masses which corresponds to the same primal metabolite?
Lipid Type
Length of chain/chains
Saturation level of chain/chains

Fig. 3. Knowledge system of CMM.

introduces a single mass, the RT is not used to trigger rules. CS is used to automatically detect the adduct. RT and CS are optional parameters.

4. **Batch Advanced Search**: it allows the user to query a batch of masses within an established tolerance with the next parameters: RT, CS, statistically non-significant experimental masses and its corresponding RT and CS, chemical alphabet, modifiers, databases, metabolite types, masses mode, ionization mode and adducts. CS is used to automatically detect the adduct.
5. **Browse Search**: it allows the user to search for metabolites based on their name and/or formula. The compounds are previously unified and the target databases as well as the type of metabolites can be filtered.

2.8. Implementation details

CMM web application uses the library PrimeFaces 6.1 for the presentation layer. The core of the application is written in Java2EE and uses an inference engine implemented using JBoss Drools [34]. To simplify the acquisition and revision of the knowledge, we use spreadsheets for the representation of the rules. Knowledge is split in different spreadsheets depending on the type of rules. They are applied sequentially: ionization rules depending on the propensity of a particular adduct formation due to the lipid class, then relationships between different adducts pertaining to the same primal metabolite and finally rules over the RT based on the length and saturation level of lipids (see Fig. 3). supplementary file S2 contains all the rules. Nowadays the maximum number of features processed by CMM is limited to 250 due to limitations in the computational resources available in the server.

The web application is an open-source project publicly available on Github (https://github.com/albertogilf/ceuMassMediator). The web application server used for deploying the Java2EE project is Apache TomEE version plume 7.0.1. The Database server used is MySql Server 5.7.16. Bash and SQL scripts are used to load data into the database. This software runs on a virtual Debian machine running under GNU/Linux. The virtual machine has 8 GB of memory and an Intel Xeon X5690 processor with 6 cores and a 3.46 GHz frequency.

2.9. LC–MS conditions to acquire the validation data

To illustrate the functionality of CMM, two data sets were used. The first data set (DS1) contains 45 experimental masses (plasma samples) previously identified by MS/MS or usage of commercially available standards (Sigma–Aldrich, Fluka Analytical). The list of these compounds is shown in supplementary file S4. These 45 experimental masses correspond to 36 compounds. Plasma samples were prepared by simple deproteinization with cold methanol/ethanol (1/1) mixture [35]. Standards were prepared in methanol with concentrations between 2 and 10 ppm. To obtain this data, samples were analyzed by a HPLC system (1200 series, Agilent Technologies) connected to an Agilent QTOF (6520) MS detector. Samples (10 μL) were applied to a reversed-phase column (Discovery HS C18 150 2.1 mm, 3 μm; Supelco) with a guard column (Discovery HS C18 20 2.1 mm, 3 μm; Supelco). Chromatographic conditions were the same as described previously [35]. Data was collected in positive and negative ESI ion modes in separate runs on a QTOF operated in the mass range from *m/z* 50 to 1000 with an acquisition rate of 1 scan per second. The capillary voltage was set to 3000 V for positive and 4000 V for negative ionization mode; the nebulizer gas flow rate was 10.5 L/min. Accurate mass measurements were obtained by means of an automated calibrant delivery system using a dual-nebulizer ESI source that continuously introduces a calibrating solution.

The second data set (DS2) contains 30 experimental masses (plasma samples) previously identified with commercially available standards (Alpha Aesar, Sigma-Aldrich, Fluka Analytical). The list of these compounds is shown in supplementary file S4. These 30 experimental masses correspond to 30 compounds. To obtain this data samples were analyzed by a capillary electrophoresis (CE) system (7100 Agilent Technologies) connected to an Agilent TOF (6224) MS detector. Samples were applied to a fused-silica capillary (100 cm × 50 μm; Agilent) under 15 mbars for 30 s. Electrophoretic separation occurs applying voltage of 30 kV under the current of 20 μA. Data were collected in positive ESI ion mode in separate ranges on a TOF operating in the mass range from *m/z* 50 to 1000 with an acquisition rate of 1 scan per second. The capillary voltage was set to 3500 V; the nebulizer gas flow rate was 10.0 L/min.

3. Results and discussion

In this section we will first present the results of compound unification based on the InChI Key. Then we will show the results of two queries: the first one performed with 45 experimental masses corresponding to 36 lipidic and non-lipidic compounds analyzed under LC–MS (DS1); and the second one performed with 30 experimental masses corresponding to 30 non-lipidic compounds analyzed under CE–MS (DS2).

Table 1
Query parameters for DS1 and DS2 in HMDB, Metlin, MB, MZ and CMM comparison.

Query parameter	HMDB	Metlin	MassBank	MZedDB	CMM
Experimental masses	45 &30				
Retention times	N/A				45 &30
Composite spectra	N/A				45 &30
Statistically non-significant masses	N/A				0
Tolerance allowed	10 ppm		0.01 Da	10 ppm	
Chemical alphabet	N/A			CHNOPS	
Modifier used	N/A				None
Databases to search	HMDB	Metlin	MB	all*	KEGG, LipidMaps, HMDB &Metlin
Restrictions	N/A	Remove peptides, drugs and toxicants	Instrument type: ESI	Apply adduct formation rules	Remove peptides
Ionization mode	Positive mode				
Adducts to search	$[M+H]^+$, $[M+Na]^+$, $[M+K]^+$, $[M+NH_4]^+$, $[M+H-H_2O]^+$ & $[M+2H]^{2+}$		N/A	$[M+H]^+$, $[M+Na]^+$, $[M+K]^+$, $[M+NH_4]^+$, $[M+H-H_2O]^+$ & $[M+2H]^{2+}$	

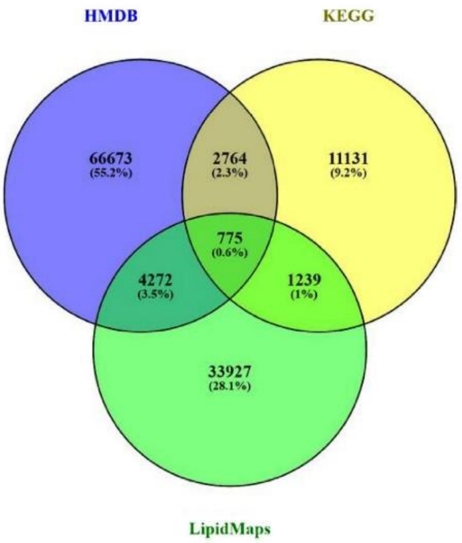


Fig. 4. Venn diagram with the results of compound unification.

Table 2
Putative annotations by CMM in each query processing step (DS1).

Step	Putative annotations	Precision	Recall
Before unification	2462	1.83%	100%
Before automatic adduct detection	1367	3.29%	100%
Before applying knowledge for scoring	905	4.97%	100%
Before manual unification of isomers	869	5.18%	100%
After manual unification of isomers	255	17.65%	100%

Table 3
Scores assigned by CMM to the real compound annotation (DS1).

Score range	Number of right annotations
[0–0.5)	0
[0.5–1)	0
[1–1.5)	1
[1.5–2]	28
No rules applied	16

Although the user only sees the final results, to show the usefulness of the different strategies used in CMM to improve the annotations (compound unification, ionization and adduct formation rules, adduct abundance rules and elution order rules) we shall analyze the results of each query step by step. These results will be compared with the results of a querying Metlin, HDMB, MassBank and MZedDB with the same experimental masses.

3.1. Compound unification

Fig. 4 is a Venn diagram which shows the results of compound unification for the databases KEGG, LipidMaps and HMDB performed based on the InChI. Percentages shown are calculated over the maximum number of compounds which could overlap (considering only compounds from the databases that provide structure information): 120,781. Compounds from other sources are not included in Fig. 4.

Although the percentages of unified compounds may not seem high, they correspond to the compounds most commonly present in biological samples. Furthermore, when evaluating the results of the unification it is necessary to bear in mind that the InChI of two compounds varies if the position or type of a bond differs. Therefore, all the isomers of a compound have different InChI. If different databases contain compounds with a single difference in the position or type of a bond, they are different compounds, and consequently they are not unified. Thus, the number of compounds overlapped among the databases is not high. While is true that during the annotation process is impossible to distinguish between such compounds, they might be differentiated in a subsequent MS^n analysis. For this reason, stereoisomeric annotations have to be kept as different entries. The small number of overlapped metabolites may also be explained by the fact that LipidMaps is exclusively devoted to lipids. Therefore, it does not cover the majority of small polar compounds present in the other databases.

3.2. Knowledge-based putative annotation scoring

A comparison of CMM results when searching for DS1 and DS2 with HMDB, Metlin, MassBank and MZedDB was performed. The first two databases were selected because they are the most similar in terms of functionality to CMM. Batch searches are allowed based on the m/z value of the features. MassBank [36] is a public data repository where users can upload their experimental data and the tool provides a database service for the identification of metabolites. MZedDB is a tool for filtering putative ionization products based on the structure of the putative annotations. The list of putative annotations is searched across the databases: aracyc, dico, HMDB, KEGG, LipidMaps, mammal, metacyc, plant and ricecyc. The last update of the data from these data sources was in 2009. The data sets were tested in HMDB, Metlin, MassBank and MZedDB and CMM with the query parameters shown in Table 1.

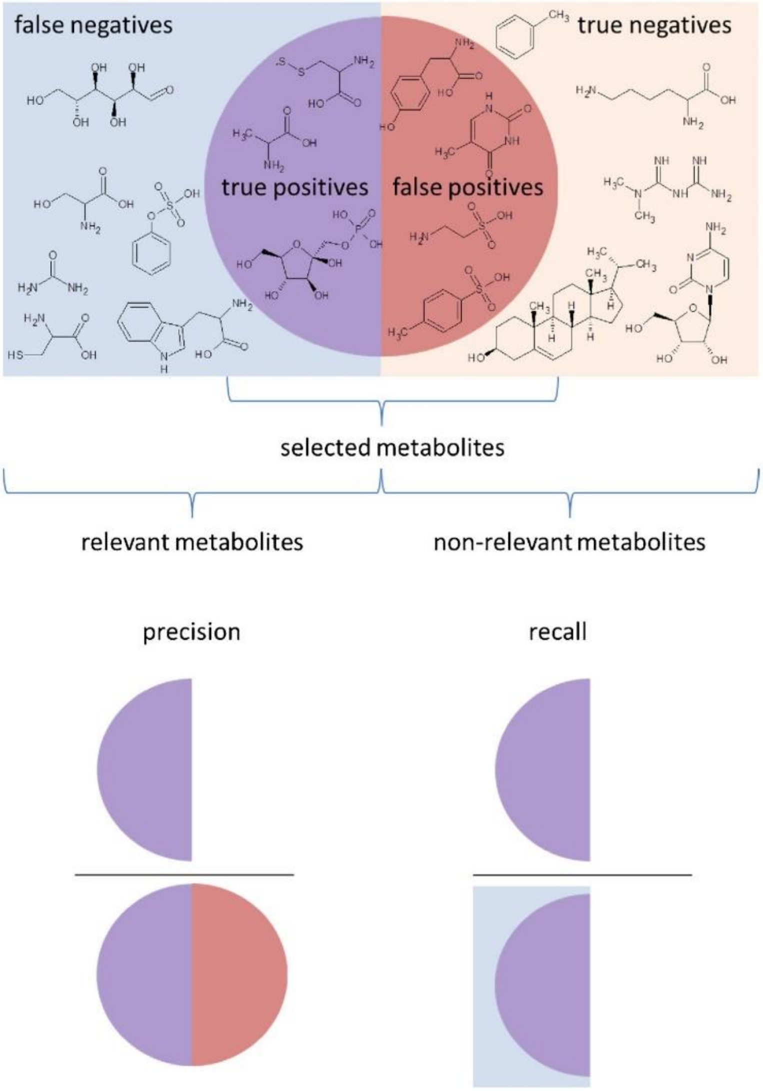


Fig. 5. Illustration of precision and recall measurements. Precision shows how many selected metabolites are relevant while recall shows how many relevant metabolites are selected.

Tables 2 and 3 summarize the results for DS1. When searching for the 45 masses a list of 2462 putative annotations was obtained from all the databases integrated in CMM (except MINE), which clearly illustrates the need for automatic filtering. Of the initial 2462 putative annotations, 1095 had been unified based on the InChI and by the manual checking process. After unification, 1367 (55.52% of the total) still remained, but CMM automatic adduct detection filtered out 463 additional annotations, leaving 905 (36.76% of the total). The last step performed by CMM is the application of the scoring rules. After the scoring step 36 putative annotations were marked with a score below 1.0, which means that the rules found some evidence pointing to that annotation being incorrect. The researcher still has to handle 869 (35.30% of the total) putative annotations, but it is important to notice that most of them correspond to different isomers of the same compound, which are not distinguishable using only MS^1 (e.g. PC(18:1/16:1) and PC(18:0/16:2)). Taking into account the different isomers corresponding to the same lipid class, the number of putative annotations retrieved decreases to 255 for the 45 initial experimental masses.

An average of 35.4 putative annotations for each input mass were filtered out. An average of 19.31 putative annotations per experimental mass were returned. 13.64 putative annotations of the 19.31 returned correspond to different isomers of the same

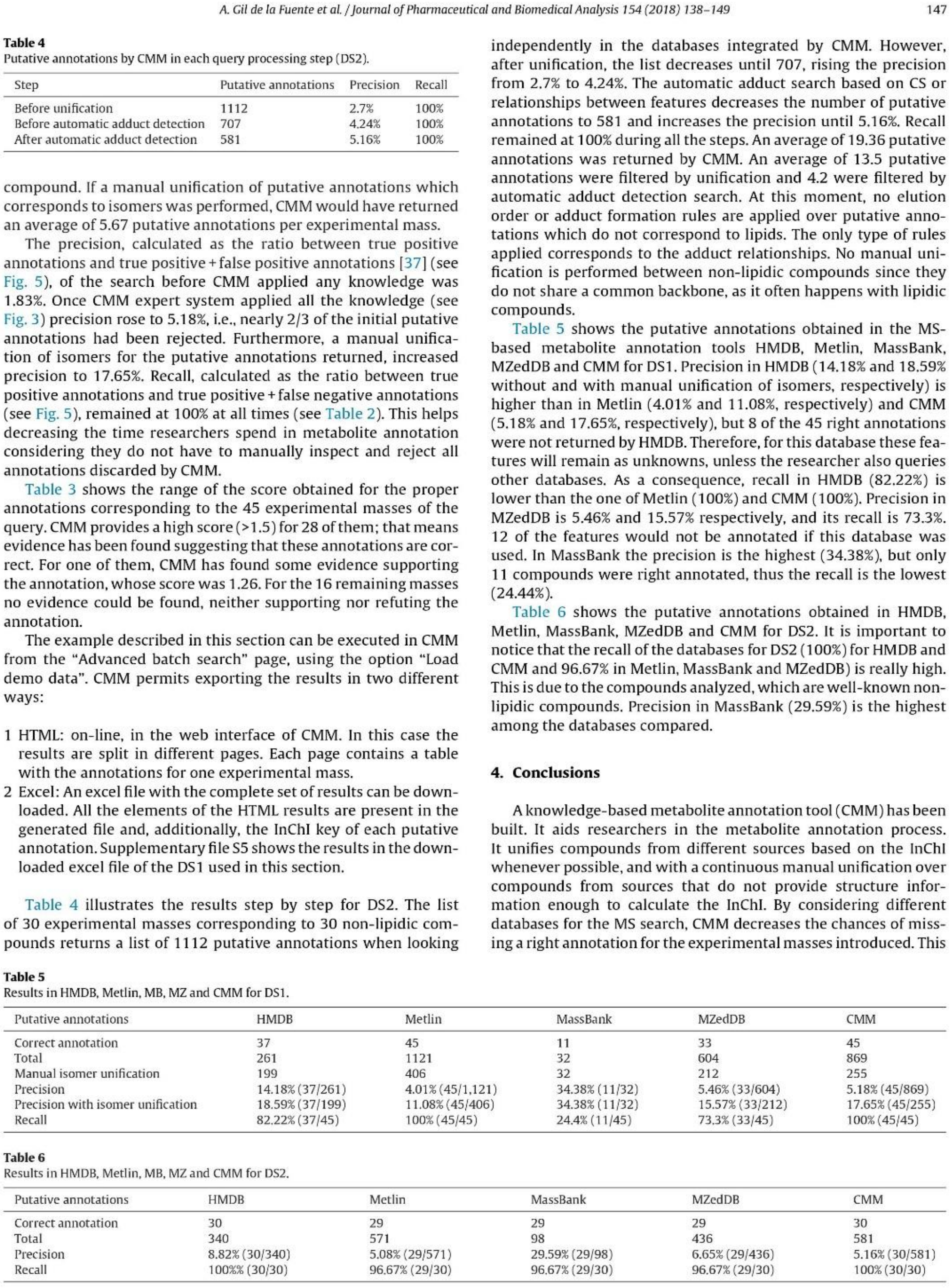


Table 4
Putative annotations by CMM in each query processing step (DS2).

Step	Putative annotations	Precision	Recall
Before unification	1112	2.7%	100%
Before automatic adduct detection	707	4.24%	100%
After automatic adduct detection	581	5.16%	100%

compound. If a manual unification of putative annotations which corresponds to isomers was performed, CMM would have returned an average of 5.67 putative annotations per experimental mass.

The precision, calculated as the ratio between true positive annotations and true positive + false positive annotations [37] (see Fig. 5), of the search before CMM applied any knowledge was 1.83%. Once CMM expert system applied all the knowledge (see Fig. 3) precision rose to 5.18%, i.e., nearly 2/3 of the initial putative annotations had been rejected. Furthermore, a manual unification of isomers for the putative annotations returned, increased precision to 17.65%. Recall, calculated as the ratio between true positive annotations and true positive + false negative annotations (see Fig. 5), remained at 100% at all times (see Table 2). This helps decreasing the time researchers spend in metabolite annotation considering they do not have to manually inspect and reject all annotations discarded by CMM.

Table 3 shows the range of the score obtained for the proper annotations corresponding to the 45 experimental masses of the query. CMM provides a high score (>1.5) for 28 of them; that means evidence has been found suggesting that these annotations are correct. For one of them, CMM has found some evidence supporting the annotation, whose score was 1.26. For the 16 remaining masses no evidence could be found, neither supporting nor refuting the annotation.

The example described in this section can be executed in CMM from the “Advanced batch search” page, using the option “Load demo data”. CMM permits exporting the results in two different ways:

1 HTML: on-line, in the web interface of CMM. In this case the results are split in different pages. Each page contains a table with the annotations for one experimental mass.
2 Excel: An excel file with the complete set of results can be downloaded. All the elements of the HTML results are present in the generated file and, additionally, the InChI key of each putative annotation. Supplementary file S5 shows the results in the downloaded excel file of the DS1 used in this section.

Table 4 illustrates the results step by step for DS2. The list of 30 experimental masses corresponding to 30 non-lipidic compounds returns a list of 1112 putative annotations when looking independently in the databases integrated by CMM. However, after unification, the list decreases until 707, rising the precision from 2.7% to 4.24%. The automatic adduct search based on CS or relationships between features decreases the number of putative annotations to 581 and increases the precision until 5.16%. Recall remained at 100% during all the steps. An average of 19.36 putative annotations was returned by CMM. An average of 13.5 putative annotations were filtered by unification and 4.2 were filtered by automatic adduct detection search. At this moment, no elution order or adduct formation rules are applied over putative annotations which do not correspond to lipids. The only type of rules applied corresponds to the adduct relationships. No manual unification is performed between non-lipidic compounds since they do not share a common backbone, as it often happens with lipidic compounds.

Table 5 shows the putative annotations obtained in the MS-based metabolite annotation tools HMDB, Metlin, MassBank, MZedDB and CMM for DS1. Precision in HMDB (14.18% and 18.59% without and with manual unification of isomers, respectively) is higher than in Metlin (4.01% and 11.08%, respectively) and CMM (5.18% and 17.65%, respectively), but 8 of the 45 right annotations were not returned by HMDB. Therefore, for this database these features will remain as unknowns, unless the researcher also queries other databases. As a consequence, recall in HMDB (82.22%) is lower than the one of Metlin (100%) and CMM (100%). Precision in MZedDB is 5.46% and 15.57% respectively, and its recall is 73.3%. 12 of the features would not be annotated if this database was used. In MassBank the precision is the highest (34.38%), but only 11 compounds were right annotated, thus the recall is the lowest (24.44%).

Table 6 shows the putative annotations obtained in HMDB, Metlin, MassBank, MZedDB and CMM for DS2. It is important to notice that the recall of the databases for DS2 (100%) for HMDB and CMM and 96.67% in Metlin, MassBank and MZedDB) is really high. This is due to the compounds analyzed, which are well-known non-lipidic compounds. Precision in MassBank (29.59%) is the highest among the databases compared.

4. Conclusions

A knowledge-based metabolite annotation tool (CMM) has been built. It aids researchers in the metabolite annotation process. It unifies compounds from different sources based on the InChI whenever possible, and with a continuous manual unification over compounds from sources that do not provide structure information enough to calculate the InChI. By considering different databases for the MS search, CMM decreases the chances of missing a right annotation for the experimental masses introduced. This

Table 5
Results in HMDB, Metlin, MB, MZ and CMM for DS1.

Putative annotations	HMDB	Metlin	MassBank	MZedDB	CMM
Correct annotation	37	45	11	33	45
Total	261	1121	32	604	869
Manual isomer unification	199	406	32	212	255
Precision	14.18% (37/261)	4.01% (45/1,121)	34.38% (11/32)	5.46% (33/604)	5.18% (45/869)
Precision with isomer unification	18.59% (37/199)	11.08% (45/406)	34.38% (11/32)	15.57% (33/212)	17.65% (45/255)
Recall	82.22% (37/45)	100% (45/45)	24.4% (11/45)	73.3% (33/45)	100% (45/45)

Table 6
Results in HMDB, Metlin, MB, MZ and CMM for DS2.

Putative annotations	HMDB	Metlin	MassBank	MZedDB	CMM
Correct annotation	30	29	29	29	30
Total	340	571	98	436	581
Precision	8.82% (30/340)	5.08% (29/571)	29.59% (29/98)	6.65% (29/436)	5.16% (30/581)
Recall	100%% (30/30)	96.67% (29/30)	96.67% (29/30)	96.67% (29/30)	100% (30/30)

is very important given the small overlap of metabolites between databases.

Having a large number of compounds in the database increases the possibility of finding the right annotation. However, it increases significantly the time needed for filtering the returned annotations. Using CMM this time is reduced, first by unification of compounds from different sources based on the InChI, and second by the automatic application of expert knowledge. CMM represents researchers' knowledge as different types of rules which are sequentially applied over the previously unified compounds. The rules available in CMM are related to ionization and adduct formation, relationships between different signals corresponding to the same feature and RT order; i.e., they are the type of rules that experienced researchers apply manually in the annotation process. Based on them, CMM filters and scores the putative annotations and allows the researchers to focus on the most plausible annotations. This approach is not only useful for MS^1, but also as a first filter when MS^n data is available.

CMM has been compared with four of the most popular tools for MS^1 searches: HMDB, Metlin, MassBank and MZedDB. The results of this comparison shows that the precision was higher in comparison to Metlin while the same recall was obtained. Regarding HMDB, the recall was higher in CMM than HMDB, since there are compounds on the data set that are not present in HMDB. Precision in HMDB is higher than in CMM. MZedDB's precision was a little higher than CMM and Metlin for DS2, but the lack of updates since 2009 makes that the compounds included in the original databases since then will not be available in this tool. Recall and precision for DS1 were lower than HMDB, Metlin and CMM. MassBank, as a public repository data, has spectra data uploaded by the users. The tool has been designed for identification with MS^n data than for annotation of compounds with MS^1 data. The information provided may be useful for users when working in similar experimental conditions than the information available there. The recall for DS1 was the lowest and most of the features would remained as unknown, concretely 34 of 45.

Summarizing, CMM searches over several databases at once, obtaining a large list of putative annotations. Subsequently, the putative annotations are filtered and scored based on expert knowledge to guide researchers through the list of results. This makes CMM a unique tool in the field of metabolites annotation for metabolomics MS-data, since it is based on researchers' knowledge instead of statistical approaches. According to our experience using CMM internally at CEMBIO for two years, thanks to the automatic scoring provided by CMM the time of manual filtering of metabolites is substantially reduced, allowing the researcher to focus on the most relevant (higher scored) annotations. CMM is publicly available on http://ceumass.eps.uspceu.es. CMM is easily extensible to incorporate new expert knowledge to score/refute metabolite annotations by creating new spreadsheets containing the new rules.

The authors will welcome any feedback or suggestions related to CMM functionality, such as new rules to incorporate in the tool, that could improve the scoring performed by CMM.

Funding

This work has been supported by grants from the Spanish Ministerio de Economía y Competitividad (grant number CTQ2014-55279-R) and USPCEU PCON10/2016. AGF acknowledges Fundación Universitaria San Pablo CEU for his PhD fellowship.

Conflict of interest

None declared.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jpba.2018.02.046.

References

[1] R. Tautenhahn, K. Cho, W. Uritboonthai, Z.J. Zhu, G.J. Patti, G. Siuzdak, An accelerated workflow for untargeted metabolomics using the metlin database, Nat. Biotechnol. 30 (9) (2012) 826–828.

[2] W.B. Dunn, A. Erban, R. Weber, D.J. Creek, M. Brown, R. Breitling, T. Hankemeier, R. Goodacre, S. Neumann, J. Kopka, M.R. Viant, Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics, Metabolomics 9 (1) (2013) S44–S66.

[3] C.H. Johnson, J. Ivanisevic, G. Siuzdak, Metabolomics: beyond biomarkers and towards mechanisms, Nat. Rev. Mol. Cell Biol. 17 (7) (2016) 451–459.

[4] C. Ruttkies, S. Wolf, S. Neumann, E.L. Schymanski, J. Hollender, Metfrag relaunched: incorporating strategies beyond in silico fragmentation, J. Cheminform. 8 (2016) 3 (1).

[5] L. Ridder, S. Verhoeven, R.V. Schaik, D.H. Van, J. Vervoort, R.C.H.D. Vos, Substructure-based annotation of high-resolution multistage msn spectral trees, Rapid Commun. Mass Spectrom. 26 (20) (2012) 2461–2471.

[6] K. Dührkop, H. Shen, M. Meusel, J. Rousu, S. Böcker, Searching molecular structure databases with tandem mass spectra using csi:fingerid, Proc. Natl. Acad. Sci. U. S. A. 112 (41) (2015) 12580–12585.

[7] A.G. de la Fuente, E.G. Armitage, A. Otero, C. Barbas, J. Godzien, Differentiating signals to make biological sense – a guide through databases for ms-based non-targeted metabolomics, Electrophoresis (2017), http://dx.doi.org/10.1002/elps.201700070 (in press).

[8] J. Godzien, E.G. Armitage, F.J. Rupérez, C. Barbas, M. Ciborowski, I. Jorge, E. Camafeita, J. Vázquez, E. Burillo, J.L. Martín-Ventura, A single in-vial dual extraction strategy for the simultaneous lipidomics and proteomics analysis of hdl and ldl fractions, J. Proteome Res. 15 (6) (2016) 1762–1775.

[9] J. Godzien, M. Ciborowski, M.P. Martínez-Alcázar, P. Samczuk, A. Kretowski, C. Barbas, Rapid and reliable identification of phospholipids for untargeted metabolomics with lc-esi-qtof-ms/ms, J. Proteome Res. 14 (8) (2015) 3204–3216.

[10] G.M. Randazzo, D. Tonoli, P. Strajhar, I. Xenarios, A. Odermatt, J. Boccard, S. Rudaz, Enhanced metabolite annotation via dynamic retention time prediction: steroidogenesis alterations as a case study, J. Chromatogr. B (2017), http://dx.doi.org/10.1016/j.jchromb.2017.04.032.

[11] M. Cao, K. Fraser, J. Huege, T. Featonby, S. Rasmussen, C. Jones, Predicting retention time in hydrophilic interaction liquid chromatography mass spectrometry and its use for peak annotation in metabolomics, Metabolomics 11 (3) (2014) 696–706.

[12] T. Hagiwara, S. Saito, Y. Ujiie, K. Imai, M. Kakuta, K. Kadota, T. Terada, K. Sumikoshi, K. Shimizu, T. Nishi, HPLC retention time prediction for metabolome analysis, Bioinformation 5 (6) (2010) 255–258.

[13] J. Stanstrup, S. Neumann, U. Vrhovsek, Predret: prediction of retention time by direct mapping between multiple chromatographic systems, Anal. Chem. 87 (18) (2015) 9421–9428.

[14] S. Milne, P. Ivanova, J. Forrester, H.A. Brown, Lipidomics: an analysis of cellular lipids by ESI-MS, Methods 39 (2006) 92–103, http://dx.doi.org/10.1016/j.ymeth.2006.05.014.

[15] J. Draper, D.P. Enot, D. Parker, M. Beckmann, S. Snowdon, W. Lin, H. Zubair, Metabolite signal identification in accurate mass metabolomics data with mzeddb, an interactive m/z annotation tool utilising predicted ionisation behaviour 'rules', BMC Bioinform. (2009).

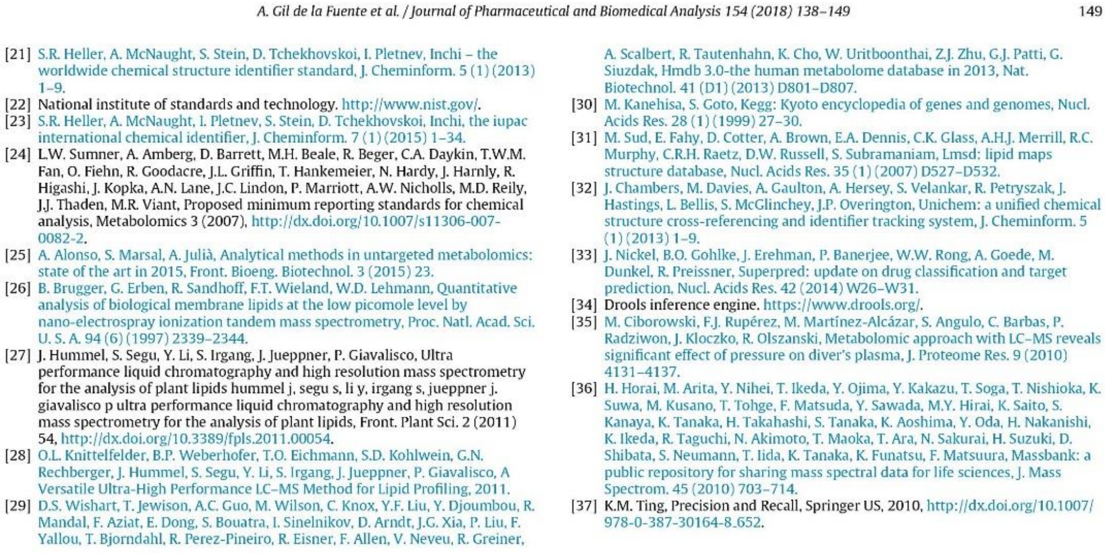
[16] K.S. Lynn, T.Y. Sung, W.L. Hsu, M.L. Cheng, A. Chen, M.S. Shiao, Y.R. Chen, C. Hsu, T.M. Lih, H.Y. Chang, C.J. Huang, W.H. Pan, Metabolite identification for mass spectrometry-based metabolomics using multiple types of correlated ion information, Anal. Chem. 87 (4) (2015) 2143–2151.

[17] C. van Rijswijk Merlijn, C. Beirnaert, M. Caron, V. Cascante, W.B. Dominguez, M.D.E. Dunn, F. Timothy, A. Giacomoni, T. Gonzalez-Beltran, K. Hankemeier, J.L. Haug, R.C. Izquierdo, F. Jimenez, N. Jourdan, M.I. Kale, O. Klapa, K. Kohlbacher, K. Koort, G.L. Kultima, P. Corguill, N.K. Moreno, S. Moschonas, C. Neumann, M. O'Donovan, P. Reczko, A. Rocca-Serra, R.M. Rosato, S.-A. Salek, V. Sansone, D. Satagopam, R. Schober, R.A. Shimmo, O. Spicer, E.A. Spjuth, M.R. Th, J.M.W. Viant, E.L. Ralf, G. Willighagen, C. Zanetti, Steinbeck, The future of metabolomics in elixir [version 2; referees: 2 approved, 1 approved with reservations], F1000Research 6 (2017) 1649, http://dx.doi.org/10.12688/f1000research.12342.2.

[18] J.G. Jeffryes, L.J. Broadbelt, K.E.J. Tyo, R.L. Colastani, C.S. Henry, M. Elbadawi-Sidhu, T. Kind, O. Fiehn, T.D. Niehaus, A.D. Hanson, MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics, J. Cheminform. 7 (2015) 44.

[19] B. Wägele, M. Witting, P. Schmitt-Kopplin, K. Suhre, Masstrix reloaded: combined analysis and visualization of transcriptome and metabolome data, PLoS ONE 7 (7) (2012) 1–10, http://dx.doi.org/10.1371/journal.pone.0039860.

[20] M. Sud, E. Fahy, D. Cotter, K. Azam, I. Vadivelu, S. Subramaniam, C. Burant, A. Edison, O. Fiehn, R. Higashi, K. Nair, S. Sumner, Metabolomics workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools, Nucleic Acids Res. 44 (2016) D463–D470.



[21] S.R. Heller, A. McNaught, S. Stein, D. Tchekhovskoi, I. Pletnev, Inchi – the worldwide chemical structure identifier standard, J. Cheminform. 5 (1) (2013) 1–9.

[22] National institute of standards and technology. http://www.nist.gov/.

[23] S.R. Heller, A. McNaught, I. Pletnev, S. Stein, D. Tchekhovskoi, Inchi, the iupac international chemical identifier, J. Cheminform. 7 (1) (2015) 1–34.

[24] L.W. Sumner, A. Amberg, D. Barrett, M.H. Beale, R. Beger, C.A. Daykin, T.W.M. Fan, O. Fiehn, R. Goodacre, J.L. Griffin, T. Hankemeier, N. Hardy, J. Harnly, R. Higashi, J. Kopka, A.N. Lane, J.C. Lindon, P. Marriott, A.W. Nicholls, M.D. Reily, J.J. Thaden, M.R. Viant, Proposed minimum reporting standards for chemical analysis, Metabolomics 3 (2007), http://dx.doi.org/10.1007/s11306-007-0082-2.

[25] A. Alonso, S. Marsal, A. Julià, Analytical methods in untargeted metabolomics: state of the art in 2015, Front. Bioeng. Biotechnol. 3 (2015) 23.

[26] B. Brugger, G. Erben, R. Sandhoff, F.T. Wieland, W.D. Lehmann, Quantitative analysis of biological membrane lipids at the low picomole level by nano-electrospray ionization tandem mass spectrometry, Proc. Natl. Acad. Sci. U. S. A. 94 (6) (1997) 2339–2344.

[27] J. Hummel, S. Segu, Y. Li, S. Irgang, J. Jueppner, P. Giavalisco, Ultra performance liquid chromatography and high resolution mass spectrometry for the analysis of plant lipids hummel j, segu s, li y, irgang s, jueppner j. giavalisco p ultra performance liquid chromatography and high resolution mass spectrometry for the analysis of plant lipids, Front. Plant Sci. 2 (2011) 54, http://dx.doi.org/10.3389/fpls.2011.00054.

[28] O.L. Knittelfelder, B.P. Weberhofer, T.O. Eichmann, S.D. Kohlwein, G.N. Rechberger, J. Hummel, S. Segu, Y. Li, S. Irgang, J. Jueppner, P. Giavalisco, A Versatile Ultra-High Performance LC–MS Method for Lipid Profiling, 2011.

[29] D.S. Wishart, T. Jewison, A.C. Guo, M. Wilson, C. Knox, Y.F. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong, S. Bouatra, I. Sinelnikov, D. Arndt, J.G. Xia, P. Liu, F. Yallou, T. Bjorndahl, R. Perez-Pineiro, R. Eisner, F. Allen, V. Neveu, R. Greiner, A. Scalbert, R. Tautenhahn, K. Cho, W. Uritboonthai, Z.J. Zhu, G.J. Patti, G. Siuzdak, Hmdb 3.0-the human metabolome database in 2013, Nat. Biotechnol. 41 (D1) (2013) D801–D807.

[30] M. Kanehisa, S. Goto, Kegg: Kyoto encyclopedia of genes and genomes, Nucl. Acids Res. 28 (1) (1999) 27–30.

[31] M. Sud, E. Fahy, D. Cotter, A. Brown, E.A. Dennis, C.K. Glass, A.H.J. Merrill, R.C. Murphy, C.R.H. Raetz, D.W. Russell, S. Subramaniam, Lmsd: lipid maps structure database, Nucl. Acids Res. 35 (1) (2007) D527–D532.

[32] J. Chambers, M. Davies, A. Gaulton, A. Hersey, S. Velankar, R. Petryszak, J. Hastings, L. Bellis, S. McGlinchey, J.P. Overington, Unichem: a unified chemical structure cross-referencing and identifier tracking system, J. Cheminform. 5 (1) (2013) 1–9.

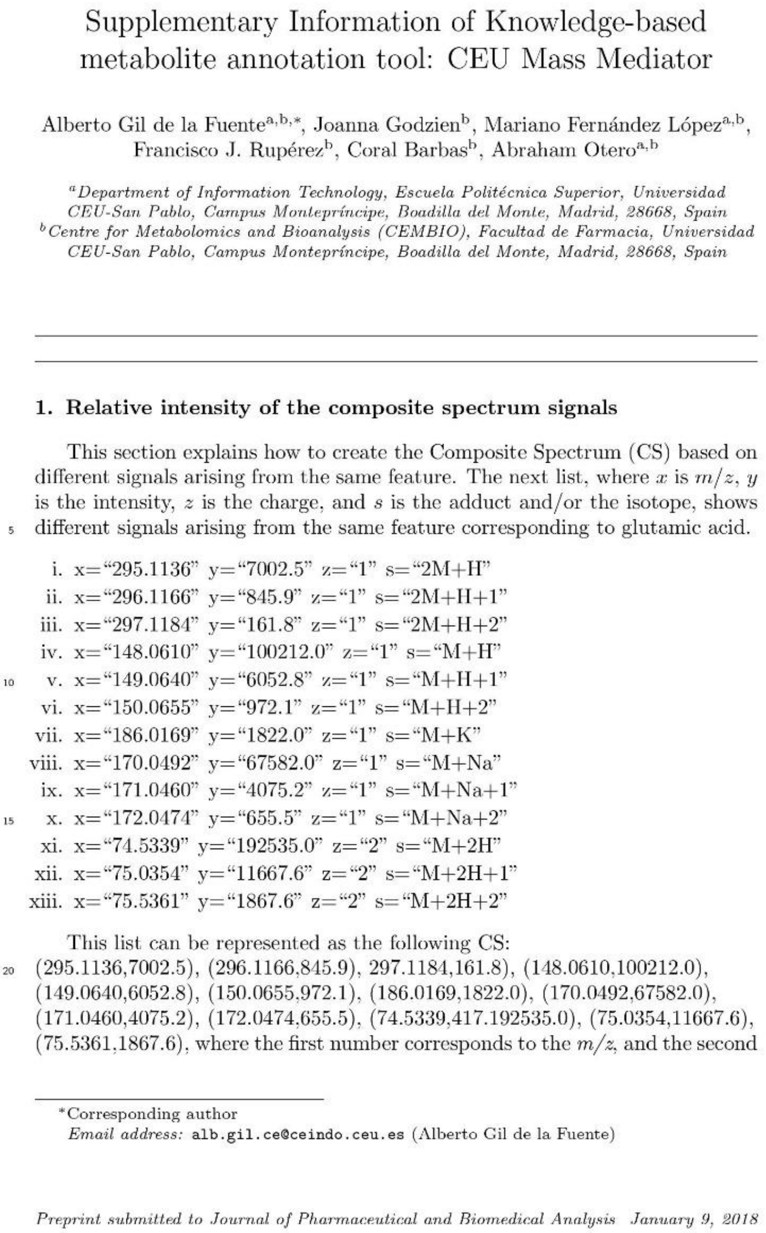
[33] J. Nickel, B.O. Gohlke, J. Erehman, P. Banerjee, W.W. Rong, A. Goede, M. Dunkel, R. Preissner, Superpred: update on drug classification and target prediction, Nucl. Acids Res. 42 (2014) W26–W31.

[34] Drools inference engine. https://www.drools.org/.

[35] M. Ciborowski, F.J. Rupérez, M. Martínez-Alcázar, S. Angulo, C. Barbas, P. Radziwon, J. Kloczko, R. Olszanski, Metabolomic approach with LC–MS reveals significant effect of pressure on diver's plasma, J. Proteome Res. 9 (2010) 4131–4137.

[36] H. Horai, M. Arita, Y. Nihei, T. Ikeda, Y. Ojima, Y. Kakazu, T. Soga, T. Nishioka, K. Suwa, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M.Y. Hirai, K. Saito, S. Kanaya, K. Tanaka, H. Takahashi, S. Tanaka, K. Aoshima, Y. Oda, H. Nakanishi, K. Ikeda, R. Taguchi, N. Akimoto, T. Maoka, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, Massbank: a public repository for sharing mass spectral data for life sciences, J. Mass Spectrom. 45 (2010) 703–714.

[37] K.M. Ting, Precision and Recall, Springer US, 2010, http://dx.doi.org/10.1007/978-0-387-30164-8_652.



Supplementary Information of Knowledge-based metabolite annotation tool: CEU Mass Mediator

Alberto Gil de la Fuente[a,b,*], Joanna Godzien[b], Mariano Fernández López[a,b], Francisco J. Rupérez[b], Coral Barbas[b], Abraham Otero[a,b]

[a] *Department of Information Technology, Escuela Politécnica Superior, Universidad CEU-San Pablo, Campus Montepríncipe, Boadilla del Monte, Madrid, 28668, Spain*
[b] *Centre for Metabolomics and Bioanalysis (CEMBIO), Facultad de Farmacia, Universidad CEU-San Pablo, Campus Montepríncipe, Boadilla del Monte, Madrid, 28668, Spain*

1. Relative intensity of the composite spectrum signals

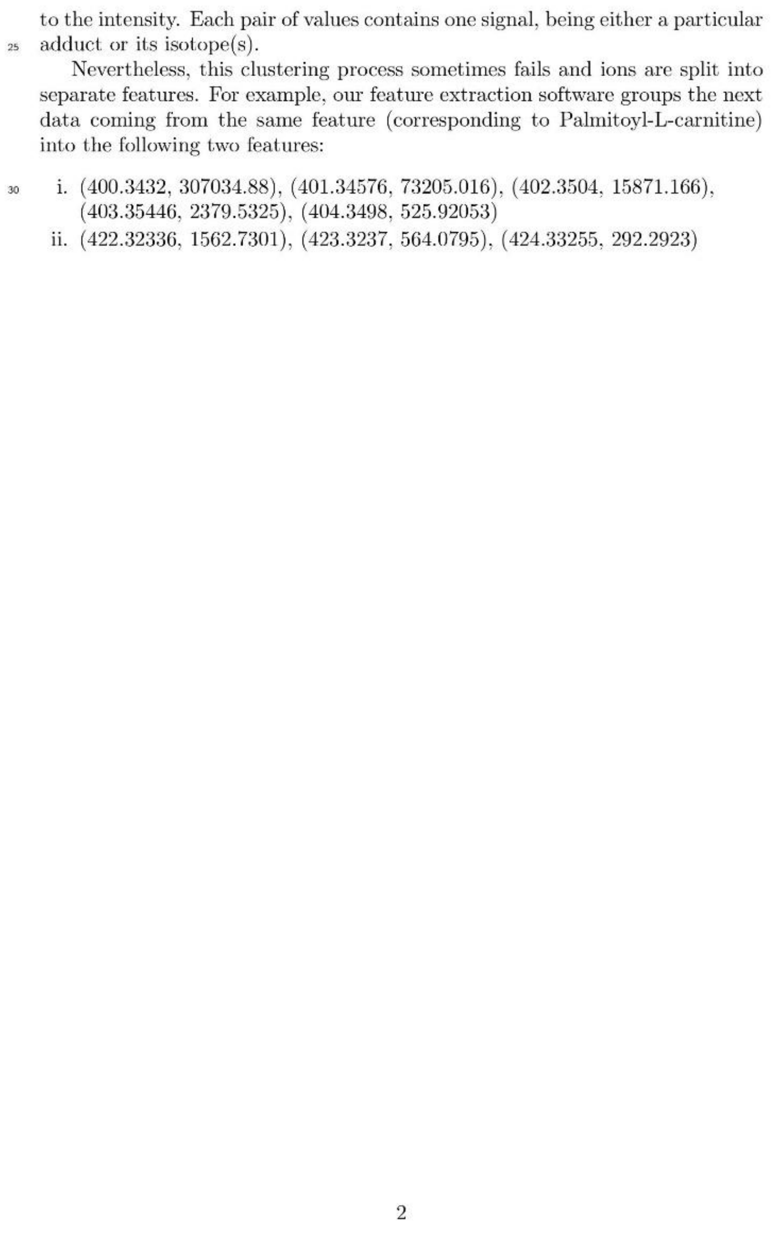
This section explains how to create the Composite Spectrum (CS) based on different signals arising from the same feature. The next list, where x is m/z, y is the intensity, z is the charge, and s is the adduct and/or the isotope, shows different signals arising from the same feature corresponding to glutamic acid.

i. x="295.1136" y="7002.5" z="1" s="2M+H"
ii. x="296.1166" y="845.9" z="1" s="2M+H+1"
iii. x="297.1184" y="161.8" z="1" s="2M+H+2"
iv. x="148.0610" y="100212.0" z="1" s="M+H"
v. x="149.0640" y="6052.8" z="1" s="M+H+1"
vi. x="150.0655" y="972.1" z="1" s="M+H+2"
vii. x="186.0169" y="1822.0" z="1" s="M+K"
viii. x="170.0492" y="67582.0" z="1" s="M+Na"
ix. x="171.0460" y="4075.2" z="1" s="M+Na+1"
x. x="172.0474" y="655.5" z="1" s="M+Na+2"
xi. x="74.5339" y="192535.0" z="2" s="M+2H"
xii. x="75.0354" y="11667.6" z="2" s="M+2H+1"
xiii. x="75.5361" y="1867.6" z="2" s="M+2H+2"

This list can be represented as the following CS: (295.1136,7002.5), (296.1166,845.9), 297.1184,161.8), (148.0610,100212.0), (149.0640,6052.8), (150.0655,972.1), (186.0169,1822.0), (170.0492,67582.0), (171.0460,4075.2), (172.0474,655.5), (74.5339,417.192535.0), (75.0354,11667.6), (75.5361,1867.6), where the first number corresponds to the m/z, and the second

[*] Corresponding author
Email address: alb.gil.ce@ceindo.ceu.es (Alberto Gil de la Fuente)

Preprint submitted to Journal of Pharmaceutical and Biomedical Analysis January 9, 2018



to the intensity. Each pair of values contains one signal, being either a particular adduct or its isotope(s).

Nevertheless, this clustering process sometimes fails and ions are split into separate features. For example, our feature extraction software groups the next data coming from the same feature (corresponding to Palmitoyl-L-carnitine) into the following two features:

i. (400.3432, 307034.88), (401.34576, 73205.016), (402.3504, 15871.166), (403.35446, 2379.5325), (404.3498, 525.92053)

ii. (422.32336, 1562.7301), (423.3237, 564.0795), (424.33255, 292.2923)

SI 2: information about ionization and adduct formation rules **using NH3** in the mobile phase modifiers

group of lipids	lipid maps nomenclature	flag	adduct formation	
			positive adducts	negative adducts
FA	Category: Fatty Acyls [FA]	expected	M+H, M+H-H2O, M+H-2H2O	M-H
	Main class: Fatty Acids and Conjugates [FA01]	possible	M+Na (only if M+H is found)	M-H-H2O
		not expected / impossible	N/A	N/A
PC	Category: Glycerophospholipids [GP]	expected	M+H	M+HCOO// M+CH3COO
	Main class: Glycerophosphocholines [GP01]	possible	M+Na, M+K (only if M+H is found)	M+Cl
	Class: Diacylglycerophosphocholines [GP0101], 1-alkyl,2-acylglycerophosphocholines [GP0102], 1-acyl,2-alkylglycerophosphocholines [GP0108], 1-(1Z-alkenyl),2-acylglycerophosphocholines [GP0103], 1-acyl,2-(1Z-alkenyl)-glycerophosphocholines [GP0109], Dialkylglycerophosphocholines [GP0104]	not expected / impossible	M+H-H2O	M-H, M-H-H2O
LPC	Category: Glycerophospholipids [GP]	expected	M+H, M+H-H2O	M+HCOO// M+CH3COO
	Main class: Glycerophosphocholines [GP01]	possible	M+Na, M+K (only if M+H is found)	M+Cl-, M-H-H2O
	Class: Monoacylglycerophosphocholines [GP0105], Monoalkylglycerophosphocholines [GP0106], 1Z-alkenylglycerophosphocholines [GP0107]	not expected / impossible	N/A	M-H

PE	Category: Glycerophospholipids [GP]	expected	M+H	M-H, M+HCOO // M+CH3COO
	Main class: Glycerophosphoethanolamines [GP02]	possible	M+Na, M+K (only if M+H is found)	M+Cl
	Diacylglycerophosphoethanolamines [GP0201], 1-alkyl,2-acylglycerophosphoethanolamines [GP0202], 1-acyl,2-alkylglycerophosphoethanolamines [GP0208], 1-(1Z-alkenyl),2-acylglycerophosphoethanolamines [GP0203], Dialkylglycerophosphoethanolamines [GP0204]	not expected / impossible	M+H-H2O	M-H-H2O
LPE	Category: Glycerophospholipids [GP]	expected	M+H	M-H, M+HCOO // M+CH3COO
	Main class: Glycerophosphoethanolamines [GP02]	possible	M+Na, M+K, (only if M+H is found) M+H-H2O	M+Cl, M-H-H2O
	Class: Monoacylglycerophosphoethanolamines [GP0205], Monoalkylglycerophosphoethanolamines [GP0206], 1Z-alkenylglycerophosphoethanolamines [GP0207]	not expected / impossible	N/A	N/A
PI	Category: Glycerophospholipids [GP]	expected	N/A	M-H
	Main class: Glycerophosphoinositols [GP06]	possible	M+Na, M+K	N/A
	Class: Diacylglycerophosphoinositols [GP0601], 1-alkyl,2-acylglycerophosphoinositols [GP0602], 1-(1Z-alkenyl),2-acylglycerophosphoinositols [GP0603], Dialkylglycerophosphoinositols [GP0604]	not expected / impossible	M+H, M+H-H2O	N/A
PG	Category: Glycerophospholipids [GP]	expected	not expected	M-H, M+HCOO // M+CH3COO
	Main class: Glycerophosphoglycerols [GP04]	possible		M+Cl
	Class: Diacylglycerophosphoglycerols [GP0401], 1-alkyl,2-acylglycerophosphoglycerols [GP0402], 1-acyl,2-alkylglycerophosphoglycerols	not expected / impossible		M-H-H2O

	[GP0411], 1-(1Z-alkenyl),2-acylglycerophosphoglycerols [GP0403], Dialkylglycerophosphoglycerols [GP0404]			
PS	Category: Glycerophospholipids [GP]	expected	M+H	M-H, M+HCOO// M+CH3COO
	Main class: Glycerophosphoserines [GP03]	possible	M+Na, M+K (only if M+H is found)	M+Cl
	Class: Diacylglycerophosphoserines [GP0301], 1-alkyl,2-acylglycerophosphoserines [GP0302], 1-(1Z-alkenyl),2-acylglycerophosphoserines [GP0303], Dialkylglycerophosphoserines [GP0304]	not expected / impossible	M+H-H2O	N/A
LPS	Category: Glycerophospholipids [GP]	expected	M+H, M+H-H2O	M-H, M+HCOO // M+CH3COO
	Main class: Glycerophosphoserines [GP03]	possible	M+Na, M+K (only if M+H is found)	M+Cl
	Class: Monoacylglycerophosphoserines [GP0305], Monoalkylglycerophosphoserines [GP0306], 1Z-alkenylglycerophosphoserines [GP0307]	not expected / impossible	N/A	N/A
PA	Category: Glycerophospholipids [GP]	expected	not expected	M-H, M+HCOO // M+CH3COO
	Main class: Glycerophosphates [GP10]	possible		M+Cl
	Class: Diacylglycerophosphates [GP1001], 1-alkyl,2-acylglycerophosphates [GP1002], 1-(1Z-alkenyl),2-acylglycerophosphates [GP1003], Dialkylglycerophosphates [GP1004]	not expected / impossible		N/A
MG	Category: Glycerolipids [GL]	expected	M+H	not expected
	Main class: Monoradylglycerols [GL01]	possible	M+Na	
		not expected / impossible	M+NH4	
DG	Category: Glycerolipids [GL]	expected	N/A	not expected

	Main class: Diradylglycerols [GL02]	possible	M+Na	
		not expected / impossible	M+NH4, M+H	
TG	Category: Glycerolipids [GL]	expected	N/A	not expected
	Main class: Triradylglycerols [GL03]	possible	M+Na	
		not expected / impossible	M+NH4, M+H	
CER	Category: Sphingolipids [SP]	expected	M+H	M-H, M+HCOO // M+CH3COO
	Main class: Ceramides [SP02]	possible	M+Na (only if M+H is found)	M+Cl
		not expected / impossible	N/A	N/A
SM	Category: Sphingolipids [SP]	expected	M+H	M+HCOO // M+CH3COO
	Main class: Phosphosphingolipids [SP03]	possible	M+Na, M+K	M+Cl
		not expected / impossible	N/A	M-H
CE	Category: Sterol Lipids [ST]	expected	M+H, M+H-H2O	not expected
	Main class: Sterols [ST01]	possible	M+Na (only if M+H is found)	
		not expected / impossible	M+NH4	

SI 2: information about ionization and adduct formation rules **not using NH3** in the mobile phase modifiers

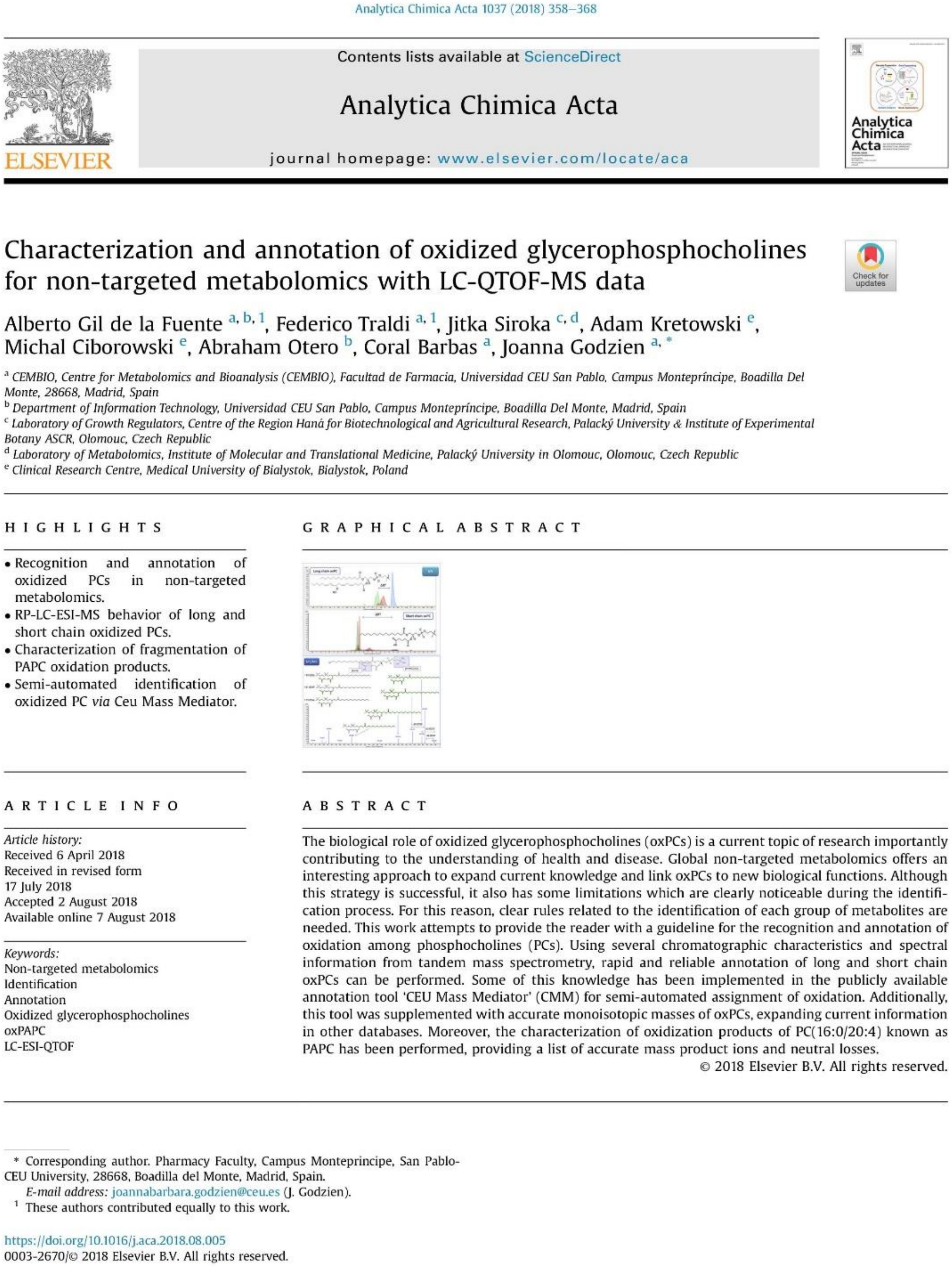
group of lipids	lipid maps nomenclature	flag	adduct formation	
			positive adducts	negative adducts
FA	Category: Fatty Acyls [FA]	expected	M+H, M+H-H2O, M+H-2H2O	M-H
	Main class: Fatty Acids and Conjugates [FA01]	possible	M+Na (only if M+H is found)	M-H-H2O
		not expected / impossible	N/A	N/A
PC	Category: Glycerophospholipids [GP]	expected	M+H	M+HCOO// M+CH3COO
	Main class: Glycerophosphocholines [GP01]	possible	M+Na, M+K (only if M+H is found)	M+Cl-
	Class: Diacylglycerophosphocholines [GP0101], 1-alkyl,2-acylglycerophosphocholines [GP0102], 1-acyl,2-alkylglycerophosphocholines [GP0108], 1-(1Z-alkenyl),2-acylglycerophosphocholines [GP0103], 1-acyl,2-(1Z-alkenyl)-glycerophosphocholines [GP0109], Dialkylglycerophosphocholines [GP0104]	not expected / impossible	M+H-H2O	M-H, M-H-H2O
LPC	Category: Glycerophospholipids [GP]	expected	M+H, M+H-H2O	M+HCOO // M+CH3COO-
	Main class: Glycerophosphocholines [GP01]	possible	M+Na, M+K (only if M+H is found)	M+Cl-, M-H-H2O
	Class: Monoacylglycerophosphocholines [GP0105], Monoalkylglycerophosphocholines [GP0106], 1Z-alkenylglycerophosphocholines [GP0107]	not expected / impossible	N/A	M-H

PE	Category: Glycerophospholipids [GP]	expected	M+H	M-H, M+HCOO // M+CH3COO
	Main class: Glycerophosphoethanolamines [GP02]	possible	M+Na, M+K (only if M+H is found)	M+Cl
	Diacylglycerophosphoethanolamines [GP0201], 1-alkyl,2-acylglycerophosphoethanolamines [GP0202], 1-acyl,2-alkylglycerophosphoethanolamines [GP0208], 1-(1Z-alkenyl),2-acylglycerophosphoethanolamines [GP0203], Dialkylglycerophosphoethanolamines [GP0204]	not expected / impossible	M+H-H2O	M-H-H2O
LPE	Category: Glycerophospholipids [GP]	expected	M+H	M-H, M+HCOO // M+CH3COO
	Main class: Glycerophosphoethanolamines [GP02]	possible	M+Na, M+K, (only if M+H is found) M+H-H2O	M+Cl, M-H-H2O
	Class: Monoacylglycerophosphoethanolamines [GP0205], Monoalkylglycerophosphoethanolamines [GP0206], 1Z-alkenylglycerophosphoethanolamines [GP0207]	not expected / impossible	N/A	N/A
PI	Category: Glycerophospholipids [GP]	expected	N/A	M-H
	Main class: Glycerophosphoinositols [GP06]	possible	M+Na, M+K	N/A
	Class: Diacylglycerophosphoinositols [GP0601], 1-alkyl,2-acylglycerophosphoinositols [GP0602], 1-(1Z-alkenyl),2-acylglycerophosphoinositols [GP0603], Dialkylglycerophosphoinositols [GP0604]	not expected / impossible	M+H, M+H-H2O	N/A
PG	Category: Glycerophospholipids [GP]	expected	not expected	M-H, M+HCOO // M+CH3COO
	Main class: Glycerophosphoglycerols [GP04]	possible		M+Cl
	Class: Diacylglycerophosphoglycerols [GP0401], 1-alkyl,2-acylglycerophosphoglycerols [GP0402], 1-acyl,2-alkylglycerophosphoglycerols	not expected / impossible		M-H-H2O

	[GP0411], 1-(1Z-alkenyl),2-acylglycerophosphoglycerols [GP0403], Dialkylglycerophosphoglycerols [GP0404]			
PS	Category: Glycerophospholipids [GP]	expected	M+H	M-H, M+HCOO// M+CH3COO
	Main class: Glycerophosphoserines [GP03]	possible	M+Na, M+K (only if M+H is found)	M+Cl
	Class: Diacylglycerophosphoserines [GP0301], 1-alkyl,2-acylglycerophosphoserines [GP0302], 1-(1Z-alkenyl),2-acylglycerophosphoserines [GP0303], Dialkylglycerophosphoserines [GP0304]	not expected / impossible	M+H-H2O	N/A
LPS	Category: Glycerophospholipids [GP]	expected	M+H, M+H-H2O	M-H, M+HCOO // M+CH3COO
	Main class: Glycerophosphoserines [GP03]	possible	M+Na, M+K (only if M+H is found)	M+Cl
	Class: Monoacylglycerophosphoserines [GP0305], Monoalkylglycerophosphoserines [GP0306], 1Z-alkenylglycerophosphoserines [GP0307]	not expected / impossible	N/A	N/A
PA	Category: Glycerophospholipids [GP]	expected	not expected	M-H, M+HCOO // M+CH3COO
	Main class: Glycerophosphates [GP10]	possible		M+Cl
	Class: Diacylglycerophosphates [GP1001], 1-alkyl,2-acylglycerophosphates [GP1002], 1-(1Z-alkenyl),2-acylglycerophosphates [GP1003], Dialkylglycerophosphates [GP1004]	not expected / impossible		N/A
MG	Category: Glycerolipids [GL]	expected	M+H, M+NH4	not expected
	Main class: Monoradylglycerols [GL01]	possible	M+Na	
		not expected / impossible	N/A	
DG	Category: Glycerolipids [GL]	expected	M`NH4	not expected

	Main class: Diradylglycerols [GL02]	possible	M+Na	
		not expected / impossible	M+H	
TG	Category: Glycerolipids [GL]	expected	M+NH4	not expected
	Main class: Triradylglycerols [GL03]	possible	M+Na	
		not expected / impossible	M+H	
CER	Category: Sphingolipids [SP]	expected	M+H	M-H, M+HCOO // M+CH3COO
	Main class: Ceramides [SP02]	possible	M+Na (only if M+H is found)	M+Cl
		not expected / impossible	N/A	N/A
SM	Category: Sphingolipids [SP]	expected	M+H	M+HCOO // M+CH3COO
	Main class: Phosphosphingolipids [SP03]	possible	M+Na, M+K	M+Cl
		not expected / impossible	N/A	M-H
CE	Category: Sterol Lipids [ST]	expected	M+NH4	not expected
	Main class: Sterols [ST01]	possible	M+H, M+H-H2O, M+Na (only if M+H is found)	
		not expected / impossible	N/A	

CHAPTER 3: CHARACTERIZATION AND ANNOTATION OF OXIDIZED GLYCEROPHOSPHOCHOLINES FOR NON-TARGETED METABOLOMICS WITH LC-QTOF-MS DATA



Analytica Chimica Acta 1037 (2018) 358–368

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Characterization and annotation of oxidized glycerophosphocholines for non-targeted metabolomics with LC-QTOF-MS data

Alberto Gil de la Fuente [a, b, 1], Federico Traldi [a, 1], Jitka Siroka [c, d], Adam Kretowski [e], Michal Ciborowski [e], Abraham Otero [b], Coral Barbas [a], Joanna Godzien [a, *]

[a] CEMBIO, Centre for Metabolomics and Bioanalysis (CEMBIO), Facultad de Farmacia, Universidad CEU San Pablo, Campus Montepríncipe, Boadilla Del Monte, 28668, Madrid, Spain
[b] Department of Information Technology, Universidad CEU San Pablo, Campus Montepríncipe, Boadilla Del Monte, Madrid, Spain
[c] Laboratory of Growth Regulators, Centre of the Region Haná for Biotechnological and Agricultural Research, Palacký University & Institute of Experimental Botany ASCR, Olomouc, Czech Republic
[d] Laboratory of Metabolomics, Institute of Molecular and Translational Medicine, Palacký University in Olomouc, Olomouc, Czech Republic
[e] Clinical Research Centre, Medical University of Bialystok, Bialystok, Poland

HIGHLIGHTS

- Recognition and annotation of oxidized PCs in non-targeted metabolomics.
- RP-LC-ESI-MS behavior of long and short chain oxidized PCs.
- Characterization of fragmentation of PAPC oxidation products.
- Semi-automated identification of oxidized PC *via* Ceu Mass Mediator.

GRAPHICAL ABSTRACT

ARTICLE INFO

Article history:
Received 6 April 2018
Received in revised form
17 July 2018
Accepted 2 August 2018
Available online 7 August 2018

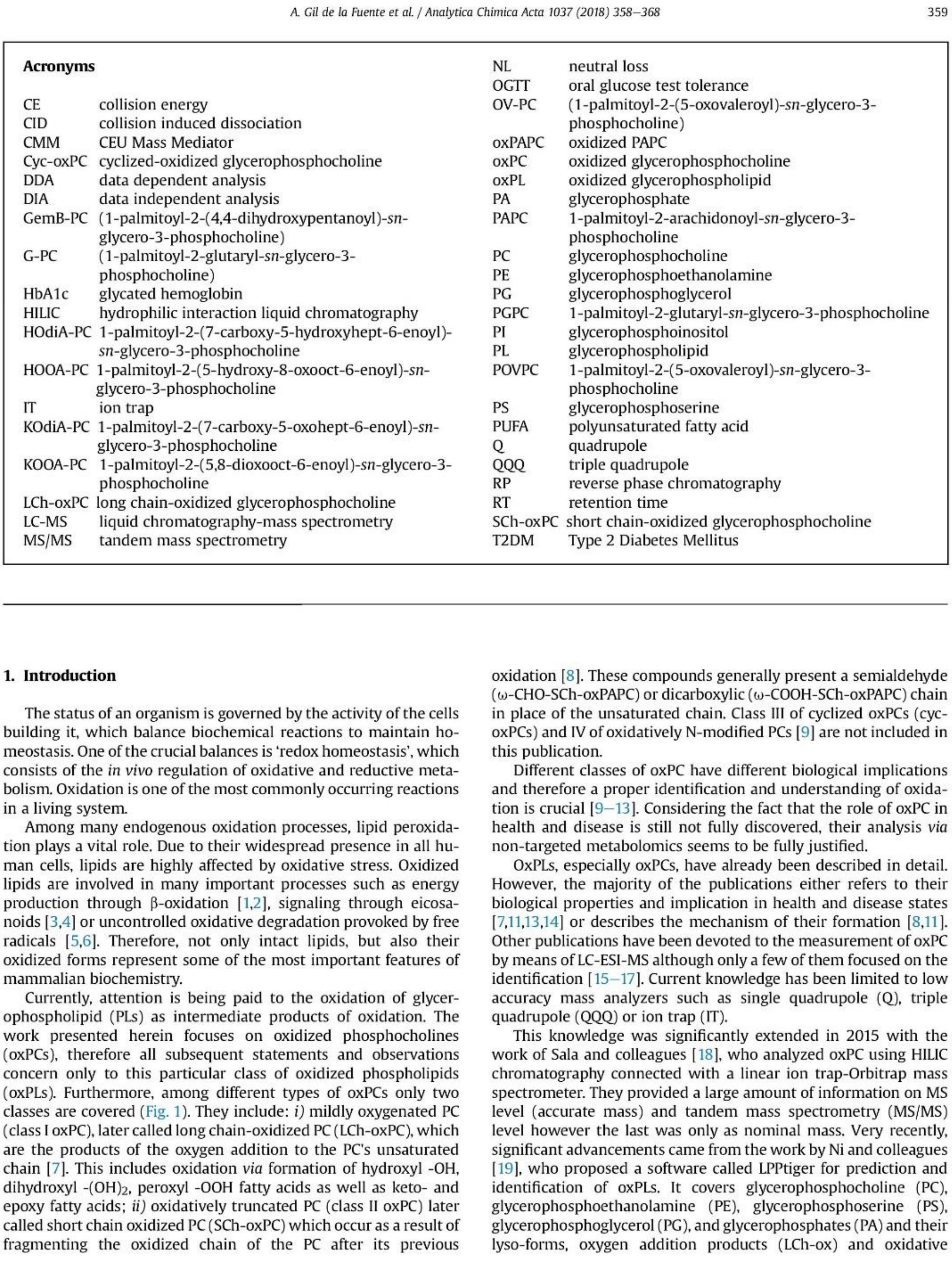
Keywords:
Non-targeted metabolomics
Identification
Annotation
Oxidized glycerophosphocholines
oxPAPC
LC-ESI-QTOF

ABSTRACT

The biological role of oxidized glycerophosphocholines (oxPCs) is a current topic of research importantly contributing to the understanding of health and disease. Global non-targeted metabolomics offers an interesting approach to expand current knowledge and link oxPCs to new biological functions. Although this strategy is successful, it also has some limitations which are clearly noticeable during the identification process. For this reason, clear rules related to the identification of each group of metabolites are needed. This work attempts to provide the reader with a guideline for the recognition and annotation of oxidation among phosphocholines (PCs). Using several chromatographic characteristics and spectral information from tandem mass spectrometry, rapid and reliable annotation of long and short chain oxPCs can be performed. Some of this knowledge has been implemented in the publicly available annotation tool 'CEU Mass Mediator' (CMM) for semi-automated assignment of oxidation. Additionally, this tool was supplemented with accurate monoisotopic masses of oxPCs, expanding current information in other databases. Moreover, the characterization of oxidization products of PC(16:0/20:4) known as PAPC has been performed, providing a list of accurate mass product ions and neutral losses.

[*] Corresponding author. Pharmacy Faculty, Campus Monteprincipe, San Pablo-CEU University, 28668, Boadilla del Monte, Madrid, Spain.
E-mail address: joannabarbara.godzien@ceu.es (J. Godzien).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.aca.2018.08.005
0003-2670/© 2018 Elsevier B.V. All rights reserved.



Acronyms

CE	collision energy
CID	collision induced dissociation
CMM	CEU Mass Mediator
Cyc-oxPC	cyclized-oxidized glycerophosphocholine
DDA	data dependent analysis
DIA	data independent analysis
GemB-PC	(1-palmitoyl-2-(4,4-dihydroxypentanoyl)-*sn*-glycero-3-phosphocholine)
G-PC	(1-palmitoyl-2-glutaryl-*sn*-glycero-3-phosphocholine)
HbA1c	glycated hemoglobin
HILIC	hydrophilic interaction liquid chromatography
HOdiA-PC	1-palmitoyl-2-(7-carboxy-5-hydroxyhept-6-enoyl)-*sn*-glycero-3-phosphocholine
HOOA-PC	1-palmitoyl-2-(5-hydroxy-8-oxooct-6-enoyl)-*sn*-glycero-3-phosphocholine
IT	ion trap
KOdiA-PC	1-palmitoyl-2-(7-carboxy-5-oxohept-6-enoyl)-*sn*-glycero-3-phosphocholine
KOOA-PC	1-palmitoyl-2-(5,8-dioxooct-6-enoyl)-*sn*-glycero-3-phosphocholine
LCh-oxPC	long chain-oxidized glycerophosphocholine
LC-MS	liquid chromatography-mass spectrometry
MS/MS	tandem mass spectrometry
NL	neutral loss
OGTT	oral glucose test tolerance
OV-PC	(1-palmitoyl-2-(5-oxovaleroyl)-*sn*-glycero-3-phosphocholine)
oxPAPC	oxidized PAPC
oxPC	oxidized glycerophosphocholine
oxPL	oxidized glycerophospholipid
PA	glycerophosphate
PAPC	1-palmitoyl-2-arachidonoyl-*sn*-glycero-3-phosphocholine
PC	glycerophosphocholine
PE	glycerophosphoethanolamine
PG	glycerophosphoglycerol
PGPC	1-palmitoyl-2-glutaryl-*sn*-glycero-3-phosphocholine
PI	glycerophosphoinositol
PL	glycerophospholipid
POVPC	1-palmitoyl-2-(5-oxovaleroyl)-*sn*-glycero-3-phosphocholine
PS	glycerophosphoserine
PUFA	polyunsaturated fatty acid
Q	quadrupole
QQQ	triple quadrupole
RP	reverse phase chromatography
RT	retention time
SCh-oxPC	short chain-oxidized glycerophosphocholine
T2DM	Type 2 Diabetes Mellitus

1. Introduction

The status of an organism is governed by the activity of the cells building it, which balance biochemical reactions to maintain homeostasis. One of the crucial balances is 'redox homeostasis', which consists of the *in vivo* regulation of oxidative and reductive metabolism. Oxidation is one of the most commonly occurring reactions in a living system.

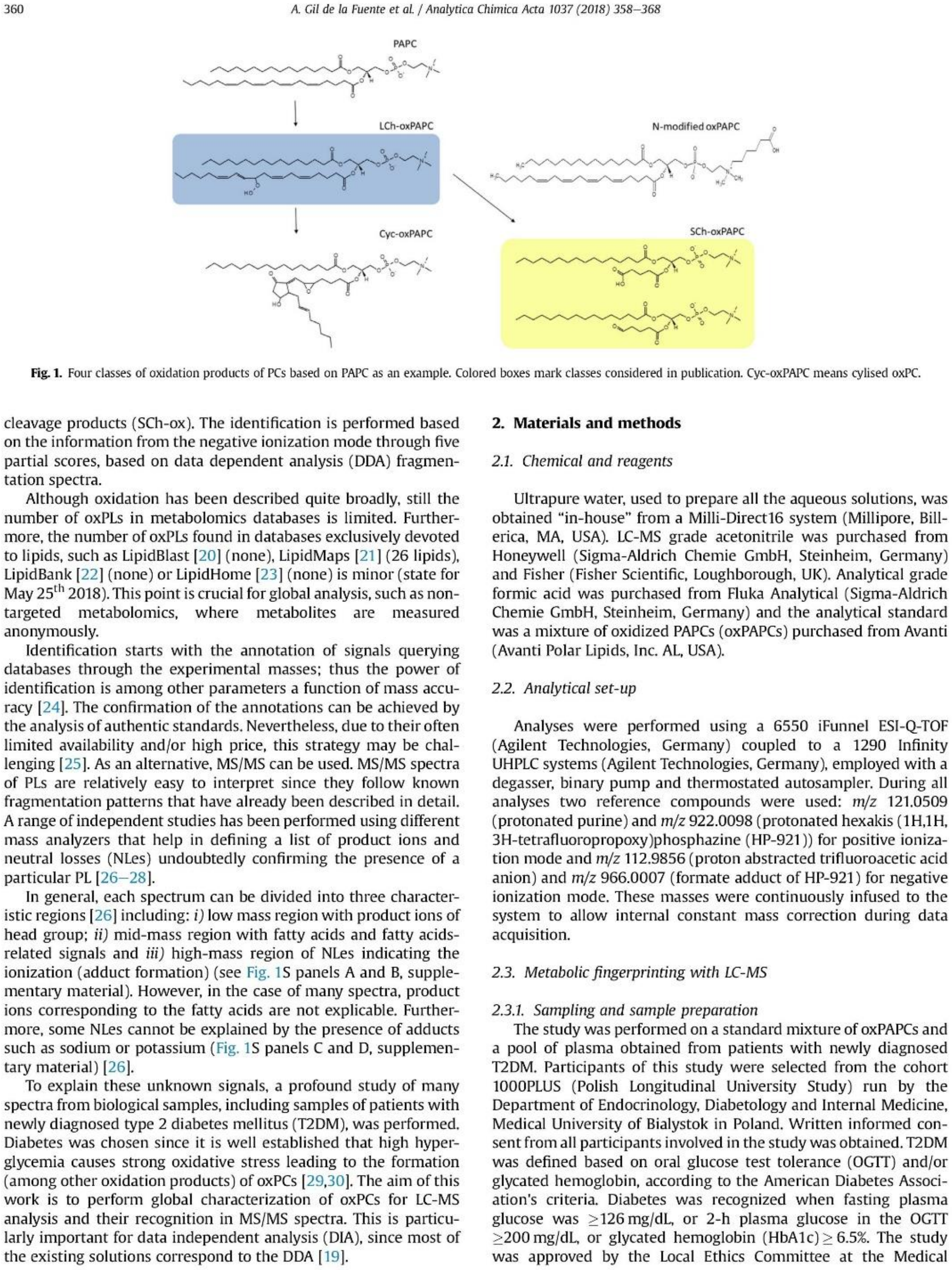
Among many endogenous oxidation processes, lipid peroxidation plays a vital role. Due to their widespread presence in all human cells, lipids are highly affected by oxidative stress. Oxidized lipids are involved in many important processes such as energy production through β-oxidation [1,2], signaling through eicosanoids [3,4] or uncontrolled oxidative degradation provoked by free radicals [5,6]. Therefore, not only intact lipids, but also their oxidized forms represent some of the most important features of mammalian biochemistry.

Currently, attention is being paid to the oxidation of glycerophospholipid (PLs) as intermediate products of oxidation. The work presented herein focuses on oxidized phosphocholines (oxPCs), therefore all subsequent statements and observations concern only to this particular class of oxidized phospholipids (oxPLs). Furthermore, among different types of oxPCs only two classes are covered (Fig. 1). They include: *i)* mildly oxygenated PC (class I oxPC), later called long chain-oxidized PC (LCh-oxPC), which are the products of the oxygen addition to the PC's unsaturated chain [7]. This includes oxidation *via* formation of hydroxyl -OH, dihydroxyl -(OH)$_2$, peroxyl -OOH fatty acids as well as keto- and epoxy fatty acids; *ii)* oxidatively truncated PC (class II oxPC) later called short chain oxidized PC (SCh-oxPC) which occur as a result of fragmenting the oxidized chain of the PC after its previous oxidation [8]. These compounds generally present a semialdehyde (ω-CHO-SCh-oxPAPC) or dicarboxylic (ω-COOH-SCh-oxPAPC) chain in place of the unsaturated chain. Class III of cyclized oxPCs (cyc-oxPCs) and IV of oxidatively N-modified PCs [9] are not included in this publication.

Different classes of oxPC have different biological implications and therefore a proper identification and understanding of oxidation is crucial [9–13]. Considering the fact that the role of oxPC in health and disease is still not fully discovered, their analysis *via* non-targeted metabolomics seems to be fully justified.

OxPLs, especially oxPCs, have already been described in detail. However, the majority of the publications either refers to their biological properties and implication in health and disease states [7,11,13,14] or describes the mechanism of their formation [8,11]. Other publications have been devoted to the measurement of oxPC by means of LC-ESI-MS although only a few of them focused on the identification [15–17]. Current knowledge has been limited to low accuracy mass analyzers such as single quadrupole (Q), triple quadrupole (QQQ) or ion trap (IT).

This knowledge was significantly extended in 2015 with the work of Sala and colleagues [18], who analyzed oxPC using HILIC chromatography connected with a linear ion trap-Orbitrap mass spectrometer. They provided a large amount of information on MS level (accurate mass) and tandem mass spectrometry (MS/MS) level however the last was only as nominal mass. Very recently, significant advancements came from the work by Ni and colleagues [19], who proposed a software called LPPtiger for prediction and identification of oxPLs. It covers glycerophosphocholine (PC), glycerophosphoethanolamine (PE), glycerophosphoserine (PS), glycerophosphoglycerol (PG), and glycerophosphates (PA) and their lyso-forms, oxygen addition products (LCh-ox) and oxidative



PAPC
LCh-oxPAPC
N-modified oxPAPC
Cyc-oxPAPC
SCh-oxPAPC

Fig. 1. Four classes of oxidation products of PCs based on PAPC as an example. Colored boxes mark classes considered in publication. Cyc-oxPAPC means cylised oxPC.

cleavage products (SCh-ox). The identification is performed based on the information from the negative ionization mode through five partial scores, based on data dependent analysis (DDA) fragmentation spectra.

Although oxidation has been described quite broadly, still the number of oxPLs in metabolomics databases is limited. Furthermore, the number of oxPLs found in databases exclusively devoted to lipids, such as LipidBlast [20] (none), LipidMaps [21] (26 lipids), LipidBank [22] (none) or LipidHome [23] (none) is minor (state for May 25th 2018). This point is crucial for global analysis, such as non-targeted metabolomics, where metabolites are measured anonymously.

Identification starts with the annotation of signals querying databases through the experimental masses; thus the power of identification is among other parameters a function of mass accuracy [24]. The confirmation of the annotations can be achieved by the analysis of authentic standards. Nevertheless, due to their often limited availability and/or high price, this strategy may be challenging [25]. As an alternative, MS/MS can be used. MS/MS spectra of PLs are relatively easy to interpret since they follow known fragmentation patterns that have already been described in detail. A range of independent studies has been performed using different mass analyzers that help in defining a list of product ions and neutral losses (NLes) undoubtedly confirming the presence of a particular PL [26–28].

In general, each spectrum can be divided into three characteristic regions [26] including: *i)* low mass region with product ions of head group; *ii)* mid-mass region with fatty acids and fatty acids-related signals and *iii)* high-mass region of NLes indicating the ionization (adduct formation) (see Fig. 1S panels A and B, supplementary material). However, in the case of many spectra, product ions corresponding to the fatty acids are not explicable. Furthermore, some NLes cannot be explained by the presence of adducts such as sodium or potassium (Fig. 1S panels C and D, supplementary material) [26].

To explain these unknown signals, a profound study of many spectra from biological samples, including samples of patients with newly diagnosed type 2 diabetes mellitus (T2DM), was performed. Diabetes was chosen since it is well established that high hyperglycemia causes strong oxidative stress leading to the formation (among other oxidation products) of oxPCs [29,30]. The aim of this work is to perform global characterization of oxPCs for LC-MS analysis and their recognition in MS/MS spectra. This is particularly important for data independent analysis (DIA), since most of the existing solutions correspond to the DDA [19].

2. Materials and methods

2.1. Chemical and reagents

Ultrapure water, used to prepare all the aqueous solutions, was obtained "in-house" from a Milli-Direct16 system (Millipore, Billerica, MA, USA). LC-MS grade acetonitrile was purchased from Honeywell (Sigma-Aldrich Chemie GmbH, Steinheim, Germany) and Fisher (Fisher Scientific, Loughborough, UK). Analytical grade formic acid was purchased from Fluka Analytical (Sigma-Aldrich Chemie GmbH, Steinheim, Germany) and the analytical standard was a mixture of oxidized PAPCs (oxPAPCs) purchased from Avanti (Avanti Polar Lipids, Inc. AL, USA).

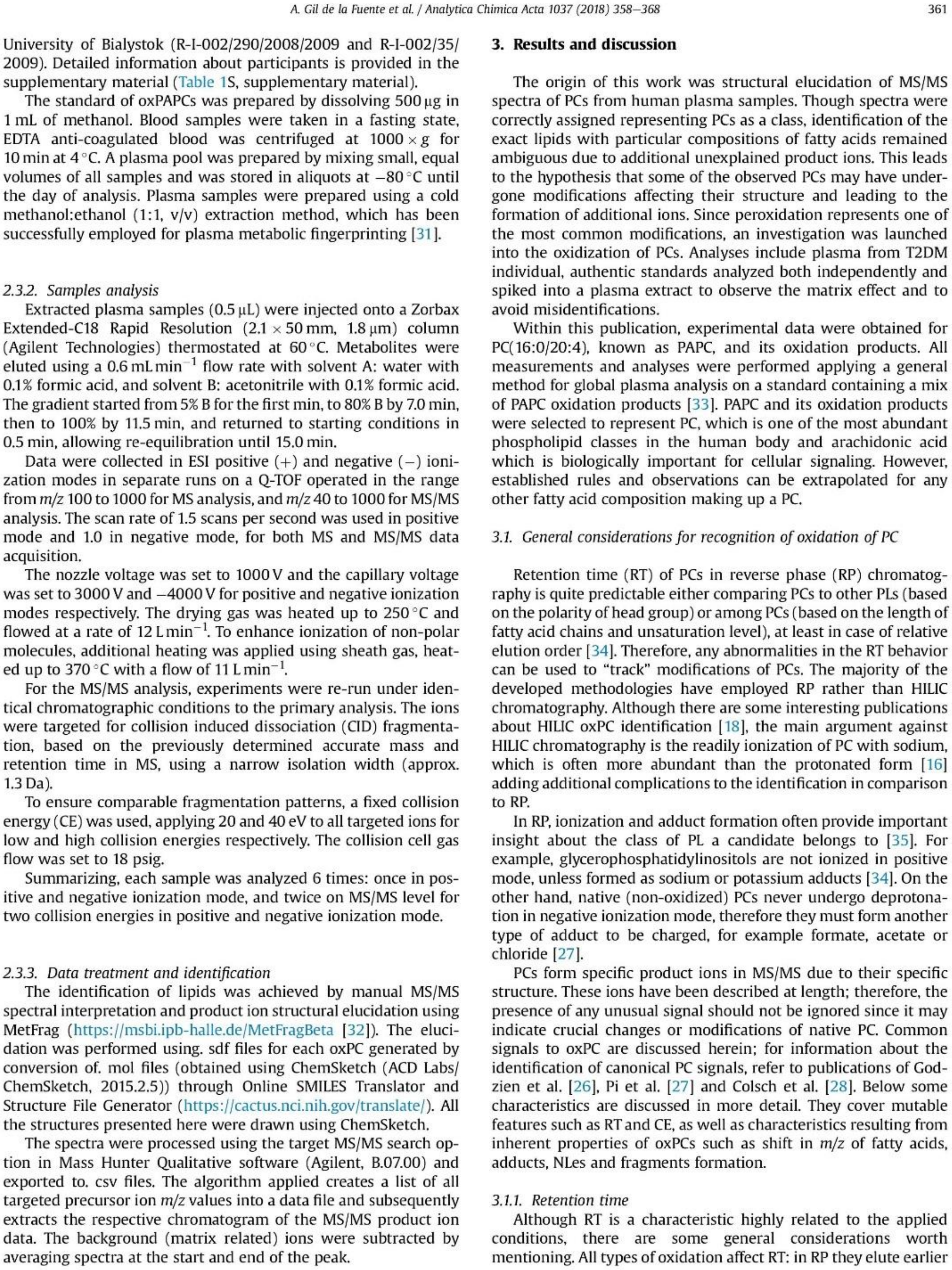
2.2. Analytical set-up

Analyses were performed using a 6550 iFunnel ESI-Q-TOF (Agilent Technologies, Germany) coupled to a 1290 Infinity UHPLC systems (Agilent Technologies, Germany), employed with a degasser, binary pump and thermostated autosampler. During all analyses two reference compounds were used: *m/z* 121.0509 (protonated purine) and *m/z* 922.0098 (protonated hexakis (1H,1H, 3H-tetrafluoropropoxy)phosphazine (HP-921)) for positive ionization mode and *m/z* 112.9856 (proton abstracted trifluoroacetic acid anion) and *m/z* 966.0007 (formate adduct of HP-921) for negative ionization mode. These masses were continuously infused to the system to allow internal constant mass correction during data acquisition.

2.3. Metabolic fingerprinting with LC-MS

2.3.1. Sampling and sample preparation

The study was performed on a standard mixture of oxPAPCs and a pool of plasma obtained from patients with newly diagnosed T2DM. Participants of this study were selected from the cohort 1000PLUS (Polish Longitudinal University Study) run by the Department of Endocrinology, Diabetology and Internal Medicine, Medical University of Bialystok in Poland. Written informed consent from all participants involved in the study was obtained. T2DM was defined based on oral glucose test tolerance (OGTT) and/or glycated hemoglobin, according to the American Diabetes Association's criteria. Diabetes was recognized when fasting plasma glucose was ≥126 mg/dL, or 2-h plasma glucose in the OGTT ≥200 mg/dL, or glycated hemoglobin (HbA1c) ≥ 6.5%. The study was approved by the Local Ethics Committee at the Medical



University of Bialystok (R-I-002/290/2008/2009 and R-I-002/35/2009). Detailed information about participants is provided in the supplementary material (Table 1S, supplementary material).

The standard of oxPAPCs was prepared by dissolving 500 μg in 1 mL of methanol. Blood samples were taken in a fasting state, EDTA anti-coagulated blood was centrifuged at 1000 × g for 10 min at 4 °C. A plasma pool was prepared by mixing small, equal volumes of all samples and was stored in aliquots at −80 °C until the day of analysis. Plasma samples were prepared using a cold methanol:ethanol (1:1, v/v) extraction method, which has been successfully employed for plasma metabolic fingerprinting [31].

2.3.2. Samples analysis

Extracted plasma samples (0.5 μL) were injected onto a Zorbax Extended-C18 Rapid Resolution (2.1 × 50 mm, 1.8 μm) column (Agilent Technologies) thermostated at 60 °C. Metabolites were eluted using a 0.6 mL $\min^{-1}$ flow rate with solvent A: water with 0.1% formic acid, and solvent B: acetonitrile with 0.1% formic acid. The gradient started from 5% B for the first min, to 80% B by 7.0 min, then to 100% by 11.5 min, and returned to starting conditions in 0.5 min, allowing re-equilibration until 15.0 min.

Data were collected in ESI positive (+) and negative (−) ionization modes in separate runs on a Q-TOF operated in the range from *m/z* 100 to 1000 for MS analysis, and *m/z* 40 to 1000 for MS/MS analysis. The scan rate of 1.5 scans per second was used in positive mode and 1.0 in negative mode, for both MS and MS/MS data acquisition.

The nozzle voltage was set to 1000 V and the capillary voltage was set to 3000 V and −4000 V for positive and negative ionization modes respectively. The drying gas was heated up to 250 °C and flowed at a rate of 12 L $\min^{-1}$. To enhance ionization of non-polar molecules, additional heating was applied using sheath gas, heated up to 370 °C with a flow of 11 L $\min^{-1}$.

For the MS/MS analysis, experiments were re-run under identical chromatographic conditions to the primary analysis. The ions were targeted for collision induced dissociation (CID) fragmentation, based on the previously determined accurate mass and retention time in MS, using a narrow isolation width (approx. 1.3 Da).

To ensure comparable fragmentation patterns, a fixed collision energy (CE) was used, applying 20 and 40 eV to all targeted ions for low and high collision energies respectively. The collision cell gas flow was set to 18 psig.

Summarizing, each sample was analyzed 6 times: once in positive and negative ionization mode, and twice on MS/MS level for two collision energies in positive and negative ionization mode.

2.3.3. Data treatment and identification

The identification of lipids was achieved by manual MS/MS spectral interpretation and product ion structural elucidation using MetFrag (https://msbi.ipb-halle.de/MetFragBeta [32]). The elucidation was performed using. sdf files for each oxPC generated by conversion of. mol files (obtained using ChemSketch (ACD Labs/ChemSketch, 2015.2.5)) through Online SMILES Translator and Structure File Generator (https://cactus.nci.nih.gov/translate/). All the structures presented here were drawn using ChemSketch.

The spectra were processed using the target MS/MS search option in Mass Hunter Qualitative software (Agilent, B.07.00) and exported to. csv files. The algorithm applied creates a list of all targeted precursor ion *m/z* values into a data file and subsequently extracts the respective chromatogram of the MS/MS product ion data. The background (matrix related) ions were subtracted by averaging spectra at the start and end of the peak.

3. Results and discussion

The origin of this work was structural elucidation of MS/MS spectra of PCs from human plasma samples. Though spectra were correctly assigned representing PCs as a class, identification of the exact lipids with particular compositions of fatty acids remained ambiguous due to additional unexplained product ions. This leads to the hypothesis that some of the observed PCs may have undergone modifications affecting their structure and leading to the formation of additional ions. Since peroxidation represents one of the most common modifications, an investigation was launched into the oxidization of PCs. Analyses include plasma from T2DM individual, authentic standards analyzed both independently and spiked into a plasma extract to observe the matrix effect and to avoid misidentifications.

Within this publication, experimental data were obtained for PC(16:0/20:4), known as PAPC, and its oxidation products. All measurements and analyses were performed applying a general method for global plasma analysis on a standard containing a mix of PAPC oxidation products [33]. PAPC and its oxidation products were selected to represent PC, which is one of the most abundant phospholipid classes in the human body and arachidonic acid which is biologically important for cellular signaling. However, established rules and observations can be extrapolated for any other fatty acid composition making up a PC.

3.1. General considerations for recognition of oxidation of PC

Retention time (RT) of PCs in reverse phase (RP) chromatography is quite predictable either comparing PCs to other PLs (based on the polarity of head group) or among PCs (based on the length of fatty acid chains and unsaturation level), at least in case of relative elution order [34]. Therefore, any abnormalities in the RT behavior can be used to "track" modifications of PCs. The majority of the developed methodologies have employed RP rather than HILIC chromatography. Although there are some interesting publications about HILIC oxPC identification [18], the main argument against HILIC chromatography is the readily ionization of PC with sodium, which is often more abundant than the protonated form [16] adding additional complications to the identification in comparison to RP.

In RP, ionization and adduct formation often provide important insight about the class of PL a candidate belongs to [35]. For example, glycerophosphatidylinositols are not ionized in positive mode, unless formed as sodium or potassium adducts [34]. On the other hand, native (non-oxidized) PCs never undergo deprotonation in negative ionization mode, therefore they must form another type of adduct to be charged, for example formate, acetate or chloride [27].

PCs form specific product ions in MS/MS due to their specific structure. These ions have been described at length; therefore, the presence of any unusual signal should not be ignored since it may indicate crucial changes or modifications of native PC. Common signals to oxPC are discussed herein; for information about the identification of canonical PC signals, refer to publications of Godzien et al. [26], Pi et al. [27] and Colsch et al. [28]. Below some characteristics are discussed in more detail. They cover mutable features such as RT and CE, as well as characteristics resulting from inherent properties of oxPCs such as shift in *m/z* of fatty acids, adducts, NLes and fragments formation.

3.1.1. Retention time

Although RT is a characteristic highly related to the applied conditions, there are some general considerations worth mentioning. All types of oxidation affect RT: in RP they elute earlier

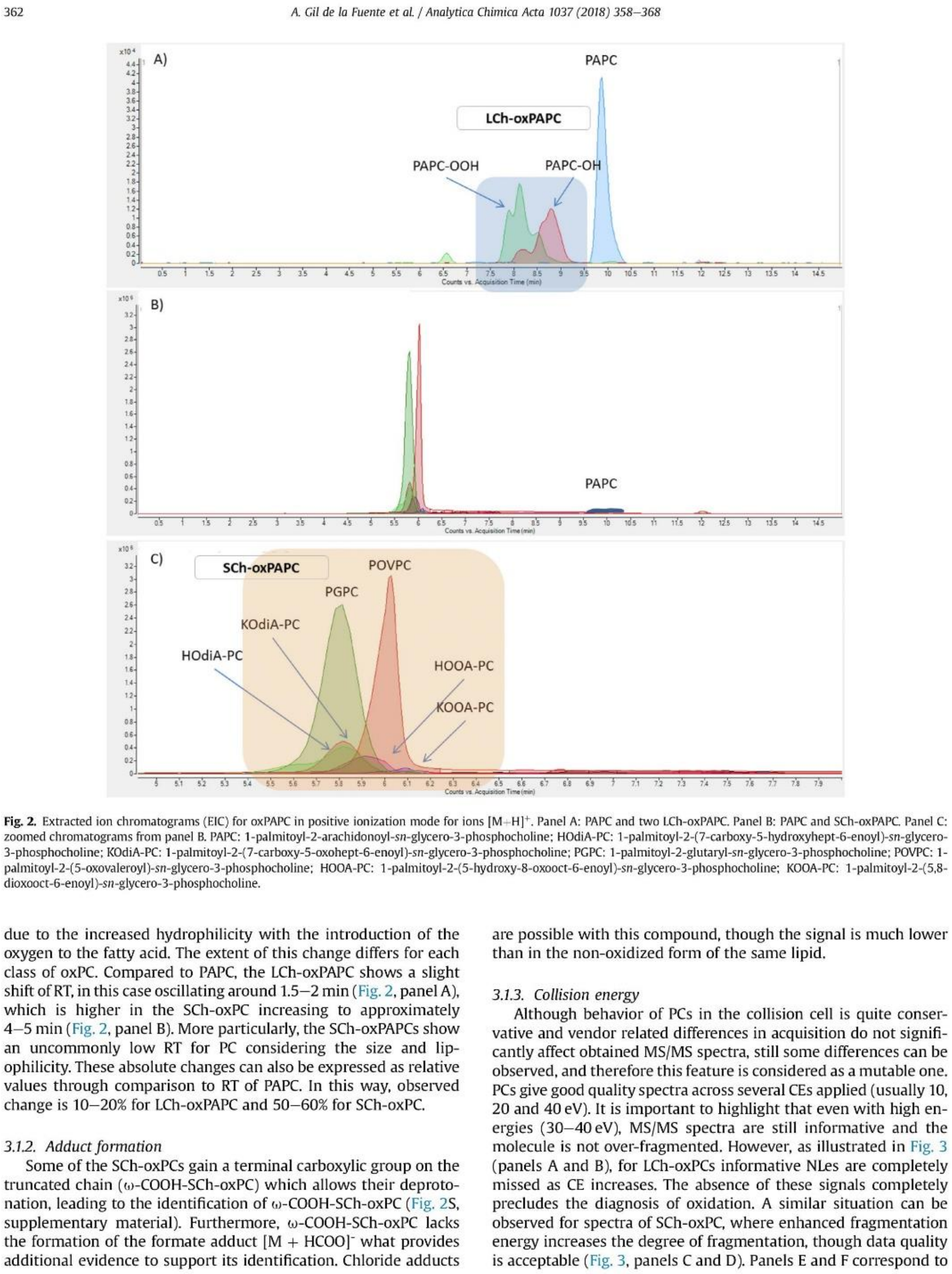


Fig. 2. Extracted ion chromatograms (EIC) for oxPAPC in positive ionization mode for ions $[M+H]^+$. Panel A: PAPC and two LCh-oxPAPC. Panel B: PAPC and SCh-oxPAPC. Panel C: zoomed chromatograms from panel B. PAPC: 1-palmitoyl-2-arachidonoyl-*sn*-glycero-3-phosphocholine; HOdiA-PC: 1-palmitoyl-2-(7-carboxy-5-hydroxyhept-6-enoyl)-*sn*-glycero-3-phosphocholine; KOdiA-PC: 1-palmitoyl-2-(7-carboxy-5-oxohept-6-enoyl)-*sn*-glycero-3-phosphocholine; PGPC: 1-palmitoyl-2-glutaryl-*sn*-glycero-3-phosphocholine; POVPC: 1-palmitoyl-2-(5-oxovaleroyl)-*sn*-glycero-3-phosphocholine; HOOA-PC: 1-palmitoyl-2-(5-hydroxy-8-oxooct-6-enoyl)-*sn*-glycero-3-phosphocholine; KOOA-PC: 1-palmitoyl-2-(5,8-dioxooct-6-enoyl)-*sn*-glycero-3-phosphocholine.

due to the increased hydrophilicity with the introduction of the oxygen to the fatty acid. The extent of this change differs for each class of oxPC. Compared to PAPC, the LCh-oxPAPC shows a slight shift of RT, in this case oscillating around 1.5–2 min (Fig. 2, panel A), which is higher in the SCh-oxPC increasing to approximately 4–5 min (Fig. 2, panel B). More particularly, the SCh-oxPAPCs show an uncommonly low RT for PC considering the size and lipophilicity. These absolute changes can also be expressed as relative values through comparison to RT of PAPC. In this way, observed change is 10–20% for LCh-oxPAPC and 50–60% for SCh-oxPC.

3.1.2. *Adduct formation*

Some of the SCh-oxPCs gain a terminal carboxylic group on the truncated chain (ω-COOH-SCh-oxPC) which allows their deprotonation, leading to the identification of ω-COOH-SCh-oxPC (Fig. 2S, supplementary material). Furthermore, ω-COOH-SCh-oxPC lacks the formation of the formate adduct $[M + HCOO]^-$ what provides additional evidence to support its identification. Chloride adducts are possible with this compound, though the signal is much lower than in the non-oxidized form of the same lipid.

3.1.3. *Collision energy*

Although behavior of PCs in the collision cell is quite conservative and vendor related differences in acquisition do not significantly affect obtained MS/MS spectra, still some differences can be observed, and therefore this feature is considered as a mutable one. PCs give good quality spectra across several CEs applied (usually 10, 20 and 40 eV). It is important to highlight that even with high energies (30–40 eV), MS/MS spectra are still informative and the molecule is not over-fragmented. However, as illustrated in Fig. 3 (panels A and B), for LCh-oxPCs informative NLes are completely missed as CE increases. The absence of these signals completely precludes the diagnosis of oxidation. A similar situation can be observed for spectra of SCh-oxPC, where enhanced fragmentation energy increases the degree of fragmentation, though data quality is acceptable (Fig. 3, panels C and D). Panels E and F correspond to

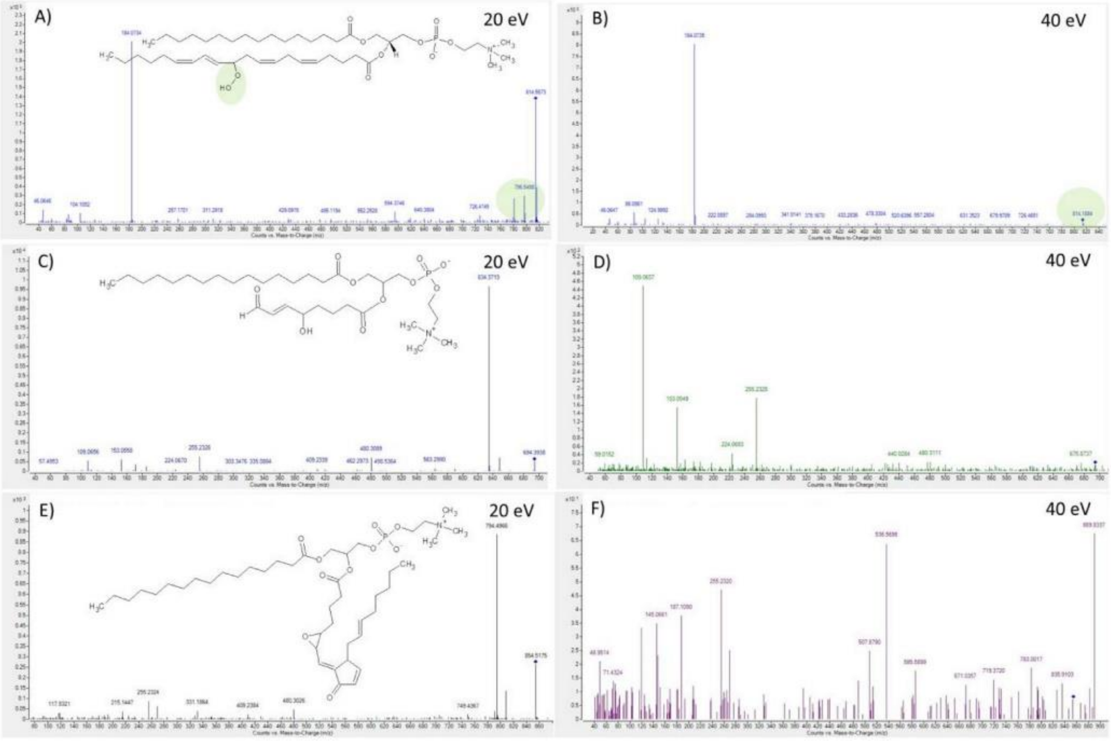


Fig. 3. Examples of MS/MS spectra acquired with different collision energies and their impact on degree of fragmentation. Panels A and B: Spectra for PC(16:0/20:4(OOH)) in 20eV and 40eV respectively (positive ionization mode); Panels C and D: PC(16:0/8:1(CHO)) in 20eV and 40eV respectively (negative ionization mode); Panels E and F: PC(16:0/epoxyA$_2$IsoP) in 20eV and 40eV respectively (negative ionization mode).

the cyc-oxPC, which although are not discussed in this publication, were intentionally brought up at this point. It is to illustrate their very different behavior at enhanced CEs causing over-fragmentation of the molecule and producing very noisy spectra. Such spectra are not suitable for the annotation process due to the low quality and lack of trustful structural information. However, this drastic change in the fragmentation degree between different CEs can be considered as an evidence about PC oxidation.

3.1.4. *Shift in m/z of fatty acid*

The exact composition of fatty acids can be easily established based on the MS/MS spectrum acquired in the negative ionization mode. The mid-mass region contains signals corresponding to the deprotonated fatty acid or demethylated lyso-form containing a particular fatty acid as well. However, sometimes one of the fatty acids is either unidentified or missing in the mid-mass region and each of these cases corresponds to a different type of oxidation: LCh-oxPC and SCh-oxPC, respectively. Fig. 4 illustrates the impact of oxidation on fatty acid for LCh-oxPC (panel A) and SCh-oxPC (panel B).

LCh-oxPCs are modified by the addition of the oxidation agent to the fatty acid, causing an increase in mass (e.g. 279.2330 Da for native fatty acid 18:2 is increased to 295.2274 Da for 18:2-OH). Consequently, oxidized fatty acids tend to remain unidentified while searching them across databases, though they may be identified manually by calculating the mass difference. To automatize this step a new functionality has been developed in CMM, as described in section 3.2. SCh-oxPC is associated with a chain shortening with oxidation, significantly reducing the mass of the fatty acid compared with the expected one (e.g. 303.2330 Da for native fatty acid 20:4 is reduced to 115.0400 Da for 5:0(CHO)). As a consequence, *m/z* of fatty acid is in a low-mass region and often is confused with head-related product ions or artefacts.

3.1.5. *Neutral loss of water*

Another important characteristic that can be used to identify oxPCs is the NL of water. NL of water is very important in the annotation of oxPCs therefore is discussed apart from other NLes. Although water loss is the most common one in MS/MS spectra, PCs normally do not lose water. However, after oxidation, one of the fatty acyl chains may present a hydroxyl group as a substituent of the carbon chain. The loss of water appears differently depending on the ESI modality used. As shown in panel A of Fig. 3S (supplementary material), the loss (18.0105 Da) is clearly visible in the high mass region of spectrum for positive ionization mode. In negative ionization mode water loss is more relevant in the mid-mass region of fatty acids (Fig. 3S panel B and C, supplementary material).

3.1.6. *Product ions and other neutral losses*

NLes and product ions for the confirmation of a particular oxidation are summarized in Table 1. The table defines the difference in the mass due to the oxidation as well as expected NLes and product ions. As can be seen for LCh-oxPC and SCh-oxPC, diagnostic NLes are observed in different polarities. To help with an

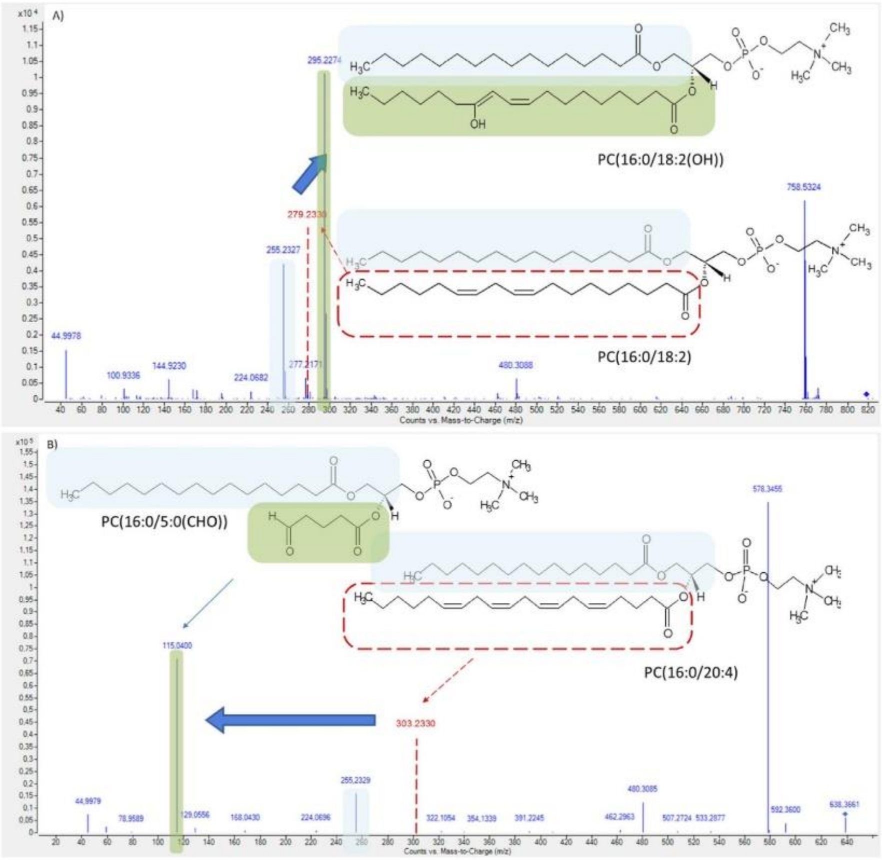


Fig. 4. Impact of oxidation on the fatty acid signal in MS/MS spectra obtained negative ionization mode. Panel A: LCh-oXPC: Example of an MS/MS spectrum of PC(16:0/18:2(OH)) (spectrum from plasma sample) and change in the oxidized fatty acid (marked with green) in comparison to the expected non-oxidized one (marked with red). Panel B: SCh-oxPC: Example of an MS/MS spectrum of PC(16:0/5:0(CHO)) (spectrum from standard) and change in the oxidized fatty acid (green) in comparison to the expected non-oxidized one (red). The second non-oxidized fatty acid is shown in blue. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

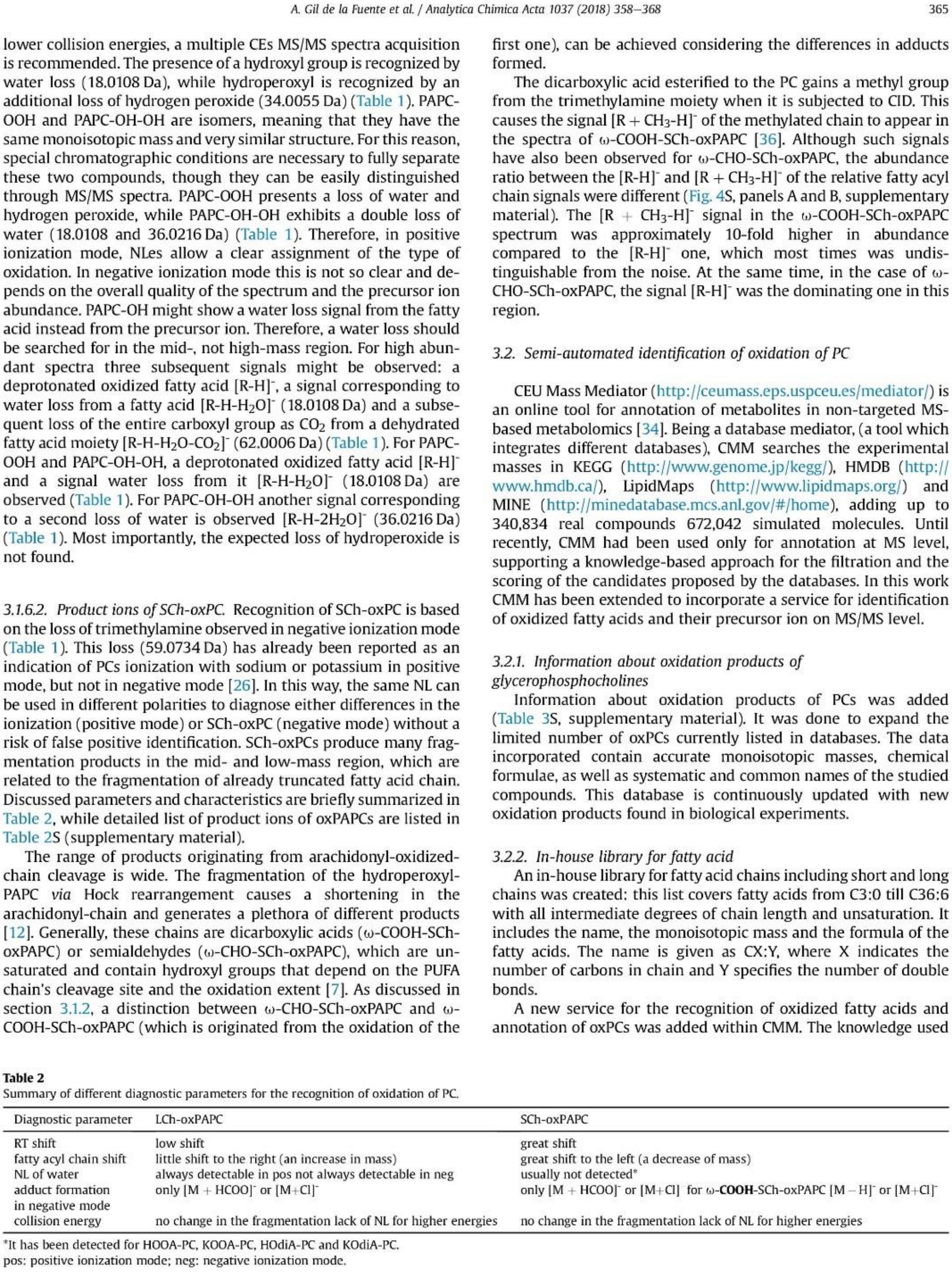
Table 1
List of diagnostic NLes and product ions for recognition of oxidation type.

oxidation type	Δm*	NL	fragment
LCh-oxPC			
-OH	+15.9949	POS: $[M+H]^+$- H_2O −18.0108 Da	NEG: $[R\text{-}H]^-$- H_2O and $[R\text{-}H]^-$- H_2O - CO_2 -18.0108 Da and −62.0216 Da
-OH −OH	+31.9898	POS: $[M+H]^+$- H_2O and $[M+H]^+$- $2H_2O$ −18.0108 Da and −36.0216 Da	NEG: $[R\text{-}H]^-$ - H_2O and $[R\text{-}H]^-$ - $2H_2O$ −18.0108 Da and −36.0216 Da
-OOH	+31.9898	POS: $[M+H]^+$- H_2O and $[M+H]^+$- OOH -18.0108 Da and −34.0049 Da	NEG: $[R\text{-}H]^-$ - H_2O −18.0108 Da
Sch-oxPC			
-CHO	+13.9793	NEG: $[M + HCOO]^-$ - $HCOO^-[M + HCOO]^-$ - $HCOOCH_3$ $[M + HCOO]^-$-$N(CH_3)_3$ - $HCOO^-$-46.0049 Da, −60.0222 Da and −105.0779 Da	–
-COOH	+29.9742	NEG: $[M - H]^-$ - $N(CH_3)_3$−59.0734 Da	–

*Δm – difference in the mass between non-oxidized and oxidized form. POS: positive ionization mode; NEG: negative ionization mode.

assignment of the composition of the fatty acid and type of oxidation, a special functionality was implemented in CMM (section 3.2).

3.1.6.1. Product ions of LCh-oxPC. In case of LCh-oxPC, all diagnostic signals are related to the presence of a particular oxidizing "agent" incorporated into the fatty acid chain. LCh-oxPCs produce more characteristic fragments in positive than in negative ionization mode to distinguish the type of oxidation developed. A typical MS/MS spectrum of LCh-oxPC contains "canonical" product ions related to the head group but also important NL which clearly determine the type of oxidation. Since these NLes are detectable using only



lower collision energies, a multiple CEs MS/MS spectra acquisition is recommended. The presence of a hydroxyl group is recognized by water loss (18.0108 Da), while hydroperoxyl is recognized by an additional loss of hydrogen peroxide (34.0055 Da) (Table 1). PAPC-OOH and PAPC-OH-OH are isomers, meaning that they have the same monoisotopic mass and very similar structure. For this reason, special chromatographic conditions are necessary to fully separate these two compounds, though they can be easily distinguished through MS/MS spectra. PAPC-OOH presents a loss of water and hydrogen peroxide, while PAPC-OH-OH exhibits a double loss of water (18.0108 and 36.0216 Da) (Table 1). Therefore, in positive ionization mode, NLes allow a clear assignment of the type of oxidation. In negative ionization mode this is not so clear and depends on the overall quality of the spectrum and the precursor ion abundance. PAPC-OH might show a water loss signal from the fatty acid instead from the precursor ion. Therefore, a water loss should be searched for in the mid-, not high-mass region. For high abundant spectra three subsequent signals might be observed: a deprotonated oxidized fatty acid [R-H]$^-$, a signal corresponding to water loss from a fatty acid [R-H-H_2O]$^-$ (18.0108 Da) and a subsequent loss of the entire carboxyl group as CO_2 from a dehydrated fatty acid moiety [R-H-H_2O-CO_2]$^-$ (62.0006 Da) (Table 1). For PAPC-OOH and PAPC-OH-OH, a deprotonated oxidized fatty acid [R-H]$^-$ and a signal water loss from it [R-H-H_2O]$^-$ (18.0108 Da) are observed (Table 1). For PAPC-OH-OH another signal corresponding to a second loss of water is observed [R-H-2H_2O]$^-$ (36.0216 Da) (Table 1). Most importantly, the expected loss of hydroperoxide is not found.

3.1.6.2. Product ions of SCh-oxPC. Recognition of SCh-oxPC is based on the loss of trimethylamine observed in negative ionization mode (Table 1). This loss (59.0734 Da) has already been reported as an indication of PCs ionization with sodium or potassium in positive mode, but not in negative mode [26]. In this way, the same NL can be used in different polarities to diagnose either differences in the ionization (positive mode) or SCh-oxPC (negative mode) without a risk of false positive identification. SCh-oxPCs produce many fragmentation products in the mid- and low-mass region, which are related to the fragmentation of already truncated fatty acid chain. Discussed parameters and characteristics are briefly summarized in Table 2, while detailed list of product ions of oxPAPCs are listed in Table 2S (supplementary material).

The range of products originating from arachidonyl-oxidized-chain cleavage is wide. The fragmentation of the hydroperoxyl-PAPC *via* Hock rearrangement causes a shortening in the arachidonyl-chain and generates a plethora of different products [12]. Generally, these chains are dicarboxylic acids (ω-COOH-SCh-oxPAPC) or semialdehydes (ω-CHO-SCh-oxPAPC), which are unsaturated and contain hydroxyl groups that depend on the PUFA chain's cleavage site and the oxidation extent [7]. As discussed in section 3.1.2, a distinction between ω-CHO-SCh-oxPAPC and ω-COOH-SCh-oxPAPC (which is originated from the oxidation of the first one), can be achieved considering the differences in adducts formed.

The dicarboxylic acid esterified to the PC gains a methyl group from the trimethylamine moiety when it is subjected to CID. This causes the signal [R + CH_3-H]$^-$ of the methylated chain to appear in the spectra of ω-COOH-SCh-oxPAPC [36]. Although such signals have also been observed for ω-CHO-SCh-oxPAPC, the abundance ratio between the [R-H]$^-$ and [R + CH_3-H]$^-$ of the relative fatty acyl chain signals were different (Fig. 4S, panels A and B, supplementary material). The [R + CH_3-H]$^-$ signal in the ω-COOH-SCh-oxPAPC spectrum was approximately 10-fold higher in abundance compared to the [R-H]$^-$ one, which most times was undistinguishable from the noise. At the same time, in the case of ω-CHO-SCh-oxPAPC, the signal [R-H]$^-$ was the dominating one in this region.

3.2. Semi-automated identification of oxidation of PC

CEU Mass Mediator (http://ceumass.eps.uspceu.es/mediator/) is an online tool for annotation of metabolites in non-targeted MS-based metabolomics [34]. Being a database mediator, (a tool which integrates different databases), CMM searches the experimental masses in KEGG (http://www.genome.jp/kegg/), HMDB (http://www.hmdb.ca/), LipidMaps (http://www.lipidmaps.org/) and MINE (http://minedatabase.mcs.anl.gov/#/home), adding up to 340,834 real compounds 672,042 simulated molecules. Until recently, CMM had been used only for annotation at MS level, supporting a knowledge-based approach for the filtration and the scoring of the candidates proposed by the databases. In this work CMM has been extended to incorporate a service for identification of oxidized fatty acids and their precursor ion on MS/MS level.

3.2.1. Information about oxidation products of glycerophosphocholines

Information about oxidation products of PCs was added (Table 3S, supplementary material). It was done to expand the limited number of oxPCs currently listed in databases. The data incorporated contain accurate monoisotopic masses, chemical formulae, as well as systematic and common names of the studied compounds. This database is continuously updated with new oxidation products found in biological experiments.

3.2.2. In-house library for fatty acid

An in-house library for fatty acid chains including short and long chains was created: this list covers fatty acids from C3:0 till C36:6 with all intermediate degrees of chain length and unsaturation. It includes the name, the monoisotopic mass and the formula of the fatty acids. The name is given as CX:Y, where X indicates the number of carbons in chain and Y specifies the number of double bonds.

A new service for the recognition of oxidized fatty acids and annotation of oxPCs was added within CMM. The knowledge used

Table 2
Summary of different diagnostic parameters for the recognition of oxidation of PC.

Diagnostic parameter	LCh-oxPAPC	SCh-oxPAPC
RT shift	low shift	great shift
fatty acyl chain shift	little shift to the right (an increase in mass)	great shift to the left (a decrease of mass)
NL of water	always detectable in pos not always detectable in neg	usually not detected*
adduct formation in negative mode	only [M + HCOO]$^-$ or [M+Cl]$^-$	only [M + HCOO]$^-$ or [M+Cl]$^-$ for ω-**COOH**-SCh-oxPAPC [M − H]$^-$ or [M+Cl]$^-$
collision energy	no change in the fragmentation lack of NL for higher energies	no change in the fragmentation lack of NL for higher energies

*It has been detected for HOOA-PC, KOOA-PC, HOdiA-PC and KOdiA-PC.
pos: positive ionization mode; neg: negative ionization mode.

was firstly hypothesized based on the fragmentation patterns observed in biological samples and then confirmed by means of authentic standards. The annotation service assumes that an unidentified (for LCh-oxPC) or missing (for SCh-oxPC) fatty acid from the mid-mass region of a negative ionization mode fragmentation spectrum is oxidized.

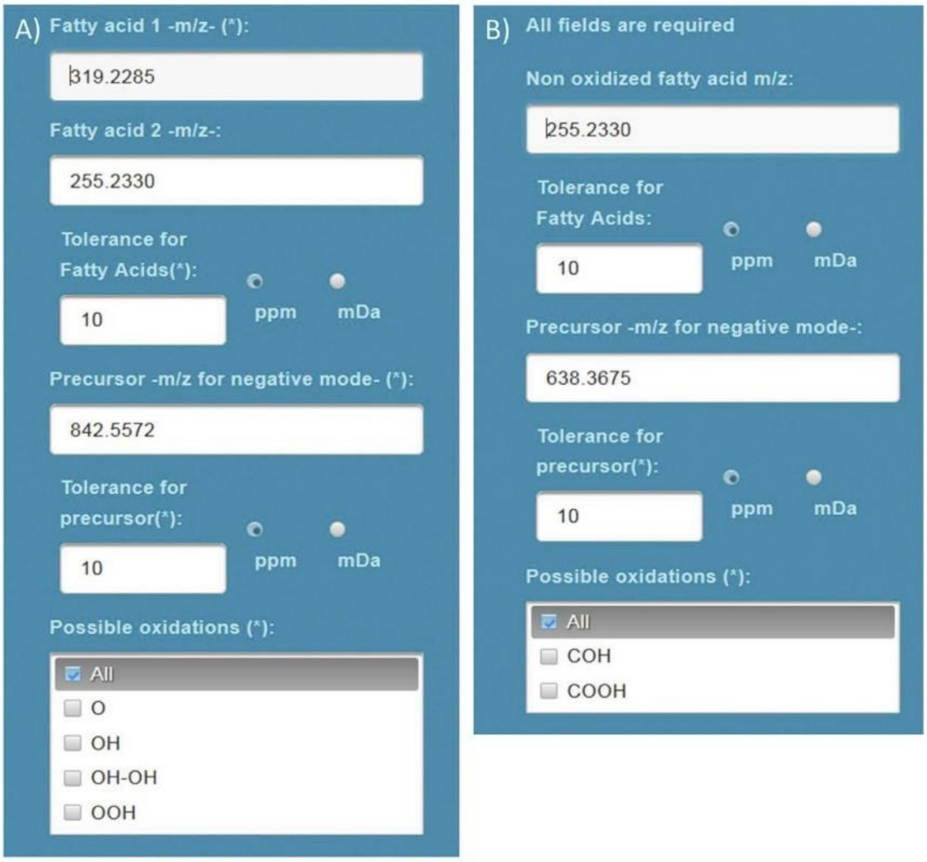
3.2.3. *Semi-automated annotation of long chain oxidized glycerophosphocholines (LCh-oxPC)*

For the annotation of LCh-oxPC, the input data includes: *m/z* of both fatty acids (oxidized and non-oxidized), *m/z* of the precursor ion, the tolerance for the mass matching of both fatty acids and the precursor ion, and the oxidation type, which can be selected between = O, -OH, -OH-OH, -OOH (Fig. 5, panel A).

The algorithm for the annotation of long chain oxidized glycerophosphocholines starts with the identification of the oxidized fatty acid (Fig. 5S, panel A, supplementary material). To this end, it subtracts the mass of the all possible oxidation types from both fatty acids and subsequently queries the in-house fatty acids dedicated library. A non-oxidized fatty acid will never obtain any candidate, while the oxidized fatty acid may return candidates for one or more oxidation types. Table 4S of supplementary material contains an example of the list of oxidations for long and short chain oxidations for particular fatty acid.

Once the oxidized and non-oxidized fatty acids are recognized and annotated, the precursor ion is used to confirm the deduced oxidation and the fatty acid's composition. It is searched for the oxidized and the non-oxidized form of the deduced PCs and to retrieve the signals arising from the diagnostic NLes that the spectra should contain. The presumed adduct of the precursor ion is formate, whose mass is subtracted from the *m/z* provided by the user.

There are two types of annotation for the precursor ion: the first refers to the oxidized form, which is searched against the list of oxPCs. This search is restricted to the oxPCs matching the mass of the precursor ion within the tolerance allowed and containing the previously annotated non-oxidized fatty acid (e.g. C16:0) and oxidized one (e.g. C20:4(OH)). The oxidized precursor ion candidates are listed in the column corresponding to the oxPC (e.g PC(16:0/20:4(OH)). However, due to the limited number of oxPCs in the databases, an alternative search is performed. The mass of the non-oxidized precursor ion, calculated by subtracting the oxidation type to the *m/z* of the oxidized precursor ion, is used to search the non-oxidized form against the general list of PCs (e.g. PC(16:0/20:4)). In this case the search is also restricted to PCs matching the mass of the non-oxidized precursor ion and containing the previously annotated fatty acids. The list of candidates for non-oxidized PC is shown in the column corresponding to non-oxidized



A) Fatty acid 1 -m/z- (*):

319.2285

Fatty acid 2 -m/z-:

255.2330

Tolerance for Fatty Acids(*): 10 ● ppm ○ mDa

Precursor -m/z for negative mode- (*):

842.5572

Tolerance for precursor(*): 10 ● ppm ○ mDa

Possible oxidations (*):

☑ All
☐ O
☐ OH
☐ OH-OH
☐ OOH

B) All fields are required

Non oxidized fatty acid m/z:

255.2330

Tolerance for Fatty Acids: 10 ● ppm ○ mDa

Precursor -m/z for negative mode-:

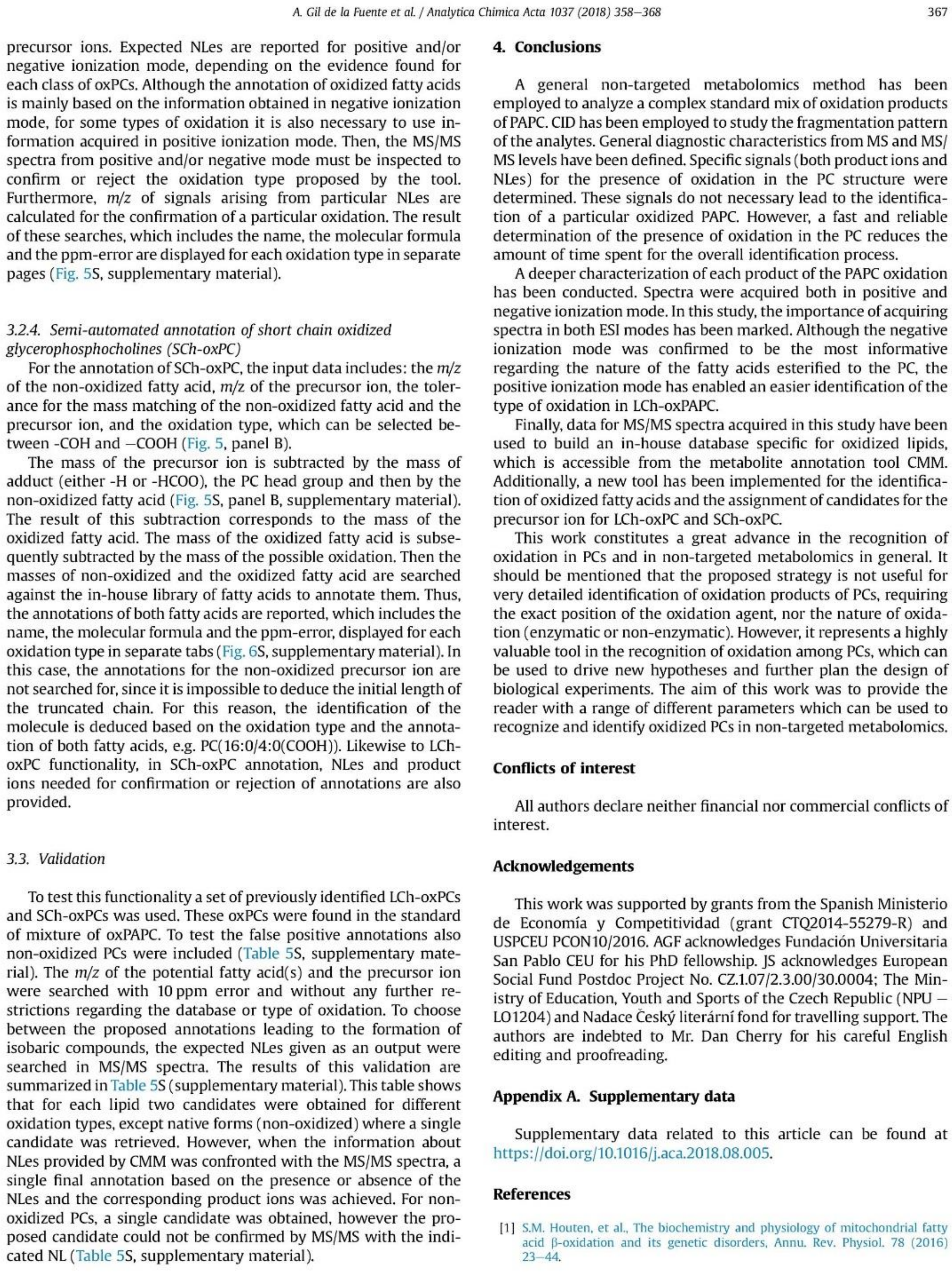
638.3675

Tolerance for precursor(*): 10 ● ppm ○ mDa

Possible oxidations (*):

☑ All
☐ COH
☐ COOH

Fig. 5. Input information with mandatory and optional fields for oxidized fatty acid identification and annotation of LCh-oxPC (panel A) andSCh-oxPC (panel B).



precursor ions. Expected NLes are reported for positive and/or negative ionization mode, depending on the evidence found for each class of oxPCs. Although the annotation of oxidized fatty acids is mainly based on the information obtained in negative ionization mode, for some types of oxidation it is also necessary to use information acquired in positive ionization mode. Then, the MS/MS spectra from positive and/or negative mode must be inspected to confirm or reject the oxidation type proposed by the tool. Furthermore, *m/z* of signals arising from particular NLes are calculated for the confirmation of a particular oxidation. The result of these searches, which includes the name, the molecular formula and the ppm-error are displayed for each oxidation type in separate pages (Fig. 5S, supplementary material).

3.2.4. Semi-automated annotation of short chain oxidized glycerophosphocholines (SCh-oxPC)

For the annotation of SCh-oxPC, the input data includes: the *m/z* of the non-oxidized fatty acid, *m/z* of the precursor ion, the tolerance for the mass matching of the non-oxidized fatty acid and the precursor ion, and the oxidation type, which can be selected between -COH and −COOH (Fig. 5, panel B).

The mass of the precursor ion is subtracted by the mass of adduct (either -H or -HCOO), the PC head group and then by the non-oxidized fatty acid (Fig. 5S, panel B, supplementary material). The result of this subtraction corresponds to the mass of the oxidized fatty acid. The mass of the oxidized fatty acid is subsequently subtracted by the mass of the possible oxidation. Then the masses of non-oxidized and the oxidized fatty acid are searched against the in-house library of fatty acids to annotate them. Thus, the annotations of both fatty acids are reported, which includes the name, the molecular formula and the ppm-error, displayed for each oxidation type in separate tabs (Fig. 6S, supplementary material). In this case, the annotations for the non-oxidized precursor ion are not searched for, since it is impossible to deduce the initial length of the truncated chain. For this reason, the identification of the molecule is deduced based on the oxidation type and the annotation of both fatty acids, e.g. PC(16:0/4:0(COOH)). Likewise to LCh-oxPC functionality, in SCh-oxPC annotation, NLes and product ions needed for confirmation or rejection of annotations are also provided.

3.3. Validation

To test this functionality a set of previously identified LCh-oxPCs and SCh-oxPCs was used. These oxPCs were found in the standard of mixture of oxPAPC. To test the false positive annotations also non-oxidized PCs were included (Table 5S, supplementary material). The *m/z* of the potential fatty acid(s) and the precursor ion were searched with 10 ppm error and without any further restrictions regarding the database or type of oxidation. To choose between the proposed annotations leading to the formation of isobaric compounds, the expected NLes given as an output were searched in MS/MS spectra. The results of this validation are summarized in Table 5S (supplementary material). This table shows that for each lipid two candidates were obtained for different oxidation types, except native forms (non-oxidized) where a single candidate was retrieved. However, when the information about NLes provided by CMM was confronted with the MS/MS spectra, a single final annotation based on the presence or absence of the NLes and the corresponding product ions was achieved. For non-oxidized PCs, a single candidate was obtained, however the proposed candidate could not be confirmed by MS/MS with the indicated NL (Table 5S, supplementary material).

4. Conclusions

A general non-targeted metabolomics method has been employed to analyze a complex standard mix of oxidation products of PAPC. CID has been employed to study the fragmentation pattern of the analytes. General diagnostic characteristics from MS and MS/MS levels have been defined. Specific signals (both product ions and NLes) for the presence of oxidation in the PC structure were determined. These signals do not necessary lead to the identification of a particular oxidized PAPC. However, a fast and reliable determination of the presence of oxidation in the PC reduces the amount of time spent for the overall identification process.

A deeper characterization of each product of the PAPC oxidation has been conducted. Spectra were acquired both in positive and negative ionization mode. In this study, the importance of acquiring spectra in both ESI modes has been marked. Although the negative ionization mode was confirmed to be the most informative regarding the nature of the fatty acids esterified to the PC, the positive ionization mode has enabled an easier identification of the type of oxidation in LCh-oxPAPC.

Finally, data for MS/MS spectra acquired in this study have been used to build an in-house database specific for oxidized lipids, which is accessible from the metabolite annotation tool CMM. Additionally, a new tool has been implemented for the identification of oxidized fatty acids and the assignment of candidates for the precursor ion for LCh-oxPC and SCh-oxPC.

This work constitutes a great advance in the recognition of oxidation in PCs and in non-targeted metabolomics in general. It should be mentioned that the proposed strategy is not useful for very detailed identification of oxidation products of PCs, requiring the exact position of the oxidation agent, nor the nature of oxidation (enzymatic or non-enzymatic). However, it represents a highly valuable tool in the recognition of oxidation among PCs, which can be used to drive new hypotheses and further plan the design of biological experiments. The aim of this work was to provide the reader with a range of different parameters which can be used to recognize and identify oxidized PCs in non-targeted metabolomics.

Conflicts of interest

All authors declare neither financial nor commercial conflicts of interest.

Acknowledgements

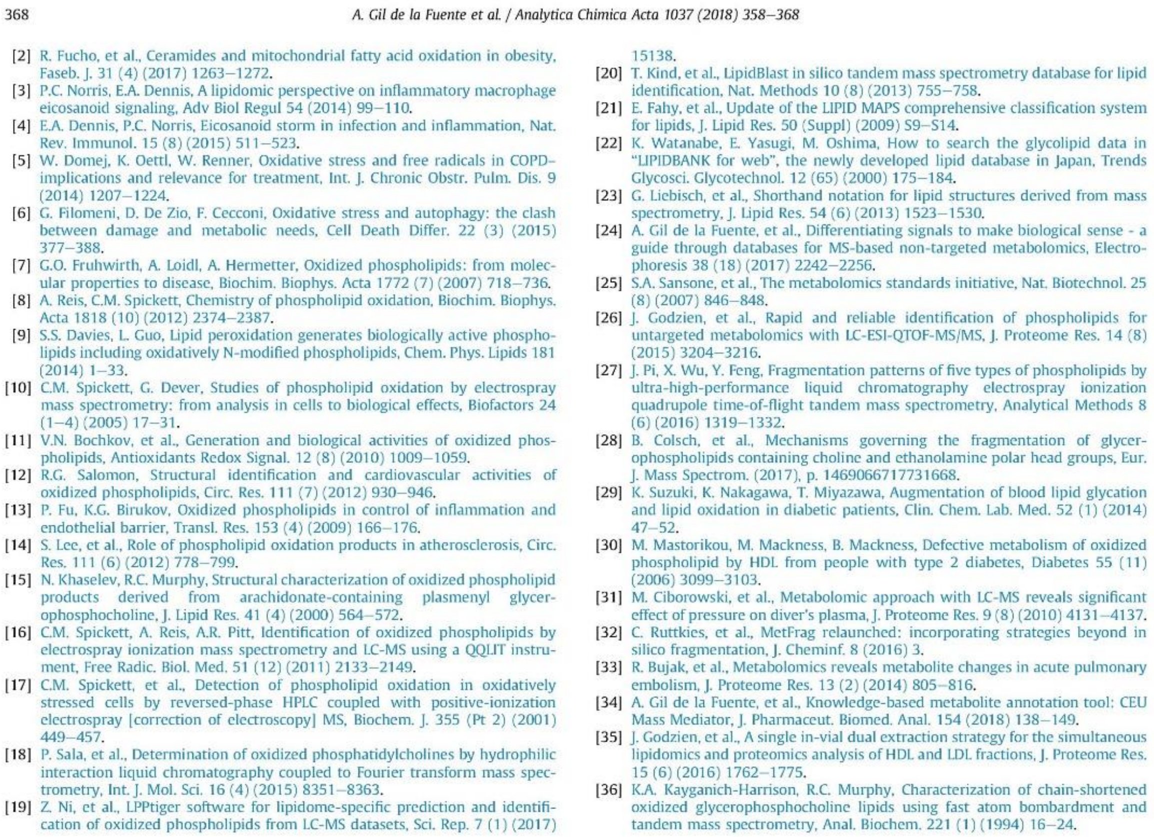
This work was supported by grants from the Spanish Ministerio de Economía y Competitividad (grant CTQ2014-55279-R) and USPCEU PCON10/2016. AGF acknowledges Fundación Universitaria San Pablo CEU for his PhD fellowship. JS acknowledges European Social Fund Postdoc Project No. CZ.1.07/2.3.00/30.0004; The Ministry of Education, Youth and Sports of the Czech Republic (NPU — LO1204) and Nadace Český literární fond for travelling support. The authors are indebted to Mr. Dan Cherry for his careful English editing and proofreading.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.aca.2018.08.005.

References

[1] S.M. Houten, et al., The biochemistry and physiology of mitochondrial fatty acid β-oxidation and its genetic disorders, Annu. Rev. Physiol. 78 (2016) 23–44.



[2] R. Fucho, et al., Ceramides and mitochondrial fatty acid oxidation in obesity, Faseb. J. 31 (4) (2017) 1263–1272.

[3] P.C. Norris, E.A. Dennis, A lipidomic perspective on inflammatory macrophage eicosanoid signaling, Adv Biol Regul 54 (2014) 99–110.

[4] E.A. Dennis, P.C. Norris, Eicosanoid storm in infection and inflammation, Nat. Rev. Immunol. 15 (8) (2015) 511–523.

[5] W. Domej, K. Oettl, W. Renner, Oxidative stress and free radicals in COPD–implications and relevance for treatment, Int. J. Chronic Obstr. Pulm. Dis. 9 (2014) 1207–1224.

[6] G. Filomeni, D. De Zio, F. Cecconi, Oxidative stress and autophagy: the clash between damage and metabolic needs, Cell Death Differ. 22 (3) (2015) 377–388.

[7] G.O. Fruhwirth, A. Loidl, A. Hermetter, Oxidized phospholipids: from molecular properties to disease, Biochim. Biophys. Acta 1772 (7) (2007) 718–736.

[8] A. Reis, C.M. Spickett, Chemistry of phospholipid oxidation, Biochim. Biophys. Acta 1818 (10) (2012) 2374–2387.

[9] S.S. Davies, L. Guo, Lipid peroxidation generates biologically active phospholipids including oxidatively N-modified phospholipids, Chem. Phys. Lipids 181 (2014) 1–33.

[10] C.M. Spickett, G. Dever, Studies of phospholipid oxidation by electrospray mass spectrometry: from analysis in cells to biological effects, Biofactors 24 (1–4) (2005) 17–31.

[11] V.N. Bochkov, et al., Generation and biological activities of oxidized phospholipids, Antioxidants Redox Signal. 12 (8) (2010) 1009–1059.

[12] R.G. Salomon, Structural identification and cardiovascular activities of oxidized phospholipids, Circ. Res. 111 (7) (2012) 930–946.

[13] P. Fu, K.G. Birukov, Oxidized phospholipids in control of inflammation and endothelial barrier, Transl. Res. 153 (4) (2009) 166–176.

[14] S. Lee, et al., Role of phospholipid oxidation products in atherosclerosis, Circ. Res. 111 (6) (2012) 778–799.

[15] N. Khaselev, R.C. Murphy, Structural characterization of oxidized phospholipid products derived from arachidonate-containing plasmenyl glycerophosphocholine, J. Lipid Res. 41 (4) (2000) 564–572.

[16] C.M. Spickett, A. Reis, A.R. Pitt, Identification of oxidized phospholipids by electrospray ionization mass spectrometry and LC-MS using a QQLIT instrument, Free Radic. Biol. Med. 51 (12) (2011) 2133–2149.

[17] C.M. Spickett, et al., Detection of phospholipid oxidation in oxidatively stressed cells by reversed-phase HPLC coupled with positive-ionization electrospray [correction of electroscopy] MS, Biochem. J. 355 (Pt 2) (2001) 449–457.

[18] P. Sala, et al., Determination of oxidized phosphatidylcholines by hydrophilic interaction liquid chromatography coupled to Fourier transform mass spectrometry, Int. J. Mol. Sci. 16 (4) (2015) 8351–8363.

[19] Z. Ni, et al., LPPtiger software for lipidome-specific prediction and identification of oxidized phospholipids from LC-MS datasets, Sci. Rep. 7 (1) (2017) 15138.

[20] T. Kind, et al., LipidBlast in silico tandem mass spectrometry database for lipid identification, Nat. Methods 10 (8) (2013) 755–758.

[21] E. Fahy, et al., Update of the LIPID MAPS comprehensive classification system for lipids, J. Lipid Res. 50 (Suppl) (2009) S9–S14.

[22] K. Watanabe, E. Yasugi, M. Oshima, How to search the glycolipid data in "LIPIDBANK for web", the newly developed lipid database in Japan, Trends Glycosci. Glycotechnol. 12 (65) (2000) 175–184.

[23] G. Liebisch, et al., Shorthand notation for lipid structures derived from mass spectrometry, J. Lipid Res. 54 (6) (2013) 1523–1530.

[24] A. Gil de la Fuente, et al., Differentiating signals to make biological sense - a guide through databases for MS-based non-targeted metabolomics, Electrophoresis 38 (18) (2017) 2242–2256.

[25] S.A. Sansone, et al., The metabolomics standards initiative, Nat. Biotechnol. 25 (8) (2007) 846–848.

[26] J. Godzien, et al., Rapid and reliable identification of phospholipids for untargeted metabolomics with LC-ESI-QTOF-MS/MS, J. Proteome Res. 14 (8) (2015) 3204–3216.

[27] J. Pi, X. Wu, Y. Feng, Fragmentation patterns of five types of phospholipids by ultra-high-performance liquid chromatography electrospray ionization quadrupole time-of-flight tandem mass spectrometry, Analytical Methods 8 (6) (2016) 1319–1332.

[28] B. Colsch, et al., Mechanisms governing the fragmentation of glycerophospholipids containing choline and ethanolamine polar head groups, Eur. J. Mass Spectrom. (2017), p. 1469066717731668.

[29] K. Suzuki, K. Nakagawa, T. Miyazawa, Augmentation of blood lipid glycation and lipid oxidation in diabetic patients, Clin. Chem. Lab. Med. 52 (1) (2014) 47–52.

[30] M. Mastorikou, M. Mackness, B. Mackness, Defective metabolism of oxidized phospholipid by HDL from people with type 2 diabetes, Diabetes 55 (11) (2006) 3099–3103.

[31] M. Ciborowski, et al., Metabolomic approach with LC-MS reveals significant effect of pressure on diver's plasma, J. Proteome Res. 9 (8) (2010) 4131–4137.

[32] C. Ruttkies, et al., MetFrag relaunched: incorporating strategies beyond in silico fragmentation, J. Cheminf. 8 (2016) 3.

[33] R. Bujak, et al., Metabolomics reveals metabolite changes in acute pulmonary embolism, J. Proteome Res. 13 (2) (2014) 805–816.

[34] A. Gil de la Fuente, et al., Knowledge-based metabolite annotation tool: CEU Mass Mediator, J. Pharmaceut. Biomed. Anal. 154 (2018) 138–149.

[35] J. Godzien, et al., A single in-vial dual extraction strategy for the simultaneous lipidomics and proteomics analysis of HDL and LDL fractions, J. Proteome Res. 15 (6) (2016) 1762–1775.

[36] K.A. Kayganich-Harrison, R.C. Murphy, Characterization of chain-shortened oxidized glycerophosphocholine lipids using fast atom bombardment and tandem mass spectrometry, Anal. Biochem. 221 (1) (1994) 16–24.

CHARACTERIZATION AND ANNOTATION OF OXIDIZED GLYCEROPHOSPHOCHOLINES FOR NON-TARGETED METABOLOMICS WITH LC-QTOF-MS DATA

SUPPLEMENTARY INFORMATION

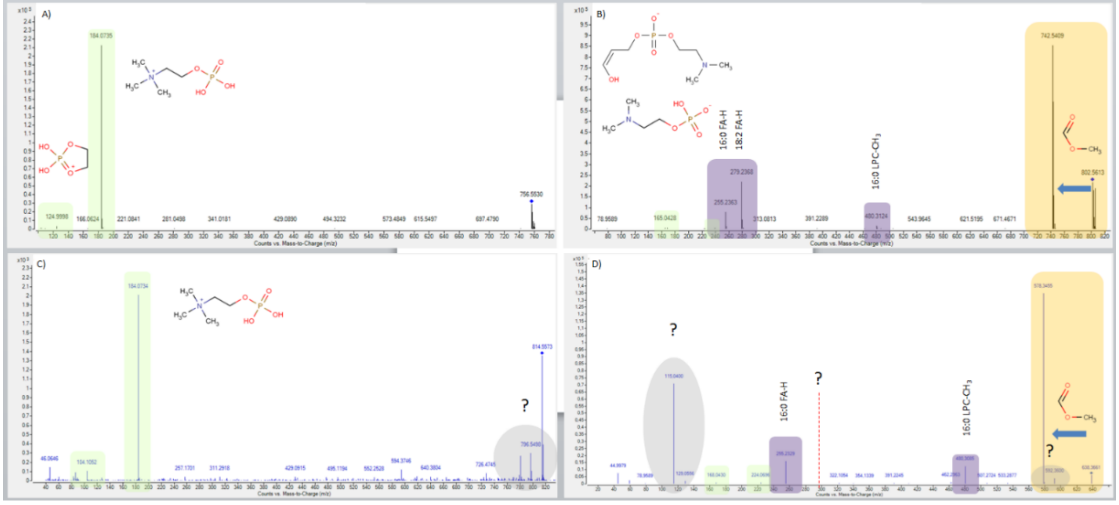


Figure 1S: an example of MS/MS spectra and their interpretation for PCs and oxPCs. Panels A and C illustrate spectra in positive mode and panels B and D illustrate spectra in negative ionization mode. Panels A and B illustrate typical spectra for non-oxidized PC, while panels C and D show spectra of oxPC. Each panel includes characteristic regions of the spectrum: low- (green), mid- (purple) and high-mass (orange). Grey color indicates unexplained signals.

Table 1S: information about participants

Gender	Age	BMI	HbA1c	Glucose	Insulin	HOMA-IR
female / male	Years ± SD	kg/m^2 ± SD	% ± SD	mg/dL ± SD	µU/mL ± SD	value ± SD
14 / 23	52.1 ± 11.3	29.2 ± 2.8	7.6 ± 2.2	143.0 ± 40.2	6.6 ± 7.9	5.1 ± 2.8

HOMA-IR: homeostasis model assessment-estimated insulin resistance index.

HbA1c: glycated hemoglobin

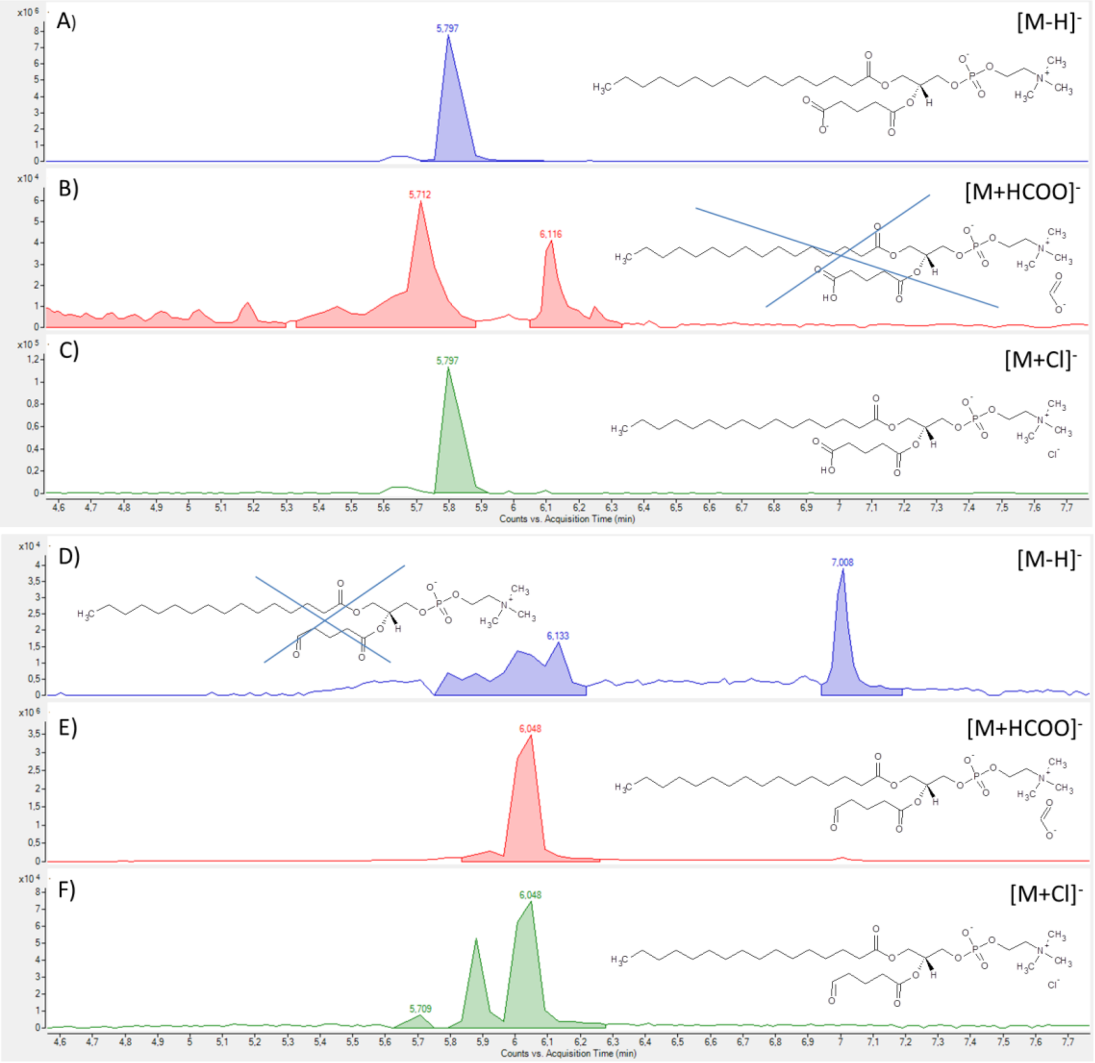


Figure 2S: EIC for different adducts of two SCh-oxPC: PC(16:0/5:0(COOH)) (panels A, B and C) and PC(16:0/5:0(CHO) (panels D, E and F). Blue peaks are for [M-H]$^-$ ions whereas red are for [M+HCOO]$^-$ and green for [M+Cl]$^-$ ions. The ω-COOH group located at the end of the chain in the *sn-2* position of PC(16:0/5:0(COOH)) allows the ionization of the compound as [M-H]$^-$, while the semialdehydic chain located in position *sn*-2 of PC(16:0/5:0(CHO) makes its ionization impossible without the addition of a modifier to the mobile phase in negative ionization mode.

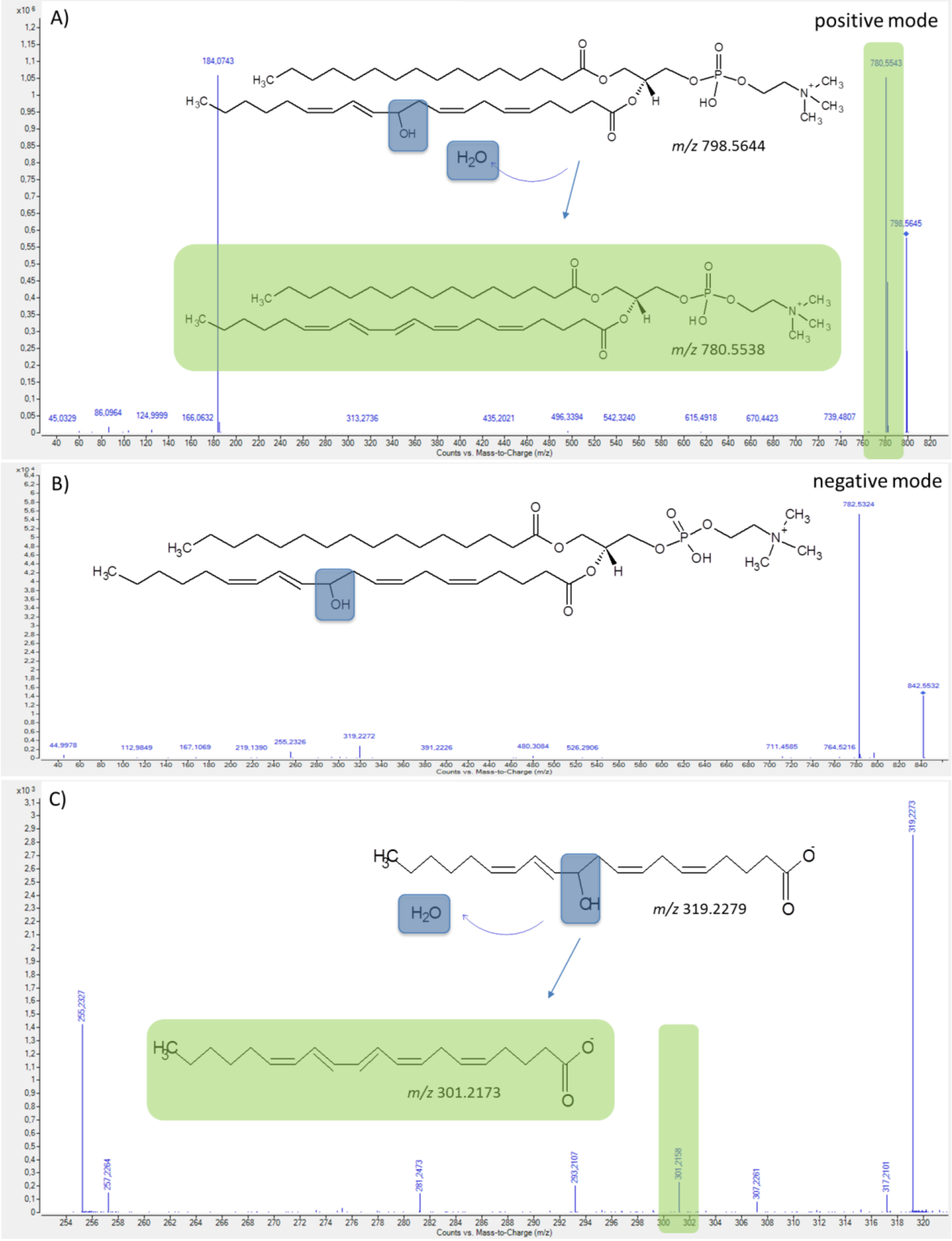


Figure 3S: Product ion spectrum for PC(16:0/20:4(OH)) in positive mode (panel A) and negative mode (panels B and C). Panel C shows a zoom of the fatty acid region for panel B.

Table 2S: Characteristic signals for SCh-oxPAPC in negative ionization mode. Low abundant signals are reported in parenthesis.

common name	composition	NLs	product ions
OB-PC	PC(16:0/4:0(CHO))	46.0055; 60.0211	624.3518; 578.3463; 564.3307; 101.0244
OV-PC	PC(16:0/5:0(CHO))	46.0055; 60.0211	638.3675; 592.3620; 578,3463; 115.0420
Hex-PC	PC(16:0/6:0(CHO))	46.0055; 60.0211	652.3831; 606.3776; 592.3620; 129.0557
Hept-PC	PC(16:0/7:0(CHO))	46.0055; 60.0211	666.3988; 620.3933; 606.3776; 143.0714
ON-PC	PC(16:0/9:0(CHO))	46.0055; 60.0211	694.4301; 648.4246; 634.4089; 171.1027
OD-PC	PC(16:0/10:0(CHO))	46.0055; 60.0211	708.4457; 662.4402; 648.4246; 185.1183
OU-PC	PC(16:0/11:0(CHO))	46.0055; 60.0211	722.4614; 676.4559; 662.4402; 199.1340
HOOA-PC	PC(16:0/8:1(CHO-OH))	46.0055; 60.0211; 43.9898; 18.0105; 14.0157;	694.3927; 676.3831; 648.3882; 634.3726; 185.0819; 171.0663; 153.0557; 109.0659;
KOOA-PC	PC(16:0/8:1(CHO-O))	46.0055; 60.0211; 43.9898; 14.0157;	692.3780; 646.3726; 632.3569; 183.0663; 169.0506; 151.0401; 139.0401; 125.0608;
S-PC	PC(16:0/4:0(COOH))	59.0735; 32.0262	594.3413; 535.2678; 131.0350; (117.0193); 99.0088;
G-PC	PC(16:0/5:0(COOH))	59.0735; 32.0262	608.3569; 549.2834; 145.0506; (131.0350); 113.0244;
Hexendia-PC	PC(16:0/6:0(COOH))	59.0735; 32.0262	620.3569; 561.2834; 157.0506; 125.0244; 113.0608; 81.0346
Heptendia-PC	PC(16:0/7:0(COOH))	59.0735; 32.0262	634.3726; 575.2991; 171.0663; 139.0401; 95.0502
HOdiA-PC	PC(16:0/8:1(COOH-OH))	59.0735; 43.9898; 32.0262	664.3831; 620.3933; 605.3096; 201.0769; (187.0612); 169.0506; 125.0608; 99.0452
KOdiA-PC	PC(16:0/8:1(COOH-O))	59.0735; 43.9898; 32.0262	662.3675; 602.2867; 199.0612; (185.0456); 167.0350; (141.0557); 123.0452; 99.0088

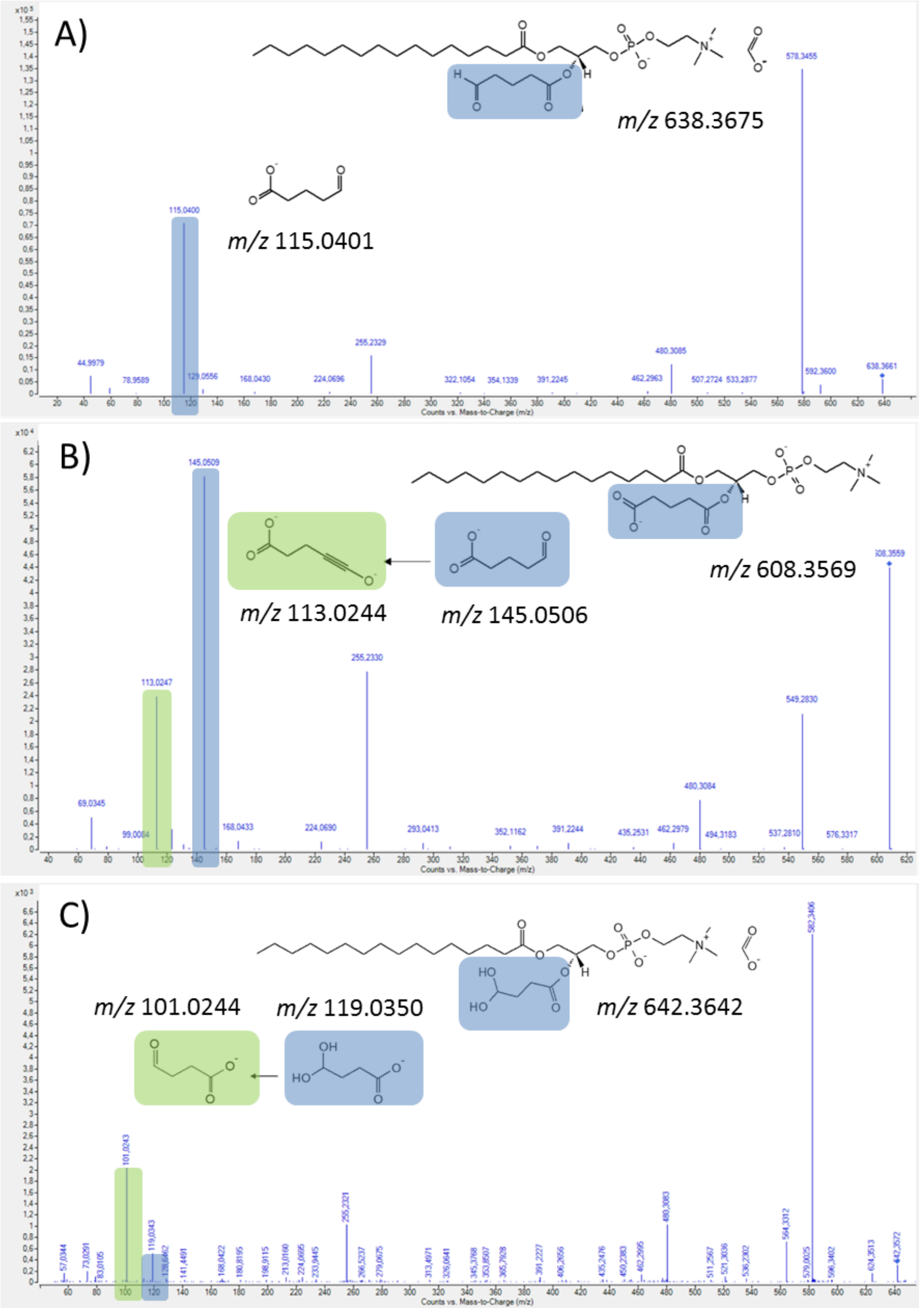
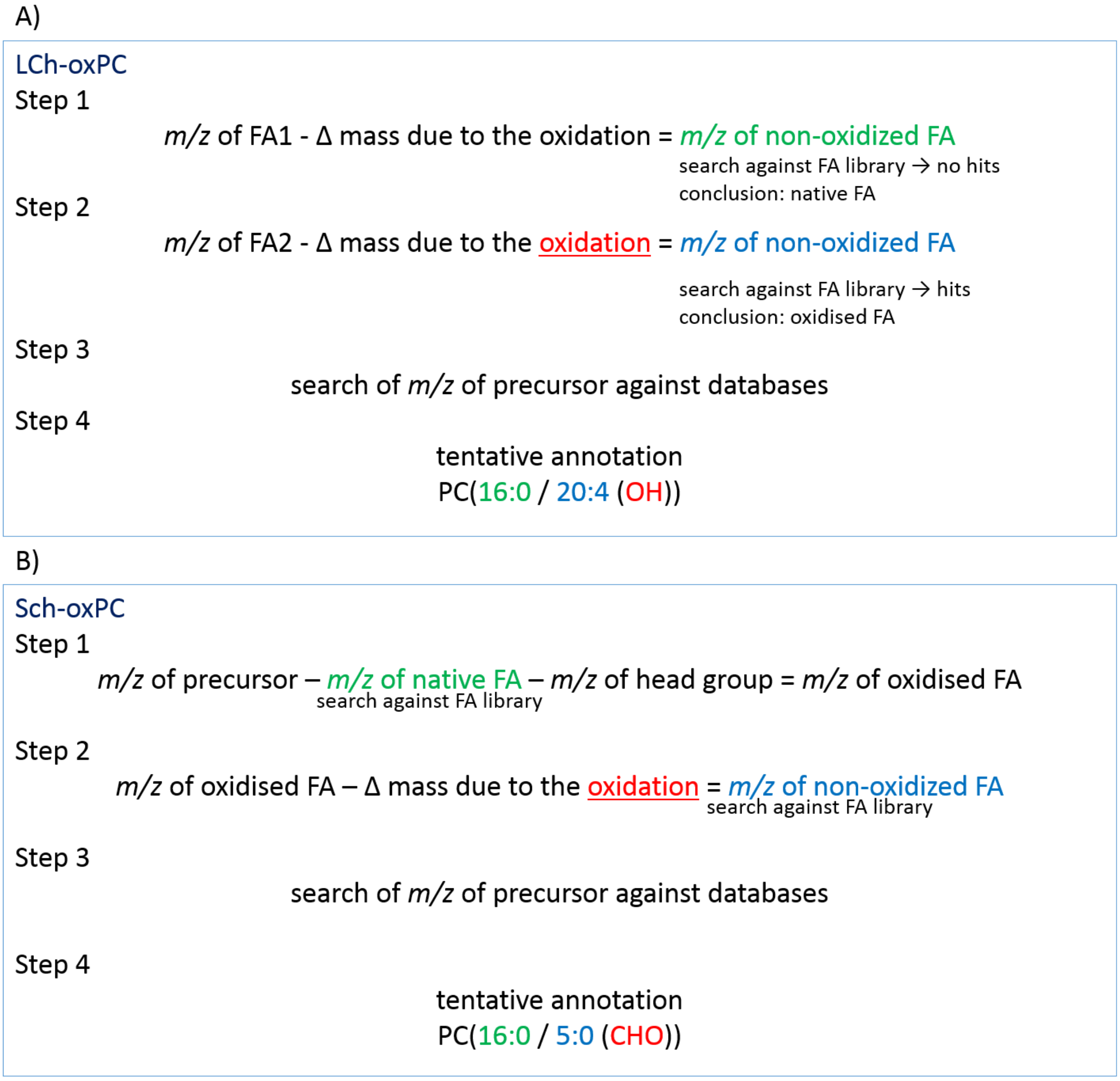


Figure 4S: Product ion spectrum of OV-PC PC(16:0/5:0(CHO)) (1-palmitoyl-2-(5-oxovaleroyl)-sn-glycero-3-phosphocholine) (Panel A), G-PC PC(16:0/5:0(COOH)) (1-palmitoyl-2-glutaryl-sn-glycero-3-phosphocholine) (Panel B) and GemB-PC PC(16:0/4:0-CHO-O) (1-palmitoyl-2-(4,4-dihydroxypentanoyl)-sn-glycero-3-phosphocholine) (Panel C). Signals in the medium-low mass region are used in order to distinguish SCh-oxPAPC with different terminal groups. OV-PC presents one dominant signal in this region, which is the [M-H]$^-$ signal of the short chain. G-PC presents two main signals instead.

Table 3S: The list of oxPCs mentioned in this publication. The list includes the common and the systematic name as well as the composition. It is followed by the accurate exact monoisotopic mass and the formula.

composition	IUPAC name	short name	molecular weight	formula	source
PC(16:0/20:4)	1-palmitoyl-2-arachidonoyl-*sn*-glycero-3-phosphocholine	PAPC	781.5621	C44H81NO8P	canonical
PC(16:0/4:0(CHO))	1-palmitoyl-4-oxobutanoyl-*sn*-glycero-3-phosphocholine	OB-PC	579.3536	C44H81NO8P	canonical
PC(16:0/4:0(COOH))	1-palmitoyl-2-succinyl-*sn*-glycero-3-phosphocholine	S-PC	595.3485	C28H54NO9P	canonical
PC(16:0/5:0(COOH))	1-palmitoyl-2-glutaryl-*sn*-glycero-3-phosphocholine	G-PC	609.3642	C28H54NO10P	canonical
PC(16:0/5:0(CHO))	1-palmitoyl-2-(5-oxovaleroyl)-*sn*-glycero-3-phosphocholine	OV-PC	593.3693	C29H56NO9P	canonical
PC(16:0/6:0(CHO))	1-palmitoyl-2-(6-oxohexanoyl)-*sn*-glycero-3-phosphocholine	Hex-PC	607.3849	C30H58NO9P	canonical
PC(16:0/6:0(COOH))	1-palmitoyl-2-(5-carboxypentanoyl)-*sn*-glycero-3-phosphocholine	Hexendia-PC	623.3798	C30H58NO10P	canonical
PC(16:0/7:0(CHO))	1-palmitoyl-2(7-oxoheptanoyl)-*sn*-glycero-3-phosphocholine	Hept-PC	621.4006	C31H60NO9P	canonical
PC(16:0/7:0(COOH))	1-palmitoyl-2-(6-carboxyhexenoyl)-*sn*-glycero-3-phosphocholine	Heptendia-PC	637.3955	C31H60NO10P	canonical
PC(16:0/9:0(COOH))	1-palmitoyl-2-azelaoyl-*sn*-glycero-3-phosphocholine	AZ-PC	665.4268	C33H64NO10P	canonical
PC(16:0/9:0(CHO))	1-palmitoyl-2-(9-oxononanoyl)-*sn*-glycero-3-phosphocholine	ON-PC	649.4319	C33H64NO9P	canonical
PC(16:0/10:0(CHO))	1-palmitoyl-2-(10-oxodecanoyl)-*sn*-glycero-3-phosphocholine	OD-PC	663.4475	C34H66NO9P	canonical
PC(16:0/11:0(CHO))	1-palmitoyl-2-(11-oxoundecanoyl)-*sn*-glycero-3-phosphocholine	OU-PC	677.4632	C35H68NO9P	canonical
PC(16:0/8:1(COOH-O))	1-palmitoyl-2-(5-keto-oct-6-ene-dioyl)-*sn*-glycero-3-phosphatidylcholine	KOdiA-PC	663.3747	C32H58NO11P	canonical
PC(16:0/8:1(CHO-OH))	1-palmitoyl-2-(5-hydroxy-8-oxooct-6-enoyl)-*sn*-glycero-3-phosphocholine	HOOA-PC	649.3955	C32H60NO10P	canonical
PC(16:0/8:1(COOH-OH))	1-palmitoyl-2-(5-hydroxy-8-oxo-oct-6-ene-dioyl)-*sn*-glycero-3-phosphocholine	HOdiA-PC	665.3903	C32H60NO11P	canonical
PC(16:0/8:1(CHO-O))	1-palmitoyl-2-(5,8-dioxo-oct-6-en-yl)-*sn*-glycero-3-phosphocholine	KOOA-PC	647.3798	C32H58NO10P	canonical
PC(16:0/4:0(ω(OH)$_2$))	1-palmitoyl-2-(4,4-dihydroxybutanoyl)-*sn*-glycero-3-phosphocholine	GemB-PC	597.3642	C28H56NO10P	canonical
PC(16:0/5:0(ω(OH)$_2$))	1-palmitoyl-2-(4,4-dihydroxypentanoyl)-*sn*-glycero-3-phosphocholine	GemP-PC	611.3798	C29H58NO10P	canonical

PC(16:0/18:2(OH))	1-hexadecanoyl-2-hydroxyoctadecadienoyl-*sn*-glycero-3-phosphocholine	POPC-OH	773.5571	C42H80NO9P	new
PC(18:0/20:4(OH))	1-octadecadienoyl-2-hydroxyarachidonyl-*sn*-glycero-3-phosphocholine	OAPC-OH	825.5584	C46H84NO9P	new
PC(18:0/18:2(OH))	1-octadecadienoyl-2-hydroxyoctadecadienoyl-*sn*-glycero-3-phosphocholine	OOPC-OH	801.5884	C44H84NO9P	new
PC(16:0/22:6(OH))	1-hexadecanoyl-2-hydroxydocosahexaenoyl-*sn*-glycero-3-phosphocholineadecadienoyl-sn-glycero-3-phosphocholine	PDPC-OH	821.5571	C46H80NO9P	new
PC(16:0/20:4(OH))	1-hexadecanoyl-2-hydroxyarachidonyl-*sn*-glycero-3-phosphocholine	PAPC-OH	797.5571	C44H80NO9P	new



A)

LCh-oxPC

Step 1

m/z of FA1 - Δ mass due to the oxidation = *m/z* of non-oxidized FA

search against FA library → no hits
conclusion: native FA

Step 2

m/z of FA2 - Δ mass due to the oxidation = *m/z* of non-oxidized FA

search against FA library → hits
conclusion: oxidised FA

Step 3

search of *m/z* of precursor against databases

Step 4

tentative annotation
PC(16:0 / 20:4 (OH))

B)

Sch-oxPC

Step 1

m/z of precursor – *m/z* of native FA – *m/z* of head group = *m/z* of oxidised FA

search against FA library

Step 2

m/z of oxidised FA – Δ mass due to the oxidation = *m/z* of non-oxidized FA

search against FA library

Step 3

search of *m/z* of precursor against databases

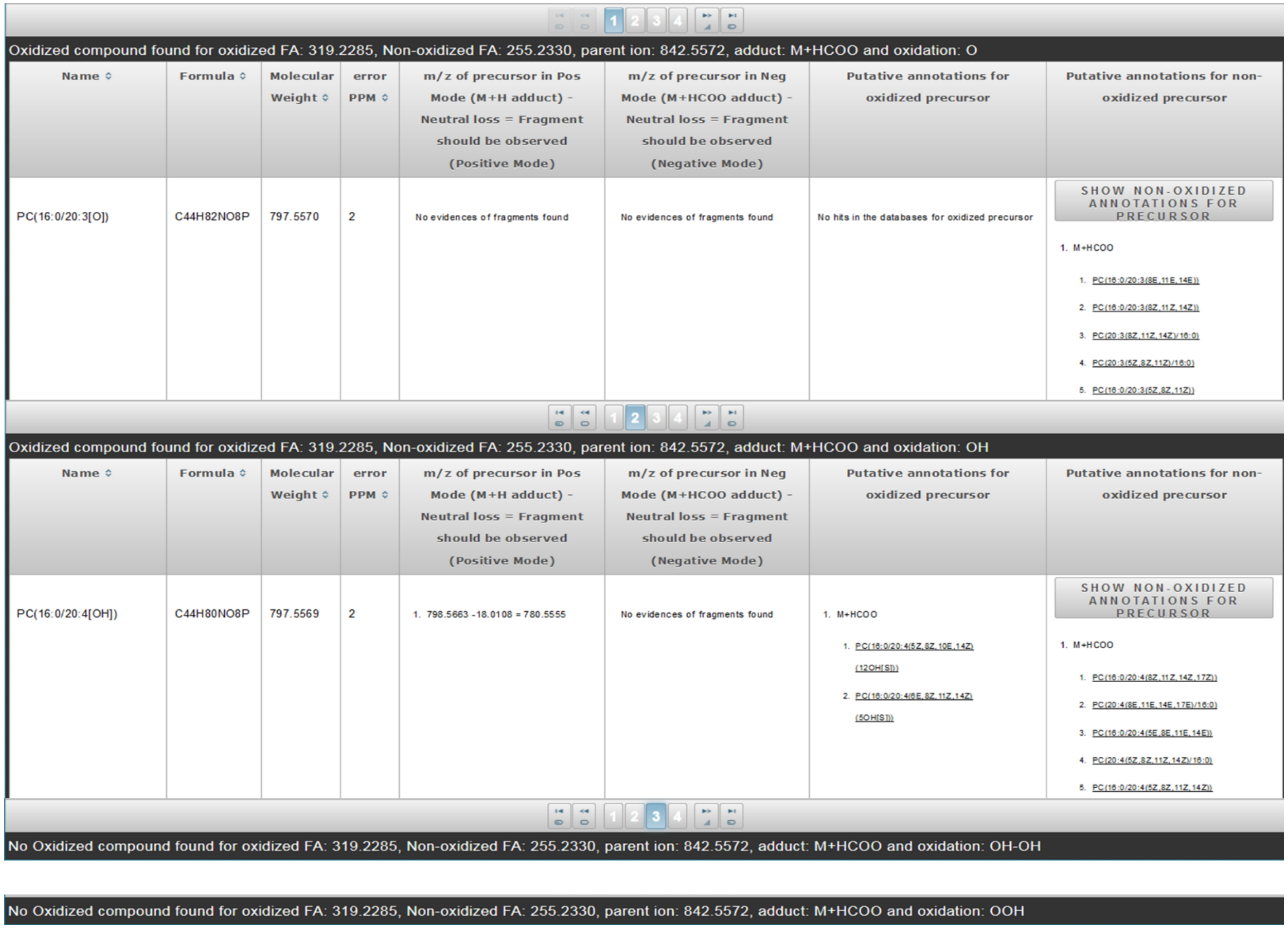
Step 4

tentative annotation
PC(16:0 / 5:0 (CHO))

Figure 5S: The scheme of the operations leading to the annotation of oxPCs: LChoxPCs (Panel A) and SChoxPCs (Panel B).

Table 4S: An example of experimental *m/z* of 294.2195 fatty acid and its possible non-oxidized masses after re-calculation of mass according to long and short chain oxidations.

m/z of oxidised fatty acid	oxidation	Δm	*m/z* of non-oxidised fatty acid	identification
Long chain oxidised glycerophosphocholines				
293.2093	**=O**	13.9793	279.2330	C18:2
293.2093	**-OH**	15.9949	277.2174	C18:3
293.2093	**-OOH**	31.9905	261.2188	-
Short chain oxidised glycerophosphocholines				
293.2093	**-CHO**	13.9793	279.2300	-
293.2093	**-COOH**	29.9742	263.2351	-



Oxidized compound found for oxidized FA: 319.2285, Non-oxidized FA: 255.2330, parent ion: 842.5572, adduct: M+HCOO and oxidation: O

Name	Formula	Molecular Weight	error PPM	m/z of precursor in Pos Mode (M+H adduct) - Neutral loss = Fragment should be observed (Positive Mode)	m/z of precursor in Neg Mode (M+HCOO adduct) - Neutral loss = Fragment should be observed (Negative Mode)	Putative annotations for oxidized precursor	Putative annotations for non-oxidized precursor
PC(16:0/20:3[O])	C44H82NO8P	797.5570	2	No evidences of fragments found	No evidences of fragments found	No hits in the databases for oxidized precursor	SHOW NON-OXIDIZED ANNOTATIONS FOR PRECURSOR 1. M+HCOO 1. PC(16:0/20:3(8E,11E,14E)) 2. PC(16:0/20:3(8Z,11Z,14Z)) 3. PC(20:3(8Z,11Z,14Z)/16:0) 4. PC(20:3(5Z,8Z,11Z)/16:0) 5. PC(16:0/20:3(5Z,8Z,11Z))

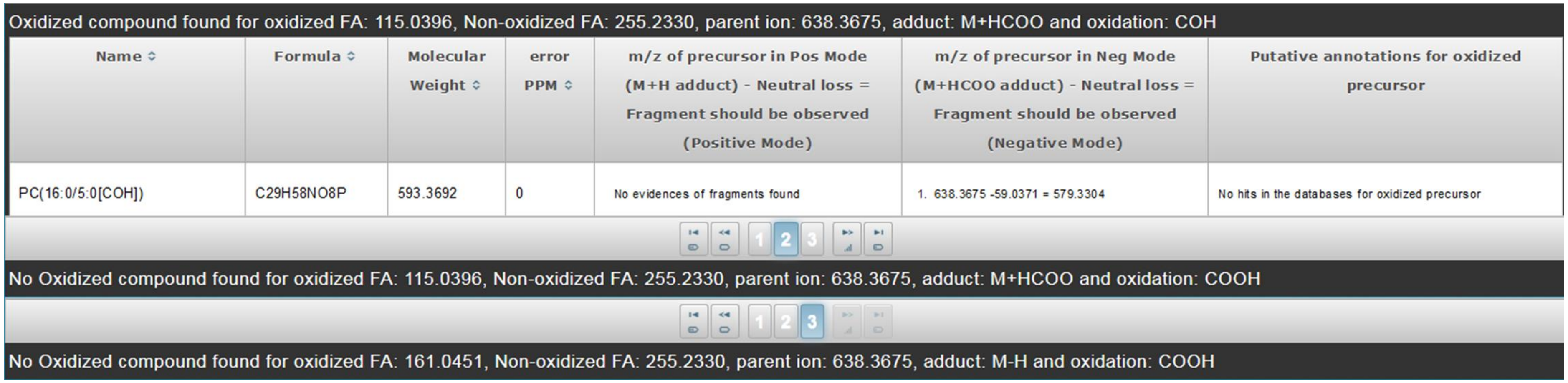
Oxidized compound found for oxidized FA: 319.2285, Non-oxidized FA: 255.2330, parent ion: 842.5572, adduct: M+HCOO and oxidation: OH

Name	Formula	Molecular Weight	error PPM	m/z of precursor in Pos Mode (M+H adduct) - Neutral loss = Fragment should be observed (Positive Mode)	m/z of precursor in Neg Mode (M+HCOO adduct) - Neutral loss = Fragment should be observed (Negative Mode)	Putative annotations for oxidized precursor	Putative annotations for non-oxidized precursor
PC(16:0/20:4[OH])	C44H80NO8P	797.5569	2	1. 798.5663 -18.0108 = 780.5555	No evidences of fragments found	1. M+HCOO 1. PC(16:0/20:4(5Z,8Z,10E,14Z)(12OH[S])) 2. PC(16:0/20:4(6E,8Z,11Z,14Z)(5OH[S]))	SHOW NON-OXIDIZED ANNOTATIONS FOR PRECURSOR 1. M+HCOO 1. PC(16:0/20:4(8Z,11Z,14Z,17Z)) 2. PC(20:4(8E,11E,14E,17E)/16:0) 3. PC(16:0/20:4(5E,8E,11E,14E)) 4. PC(20:4(5Z,8Z,11Z,14Z)/16:0) 5. PC(16:0/20:4(5Z,8Z,11Z,14Z))

No Oxidized compound found for oxidized FA: 319.2285, Non-oxidized FA: 255.2330, parent ion: 842.5572, adduct: M+HCOO and oxidation: OH-OH

No Oxidized compound found for oxidized FA: 319.2285, Non-oxidized FA: 255.2330, parent ion: 842.5572, adduct: M+HCOO and oxidation: OOH

Figure 5S: Output of identification of oxidized fatty acids for four possible oxidations for LChoxPC functionality.



Oxidized compound found for oxidized FA: 115.0396, Non-oxidized FA: 255.2330, parent ion: 638.3675, adduct: M+HCOO and oxidation: COH

Name	Formula	Molecular Weight	error PPM	m/z of precursor in Pos Mode (M+H adduct) - Neutral loss = Fragment should be observed (Positive Mode)	m/z of precursor in Neg Mode (M+HCOO adduct) - Neutral loss = Fragment should be observed (Negative Mode)	Putative annotations for oxidized precursor
PC(16:0/5:0[COH])	C29H58NO8P	593.3692	0	No evidences of fragments found	1. 638.3675 -59.0371 = 579.3304	No hits in the databases for oxidized precursor

No Oxidized compound found for oxidized FA: 115.0396, Non-oxidized FA: 255.2330, parent ion: 638.3675, adduct: M+HCOO and oxidation: COOH

No Oxidized compound found for oxidized FA: 161.0451, Non-oxidized FA: 255.2330, parent ion: 638.3675, adduct: M-H and oxidation: COOH

Figure 6S: Output of identification of oxidized fatty acids for two possible oxidations with different ionisation possibilities for SChoxPC functionality in CMM.

Table 5S: Results of validation of CMM for identification of oxPC. Bolded text indicates final correct result.

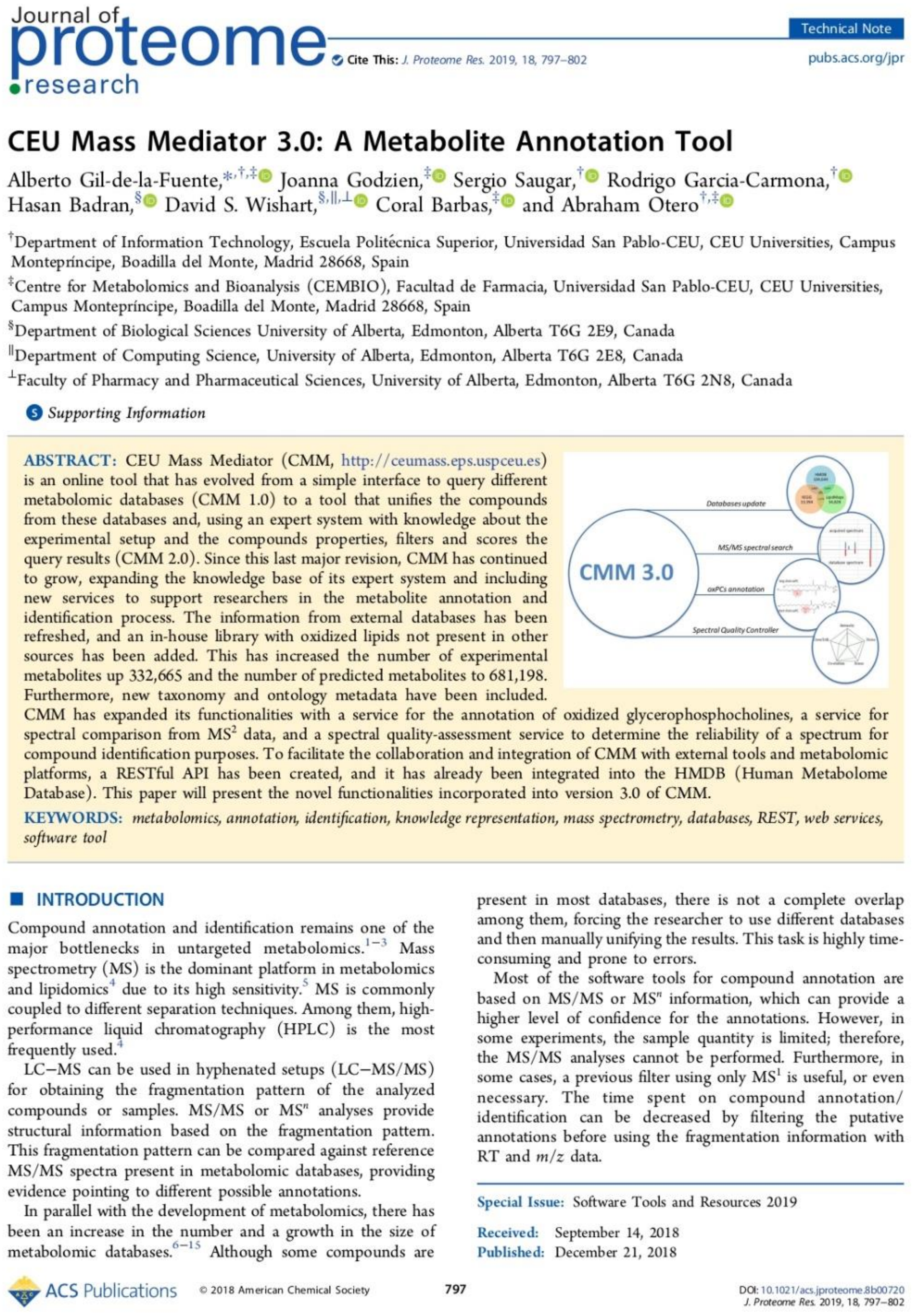
oxPC	m/z of the precursor ion	m/z of the fatty acid	oxidation type	fatty acid	evidence	oxidised precursor	non-oxidised precursor
PC(16:0/20:4(OH))	842.5572	319.2285 255.2331	=O	C20:3	no evidence	no hit	PC(16:0/20:3)
			-OH	**C20:4**	**NL of 18:0108 producing fragment 780.5554**	**PC(16:0/20:4(OH))**	**PC(16:0/20:4)**
			-OH-OH	no hit			
			-OOH	no hit			
PC(16:0/22:6(OH))	866.5564	343.2282 255.2333	=O	C22:5	no evidence	no hit	PC(16:0/22:5)
			-OH	**C22:6**	**NL of 18:0108 producing fragment 804.5546**	**no hit**	**PC(16:0/22:6)**
			-OH-OH	no hit			
			-OOH	no hit			
PC(16:0/18:2(OH))	818.5572 255.2330	295.2280	=O	C18:1	no evidence	no hit	PC(16:0/18:1)
			-OH	**C18:2**	**NL of 18:0108 producing fragment 756.5554**	**no hit**	**PC(16:0/18:2)**
			-OH-OH	no hit			
			-OOH	no hit			
PC(16:0/18:2(O))	816.5405	293.2140 255.2331	**=O**	**C18:2**	**no evidence**	**no hit**	**PC(16:0/18:2)**
			-OH	C18:3	NL of 18:0108 producing fragment 754.5387	no hit	PC(16:0/18:3)
			-OH-OH	no hit			
			-OOH	no hit			
PC(16:0/20:4(OOH))	886.5825	335.2237 255.2331	=O	no hit			
			-OH	no hit			
			-OH-OH	C20:4	NL of 18:0108 producing fragment 824.5807 and NL of 36.0216 producing fragment 806.5699	no hits	PC(18:0/20:4)

			-OOH	**C20:4**	**NL of 18:0108 producing fragment 824.5807 and NL of 34.0049 producing fragment 808.5866**	**no hits**	**PC(18:0/20:4)**
PC(16:0/18:2(OOH))	834.5520	311.2242 255.2331	=O	no hit			
			-OH	no hit			
			-OH-OH	C18:2	NL of 18:0108 producing fragment 772.5502 and NL of 36.0216 producing fragment 754.5394	no hits	PC(16:0/18:2)
			-OOH	**C18:2**	**NL of 18:0108 producing fragment 772.5502 and NL of 34.0049 producing fragment 756.5561**	**no hits**	**PC(16:0/18:2)**
PC(16:0/4:0(CHO))	624.3518	255.2332	**-CHO**	**C4:0**	**NL of 59.3147 producing fragment 565.3147**	**PC(16:0/4:0(CHO)**	**NA**
			-COOH [M+HCOO]-	no hits			
			-COOH [M-H]-	no hits			
PC(16:0/6:0(CHO))	652.3831	255.2331	-CHO	C6:0	NL of 59.3147 producing fragment 593.3460	PC(16:0/6:0(CHO))	NA
			-COOH [M+HCOO]-	no hits			
			-COOH [M-H]-	no hits			
PC(16:0/7:0(CHO))	666.3988	255.2331	**-CHO**	**C7:0**	**NL of 59.3147 producing fragment 607.3617**	**PC(16:0/7:0(CHO))**	**NA**

			-COOH [M+HCOO]-	no hits			
			-COOH [M-H]-	no hits			
PC(16:0/4:0(COOH))	594.3412	255.2333	-CHO	no hits			
			-COOH [M+HCOO]-	no hits			
			-COOH [M-H]-	**C4:0**	**NL of 59.3147 producing fragment 535.3041**	**PC(16:0/4:0(COOH))**	**NA**
PC(16:0/6:0(COOH))	668.3778	255.2332	-CHO	no hits			
			-COOH [M+HCOO]-	**C6:0**	**NL of 59.3147 producing fragment 609.3407**	**PC(16:0/6:0(COOH))**	**NA**
			-COOH [M-H]-	no hits			
PC(16:0/7:0(COOH))	636.3882	255.2332	-CHO	no hits			
			-COOH [M+HCOO]-	no hits			
			-COOH [M-H]-	**C7:0**	**NL of 59.3147 producing fragment 577.3511**	**PC(16:0/7:0(COOH))**	**NA**
PC(16:0/20:4)	826.5603	255.2333 303.2335	=O	no hits			
			-OH	no hits			
			-OH-OH	no hits			
			-OOH	no hits			
			-CHO	no hits			
			-COOH [M+HCOO]-	no hits			
			-COOH [M-H]-	21:3	NL of 59.3147 producing	PC(16:0/21:3(COOH))	NA

					fragment 767.5232		
PC(18:1/16:0)	804.5771	281.2492 255.2333	=O	no hits			
			-OH	no hits			
			-OH-OH	no hits			
			-OOH	no hits			
			-CHO	no hits			
			-COOH [M+HCOO]-	no hits			
			-COOH [M-H]-	19:0	NL of 59.3147 producing fragment 745.5400	PC(16:0/19:0(COOH))	NA
PC(20:4/20:0)	882.6232	303.2340 311.2958	=O	no hits			
			-OH	no hits			
			-OH-OH	no hits			
			-OOH	no hits			
			-CHO	no hits			
			-COOH [M+HCOO]-	no hits			
			-COOH [M-H]-	25:3	NL of 59.3147 producing fragment 823.5861	PC(16:0/25:3(COOH))	NA

CHAPTER 4: CEU MASS MEDIATOR 3.0: A METABOLITE ANNOTATION TOOL



Journal of proteome research

Cite This: J. Proteome Res. 2019, 18, 797−802

Technical Note

pubs.acs.org/jpr

CEU Mass Mediator 3.0: A Metabolite Annotation Tool

Alberto Gil-de-la-Fuente,*[†,‡] Joanna Godzien,[‡] Sergio Saugar,[†] Rodrigo Garcia-Carmona,[†] Hasan Badran,[§] David S. Wishart,[§,‖,⊥] Coral Barbas,[‡] and Abraham Otero[†,‡]

[†]Department of Information Technology, Escuela Politécnica Superior, Universidad San Pablo-CEU, CEU Universities, Campus Montepríncipe, Boadilla del Monte, Madrid 28668, Spain

[‡]Centre for Metabolomics and Bioanalysis (CEMBIO), Facultad de Farmacia, Universidad San Pablo-CEU, CEU Universities, Campus Montepríncipe, Boadilla del Monte, Madrid 28668, Spain

[§]Department of Biological Sciences University of Alberta, Edmonton, Alberta T6G 2E9, Canada

[‖]Department of Computing Science, University of Alberta, Edmonton, Alberta T6G 2E8, Canada

[⊥]Faculty of Pharmacy and Pharmaceutical Sciences, University of Alberta, Edmonton, Alberta T6G 2N8, Canada

Supporting Information

ABSTRACT: CEU Mass Mediator (CMM, http://ceumass.eps.uspceu.es) is an online tool that has evolved from a simple interface to query different metabolomic databases (CMM 1.0) to a tool that unifies the compounds from these databases and, using an expert system with knowledge about the experimental setup and the compounds properties, filters and scores the query results (CMM 2.0). Since this last major revision, CMM has continued to grow, expanding the knowledge base of its expert system and including new services to support researchers in the metabolite annotation and identification process. The information from external databases has been refreshed, and an in-house library with oxidized lipids not present in other sources has been added. This has increased the number of experimental metabolites up 332,665 and the number of predicted metabolites to 681,198. Furthermore, new taxonomy and ontology metadata have been included. CMM has expanded its functionalities with a service for the annotation of oxidized glycerophosphocholines, a service for spectral comparison from MS^2 data, and a spectral quality-assessment service to determine the reliability of a spectrum for compound identification purposes. To facilitate the collaboration and integration of CMM with external tools and metabolomic platforms, a RESTful API has been created, and it has already been integrated into the HMDB (Human Metabolome Database). This paper will present the novel functionalities incorporated into version 3.0 of CMM.

KEYWORDS: *metabolomics, annotation, identification, knowledge representation, mass spectrometry, databases, REST, web services, software tool*

INTRODUCTION

Compound annotation and identification remains one of the major bottlenecks in untargeted metabolomics.[1−3] Mass spectrometry (MS) is the dominant platform in metabolomics and lipidomics[4] due to its high sensitivity.[5] MS is commonly coupled to different separation techniques. Among them, high-performance liquid chromatography (HPLC) is the most frequently used.[4]

LC−MS can be used in hyphenated setups (LC−MS/MS) for obtaining the fragmentation pattern of the analyzed compounds or samples. MS/MS or MS^n analyses provide structural information based on the fragmentation pattern. This fragmentation pattern can be compared against reference MS/MS spectra present in metabolomic databases, providing evidence pointing to different possible annotations.

In parallel with the development of metabolomics, there has been an increase in the number and a growth in the size of metabolomic databases.[6−15] Although some compounds are present in most databases, there is not a complete overlap among them, forcing the researcher to use different databases and then manually unifying the results. This task is highly time-consuming and prone to errors.

Most of the software tools for compound annotation are based on MS/MS or MS^n information, which can provide a higher level of confidence for the annotations. However, in some experiments, the sample quantity is limited; therefore, the MS/MS analyses cannot be performed. Furthermore, in some cases, a previous filter using only MS^1 is useful, or even necessary. The time spent on compound annotation/identification can be decreased by filtering the putative annotations before using the fragmentation information with RT and m/z data.

Special Issue: Software Tools and Resources 2019

Received: September 14, 2018
Published: December 21, 2018

DOI: 10.1021/acs.jproteome.8b00720
J. Proteome Res. 2019, 18, 797−802

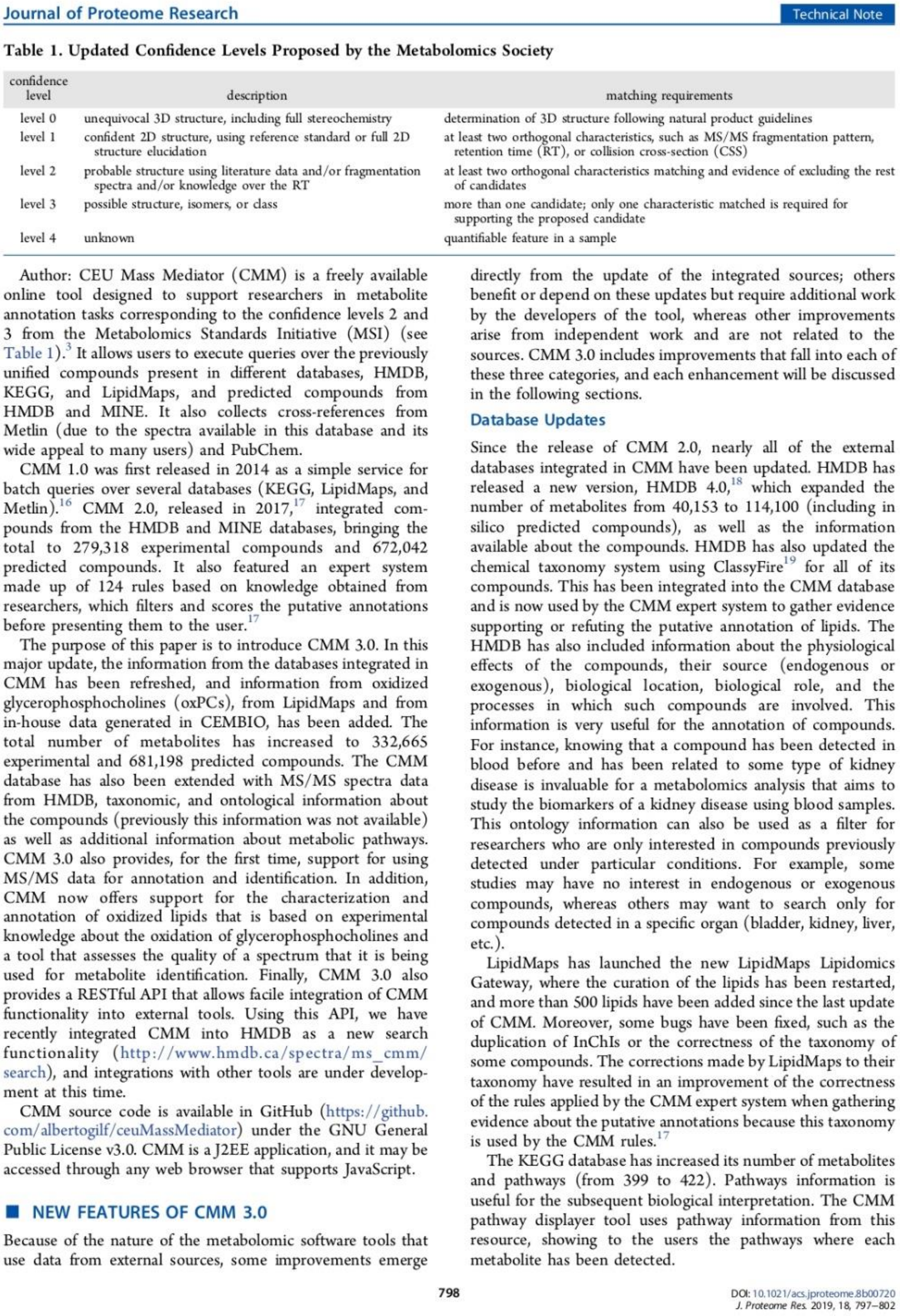


Table 1. Updated Confidence Levels Proposed by the Metabolomics Society

confidence level	description	matching requirements
level 0	unequivocal 3D structure, including full stereochemistry	determination of 3D structure following natural product guidelines
level 1	confident 2D structure, using reference standard or full 2D structure elucidation	at least two orthogonal characteristics, such as MS/MS fragmentation pattern, retention time (RT), or collision cross-section (CSS)
level 2	probable structure using literature data and/or fragmentation spectra and/or knowledge over the RT	at least two orthogonal characteristics matching and evidence of excluding the rest of candidates
level 3	possible structure, isomers, or class	more than one candidate; only one characteristic matched is required for supporting the proposed candidate
level 4	unknown	quantifiable feature in a sample

Author: CEU Mass Mediator (CMM) is a freely available online tool designed to support researchers in metabolite annotation tasks corresponding to the confidence levels 2 and 3 from the Metabolomics Standards Initiative (MSI) (see Table 1).[3] It allows users to execute queries over the previously unified compounds present in different databases, HMDB, KEGG, and LipidMaps, and predicted compounds from HMDB and MINE. It also collects cross-references from Metlin (due to the spectra available in this database and its wide appeal to many users) and PubChem.

CMM 1.0 was first released in 2014 as a simple service for batch queries over several databases (KEGG, LipidMaps, and Metlin).[16] CMM 2.0, released in 2017,[17] integrated compounds from the HMDB and MINE databases, bringing the total to 279,318 experimental compounds and 672,042 predicted compounds. It also featured an expert system made up of 124 rules based on knowledge obtained from researchers, which filters and scores the putative annotations before presenting them to the user.[17]

The purpose of this paper is to introduce CMM 3.0. In this major update, the information from the databases integrated in CMM has been refreshed, and information from oxidized glycerophosphocholines (oxPCs), from LipidMaps and from in-house data generated in CEMBIO, has been added. The total number of metabolites has increased to 332,665 experimental and 681,198 predicted compounds. The CMM database has also been extended with MS/MS spectra data from HMDB, taxonomic, and ontological information about the compounds (previously this information was not available) as well as additional information about metabolic pathways. CMM 3.0 also provides, for the first time, support for using MS/MS data for annotation and identification. In addition, CMM now offers support for the characterization and annotation of oxidized lipids that is based on experimental knowledge about the oxidation of glycerophosphocholines and a tool that assesses the quality of a spectrum that it is being used for metabolite identification. Finally, CMM 3.0 also provides a RESTful API that allows facile integration of CMM functionality into external tools. Using this API, we have recently integrated CMM into HMDB as a new search functionality (http://www.hmdb.ca/spectra/ms_cmm/search), and integrations with other tools are under development at this time.

CMM source code is available in GitHub (https://github.com/albertogilf/ceuMassMediator) under the GNU General Public License v3.0. CMM is a J2EE application, and it may be accessed through any web browser that supports JavaScript.

NEW FEATURES OF CMM 3.0

Because of the nature of the metabolomic software tools that use data from external sources, some improvements emerge directly from the update of the integrated sources; others benefit or depend on these updates but require additional work by the developers of the tool, whereas other improvements arise from independent work and are not related to the sources. CMM 3.0 includes improvements that fall into each of these three categories, and each enhancement will be discussed in the following sections.

Database Updates

Since the release of CMM 2.0, nearly all of the external databases integrated in CMM have been updated. HMDB has released a new version, HMDB 4.0,[18] which expanded the number of metabolites from 40,153 to 114,100 (including in silico predicted compounds), as well as the information available about the compounds. HMDB has also updated the chemical taxonomy system using ClassyFire[19] for all of its compounds. This has been integrated into the CMM database and is now used by the CMM expert system to gather evidence supporting or refuting the putative annotation of lipids. The HMDB has also included information about the physiological effects of the compounds, their source (endogenous or exogenous), biological location, biological role, and the processes in which such compounds are involved. This information is very useful for the annotation of compounds. For instance, knowing that a compound has been detected in blood before and has been related to some type of kidney disease is invaluable for a metabolomics analysis that aims to study the biomarkers of a kidney disease using blood samples. This ontology information can also be used as a filter for researchers who are only interested in compounds previously detected under particular conditions. For example, some studies may have no interest in endogenous or exogenous compounds, whereas others may want to search only for compounds detected in a specific organ (bladder, kidney, liver, etc.).

LipidMaps has launched the new LipidMaps Lipidomics Gateway, where the curation of the lipids has been restarted, and more than 500 lipids have been added since the last update of CMM. Moreover, some bugs have been fixed, such as the duplication of InChIs or the correctness of the taxonomy of some compounds. The corrections made by LipidMaps to their taxonomy have resulted in an improvement of the correctness of the rules applied by the CMM expert system when gathering evidence about the putative annotations because this taxonomy is used by the CMM rules.[17]

The KEGG database has increased its number of metabolites and pathways (from 399 to 422). Pathways information is useful for the subsequent biological interpretation. The CMM pathway displayer tool uses pathway information from this resource, showing to the users the pathways where each metabolite has been detected.

CMM has been supplemented with information from oxPCs from an in-house library, expanding the current information present in the other databases. There are 248 oxidized glycerophospholipids (oxGPs) currently in LipidMaps and 48 oxPCs. The CMM library contains 24 new oxPCs that are not present in any of the databases integrated.[20] Table 2 illustrates the information present in CMM from each source.

Table 2. Number of Compounds and Type of Information Available in the Different Versions of CMM

database	CMM 1.0	CMM 2.0	CMM 3.0
HMDB	0	74,484	114,065
	N/A	structure	structure, taxonomy, pathways, ontology
KEGG	13,526	15,909	18,293
	N/A	structure, pathways	structure, pathways
LipidMaps	37,576	40,213	42,555
	N/A	structure, taxonomy	structure, taxonomy (corrected)
MINE	0	672,042	672,042
	N/A	structure	structure
in-house	0	0	24
	N/A	N/A	structure, taxonomy

Figure 1 shows the overlap of the metabolite coverage between the different databases integrated in CMM 2.0 and CMM 3.0. The comparison only covers those compounds for which there is sufficient information available to perform unequivocal compound unification (i.e., those with 3D structure, the InChI, the InChI key, or the canonical SMILES available). The reason why there seems to be a reduction (from 1239 to 1175 metabolites) in the overlap between LipidMaps and KEGG in CMM 3.0 (despite the addition of more compounds to both databases) is the previously discussed errata present in some LipidMaps compounds, which sometimes caused the incorrect unifications in CMM 2.0. Although the global number of metabolites increased, the overlap of metabolites among the databases is still low.

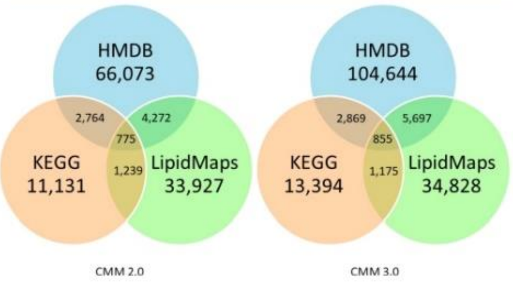


Figure 1. Venn diagram with the coverage of metabolites between different databases integrated in CMM 2.0 and CMM 3.0.

Annotation of Oxidized Lipids

The biological role of oxidized lipids is currently an active research topic, notably contributing to the understanding of health and disease. Within human metabolomic studies, it was observed that lipids are significantly affected by oxidative stress.[21,22] However, until now, the number of tools to support the annotation of oxidized lipids has been small.[23] Nowadays, this task usually starts with the annotation of the signals by searching for experimental m/z matches in the databases. However, the number of oxidized lipids currently present in such databases is low. LipidMaps, the reference database for lipids, contains only 248 oxidized lipids for all types of heads. This makes the annotation of oxidized phospholipids challenging, especially when the target compound is not present in the database, and it increases the likelihood of the compound being assigned as an unknown. There are some patterns that can provide the researcher with clues about the possibility that a feature may correspond to an oxidized lipid, for example, a lower RT for reversed-phase (RP) chromatography of the shortened chains due to the lower hydrophobicity, a high level of fragmentation with 40 eV, the ionization through deprotonation ($[M - H]^-$) and the presence of the neutral loss of water (usually detectable in positive ionization mode).[20] CMM has developed a service to annotate and characterize oxidized lipids. The full process is explained in SI1.

MS/MS Search

There is a large number of MS/MS-based annotation tools using different approaches to spectral matching or compound identification. Until now, CMM had only supported the annotation of MS^1 data. CMM 3.0 has integrated the MS/MS information present in HMDB, including experimental and predicted spectra, with the aim of providing support to compare experimental MS/MS spectra against the HMDB reference data. The large number of existing algorithms to calculate spectral similarity and their robustness leads us to select the three most popular promising ones and to perform an independent evaluation instead of proposing a new solution.[8,10,24−27] The results of the evaluation are shown in SI2.

To use the MS/MS search service of CMM, the researcher should introduce the precursor ion mass, the list of pairs of m/z and intensity of the product ions, and the tolerance allowed for the precursor ion and the product ions (in Da or ppm). The intensity can be normalized or absolute (CMM normalizes the values if needed). The ionization mode used and voltage applied are also necessary to restrict the search over the corresponding experimental setup. The researcher can choose if the MS/MS spectra comparison should be performed against experimental or predicted spectra. Once all of the information is submitted, the CMM comparison algorithm performs an initial filter of the putative annotations based on the precursor ion, the precursor ion tolerance, the ionization mode, the fragmentation voltage and the type of spectra. The compounds with spectra available under these conditions are then scored to determine the similarity between the input spectra and the putative annotations.

Spectral Quality Assessment for Identification Purposes

The success of an untargeted metabolomics study depends on the correctness of the identification process. Decreasing the number of unknowns and misidentifications is key to having a biologically significant finding. Nevertheless, the time to perform a study is restricted, and the low availability of reference standards for clear compound identification often hinders this task in untargeted approaches. Consequently, a high-quality MS/MS spectrum is paramount to improve the annotation rate of compounds, whereas a low-quality spectrum increases the risk of misidentification.

However, assessing spectral quality can be difficult. To provide researchers with a systematic method to evaluate the quality of a MS/MS spectrum, CMM has created a pentagonal-point evaluation system that takes into account: (1) the quality of the overall intensity, (2) the impact of the noise, (3) the number of MS/MS scans obtained, (4) the presence of different precursor ions in the collision cell at the same time, and (5) the presence of delayed ions from the previous scans, a phenomena known as cross-talk. In Figure 2, we can see a graphic representation of this evaluation system. The pentagon on the left shows the evaluation result for an excellent spectrum (green lines), whereas the pentagon in the right corresponds to an inadequate spectrum (red lines). The closer the lines are to the pentagon vertex, the better the spectrum is. A full explanation about the principles used and how the spectral quality assessment has been developed can be found in SI3.[28]

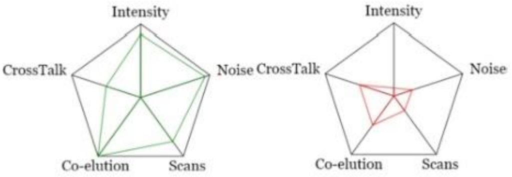


Figure 2. Graphic representation of the pentagonal-point evaluation system used in CMM 3.0 to asses the quality of the spectrum introduced by the user. On the left, a good quality spectrum. On the right, a poor quality spectrum.

RESTful API

The metabolomics field is continuously growing and so is the number of tools available to assist in metabolomic data analysis. Typically, metabolomic tools do not provide all of the functionalities that a metabolomics workflow requires. Therefore, researchers often have to use different tools to carry out their analyses and are forced to use results from one tool in another, a task that is not always trivial. To mitigate this disadvantage, several platforms that try to integrate different external tools into a single pipeline have emerged. Two of the most popular open platforms are the Workflow4Metabolomics[29] and PhenoMeNal.[30] The Elixir metabolomics community aims to share the data and the software tools with these frameworks because they have proven to be useful, and they can improve the reproducibility of the data analysis.[31]

CMM provides functionality that is not available in other metabolomics tools: its knowledge-based approach to filtering and scoring putative annotations as well as its support for the identification of oxidized lipids using experimental knowledge. The compound unification from multiple external databases also provides great value. Because of these unique features, there has been a growing number of requests to integrate CMM features in external tools through an application programming interface (API). This RESTful (REpresentational State Transfer architectural style compliant) API for CMM has been already integrated into the HMDB environment (http://www.hmdb.ca/spectra/ms_cmm/search), where users can take advantage of the CMM filtering and scoring functionality directly from the HMDB web interface. The details of this API can be found in SI4.

■ CONCLUSIONS

We have presented version 3.0 of CMM, a free online tool to support many of the needs of researchers in the annotation of metabolites. CMM integrates and unifies experimental and predicted metabolites from several databases, including HMDB, KEGG, LipidMaps, and MINE, allowing the user to query in all of them through a single interface. In addition, CMM uses an expert system to filter and score putative annotations, allowing researchers to focus on those annotations that are more plausible.

After the data refreshing performed in version 3.0 and the integration of an in-house library of oxidized lipids, the total number of available experimental metabolites in CMM 3.0 is 332,665 and the total number of predicted metabolites is 681,198. Taxonomy and ontology information from HMDB and LipidMaps has been added to the CMM database, being used by its expert system. Novel functionalities that have been added to CMM 3.0 include MS/MS search support, a service for the annotation of oxidized glycerophospholipids and a spectral quality-assessment tool to measure the quality of the MS/MS spectra. Furthermore, in CMM 3.0, the search services are now available through a RESTful API that has already been used to integrate CMM functionalities into the HMDB. For future work, we intend to further exploit the taxonomy and ontology information that is available now in the CMM database to enhance the filtering and scoring performed by our expert system. Some of this information is already used by the 124 rules currently available in the expert system.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.8b00720.

> SI1: Annotation of oxidized lipids. SI2: Independent evaluation of MS/MS search. SI3: Spectral quality assessment development. SI4: CMM RESTful API (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: alb.gil.ce@ceindo.ceu.es. Tel: +34 913724711.

ORCID

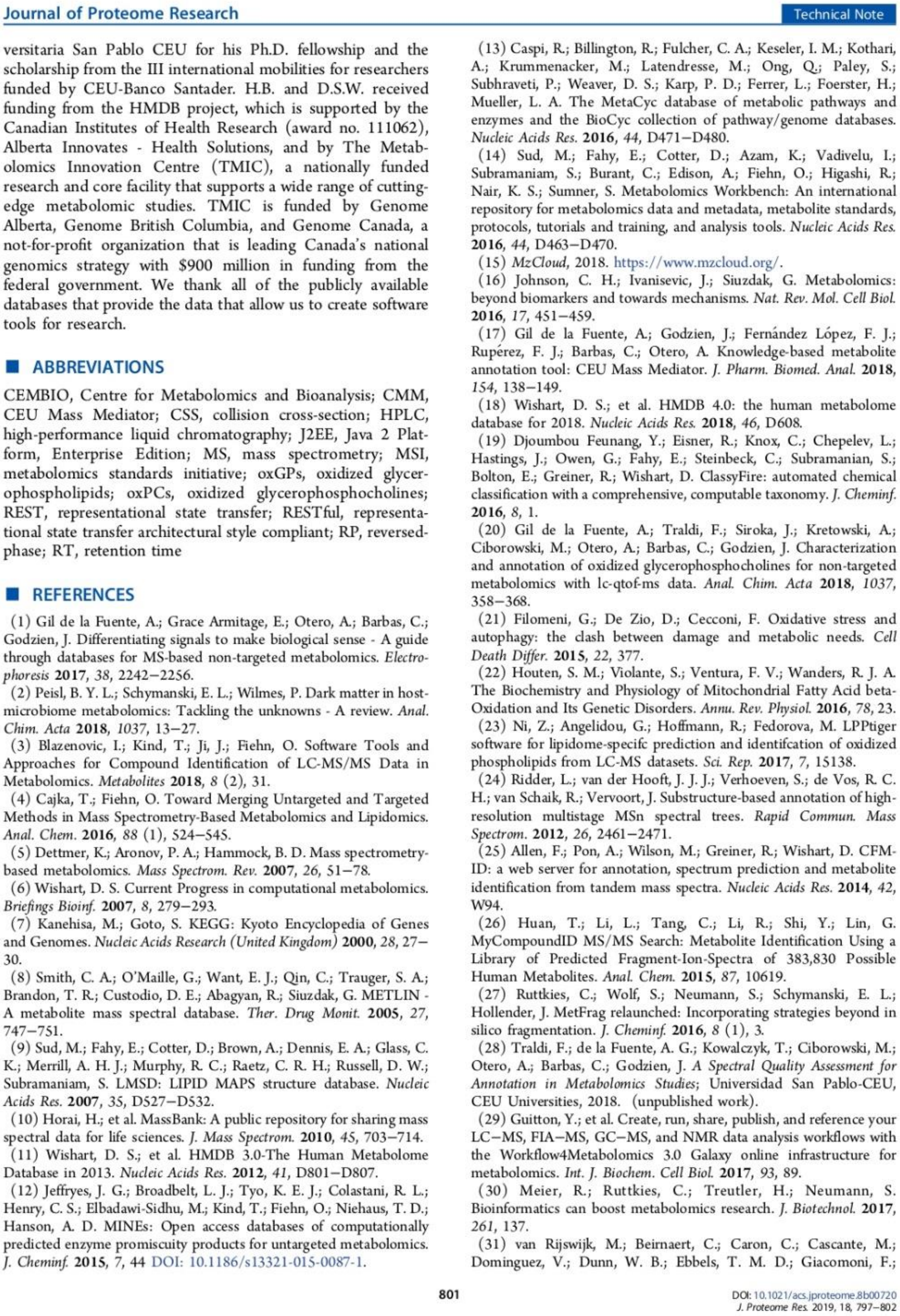
Alberto Gil-de-la-Fuente: 0000-0002-5951-1601
Joanna Godzien: 0000-0002-9477-057X
Sergio Saugar: 0000-0002-4216-9256
Rodrigo Garcia-Carmona: 0000-0003-4427-9579
Hasan Badran: 0000-0002-8072-5233
David S. Wishart: 0000-0002-3207-2434
Coral Barbas: 0000-0003-4722-491X
Abraham Otero: 0000-0003-4568-2933

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

A.G.-F., J.G., C.B., and A.O. received funding from the Spanish Ministerio de Economía y Competitividad (grant CTQ2014-55279-R). A.G.-F., C.B., and A.O. received funding from USPCEU PCON10/2016. A.G.-F. thanks Fundación Uni-



versitaria San Pablo CEU for his Ph.D. fellowship and the scholarship from the III international mobilities for researchers funded by CEU-Banco Santader. H.B. and D.S.W. received funding from the HMDB project, which is supported by the Canadian Institutes of Health Research (award no. 111062), Alberta Innovates - Health Solutions, and by The Metabolomics Innovation Centre (TMIC), a nationally funded research and core facility that supports a wide range of cutting-edge metabolomic studies. TMIC is funded by Genome Alberta, Genome British Columbia, and Genome Canada, a not-for-profit organization that is leading Canada's national genomics strategy with $900 million in funding from the federal government. We thank all of the publicly available databases that provide the data that allow us to create software tools for research.

ABBREVIATIONS

CEMBIO, Centre for Metabolomics and Bioanalysis; CMM, CEU Mass Mediator; CSS, collision cross-section; HPLC, high-performance liquid chromatography; J2EE, Java 2 Platform, Enterprise Edition; MS, mass spectrometry; MSI, metabolomics standards initiative; oxGPs, oxidized glycerophospholipids; oxPCs, oxidized glycerophosphocholines; REST, representational state transfer; RESTful, representational state transfer architectural style compliant; RP, reversed-phase; RT, retention time

REFERENCES

(1) Gil de la Fuente, A.; Grace Armitage, E.; Otero, A.; Barbas, C.; Godzien, J. Differentiating signals to make biological sense - A guide through databases for MS-based non-targeted metabolomics. *Electrophoresis* **2017**, *38*, 2242−2256.

(2) Peisl, B. Y. L.; Schymanski, E. L.; Wilmes, P. Dark matter in host-microbiome metabolomics: Tackling the unknowns - A review. *Anal. Chim. Acta* **2018**, *1037*, 13−27.

(3) Blazenovic, I.; Kind, T.; Ji, J.; Fiehn, O. Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics. *Metabolites* **2018**, *8* (2), 31.

(4) Cajka, T.; Fiehn, O. Toward Merging Untargeted and Targeted Methods in Mass Spectrometry-Based Metabolomics and Lipidomics. *Anal. Chem.* **2016**, *88* (1), 524−545.

(5) Dettmer, K.; Aronov, P. A.; Hammock, B. D. Mass spectrometry-based metabolomics. *Mass Spectrom. Rev.* **2007**, *26*, 51−78.

(6) Wishart, D. S. Current Progress in computational metabolomics. *Briefings Bioinf.* **2007**, *8*, 279−293.

(7) Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research (United Kingdom)* **2000**, *28*, 27−30.

(8) Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G. METLIN - A metabolite mass spectral database. *Ther. Drug Monit.* **2005**, *27*, 747−751.

(9) Sud, M.; Fahy, E.; Cotter, D.; Brown, A.; Dennis, E. A.; Glass, C. K.; Merrill, A. H. J.; Murphy, R. C.; Raetz, C. R. H.; Russell, D. W.; Subramaniam, S. LMSD: LIPID MAPS structure database. *Nucleic Acids Res.* **2007**, *35*, D527−D532.

(10) Horai, H.; et al. MassBank: A public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **2010**, *45*, 703−714.

(11) Wishart, D. S.; et al. HMDB 3.0-The Human Metabolome Database in 2013. *Nucleic Acids Res.* **2012**, *41*, D801−D807.

(12) Jeffryes, J. G.; Broadbelt, L. J.; Tyo, K. E. J.; Colastani, R. L.; Henry, C. S.; Elbadawi-Sidhu, M.; Kind, T.; Fiehn, O.; Niehaus, T. D.; Hanson, A. D. MINEs: Open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. *J. Cheminf.* **2015**, *7*, 44 DOI: 10.1186/s13321-015-0087-1.

(13) Caspi, R.; Billington, R.; Fulcher, C. A.; Keseler, I. M.; Kothari, A.; Krummenacker, M.; Latendresse, M.; Ong, Q.; Paley, S.; Subhraveti, P.; Weaver, D. S.; Karp, P. D.; Ferrer, L.; Foerster, H.; Mueller, L. A. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **2016**, *44*, D471−D480.

(14) Sud, M.; Fahy, E.; Cotter, D.; Azam, K.; Vadivelu, I.; Subramaniam, S.; Burant, C.; Edison, A.; Fiehn, O.; Higashi, R.; Nair, K. S.; Sumner, S. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* **2016**, *44*, D463−D470.

(15) *MzCloud*, 2018. https://www.mzcloud.org/.

(16) Johnson, C. H.; Ivanisevic, J.; Siuzdak, G. Metabolomics: beyond biomarkers and towards mechanisms. *Nat. Rev. Mol. Cell Biol.* **2016**, *17*, 451−459.

(17) Gil de la Fuente, A.; Godzien, J.; Fernández López, F. J.; Rupérez, F. J.; Barbas, C.; Otero, A. Knowledge-based metabolite annotation tool: CEU Mass Mediator. *J. Pharm. Biomed. Anal.* **2018**, *154*, 138−149.

(18) Wishart, D. S.; et al. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **2018**, *46*, D608.

(19) Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Greiner, R.; Wishart, D. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminf.* **2016**, *8*, 1.

(20) Gil de la Fuente, A.; Traldi, F.; Siroka, J.; Kretowski, A.; Ciborowski, M.; Otero, A.; Barbas, C.; Godzien, J. Characterization and annotation of oxidized glycerophosphocholines for non-targeted metabolomics with lc-qtof-ms data. *Anal. Chim. Acta* **2018**, *1037*, 358−368.

(21) Filomeni, G.; De Zio, D.; Cecconi, F. Oxidative stress and autophagy: the clash between damage and metabolic needs. *Cell Death Differ.* **2015**, *22*, 377.

(22) Houten, S. M.; Violante, S.; Ventura, F. V.; Wanders, R. J. A. The Biochemistry and Physiology of Mitochondrial Fatty Acid beta-Oxidation and Its Genetic Disorders. *Annu. Rev. Physiol.* **2016**, *78*, 23.

(23) Ni, Z.; Angelidou, G.; Hoffmann, R.; Fedorova, M. LPPtiger software for lipidome-specifc prediction and identifcation of oxidized phospholipids from LC-MS datasets. *Sci. Rep.* **2017**, *7*, 15138.

(24) Ridder, L.; van der Hooft, J. J. J.; Verhoeven, S.; de Vos, R. C. H.; van Schaik, R.; Vervoort, J. Substructure-based annotation of high-resolution multistage MSn spectral trees. *Rapid Commun. Mass Spectrom.* **2012**, *26*, 2461−2471.

(25) Allen, F.; Pon, A.; Wilson, M.; Greiner, R.; Wishart, D. CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Res.* **2014**, *42*, W94.

(26) Huan, T.; Li, L.; Tang, C.; Li, R.; Shi, Y.; Lin, G. MyCompoundID MS/MS Search: Metabolite Identification Using a Library of Predicted Fragment-Ion-Spectra of 383,830 Possible Human Metabolites. *Anal. Chem.* **2015**, *87*, 10619.

(27) Ruttkies, C.; Wolf, S.; Neumann, S.; Schymanski, E. L.; Hollender, J. MetFrag relaunched: Incorporating strategies beyond in silico fragmentation. *J. Cheminf.* **2016**, *8* (1), 3.

(28) Traldi, F.; de la Fuente, A. G.; Kowalczyk, T.; Ciborowski, M.; Otero, A.; Barbas, C.; Godzien, J. *A Spectral Quality Assessment for Annotation in Metabolomics Studies*; Universidad San Pablo-CEU, CEU Universities, 2018. (unpublished work).

(29) Guitton, Y.; et al. Create, run, share, publish, and reference your LC−MS, FIA−MS, GC−MS, and NMR data analysis workflows with the Workflow4Metabolomics 3.0 Galaxy online infrastructure for metabolomics. *Int. J. Biochem. Cell Biol.* **2017**, *93*, 89.

(30) Meier, R.; Ruttkies, C.; Treutler, H.; Neumann, S. Bioinformatics can boost metabolomics research. *J. Biotechnol.* **2017**, *261*, 137.

(31) van Rijswijk, M.; Beirnaert, C.; Caron, C.; Cascante, M.; Dominguez, V.; Dunn, W. B.; Ebbels, T. M. D.; Giacomoni, F.;

Gonzalez-Beltran, A.; Hankemeier, T.; et al. The future of metabolomics in ELIXIR. *F1000Research* **2017**, *6*, 1649 DOI: 10.12688/f1000research.12342.1.

Supporting Information. CEU Mass Mediator 3.0: A Metabolite Annotation Tool

Alberto Gil-de-la-Fuente,[*,†,‡] Joanna Godzien,[‡] Sergio Saugar,[†] Rodrigo Garcia-Carmona,[†] Hasan Badran,[¶] David S. Wishart,[¶,§,‖] Coral Barbas,[‡] and Abraham Otero[†,‡]

†*Department of Information Technology, Escuela Politécnica Superior, Universidad San Pablo-CEU, CEU Universities, Campus Montepríncipe, Boadilla del Monte, Madrid, 28668, Spain*

‡*Centre for Metabolomics and Bioanalysis (CEMBIO), Facultad de Farmacia, Universidad San Pablo-CEU, CEU Universities, Campus Montepríncipe, Boadilla del Monte, Madrid, 28668, Spain*

¶*Department of Biological Sciences University of Alberta, Edmonton, Alberta T6G 2E9, Canada*

§*Department of Computing Science, University of Alberta, Edmonton, Alberta T6G 2E8, Canada*

‖*Faculty of Pharmacy and Pharmaceutical Sciences, University of Alberta, Edmonton, Alberta T6G 2N8, Canada*

E-mail: alb.gil.ce@ceindo.ceu.es

Phone: +34 913724711

Table of contents

SI1: Annotation of oxidized lipids

Three different regions can be recognized in the MS/MS spectrum in the negative ionization mode of a lipid (see Figure S-1). The high-mass (yellow) region corresponds to the precursor ion and the neutral losses. The m/z 802.56128 corresponds to the precursor ion PC(16:0/18:2), and the m/z 742.54089 to the precursor ion with a neutral loss of $C_2H_4O_2$ (60.02 Da). The mid-mass (blue) region corresponds to the fatty acids (FAs) if they have not suffered an oxidation that shortened the chain. In Figure S-1 the two FAs 16:0 and 18:2 are represented by the m/z 255.23630 and 279.23679, respectively, with corresponding lyso-forms, either de-methylated or de-methylated and de-hydrated. Finally, the green region corresponds to the product ions and to the shortened oxidized chains (if present). There are no shortened oxidized chains in the Figure S-1, since oxidation has not occurred, and there are product ions from the head group (PC). The m/z 224.07240 is formed by $C_7H_{15}NPO_5$ and m/z 168.04520 from $C_4H_{11}NPO_4$, both being characteristic product ions for PCs. The researcher can hypothesize or perform a structural elucidation for the annotation by assigning the experimental masses to new structures formed by the transformation of the precursor ion. Subsequently they could assign the peaks to the structures based on the m/z and the biological transformations, but this task is difficult if the annotation of the precursor ion is not correct. To mitigate this disadvantage, CMM includes a service to support the annotation of oxidized glycerophosphocholines (oxPCs).

CMM identifies the oxidized and the non-oxidized FAs for long chain oxPCs by matching against a FA database containing all the FAs from 3:0 to 36:6 and their possible oxidations products. The mass of these FAs and their alterations has been generated algorithmically. If this search returns matches, then CMM searches for oxPCs in its database that correspond to the previously identified FAs. For short chain oxPCs, the short FA is not present in the mid-mass region; therefore it is difficult to identify which product ion from the low-mass region corresponds to the short chain oxidized FA. However, the mass of this product ion can be calculated by subtracting the non-oxidized FA from the precursor ion. If the

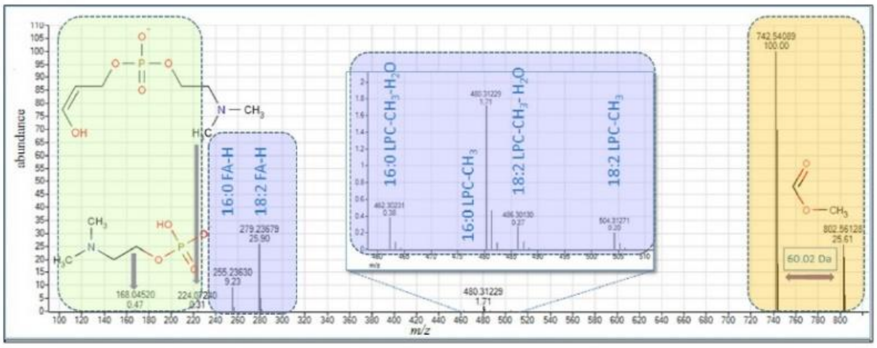


Figure S-1: Peaks regions for MS/MS analysis of oxPCs. High-mass (yellow) region: precursor ion and neutral losses; mid-mass (blue) region: non-shortened FAs; low-mass (green) region: product ions and shortened FAs (if present).

mass calculated corresponds to an oxidized FA present in the algorithmically generated database, the researcher can check if the product ion is present in the MS/MS spectrum and, subsequently CMM performs a query to see if there is any hit in the databases. Nevertheless, CMM computes the putative annotation regardless there is any hit in the databases with the aim of increasing the coverage of the oxPCs.

This service includes knowledge about the fragmentation pattern and a list of 24 oxidized lipids from an in-house library, some of which are not present in LipidMaps. The flowchart in Figure S-2 shows the annotation of long chain and short chain oxPCs. For the long chain oxidation, the algorithm receives the m/z of the two FAs (FA1 and FA2) and the precursor ion. It detects and annotates the oxidized FA by matching against a database of FAs, and then annotates the other FA, the non-oxidized one. Once the two FAs are annotated, it creates a tentative annotation (see step 4) and checks to see if there is any hit in the databases. The annotation is always returned, no matter if the oxPC is present in the databases or not, to overcome the limited number of oxPCs present in the databases. For the short chain oxidation, only the m/z of the non-oxidized FA is visible in the mid-mass region, but based on the m/z of the precursor ion, the mass of the PC head group and the non-oxidized FA, it can calculate the m/z of the oxidized FA. Once this is done, the user can look through the entire MS/MS spectrum to see if this mass is present, and process the annotation in an analogous way to the annotation of the long chain oxPCs.

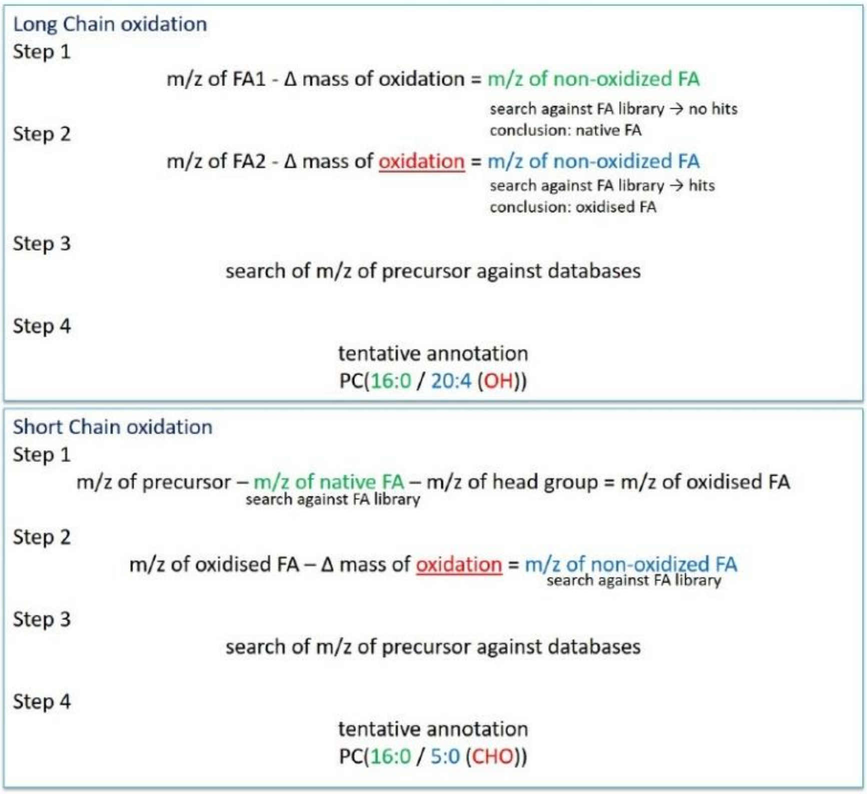

Long Chain oxidation
Step 1
m/z of FA1 - Δ mass of oxidation = m/z of non-oxidized FA
search against FA library → no hits
conclusion: native FA
Step 2
m/z of FA2 - Δ mass of oxidation = m/z of non-oxidized FA
search against FA library → hits
conclusion: oxidised FA
Step 3
search of m/z of precursor against databases
Step 4
tentative annotation
PC(16:0 / 20:4 (OH))
Short Chain oxidation
Step 1
m/z of precursor – m/z of native FA – m/z of head group = m/z of oxidised FA
search against FA library
Step 2
m/z of oxidised FA – Δ mass of oxidation = m/z of non-oxidized FA
search against FA library
Step 3
search of m/z of precursor against databases
Step 4
tentative annotation
PC(16:0 / 5:0 (CHO))

Figure S-2: Flowchart for annotation of long and short chaind oxPCs.

SI2: Independent evaluation of MS/MS search

CMM has carried out an independent evaluation of three different: the euclidean distance (see equation S-1), the dot product used previously by MyCompoundID[1] (see equation S-2) and a weighted dot product used previously by Metfrag[2] but penalizing the peaks from the acquired spectra not present in the reference one (see equation S-3).

$$
\begin{gathered}
Euclidean\,distance\,score = \sum_{i=1}^{\substack{number\,of\\matched\,peaks}} \frac{1}{\sqrt{(IP_i - LP_i)^2}} = \\
\sum_{i=1}^{\substack{number\,of\\matched\,peaks}} \frac{1}{\sqrt{(IPMZ_i - LPMZ_i)^2 + (IPIntenstiy_i - LPIntensity_i)^2}}
\end{gathered}
\tag{eqn S-1}
$$

$$
\begin{gathered}
dot\,product\,penalised = \frac{\sum_{i=1}^{\substack{number\,of\\matched\,peaks}} (IP_i * LP_i)}{\sum_{j=1}^{\substack{number\,of\\input\,peaks}} (IP_j * IP_j)} = \\
\frac{\sum_{i=1}^{\substack{number\,of\\matched\,peaks}} (IPMZ_i * LPMZ_i) + (IPIntensity_i * LPIntensity_i)}{\sum_{j=1}^{\substack{number\,of\\input\,peaks}} (IPMZ_j * IPMZ_j) + (IPIntensity_j * IPIntensity_j)}
\end{gathered}
\tag{eqn S-2}
$$

$$
\substack{weighted\,dot\\product\,penalized} = \frac{\sum_{i=1}^{\substack{number\,of\\matched\,peaks}} (IPMZ_i * LPMZ_i) * 3 + (IPIntensity_i * LPIntensity_i)^{0.6}}{\sum_{j=1}^{\substack{number\,of\\input\,peaks}} (IPMZ_j * IPMZ_j)^3 + (IPIntensity_j * IPIntensity_j)^{0.6}}
\tag{eqn S-3}
$$

In the three equations, $IP = experimental\,peak$ and $LP = library\,peak$. The equation S-3 gives more weight to the precursor ion (intensity and m/z) matching than to the product ions (3 *vs* 0.6).

We developed our own implementation of these three algorithms, and they were tested with 30 spectra publicly available in MassBank repository.[3] To compare the results of our implementation with the original implementations of these algorithms, we report here also the results of searching for these 30 compounds with HMDB and MyCompoundId. In these tests we evaluated if the correct annotation appeared in the search results in the top-1, top-5 or among all the putative annotations returned by the tool. The Figure S-3 shows the results for the evaluation. The euclidean distance (see the equation S-1) demonstrates a slightly better performance than the other two algorithms developed, therefore this has been the one chosen for the production version of CMM 3.0.

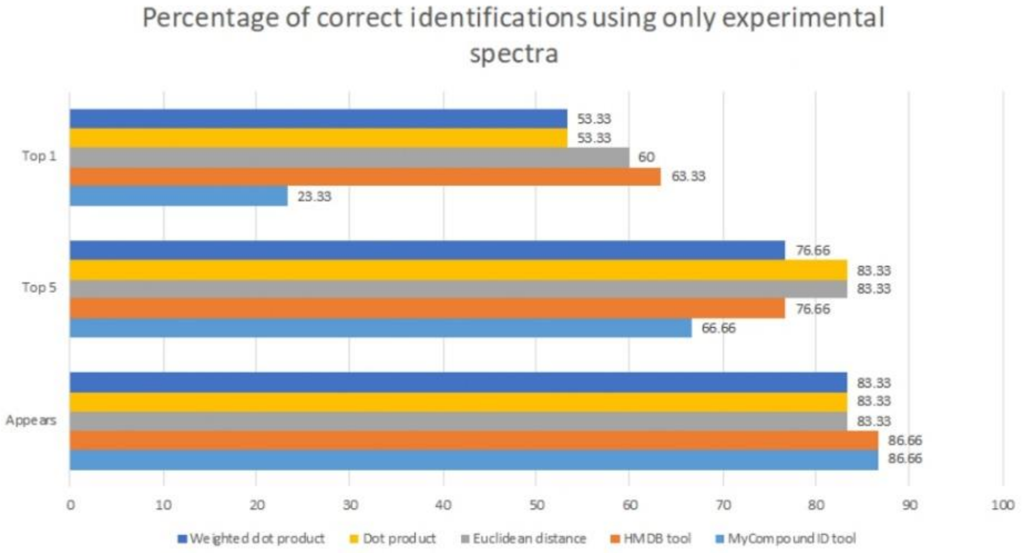


Figure S-3: Percentage of correct annotation from MS/MS search using only experimental spectra.

SI3: Spectral quality assessment development

To provide researchers with a systematic method to evaluate the quality of a MS/MS spectrum CMM has created a pentagonal-point evaluation system that takes into account:

1. The quality of the overall intensity.
2. The impact of the noise.
3. The number of MS/MS scans obtained.
4. The presence of different precursor ions in the collision cell at the same time.
5. The presence of delayed ions from the previous scans, a phenomena known as cross-talk.

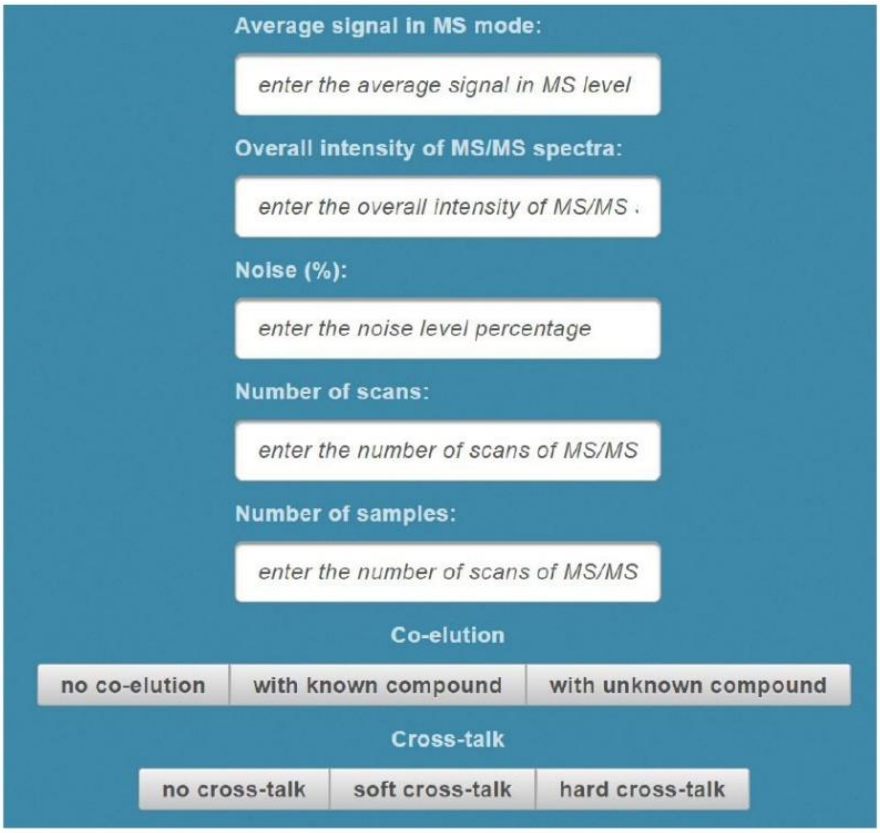
The correct detection of the product ions is key for the identification of compounds via a MS/MS spectrum. If the intensity of the mass fragments is high enough, the product ions can be better identified. To score the overall intensity, CMM takes into consideration the average signal in the MS^1 mode; if the average signal is low, the intensity of the compound needed for a good quality spectrum is lower than in experiments with a higher average signal. The values for the intensity score are shown in the Table S-1. The values for an acceptable spectrum vary linearly from 0 (inadequate intensity) to 1 (optimum intensity) depending on the average intensity in the MS^1 spectrum and the intensity of the compound analyzed in the MS/MS spectrum. The noise influences the product ions detection as well, since a high noise requires a higher intensity for the reliability of the product ion detection, and a low noise permits the detection of product ions with low intensities. The noise is measured as a percentage (from 0 to 100). A noise $\leq 5\%$ has a maximum score of 1, a noise $\geq 20\%$ has a score of 0, and a noise in the range (5-20) has a score which varies linearly between (1-0). If the noise is $\leq 5\%$, the intensity score rises until 0.5 for low intesity spectrum (intensity score ≤ 0.5), since having a low level of noise enables the identification of lower intensity product ions.

Table S-1: Scoring system of the MS/MS intensity depending on the average signal intensity in MS analysis.

Average intensity in MS analysis	Inadequate spectra (0)	Acceptable spectra (0-1)	Excellent spectra (1)
$\leq 10^5$	$\leq 10^2$	10^2 - 10^3	$\geq 10^3$
10^5 - 10^7	$\leq 10^3$	10^3 - 10^4	$\geq 10^4$
10^7 - 10^8	$\leq 10^4$	10^4 - 10^5	$\geq 10^5$
$\geq 10^8$	$\leq 10^5$	10^5 - 10^6	$\geq 10^6$

The third aspect that biases the confidence of the MS/MS annotation is the number of measurements (scans) performed during the MS/MS analysis. A higher number of scans increases the reliability of the measured ions. The score for the number of scans is 0 if there are only 1 or 2 scans, 0.25 if there are 3, 0.5 for 4, 0.75 for 5, and 1 for > 5. However, the researcher can use the concordance between different samples to make the spectrum more reliable if the number of scans is low by analyzing different samples to obtain their MS/MS spectrum and then verifying the fragmentation in two or more samples.

The presence of more than one compound in the collision cell at the same time significantly hampers the assignment of the product ions to the different precursor ions present. Therefore, the presence of more than one compound makes the identification impossible if the compounds and their fragmentation pattern are not previously known. The score of a spectrum when co-elution occurs with an unknown compound is 0, independently of other parameters. The score of the co-elution is 0.5 when co-elution with a known compound that has a known fragmentation pattern occurs, and 1 if co-elution did not happen in the collision cell. The last aspect that the score takes into account is the presence of cross talk. This phenomenon hinders the product ion detection, since it increases the possibility of assigning delayed signals as product ions. If this phenomenon has happened, the researcher can see m/zs higher than the precursor ion. If the intensity of these m/zs is high, the score is 0; if it is low, is 0.5; and if it does not exist, then the score is 1. CMM user interface to enter this information is shown in Figure S-4.



Average signal in MS mode:

enter the average signal in MS level

Overall intensity of MS/MS spectra:

enter the overall intensity of MS/MS .

Noise (%):

enter the noise level percentage

Number of scans:

enter the number of scans of MS/MS

Number of samples:

enter the number of scans of MS/MS

Figure S-4: Web interface of the spectral quality assessment tool.

An overall score is calculated as the sum of each partial score. The spectrum is ranked as excellent if the overall score is ≥ 3.5, acceptable if it is between 2.0 and 3.5, and inadequate if it is below 2.0. The overall score is shown numerically and the color of the lines are red, yellow or green according to the spectral quality (inadequate, acceptable and excellent respectively).

SI4: CMM RESTful API

An API enables two software tools to communicate with each other by establishing a contract (which services can be requested, how to invoke them and what outcome is expected). As CMM is a web application, a web service has been created to expose its functionality (i.e. its API). Currently, there are two different architectural approaches for designing a web service. The traditional approach, commonly known as Web Services, is based on a Service Oriented Architecture and tied to a well defined protocol stack (SOAP, WSDL, UDDI and so on). The other approach, based on the REST (REpresentational State Transfer) architectural style for hypermedia distributed systems (the style underlying the World Wide Web), is called RESTful Web Services. Unlike the traditional approaches that only use the HTTP protocol as a transport layer, RESTful Web Services take advantages of all web related protocols (HTTP, URI, Mime Types) and are easier to use than the traditional ones. Furthermore, the use of ubiquitous technologies like HTTP, JSON or XML supported by most of the programming languages make this approach very suitable for integrating those services in an effortless way.

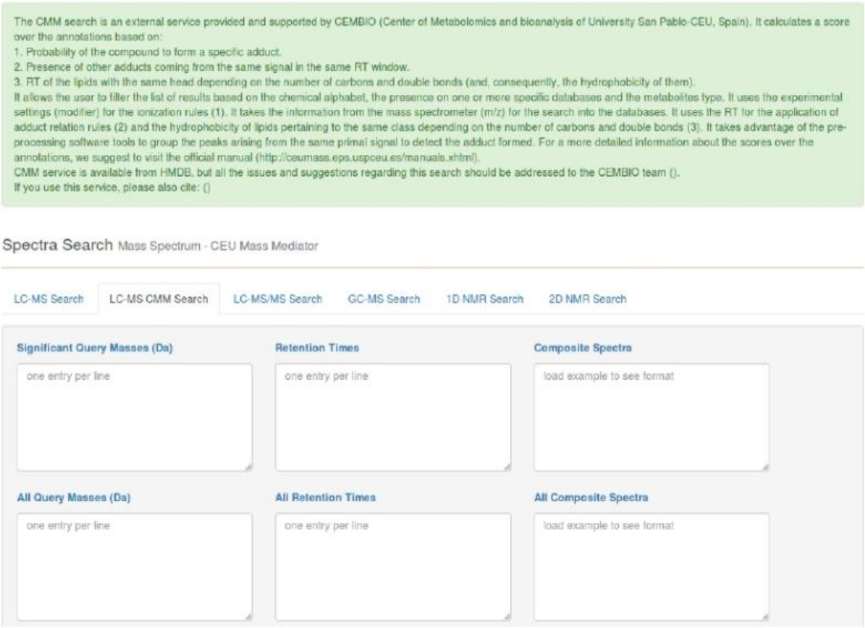
CMM 3.0 features a RESTful API that allows its integration with external tools. This API is structured around two resources providing the batch and advanced batch CMM services. The URIs of these resources follow the nomenclature:
http://ceumass.eps.uspceu.es/api/v3/<serviceName>. The concrete URIs for the provided services are:

1. batch(http://ceumass.eps.uspceu.es/api/v3/batch).
2. advancedbatch (http://ceumass.eps.uspceu.es/api/v3/advancedbatch).

Although the results of the requests seem like a query and a GET method might appear to be more appropriate at first sight, the amount of data potentially needed for performing the request precludes the use of this method: the number of m/zs, RTs and composite spectra can exceed the standard length of a parametrized URI, the method used to send the parameters in the GET method. Thus, both resources support the POST method to return the results of the request. Consequently, the body of the POST method must include all necessary input parameters needed for each service. As JSON has become the main format for data exchange over the web and it is supported by most of the tools and technologies, CMM uses this format for the communication through the RESTful API. Thus, both input parameters and service outcomes, must be encoded using this media type (see S1 for further details about service request and response specifications). The complete manual of the application is available at `http://ceumass.eps.uspceu.es/manuals.xhtml`.

This RESTful API for CMM has been already integrated into the HMDB environment (http://www.hmdb.ca/spectra/ms_cmm/search), where users can take advantage of the CMM filtering and scoring functionality directly from the HMDB web interface (see Figures S-5 and S-6). The integration with HMDB has been configured to use only the data from its own database, but it can take advantage of the filtering and scoring performed by CMM. We also have plans to integrate CMM with the Workflow4Metabolomics. Such service integration lets users perform all the steps of the metabolomics work-flow in a simple way.

The following sections show detailed information on how to invoke these services using the new API.



The CMM search is an external service provided and supported by CEMBIO (Center of Metabolomics and bioanalysis of University San Pablo-CEU, Spain). It calculates a score over the annotations based on:

1. Probability of the compound to form a specific adduct.
2. Presence of other adducts coming from the same signal in the same RT window.
3. RT of the lipids with the same head depending on the number of carbons and double bonds (and, consequently, the hydrophobicity of them).

It allows the user to filter the list of results based on the chemical alphabet, the presence on one or more specific databases and the metabolites type. It uses the experimental settings (modifier) for the ionization rules (1). It takes the information from the mass spectrometer (m/z) for the search into the databases. It uses the RT for the application of adduct relation rules (2) and the hydrophobicity of lipids pertaining to the same class depending on the number of carbons and double bonds (3). It takes advantage of the pre-processing software tools to group the peaks arising from the same primal signal to detect the adduct formed. For a more detailed information about the scores over the annotations, we suggest to visit the official manual (http://ceumass.eps.uspceu.es/manuals.xhtml).

CMM service is available from HMDB, but all the issues and suggestions regarding this search should be addressed to the CEMBIO team ().

If you use this service, please also cite: ()

Spectra Search Mass Spectrum - CEU Mass Mediator

Significant Query Masses (Da): one entry per line

Retention Times: one entry per line

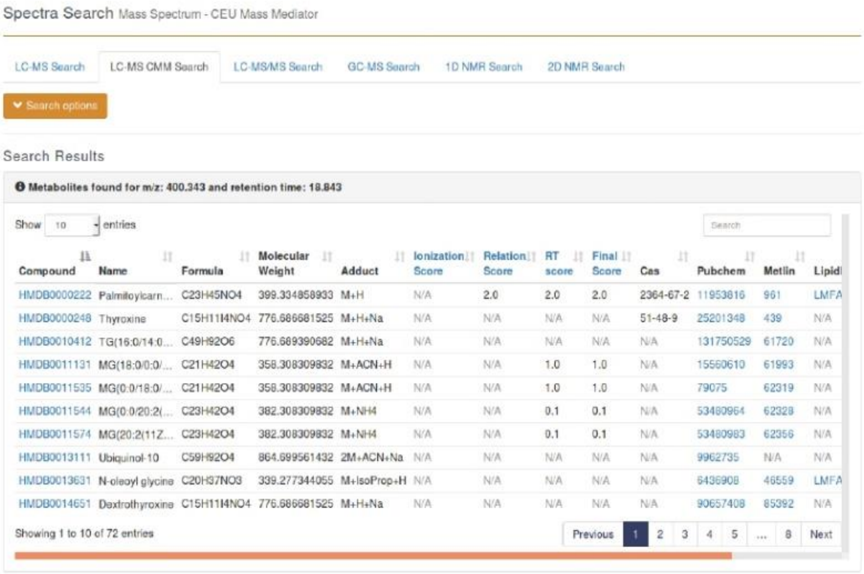
Composite Spectra: load example to see format

All Query Masses (Da): one entry per line

All Retention Times: one entry per line

All Composite Spectra: load example to see format

Figure S-5: Web interface of the CMM search input form in HMDB.



Spectra Search Mass Spectrum - CEU Mass Mediator

Search Results

Metabolites found for m/z: 400.343 and retention time: 18.843

Show 10 entries

Compound	Name	Formula	Molecular Weight	Adduct	Ionization Score	Relation Score	RT score	Final Score	Cas	Pubchem	Metlin	Lipidl
HMDB0000222	Palmitoylcarn...	C23H45NO4	399.334858933	M+H	N/A	2.0	2.0	2.0	2364-67-2	11953816	961	LMFA
HMDB0000248	Thyroxine	C15H11I4NO4	776.686681525	M+H+Na	N/A	N/A	N/A	N/A	51-48-9	25201348	439	N/A
HMDB0010412	TG(16:0/14:0...	C49H92O6	776.689390682	M+H+Na	N/A	N/A	N/A	N/A	N/A	131750529	61720	N/A
HMDB0011131	MG(18:0/0:0/...	C21H42O4	358.308309832	M+ACN+H	N/A	N/A	1.0	1.0	N/A	15560610	61993	N/A
HMDB0011535	MG(0:0/18:0/...	C21H42O4	358.308309832	M+ACN+H	N/A	N/A	1.0	1.0	N/A	79075	62319	N/A
HMDB0011544	MG(0:0/20:2(...	C23H42O4	382.308309832	M+NH4	N/A	N/A	0.1	0.1	N/A	53480964	62328	N/A
HMDB0011574	MG(20:2(11Z...	C23H42O4	382.308309832	M+NH4	N/A	N/A	0.1	0.1	N/A	53480983	62356	N/A
HMDB0013111	Ubiquinol-10	C59H92O4	864.699561432	2M+ACN+Na	N/A	N/A	N/A	N/A	N/A	9962735	N/A	N/A
HMDB0013631	N-oleoyl glycine	C20H37NO3	339.277344055	M+IsoProp+H	N/A	N/A	N/A	N/A	N/A	6436908	46559	LMFA
HMDB0014651	Dextrothyroxine	C15H11I4NO4	776.686681525	M+H+Na	N/A	N/A	N/A	N/A	N/A	90657408	85392	N/A

Showing 1 to 10 of 72 entries

Figure S-6: Search results, including filtering and scoring performed by CMM, displayed in HMDB.

Batch Search Service

Batch search enables the user to find metabolites through the *m/z* or the *neutral masses*.

The service is accessed through the following URI:

http://ceumass.eps.uspceu.es/api/v3/batch

To perform a query, the user must send a *POST* request. This request must include:

- A *Content-type* header set to *application/json*.
- A request body with a JSON object that includes all data needed for the query: *masses* to search in CMM, *tolerance* allowed for the putative annotations regarding the masses, *metabolite types* to search, *masses mode*, *ionization mode*, possible *adducts* formed when running the experiment and *databases* that will be included in the search.

The query's attributes, its name, type, default value (the value which will be used if the user does not specify the attribute) and optativity are defined in table S-2. As the value of some attributes is restricted to a range of literals, table S-3 shows the defined enumeration types.

Table S-2: Batch Search service - Request - Query

Name		Type	Default value
mandatory	masses	array of doubles	-
	tolerance	double (Range: [0..100])	10
	tolerance_mode	tolerance_mode_enum	“ppm”
	databases	array of database_enum	“all-except-mine”
	metabolites_type	metabolites_type_enum	“all-except-peptides”
	masses_mode	masses_mode_enum	“mz”
	ion_mode	ion_mode_enum	“positive”
	adducts	array of positive_enum	[“M+H”, “M+2H”, “M+Na”, “M+K”, “M+NH4”, “M+H-H2O”]
		array of negative_enum	[“M-H”, “M+Cl”, “M+FA-H”, “M-H-H2O”]
		array of neutral_enum	[“M”]

The following example shows a query to the Batch Search service:

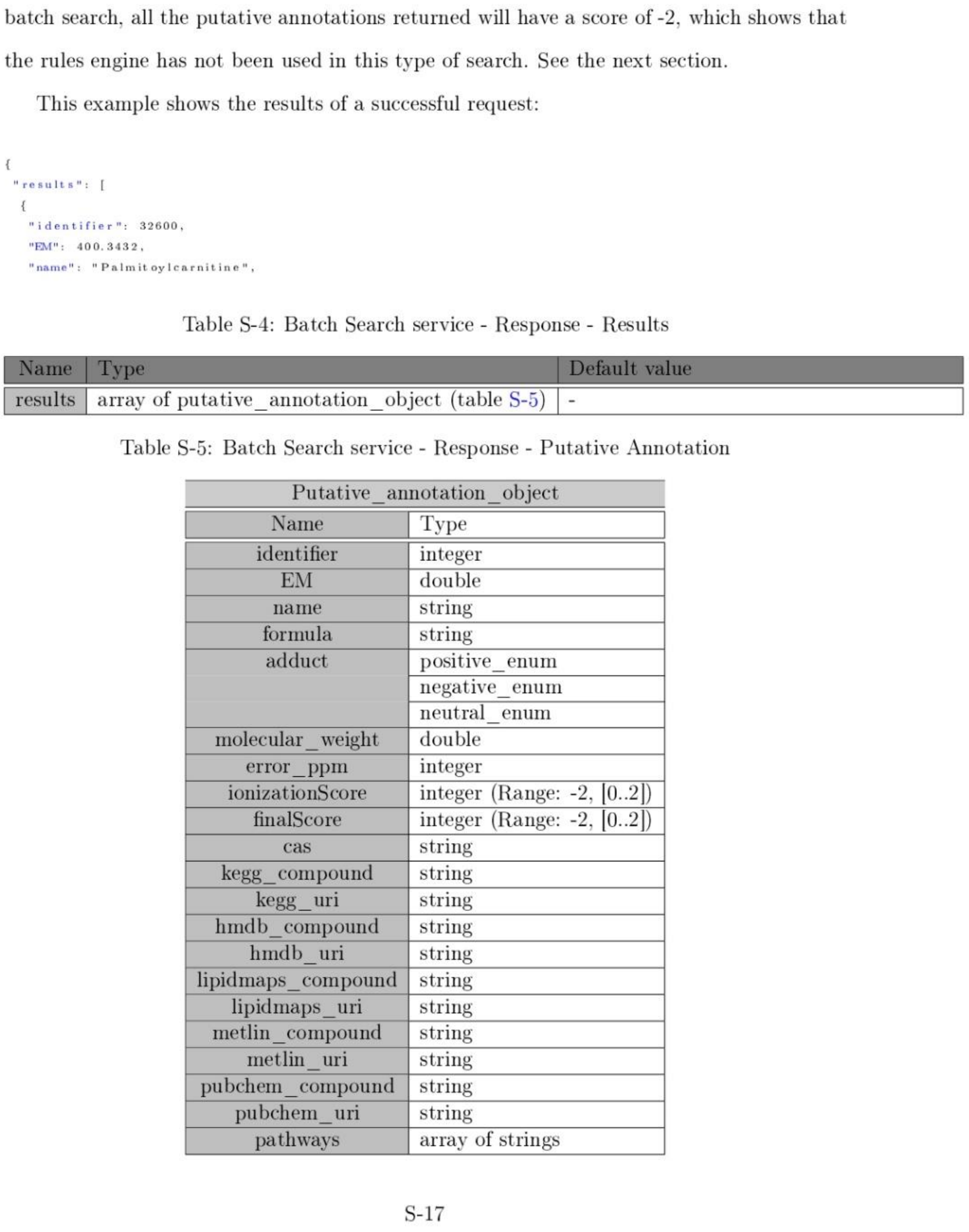
```
{
 "metabolites_type": "all-except-peptides",
 "databases": ["hmdb"],
 "masses_mode": "mz",
 "ion_mode": "positive",
 "adducts": ["all"],
 "tolerance": 10.0,
 "tolerance_mode": "ppm",
 "masses": [400.3432, ..., 288.2174]
}
```

If the request contains no errors and is therefore correctly processed, the service returns a set (table S-4) of putative annotations for the masses submitted. Each putative annotation structure (table S-5) contains the *name* of the putative annotation compound, its *formula*, its *molecular weight*, the difference between the molecular weight and the corresponding experimental mass, and references of the compound in external *databases*.

While some of these attributes are related with score rules, please bear in mind that rules are only applied when using the Batch Advanced Search service. Therefore, when using the

Table S-3: Batch Search service - Enumeration types

Name	Values
tolerance_mode_enum	"ppm", "mDa"
database_enumeration	"all", "all-except-mine", "HMDB", "LipidMaps", "Metlin", "Kegg", "in-house", "mine"
metabolites_type_enum	"all-except-peptides", "only-lipids", "all-including-peptides"
masses_mode_enum	"neutral", "mz"
ion_mode_enum	"neutral", "positive", "negative",
positive_enum	"M+H", "M+2H", "M+Na", "M+K", "M+NH4", "M+H-H2O", "M+H+NH4", "2M+H", "2M+Na", "M+H+HCOONa", "2M+H-H2O", "M+3H", "M+2H+Na", "M+H+2K", "M+H+2Na", "M+3Na", "M+H+Na", "M+H+K", "M+ACN+2H", "M+2Na", "M+2ACN+2H", "M+3ACN+2H", "M+CH3OH+H", "M+ACN+H", "M+2Na-H", "M+IsoProp+H", "M+ACN+Na", "M+2K-H", "M+DMSO+H", "M+2ACN+H", "M+IsoProp+Na+H", "2M+NH4", "2M+K", "2M+ACN+H", "2M+ACN+Na"
negative_enum	"M-H", "M+Cl", "M+FA-H", "M-H-H2O", "M-H+HCOONa", "2M-H", "M-3H", "M-2H", "M+Na-2H", "M+K-2H", "M+Hac-H", "M+Br", "M+TFA-H", "2M+FA-H", "2M+Hac-H", "3M-H"
neutral_enum	"M"



batch search, all the putative annotations returned will have a score of -2, which shows that the rules engine has not been used in this type of search. See the next section.

This example shows the results of a successful request:

```
{
 "results": [
  {
   "identifier": 32600,
   "EM": 400.3432,
   "name": "Palmitoylcarnitine",
```

Table S-4: Batch Search service - Response - Results

Name	Type	Default value
results	array of putative_annotation_object (table S-5)	-

Table S-5: Batch Search service - Response - Putative Annotation

Putative_annotation_object	
Name	Type
identifier	integer
EM	double
name	string
formula	string
adduct	positive_enum
	negative_enum
	neutral_enum
molecular_weight	double
error_ppm	integer
ionizationScore	integer (Range: -2, [0..2])
finalScore	integer (Range: -2, [0..2])
cas	string
kegg_compound	string
kegg_uri	string
hmdb_compound	string
hmdb_uri	string
lipidmaps_compound	string
lipidmaps_uri	string
metlin_compound	string
metlin_uri	string
pubchem_compound	string
pubchem_uri	string
pathways	array of strings

```
      "formula": "C23H45NO4",
      "adduct": "M+H",
      "molecular_weight": 399.334858933,
      "error_ppm": 3,
      "ionizationScore": -2,
      "finalScore": -2,
      "cas": "2364-67-2",
      "kegg_compound": "C02990",
      "kegg_uri": "http://www.genome.jp/dbget-bin/www_bget?cpd:C02990",
      "hmdb_compound": "HMDB0000222",
      "hmdb_uri": "http://www.hmdb.ca/metabolites/HMDB0000222",
      "lipidmaps_compound": "LMFA07070004",
      "lipidmaps_uri": "http://www.lipidmaps.org/data/LMSDRecord.php?LMID=LMFA07070004",
      "metlin_compound": "961",
      "metlin_uri": "https://metlin.scripps.edu/metabo_info.php?molid=961",
      "pubchem_compound": "11953816",
      "pubchem_uri": "https://pubchem.ncbi.nlm.nih.gov/compound/11953816",
      "pathways": []
    },
    ...
  ]
}
```

Batch Advanced Search Service

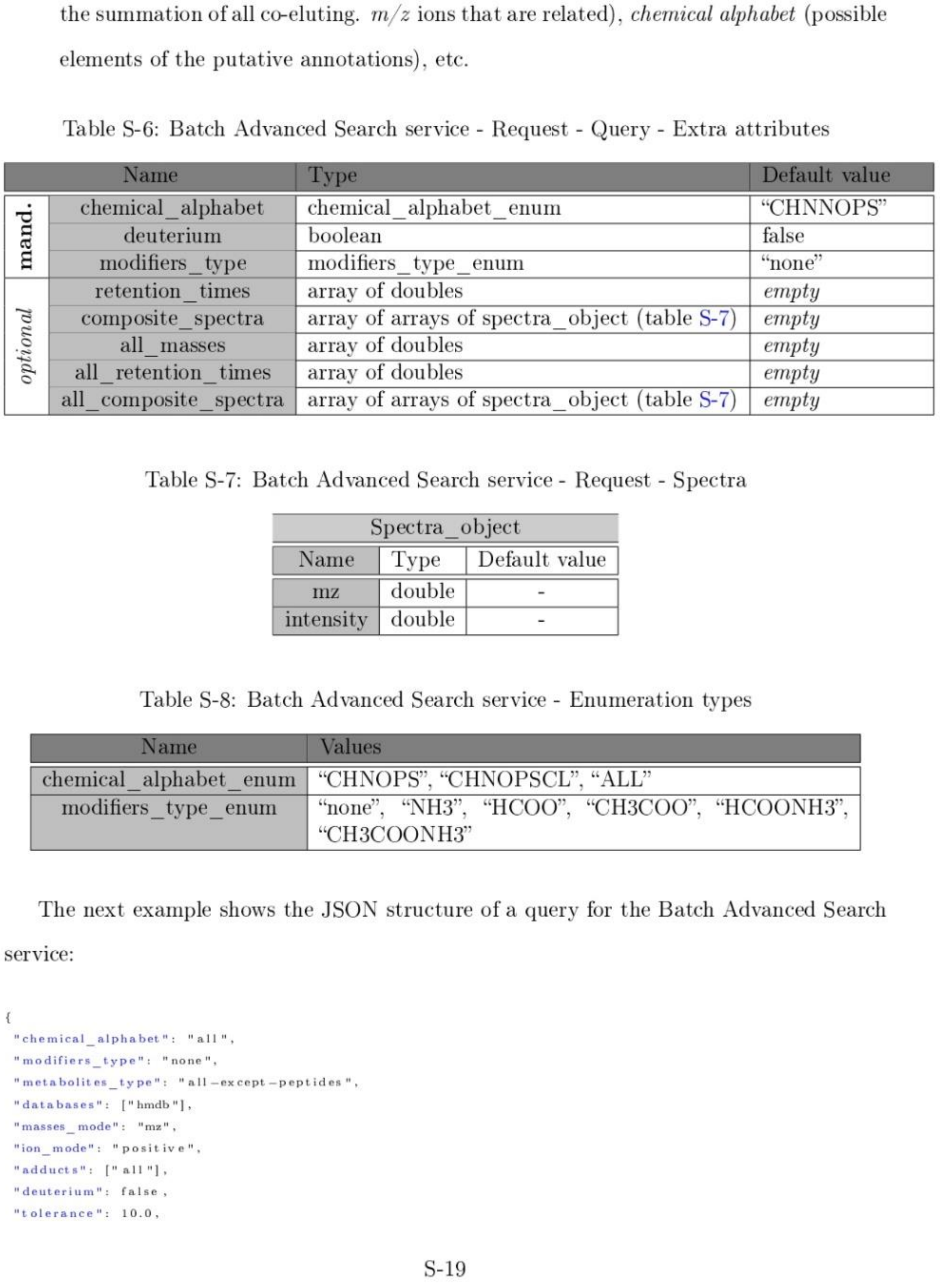
Batch advanced search also enables the user to find metabolites through the *m/z* or the *neutral masses* query parameters. But, in contrast with the previous service, it uses additional information devoted to biomarker discovery experiments.

The service is accessed through the following URI:

`http://ceumass.eps.uspceu.es/api/v3/advancedbatch`

To perform a query, the user must send a *POST* request. This request must include:

- A *Content-type* header set to *application/json*.
- A request body with a JSON object that includes all data needed for the query. In this case, the query is just an extension of the Batch Search query. Therefore, it must include all attributes described in table S-2 and, on top of that, provide the additional information shown in table S-6: *retention times*, *composite spectra* (spectra created by



the summation of all co-eluting. m/z ions that are related), *chemical alphabet* (possible elements of the putative annotations), etc.

Table S-6: Batch Advanced Search service - Request - Query - Extra attributes

	Name	Type	Default value
mand.	chemical_alphabet	chemical_alphabet_enum	“CHNNOPS”
	deuterium	boolean	false
	modifiers_type	modifiers_type_enum	“none”
optional	retention_times	array of doubles	*empty*
	composite_spectra	array of arrays of spectra_object (table S-7)	*empty*
	all_masses	array of doubles	*empty*
	all_retention_times	array of doubles	*empty*
	all_composite_spectra	array of arrays of spectra_object (table S-7)	*empty*

Table S-7: Batch Advanced Search service - Request - Spectra

Spectra_object		
Name	Type	Default value
mz	double	-
intensity	double	-

Table S-8: Batch Advanced Search service - Enumeration types

Name	Values
chemical_alphabet_enum	“CHNOPS”, “CHNOPSCL”, “ALL”
modifiers_type_enum	“none”, “NH3”, “HCOO”, “CH3COO”, “HCOONH3”, “CH3COONH3”

The next example shows the JSON structure of a query for the Batch Advanced Search service:

```
{
 "chemical_alphabet": "all",
 "modifiers_type": "none",
 "metabolites_type": "all-except-peptides",
 "databases": ["hmdb"],
 "masses_mode": "mz",
 "ion_mode": "positive",
 "adducts": ["all"],
 "deuterium": false,
 "tolerance": 10.0,
 "tolerance_mode": "ppm",
 "masses": [400.3432, ..., 288.2174],
 "all_masses": [],
 "retention_times": [18.842525, ..., 4.021555],
 "all_retention_times": [],
 "composite_spectra": [
  [{
    "mz": 400.3432,
    "intensity": 307034.88
   },
   ...,
   {
    "mz": 311.20145,
    "intensity": 400.03336
   },
   ...
  ]
 ]
}
```

When using the Batch Advance Search service, CMM scores the putative annotations based on expert knowledge. Thus, the response structure of this service contains all attributes already defined in table S-5, plus some other attributes defined in table S-9.

Table S-9: Batch Advanced Search service - Response - Putative Annotation - Extra Attributes

Putative_annotation_object - additional attributes	
Name	Type
RT	double
adductRelationScore	integer (Range: -2, [0..2])
RTscore	integer (Range: -2, [0..2])
finalScore	integer (Range: -2, [0..2])

This example shows the results of a successful request:

```
{
 "results": [
  {
   "RT": 8.144917,
   "adductRelationScore": -2,
   "RTscore": 2,
   "identifier": 111123,
   "EM": 338.2299,
   "name": "MG(0:0/i-12:0/0:0)",
   "formula": "C15H30O4",
   "adduct": "M+ACN+Na",
    "molecular_weight": 274.214409446,
    "error_ppm": 1,
    "ionizationScore": -2,
    "finalScore": 2,
    "kegg_compound": "",
    "kegg_uri": "",
    "hmdb_compound": "HMDB0072858",
    "hmdb_uri": "http://www.hmdb.ca/metabolites/HMDB0072858",
    "lipidmaps_compound": "",
    "lipidmaps_uri": "",
    "metlin_compound": "",
    "metlin_uri": "",
    "pubchem_compound": "131779644",
    "pubchem_uri": "https://pubchem.ncbi.nlm.nih.gov/compound/131779644",
    "pathways": []
   },
   ...
 ]
}
```

References

(1) Huan, T.; Li, L.; Tang, C.; Li, R.; Shi, Y.; Lin, G. MyCompoundID MS/MS Search: Metabolite Identification Using a Library of Predicted Fragment-Ion-Spectra of 383,830 Possible Human Metabolites. *Analytical Chemistry* **2015**, *87*, 10619.

(2) Ruttkies, C.; Wolf, S.; Neumann, S.; Schymanski, E. L.; Hollender, J. MetFrag relaunched: Incorporating strategies beyond in silico fragmentation. *Journal of Cheminformatics* **2016**, *8 (1)*, 3.

(3) Horai, H. et al. MassBank: A public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry* **2010**, *45*, 703–714.

SUMMARY, GLOBAL DISCUSSION, AND FUTURE PERSPECTIVES

7.1 Summary

CMM was created in 2012 with the goal of providing a single interface to query distinct metabolomic databases. In 2017, a first major revision was released to assist in metabolite annotation with new functionalities using a knowledge-based approach to filter and score the putative annotations obtained by querying them. In 2018, a second major revision was published containing relevant changes such as the update of data sources, a MS/MS search service, a dedicated service for oxPCs identification and a spectra quality controller.

Currently, CMM integrates 332,665 experimental compounds from the metabolomic databases HMDB, KEGG, LipidMaps, Metlin and an in-house library containing oxPCs, and 681,198 predicted compounds from MINE. CMM allows the user to simultaneously query these sources. It scores the annotations based on the probability of ionization and adduct formation, the presence or absence of other expected adducts originating from the same signal, and the elution order of lipids belonging to the same class when working in reversed-phase (RP) mode. CMM is a free an open source J2EE (Java 2 Platforms, Enterprise Edition) application (https://github.com/albertogilf/ceuMassMediator) currently running on TomEE 7.0.2 and MySQL server 5.7.24 that can be accessed through web browsers supporting JavaScript (JS) (http://ceumass.eps.uspceu.es) or through its REST API (http://ceumass.eps.uspceu.es/mediator/api/v3). CMM updates the data from the original sources approximately every 6 months and provides a JavaScript Object Notation (JSON) based REST API for all its services to facilitate communication with other tools in an automated way (see Figure 4). The following subsections summarize with more detail the main contributions made in this thesis.

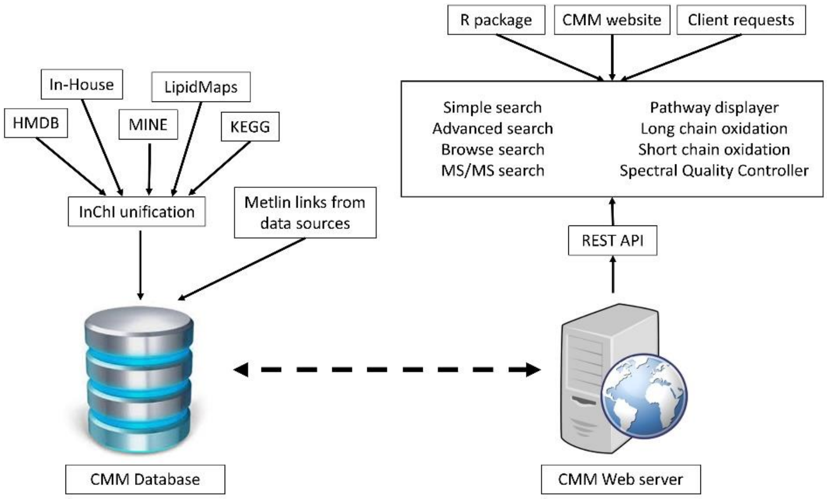


Figure 4 CEU Mass Mediator architecture and list of the services (end-points) available.

7.1.1 The expert system using MS^1 annotation

CMM simple and batch search options allow users to find the putative annotation for the *m/z* values acquired using any type of accurate mass spectrometer. It also enables filtering the compounds based on the data source and/or the type of metabolite searched. The advanced search is designed specifically for ESI-MS (see Figure 5).

CMM can distinguish between the statistically significant signals between the compared groups and the complete set of acquired signals. All measured signals can provide information to support or refute putative annotations of the statistically significant signals, signals which, as potential biomarkers, are the main target of the annotation process. If the user provides the complete signal matrix, CMM will try to extract evidence from it to achieve confidence level 2 in the annotation of significant signals. CMM also exploits information from the Composite Spectrum (CS), the set of all related co-eluting *m/z* ions, including isotopes, adducts, charges, multimers and IPs formed by in-source fragmentation or neutral losses

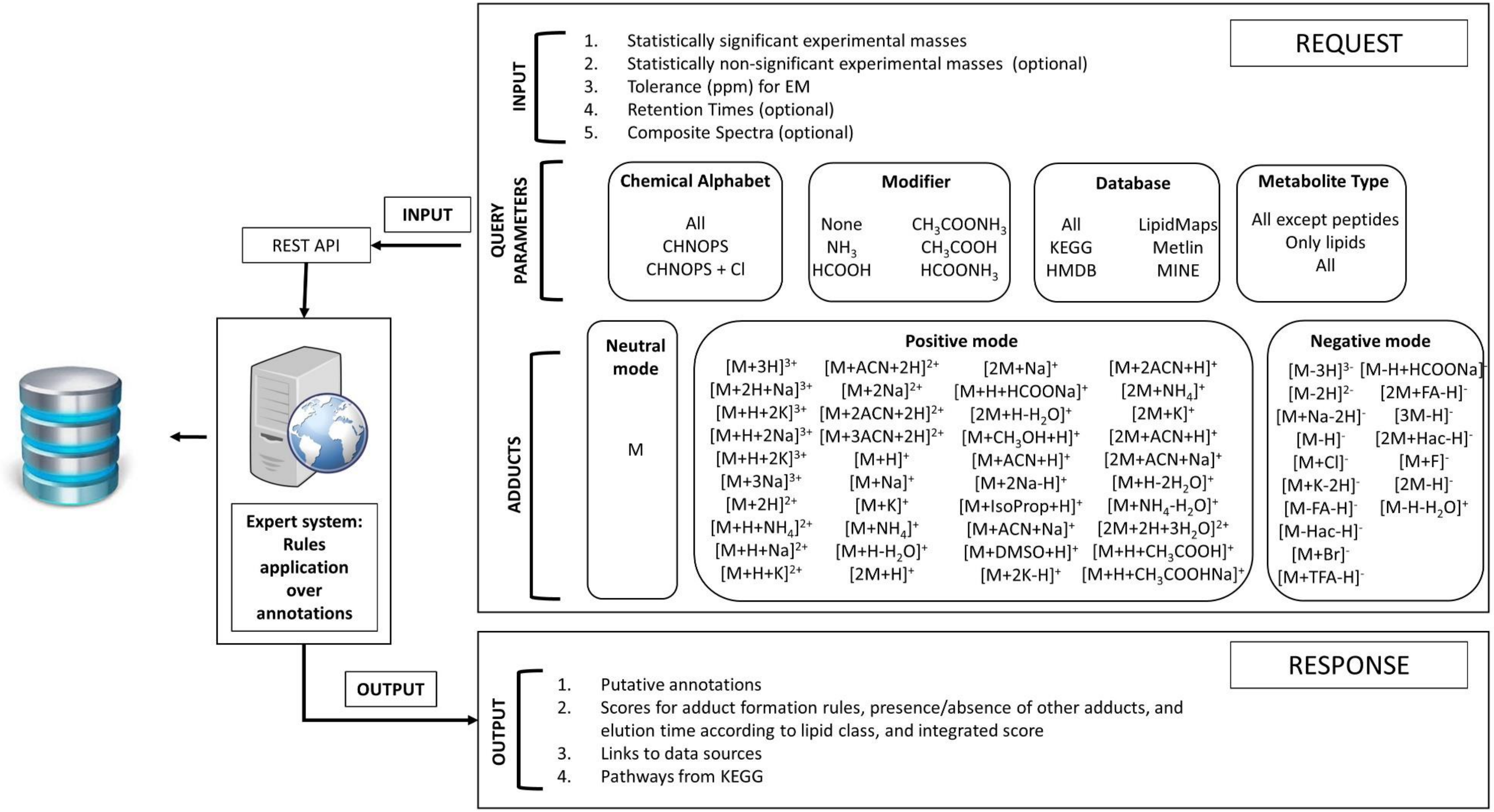


Figure 5 Workflow of CEU Mass Mediator MS^1 batch advanced search.

Users can restrict the chemical elements of the putative annotations, such as deuterated compounds, based on the Chemical Alphabet. Information about the mobile phase modifier used in the experiments can be added to restrict the formation of possible adducts to only the expected ones. The possible adducts supported are shown in Figure 5.

Once the query has been performed according to the user input, CMM incorporates an expert system that scores the putative compounds (see Figure 6). This expert system uses 122 rules divided in three main groups: (1) probability of the compounds forming a specific adduct (score χ_1), (2) presence or absence of other adducts coming from the same signal (determined for co-eluting signals within a defined RT window) (score χ_2), and (3) elution order of lipids belonging to the same class when working in RP (score χ_3). These three scores are integrated into a single overall score by computing their weighted geometric mean:

$$\chi = exp\left(\frac{\sum_{i=1}^{3} \omega_i \cdot ln\chi_i}{\sum_{i=1}^{3} \omega_i}\right)$$

where ω_i is the weight of each score; ω_1 = 1, ω_2 = 1, and ω_3 ϵ [0, 2]. ω_3 weight depends on the number of rules that were applied for lipid elution time (this number is variable and depends on how many other lipids could be used in the lipid elution order).

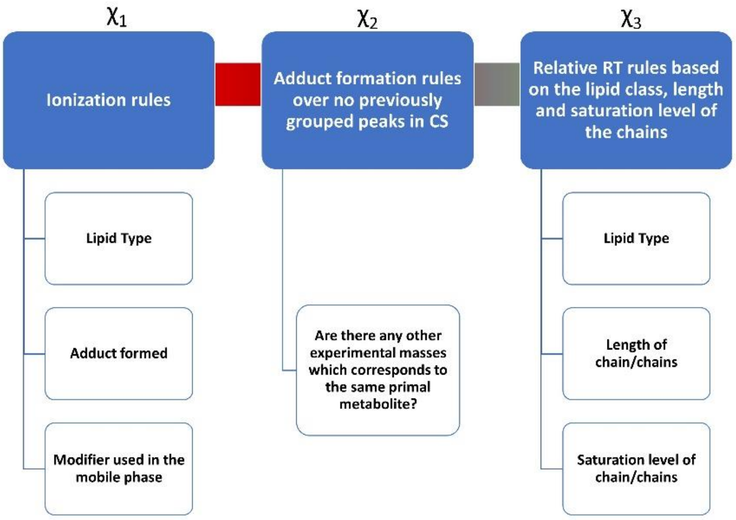
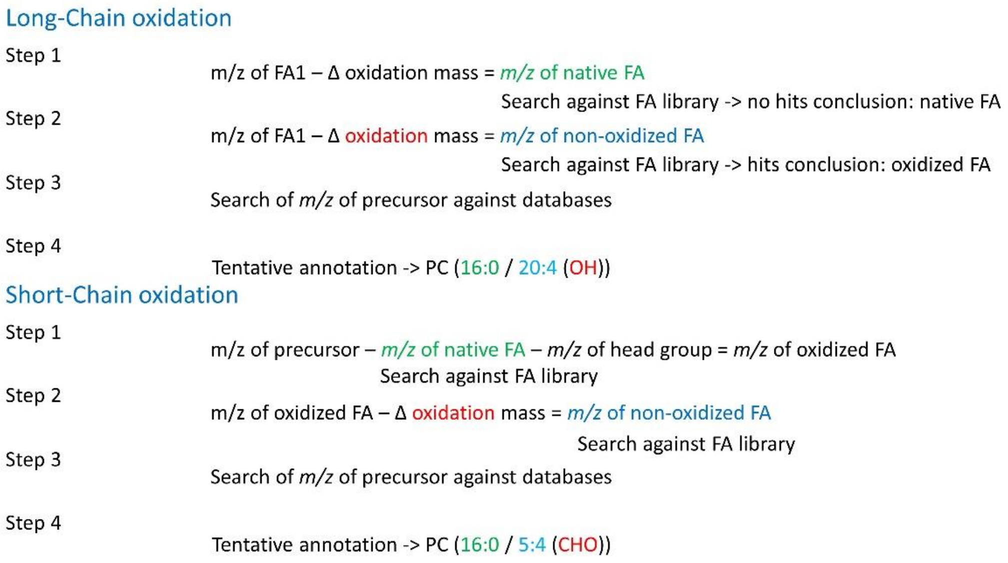

χ1
χ2
χ3
Ionization rules
Adduct formation rules over no previously grouped peaks in CS
Relative RT rules based on the lipid class, length and saturation level of the chains
Lipid Type
Adduct formed
Modifier used in the mobile phase
Are there any other experimental masses which corresponds to the same primal metabolite?
Lipid Type
Length of chain/chains
Saturation level of chain/chains

Figure 6 Rules of CEU Mass Mediator expert system.

7.1.2 A semi-automated tool for oxPCs identification

CMM also provides support for the identification of oxPCs. Recently, these compounds have been characterized as relevant biomarkers of health and diseases status, driving the interest in specific tools to support their identification and understanding their biological function. CMM aids in the identification of oxPCs from LC-ESI-MS/MS experimental data. It integrates knowledge about fragmentation of oxPCs as either long- or short-chain oxidized lipids, characterized by different oxidation and fragmentation processes, as well as a different handling of their oxidized derivatives. Based on the fragmentation patterns of oxPCs, the procedure compares the input spectrum introduced by the user with an internal database of oxPCs containing both curated and computationally generated records. The steps for the identification of short-chain and long-chain oxPCs are shown in Figure 7.



Long-Chain oxidation

Step 1
m/z of FA1 – Δ oxidation mass = *m/z* of native FA
Search against FA library -> no hits conclusion: native FA

Step 2
m/z of FA1 – Δ oxidation mass = *m/z* of non-oxidized FA
Search against FA library -> hits conclusion: oxidized FA

Step 3
Search of *m/z* of precursor against databases

Step 4
Tentative annotation -> PC (16:0 / 20:4 (OH))

Short-Chain oxidation

Step 1
m/z of precursor – *m/z* of native FA – *m/z* of head group = *m/z* of oxidized FA
Search against FA library

Step 2
m/z of oxidized FA – Δ oxidation mass = *m/z* of non-oxidized FA
Search against FA library

Step 3
Search of *m/z* of precursor against databases

Step 4
Tentative annotation -> PC (16:0 / 5:4 (CHO))

Figure 7 Flowchart of oxidized lipids identification in CEU Mass Mediator.

The oxPCs elute earlier than PCs in RP and later in HILIC (Hydrophilic Interaction Liquid Chromatography) due to an increment in their hydrophilicity. The oxPCs tend to form the adducts $[M-H]^-$ and $[M+HCOO]^-$

in short chain oxidations and only the adduct $[M-H]^-$ in long chain oxidations as well as the neutral loss of water, that usually appears in oxPCs while it is very uncommon in non-oxidized PCs. Different collision energy was applied to the oxPCs to observe the changes in the IPs formed. Although sometimes the information available is not enough to the unequivocal identification of a structure, this method is fast and reliable to determine the presence of an oxPC, thus reducing the false putative annotations and the amount of time spent by the researchers.

7.1.3 Use of non-analytical information

CMM enables sorting the compounds for their subsequent biological interpretation based on the number of compounds from a specific pathway present in the experimental data and the compound's relevance for a given pathway. Relevance is determined by the number of pathways in which a compound is present. Water is an example of a compound with low relevance because, due to its ubiquity in most pathways, its presence does not yield specific biological relevance.

7.1.4 MS^n support

CMM provides also a MS/MS search that supports metabolite identification using MS/MS data. This search is based on spectral similarity measurements between experimental spectra and library spectra of standards and/or predicted spectra contained in HMDB.

Another unique functionality of CMM is its ability to calculate the quality of a MS/MS spectrum for identification purposes. Experimental conditions are key to obtain a clear spectrum that enables a more reliable identification. A spectrum of inferior quality usually leads to too many unknowns or, even worse, to misidentifications. CMM ranks the quality of a spectrum considering the intensity of the signal in both MS and MS/MS analysis, noise level, number of scans performed to acquire the spectrum, number of samples analyzed (correspondence between different samples provides more confidence to the fragments obtained), presence of more than one compound in the collision cell (arising from chromatographic co-elution), crosstalk, and any spectrum contamination by ions present in the collision cell but originating from previous scans (see Figure 8).

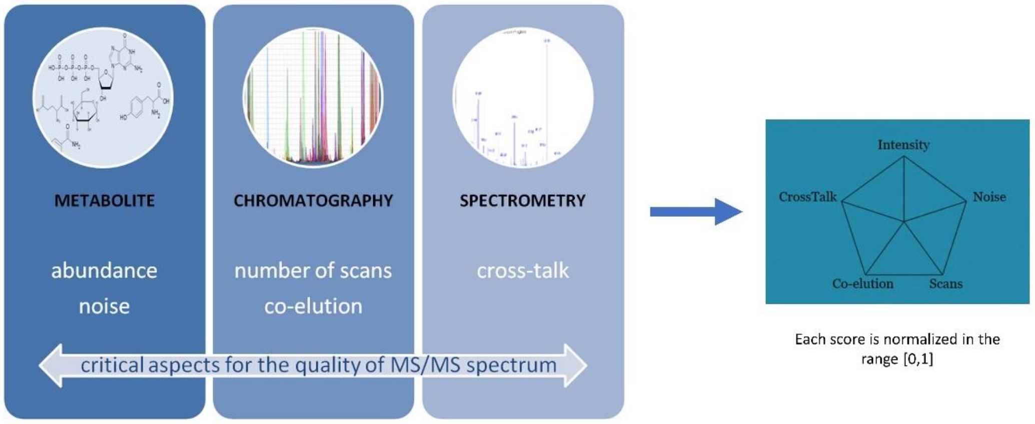


Figure 8 Aspects measured by the CMM spectral quality controller and pentagonal representation of the score of each aspect.

7.1.5 Restful API

All these services offered by CMM can be accessed through a REST API, enabling users to include them in their data analysis and workflows. Currently CMM is already integrated in the most cited metabolomic database: HMDB. The HMDB users can perform MS^1 searches with the filters and the expert system provided by CMM and can obtain the scores for the putative annotations calculated by CMM. The service can be used with no need of accessing CMM webpage and it avoids the process of learning a new environment. It is publicly accessible through the URL: http://www.hmdb.ca/spectra/ms_cmm/search.

CMM is also accessible through an R package available in the Comprehensive **R** Archive Network (CRAN) collection. The users with experience in R can use all the CMM services through the package https://rdrr.io/github/lzyacht/cmmr/. The R package has been developed by Yaoxiang Li, a current collaborator of CMM.

7.2 Global discussion

Untargeted Metabolomics is a promising research area for different purposes. The idea of observing the biochemical changes without any previous hypothesis is innovative and breaks with the classical paradigm, consisting on observing a phenomenon to create hypotheses that are subsequentially checked. Although the hypothesis-based approaches have proved successful in many cases, it fails in some situations where lack of knowledge prevents researchers from formulating the right prior hypotheses, or where controlling the high number of variables interacting in living organisms is not possible. Untargeted metabolomics opens the investigation to unexpected findings. There are two main objectives in untargeted metabolomics, that are compatible: i) finding biomarkers for a given situation, pathology, treatment, etc and ii) generating a hypothesis about the mechanisms altered in a defined physiopathological situation.

Untargeted metabolomics main challenge is the metabolite identification and the subsequent biological interpretation, which depends in a large extent on the metabolite identification. Increasing the metabolite coverage will provide with a better picture of the processes concurring in living organisms, and it will increase the chances to achieve the right biological interpretation.

The metabolomic workflow can be improved by enhancing the sample preparation and the data acquisition steps, therefore increasing the data quality and easing the metabolite identification (analytical aspects). But the metabolite identification itself can be improved by expanding the metabolome completeness or developing more robust and reliable methods to perform identification, using both computational and methodological aspects. Currently, there are a high number of researchers working in this field and providing different solutions to increase the metabolite coverage from both analytical and computational sides.

CMM utilizes analytical and nonanalytical data to process the matrix of features obtained after the data processing, including chromatographic, information, pathway information, and it plans to use taxonomy and ontology information that it is already available in the integrated data sources. CMM has novel functionality not present in other tools such as the application of rules based on RT principles for the annotation of features obtained by MS[1], the identification of oxPCs using experimental knowledge or a spectra quality controller to help focusing efforts on the features most likely to be identified. While new tool updates and new tools are currently using filters based on the RT of the compounds, CMM was the first tool to exploit this information.

CMM has been conceived to overcome the current and future challenges in the metabolite annotation and identification, and these challenges may come from our laboratory, CEMBIO, or from other external groups. When CMM was born in 2012, it was used mainly at CEMBIO, but since then it has significantly expanded its user base. Figure 9 shows the CMM monthly users since the beginning of CMM 2.0 in July 2016 until the present (June 2019). It can be observed that the number of users keeps an upward trend. During the last 6 months (2019), the average number of monthly users is 264, being about 75% of them from outside Spain. These users on average have spent over 15 minutes using CMM and have seen 8.5 pages per session. In a traditional website, such as an online newspaper, users usually spend on average between 1 or 2 minutes on the web, and see about 1.5 pages per session; i.e., CMM users are doing real work with the tool. Analyzing this data, it can be concluded that CMM has had a very good acceptance by the metabolomic community.

7.3 Future perspectives

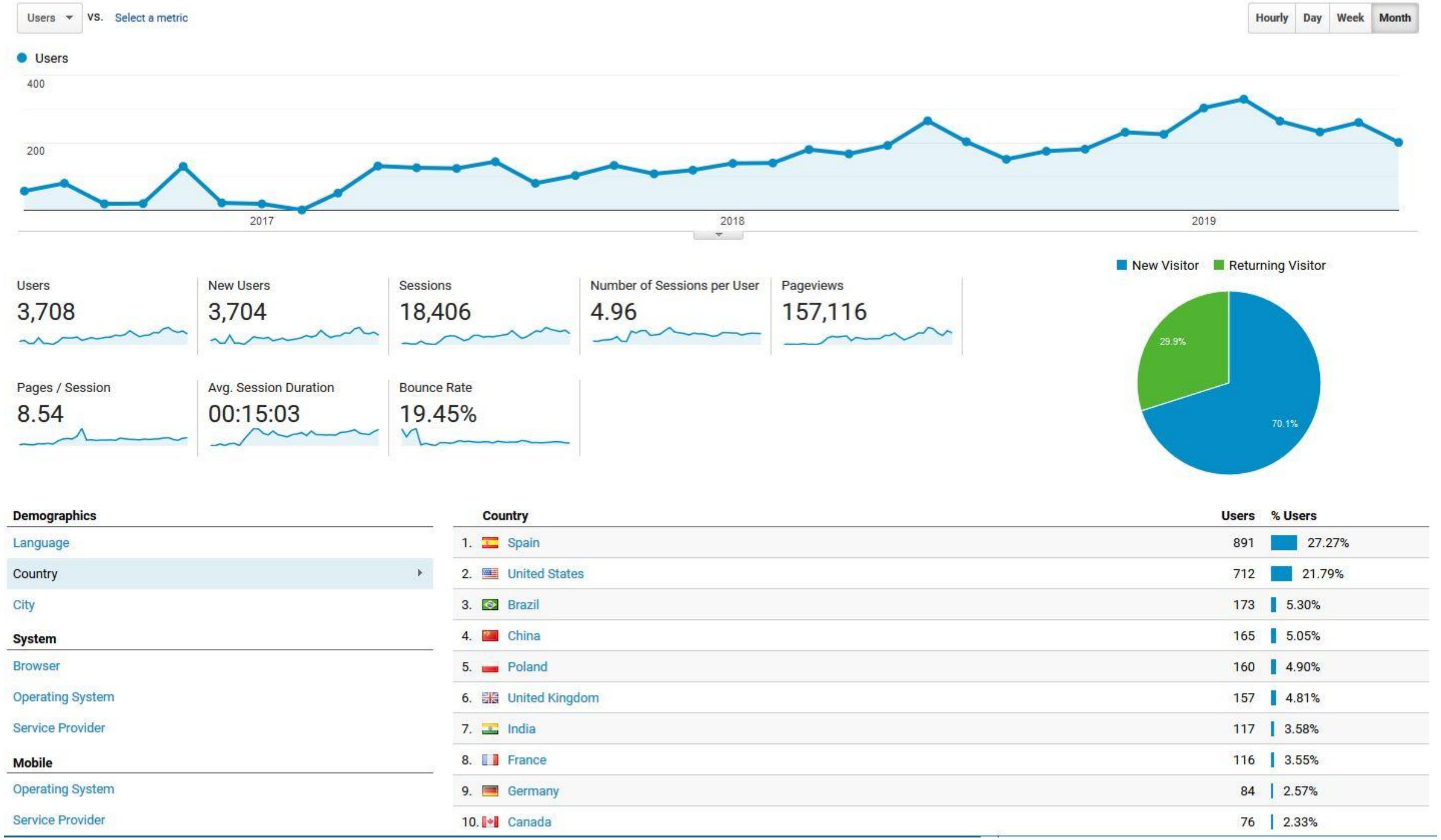
Applying analytical and non-analytical information yields a more reliable and less cumbersome approximation for metabolite identification and annotation. As more information becomes available in public resources, more opportunities to exploit it can be developed, and a higher confidence level can be reached for the putative annotations.

Regarding analytical information, currently in CMM there is a lack of information about CCS. The CCS is a promising technique to distinguish between isomers. There is also limited information about MT of compounds using CE-ESI-MS. This experimental information can provide a better support for the identification of compounds analyzed under this technique. Concerning non-analytical information, we consider that the exploiting of the taxonomy information of the compounds during the identification process, such as the information about endogenous or exogenous compounds can be substantively extended.

CMM metabolite identification can be improved in two main ways: providing general methods for supporting the identification regardless the analytical set up used and creating specific methods to improve the precision of the metabolite identification when using a concrete set up or technique. The creation of such specific services is a promising strategy for already well-established analytical methods. We plan to create a service for the metabolite identification using CE-ESI-MS and exploiting the efficient mobility and the relative migration time regarding a specific background electrolyte.

CMM has been already integrated into HMDB. However, we plan to integrate it into existing workflow tools such as Workflow4Metabolomics, KNIME or Taverna. Workflow4Metabolomics is particularly interesting because it provides data repositories and processing services for all the stages of the metabolomics workflow. They believe in the idea of providing a single interface to use all the necessary tools during all these metabolomic workflow stages. We fully agree with the idea that a single interface to use a set of different tools will be appreciated by the researchers.

CMM was born to fulfill the CEMBIO needs, but it has grown with feedback of internal and external users that continuously provide new ideas to improve the tool. Some of the users that have provided feedback come from Canada, the United States, Colombia, Brazil, Poland or United Kingdom. We encourage the community continue providing feedback and ideas about how to improve CMM, that, without doubt, shall be extremely valuable to guide the future development of CMM.



Users	3,708
New Users	3,704
Sessions	18,406
Number of Sessions per User	4.96
Pageviews	157,116
Pages / Session	8.54
Avg. Session Duration	00:15:03
Bounce Rate	19.45%

	Country	Users	% Users
1.	Spain	891	27.27%
2.	United States	712	21.79%
3.	Brazil	173	5.30%
4.	China	165	5.05%
5.	Poland	160	4.90%
6.	United Kingdom	157	4.81%
7.	India	117	3.58%
8.	France	116	3.55%
9.	Germany	84	2.57%
10.	Canada	76	2.33%

Figure 9 Monthly users of CMM obtained with google analytics during the period July 1st, 2016 to June 30th, 2019.

SCIENTIFIC PUBLICATIONS

Journal publications

1) **Gil-de-la-Fuente, A.**, Armitage, E.G., Otero, A., Barbas, C. and Godzien, J.; *Differentiating signals to make biological sense - a guide through databases for MS-based non-targeted metabolomics*; **Electrophoresis**, 2017, 38(18), 2242-2256
   - Impact factor: 2.744, Q2 (Analytical chemistry; Biochemical research methods)

2) **Gil-de-la-Fuente, A.**, Godzien, J., Fernández López, M., Rupérez, F.J., Barbas, C. and Otero, A.; *Knowledge-based metabolite annotation tool: CEU Mass Mediator*; **Journal of Pharmaceutical and Biomedical Analysis**, 2018, 154, 138-149
   - Impact factor 3.255, Q1 (Analytical chemistry; Pharmacology and pharmacy)

3) **Gil-de-la-Fuente, A.**, Traldi, F., Siroka, J., Kretowski, A., Ciborowski, M., Otero, A., Barbas, C. and Godzien, J.; *Characterization and annotation of oxidized glycerophosphocholines for non-targeted metabolomics with LC-QTOF-MS data;* **Analytica Chimica Acta**, 2018, 1037(11), 358-368
   - Impact factor 5.123, Q1 (Analytical chemistry)

4) **Gil-de-la-Fuente, A.**, Godzien, J., Saugar, S., Garcia-Carmona, R., Badran, H., Wishart, D.S., Barbas, C. and Otero, A.; *CEU Mass Mediator 3.0: A Metabolite Annotation Tool*; **Journal of Proteome Research**, 2019, 18(2), 797-802
   - Impact factor 3.950, Q1 (Biochemical research methods)
5) Fernández-López, M., Gil-de-la-Fuente, A., Godzien, J., Rupérez, F.J., Barbas, C. and Otero, A.; LAS: A Lipid Annotation Service Capable of Explaining the Annotations it Generates; **Computational and Structural Biotechnology Journal**, 2019, 17, 1113-1122
   - Impact Factor: 4.720 Q1 (Biochemistry and molecular biology)

Book chapters

1) Godzien, J., **Gil-de-la-Fuente, A.**, Otero, A., Barbas, C.; *Data Analysis for Omic Sciences: Methods and Applications. Chapter 15: Metabolite Annotation and Identification*; **Elsevier**, 2018, ISBN: 978-0-444-64044-4
2) **Gil-de-la-Fuente, A.**, Godzien, J., Otero, A. and Barbas, C. *Processing Metabolomics and Proteomics Data with Open Software. Chapter 11: Metabolite annotation with CEU Mass Mediator*, **Royal Society of Chemistry**, 2019, **Under Review**
3) **Gil-de-la-Fuente, A.**, Godzien, J., Otero, A. and Barbas, C. *Processing Metabolomics and Proteomics Data with Open Software. Chapter 12: Metabolite annotation using in-silico generated compounds: MINE and BioTransformer*, **Royal Society of Chemistry**, 2019, **Under Review**

International Conferences

1) **Gil-de-la-Fuente, A.**, Godzien, J., Barbas, C., Otero, A.; *CEU Mass Mediator: a knowledge-based tool for metabolite annotation*; **28th Pharmaceutical and Biomedical Analysis Conference**, 2019, Jul 2-5, Oral presentation presented by **Gil-de-la-Fuente, A.**
2) **Gil-de-la-Fuente, A.**, Traldi, F., Kowalczyk, T., Ciborowski, M., Otero, A., Barbas, C., Godzien, J.; *A fast and reliable spectral quality assessment in metabolomics studies*; **14th Annual Conference of the Metabolomics Society**, Jun 24-28, Oral presentation presented by **Gil-de-la-Fuente, A.**
3) **Gil-de-la-Fuente, A.**, Barbas, C., Godzien, J., Neumann S., Sumner, L.W.; *CEU Mass Mediator: a knowledge-based tool for metabolite annotation*; **14th Annual Conference of the Metabolomics Society**, 2019, Jun 24-28, Presented by Neumann S., Sumner, L.W., Gil de la Fuente, A., Godzien, J., Barbas, C.