UNIVERSITAT DE BARCELONA

Development and application of analytical and chemometric methodology for environmental metabolomic studies based on one- and two-dimensional liquid chromatography coupled to mass spectrometry

Miriam Carolina Pérez Cova

Aquesta tesi doctoral està subjecta a la llicència ***Reconeixement- NoComercial – CompartirIgual 4.0. Espanya de Creative Commons***.

Esta tesis doctoral está sujeta a la licencia ***Reconocimiento - NoComercial – CompartirIgual 4.0. España de Creative*** **Commons**.

This doctoral thesis is licensed under the ***Creative Commons Attribution-NonCommercial-ShareAlike 4.0. Spain License***.

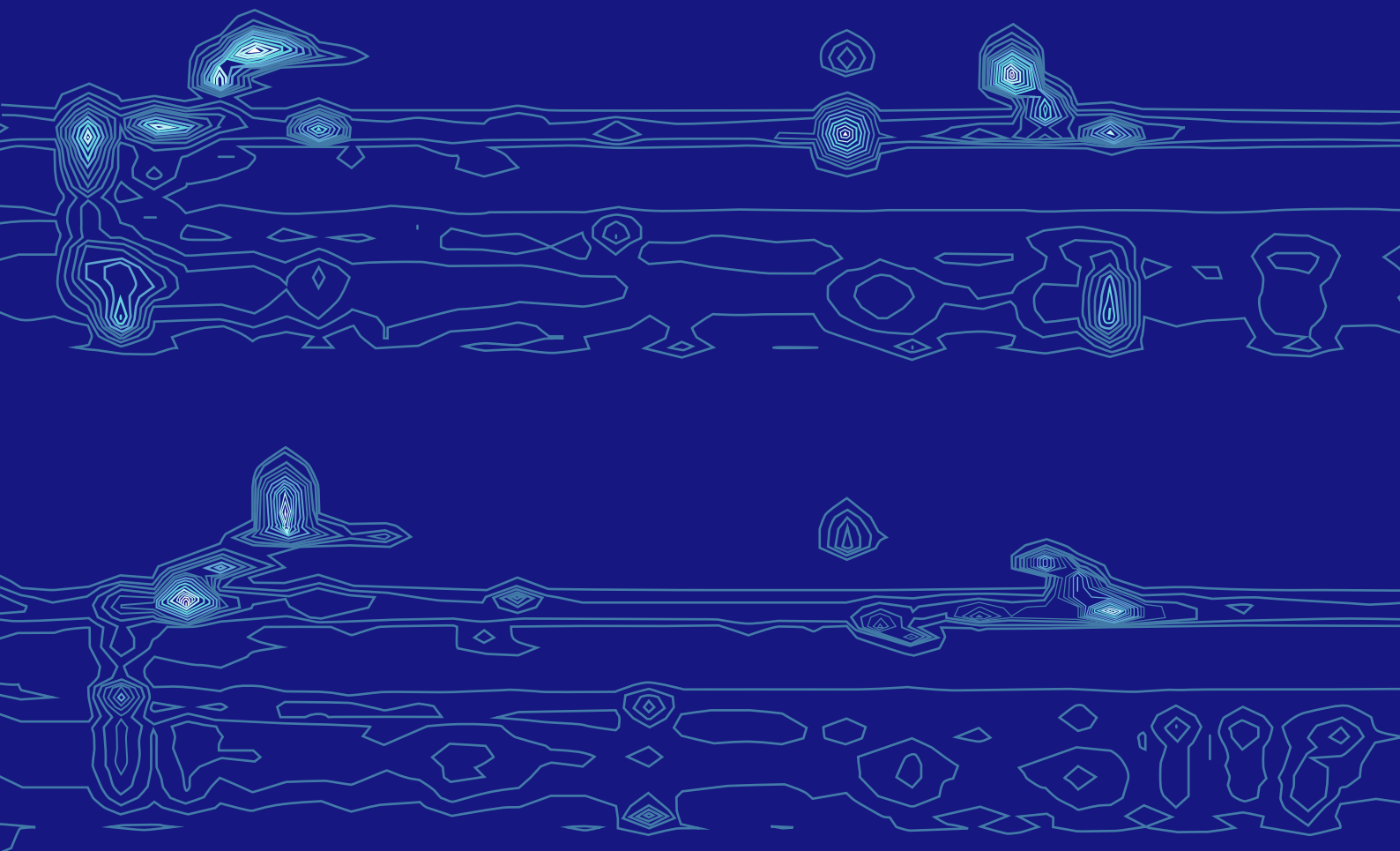
UNIVERSITAT DE
BARCELONA

Development and application of analytical and chemometric methodology for environmental metabolomic studies based on one-and two-dimensional liquid chromatography coupled to mass spectrometry

Miriam Carolina Pérez Cova

Development and application of analytical and chemometric methodology for environmental metabolomic studies based on one- and two-dimensional liquid chromatography coupled to mass spectrometry

Miriam Carolina Pérez Cova



UNIVERSITAT DE BARCELONA

CSIC
Consejo Superior de Investigaciones Científicas

UNIVERSITAT DE BARCELONA

CSIC

CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

Development and application of analytical and chemometric methodology for environmental metabolomic studies based on one-and two-dimensional liquid chromatography coupled to mass spectrometry

Miriam Carolina Pérez Cova

Doctoral programme: “Química Analítica i Medi Ambient”

Development and application of analytical and chemometric methodology for environmental metabolomic studies based on one-and two-dimensional liquid chromatography coupled to mass spectrometry

A Thesis submitted for the degree of Doctor in Analytical Chemistry by:

Miriam Carolina Pérez Cova

Supervisors:

Dr. Romà Tauler Ferré

Dr. Joaquim Jaumot Soler

Department of Environmental Chemistry
Institute of Environmental Assessment and Water Research (IDAEA)
Spanish National Research Council (CSIC)

Tutor:

Dr. Anna Maria De Juan Capdevila

Department of Chemical Engineering and Analytical Chemistry (UB)

Barcelona, June of 2022

Dr. Romà Tauler Ferré, Professor of the Department of Environmental Chemistry of the Institute of Environmental Assessment and Water Research, and, **Dr. Joaquim Jaumot Soler,** Research Scientist of the same Department,

STATE THAT:

the current PhD report entitled "**Development and application of analytical and chemometric methodology for environmental metabolomic studies based on one-and two-dimensional liquid chromatography coupled to mass spectrometry**" has been elaborated under our supervision by **Ms. Miriam Carolina Pérez Cova** in the Department of Environmental Chemistry of the Institute of Environmental Assessment and Water Research, and also that all the results presented in this manuscript are consequence of the research work of the hereby mentioned doctoral student.

And in order to make it certain, we sign the current certificate.

Barcelona, June of 2022

Dr. Romà Tauler Ferré **Dr. Joaquim Jaumot Soler**

A mis finshüns,

“I was taught that the way of progress was neither swift nor easy”

Marie Curie

Acknowledgements

Llega la hora de cerrar esta etapa, y me faltan las palabras para expresar la gratitud tan grande que siento por todas las personas que me han acompañado en este arduo viaje. Han sido cinco años especialmente duros, de gran dedicación, constancia, sacrificio y de mucho crecimiento tanto en el plano profesional como en el personal. Al echar la vista atrás, lo hago con orgullo, viendo que todo el esfuerzo, sudor y lágrimas, han valido la pena. Y precisamente por ello, me gustaría agradecer a todos los que de una manera u otra han estado conmigo durante esta etapa.

En primer lugar, a mis directores de Tesis, Dr. Joaquim Jaumot y Dr. Romà Tauler. Aquesta Tesi no hagués estat possible sense vosaltres. Gracias por apostar por una desconocida venida de tierras lejanas. Espero haber cumplido con vuestras expectativas. Joaquim, muchas gracias por creer en mí incluso en los momentos en los que ni yo misma creía, por apoyarme y ayudarme a levantar tras cada caída, por intentar ver siempre el lado positivo y por buscar soluciones conmigo, una y otra vez. Espero que te sientas tan orgulloso de esta Tesis como yo, porque ha salido a delante gracias a nuestro grandísimo esfuerzo y dedicación. No tengo palabras para agradecértelo lo suficiente. Romà, muchísimas gracias por el inestimable apoyo durante estos años, por abrirme puertas, por animarme a cruzar el charco, por alegrarte de mis logros y, en especial, por la ayuda este último año de trabajo frenético.

También quisiera agradecer a mi tutora, la Dra. Anna de Juan Capdevila, por toda la ayuda y apoyo brindados. Anna, has sido mucho más que una tutora, también una docente y mentora extraordinaria, todo un ejemplo para mí. Gràcies per tot.

Quisiera hacer una mención especial a Susana, por todo su apoyo en la sombra durante todo este tiempo. Gracias, de corazón.

Por otra parte, quisiera agradecer a todos los que, de una forma u otra, han formado parte mi grupo de investigación (*Chemometrics for Environmental Omics*, Ch4EO) estos años. Carma, moltes gràcies, por estar siempre ahí para cualquier consulta y por apuntarte a cualquier actividad de divulgación que te propusiera. Haber tenido la oportunidad de trabajar contigo este último año codo con codo simplemente no tiene precio. Gracias por todo lo que he has enseñado. Eres un referente para mí, tanto en lo profesional como en lo personal. Stefan, muchas gracias por haber acudido a nuestro rescate este último año. Ha sido un verdadero placer trabajar contigo, y gracias también por los buenos ratos en las comidas (pre-pandemia) y por ese fin de semana en Sant Carles. Marc, has sido mi compañero de Tesis todo este tiempo. Hemos tenido la suerte de empezar y acabar muy próximos. Gracias por ayudarme cuando estaba más perdida que un pulpo en un garaje, y gracias por esos viajes que hemos compartido. A mi equipo de "soporte informático" personal e incondicional, por estar siempre dispuestos a comprobar si mi ordenador se había apagado o si estaba haciendo de las suyas. Carlos, sabes que no hubiera sido lo mismo sin ti. Gracias por aparecer justo cuando lo necesitaba, por nuestras conversaciones profundas con Nestea de por medio, por apoyarnos mutuamente cuando las cosas se torcían, y por todas las experiencias y planes que hemos hecho en estos dos añitos, que no son pocas. Estoy convencida de que dejarás el pabellón bien alto con tu doctorado. Albert, gracias también por todas esas comidas pre y post pandemia, por los buenos ratos, las risas, los planes y las conversaciones. Estoy segura de que dejo mi relevo en las mejores manos. Aina, gracias por apuntarte a nuestras locuras de planes, por poner cabeza en la oficina y en el lab en mi ausencia, y por estar ahí cuando lo he necesitado. Sabes que es recíproco, y que puedes contar conmigo para lo que necesites. A Paula y David, por haber venido al lab a darle vidilla, y por apuntaros a las quedadas, del tipo que fueran. Jaume,

gracias por tus preguntas. Poder ayudarte con tus proyectos ha sido un verdadero placer, y espero que hayas aprendido de la experiencia al menos tanto como yo. Al resto de los TFGs y TFMs que han pasado por el lab estos años, por todo lo que me habéis enseñado, en especial a Cora y Miriam Condeminas. Y por supuesto, a Laia López. Eres de lo mejorcito que me llevo del doctorado. Quién hubiera dicho que tu TFM también nos traería esta GRAN amistad. Mi compi de dulces, de excursiones, de viajes, de playa... Mil gracias por estos cinco años, por estar siempre ahí, por abrirme las puertas de tu casa y por apoyarnos mutuamente en nuestros doctorados.

A todos aquellos que formaron parte de CHEMAGEB y que conocí en mi primer año de doctorado. Muchas gracias por abrirme el camino. En especial, a Meri, por enseñarme a usar el montaje de 2DLC. Y evidentemente a Francesc. Fuiste la primera persona que me tendió la mano en Barcelona, que me incluyó en sus planes, que me hizo querer la ciudad. Moltíssimes gràcies, sobre todo por tu ayuda y tus consejos durante todos estos años, incluso desde la distancia. Eres mi referente de lucha y superación.

To all international students and researchers that came to our lab during this period, thank you for all the memories shared, especially to Jamile, Mehrnoosh, Andrés, Athena and Flavia.

Quiero hacer un inestimable reconocimiento al Servicio de Espectrometría de Masas del IDAEA, a Roser, Alex y Dori. Sin vosotros, esta Tesis no hubiera sido posible, de manera literal. Muchísimas gracias por toda la paciencia y el apoyo durante estos años, por ayudarme a solucionar todos los problemas con los equipos, y por darme ánimos. Hacéis un trabajo esencial en el centro.

Durante la mitad de mi doctorado fui *pseudo*-miembro del grupo de Toxicología ambiental, de tantas horas que pasaba por la 5ª planta y en el Estabulario. Muchísimas gracias a Benjamí Piña y a Laia Navarro por haber hecho posible el proyecto con los peces cebra. Era un campo totalmente nuevo para mí y ha supuesto un gran reto, especialmente por el volumen tan grande del estudio y su complejidad, pero ha valido la pena. Marta, muchísimas gracias, no solo por enseñarme a hacer ensayos con levadura y peces, sino sobre todo por tu apoyo incondicional estos cinco años, por animarme cuando lo veía todo negro y por tu paciencia infinita. Claudia, mi compañera de sufrimiento *doctoril*, gracias por tantos momentos compartidos, por las horas en el Estabulario, y por contarme tus increíbles anécdotas en alta mar. También gracias a los antiguos miembros del grupo de Ecotox que me brindaron ayuda en mi etapa inicial, sobre todo a Inma, Anna y Eli. Y por supuesto, a Rubén. Has sido uno de los pilares fundamentales de esta Tesis. Mil gracias por todo. Por tu inestimable ayuda, por apoyarme incondicionalmente, por tratar de animarme en los momentos valle y compartir conmigo los momentos cresta, por tantos y tantos planes y experiencias.

Gracias también a Eva y Alejandro, del Estabulario, por su paciencia y apoyo con los experimentos con peces cebra.

Al increíble grupo de Comunicación del CID. Gracias por vuestra inestimable labor, muchas veces infravalorada. En especial, a Ale, mi amigo y vecino. Has sido un apoyo fundamental este último año. Mil gracias por echarme una mano con la portada de esta Tesis, por apuntarte a todos mis planes, por extravagantes que sean, y por nuestras largas conversaciones sobre cultura y psicología. Sin ti, mi salud mental este último año no hubiera sobrevivido.

Gracias a todos los profesores con los que compartí docencia estos años, porque todos me ayudasteis a ser mejor docente.

I would like to thank Dr. Dwight Stoll for giving me the chance to fulfill one of my dreams: doing a research stay in your lab, in the US. It was a very enriching experience, both from professional and

personal perspectives. Thanks to Gabriel Leme, for his patience, support and the friendship that we shared. Thanks to all the Gusties' students and professors I cross my path with, especially thanks to Tina, Amanda and my yoga mates.

Querría dar las gracias también al Departamento de Posgrado y Especialización del CSIC, por otorgarme el primer premio nacional por mi vídeo de divulgación, gracias al cual pude realizar mi segunda estancia predoctoral.

I would like to thank Dr. Craig Wheelock for allowing me to come to Sweden, to his lab in times of pandemic. It was a wonderful opportunity in all ways. Thank you for letting me be a part of BAMSE project for a month; it was enlightening. Thanks to all IMP members for welcoming me. Tack så mycket!! En especial a Toni Checa, por acogerme, por tu paciencia, por todo lo que me enseñaste, por animarme a redirigir mi camino sin miedo. Puedo decir que la estancia en Suecia marcó mi vida de muchas maneras, así que muchas gracias por haberla hecho posible. C'est quand même curieux que je sois allée en Suède et j'ai fini par parler plus de français que d'anglais. Merci Isabel, pour me découvrir MS-DIAL, le logiciel qui a sauvé ma thèse. Merci aussi pour nos promenades au bord des lacs, pour les excursions au centre-ville et pour nos réflexions sur comment continuer après le doctorat. Et sur tout, merci Johanna, pour faire de mon stage une de plus merveilleuses aventures de ma vie. Merci pour les pic-nics aux parcs, pour aller nager dans le lacs, pour faire du tourisme ensemble, pour me présenter à tes potes aussi, et pour me faire reconnecter avec la France.

A mis amigos. Al grupo de canarios de Barcelona, que dio un giro de 180º a mi vida aquí. A mi sevillana favorita, Esther. Ya sabes que en Barcelona y en Tenerife tienes casa. Y a mis amigos de Tenerife, en especial a Melissa, Diana y Ale, que han estado apoyándome durante todo este tiempo, escuchándome quejarme de la Tesis año tras año, y que me han animado a seguir adelante a pesar de todo.

Por supuesto, a mi familia. Mi gran apoyo incondicional a lo largo de mi vida. Gracias, gracias y gracias. Quisiera hacer una mención especial a mi tía Lucy y a mi tía Marisa. También a los que ya no están con nosotros, pero que estoy segura de que estarían muy orgullosos de lo que he conseguido.

Y, por último, a mis padres y mi hermana Carla, mi trípode. Ni esta Tesis (ni nada de lo que logrado en la vida) hubiera salido adelante si no los tuviera a ustedes conmigo, a mi lado, apoyándome SIEMPRE, en las buenas, pero sobre todo en las malas. Los quiero infinito. No tengo palabras para expresar lo importantes que son para mí.

This thesis was supported by a FPU predoctoral fellow from the Spanish Ministry of Universities (ref. FPU 16/02640), in addition to the grants CTQ2017-82598-P and CEX2018-000794-S funded by MCIN/AEI and the AGAUR Grant 2017SGR753.

INDEX

Abstract

Chemical exposure to emerging contaminants (ECs) is a major concern nowadays. These ECs have recently become a global environmental threat, and an in-depth characterization of their occurrence and toxic impact is needed. In this context, omic sciences have arisen as powerful tools to shed some light on the biological mechanisms affected by exposure to these chemicals. Particularly, metabolomics and lipidomics can provide a snapshot of what is actually happening at the molecular level, pointing to metabolic pathways affected by the contaminants. New analytical methodologies are required to extract the sought information in more complex biological matrices (from single cells to whole organisms). Hence, a major emphasis has been put on developing multidimensional separations and multiplatform approaches to increase the metabolome coverage. However, these novel approaches bring about massive datasets, and the complexity of the data analysis augments considerably. Therefore, chemometric strategies are a perfect match to get through this bottleneck and provide useful tools to obtain the most from the data collected.

In this PhD Thesis, the focus was set on developing analytical protocols, especially using two-dimensional liquid chromatography coupled to mass spectrometry (LC × LC-MS), as well as chemometric data analysis strategies applicable to environmental metabolomic studies. On the one hand, LC × LC-MS methods have been developed for both untargeted and targeted analyses. Active modulation strategies have been also successfully implemented in the multidimensional chromatographic separation of lipids. On the other hand, the Regions Of Interest (ROI) approach for compression and filtering has been validated for LC × LC-MS analyses. Regarding chemometric resolution methods (i.e., which allow obtaining quantitative and qualitative information from the sample constituents), and due to deviations from an ideal trilinear behavior presented by LC × LC datasets, the use of the Multivariate Curve Resolution Alternating Least Squares (MCR-ALS) method has been preferred. Different quantification strategies have been tested based on the Regions Of Interest Multivariate Curve Resolution (ROIMCR) approach. In addition, several multivariate statistical methods based on the analysis of variance (ANOVA) have been compared for metabolomic studies. As a result, a combination of ANOVA-simultaneous component analysis (ASCA) and partial least squares discriminant analysis (PLS-DA) has been selected for statistical analysis and variable (metabolite) selection, respectively. All in all, different metabolomic workflows have been validated for the assessment of emerging contaminants in model biosystems.

Resumen

Actualmente, una de las principales amenazas medioambientales globales reside en la exposición química a contaminantes emergentes, lo que hace necesaria una caracterización tanto de su presencia como de su toxicidad. En este contexto, las ciencias ómicas han aparecido como herramientas muy útiles para arrojar luz sobre los mecanismos biológicos y rutas metabólicas que se ven afectados debido a las exposiciones a estos compuestos. En concreto, la metabolómica y la lipidómica proporcionan información de lo que está ocurriendo a nivel molecular. Por tanto, se requieren nuevas metodologías que sean capaces de extraer dicha información en matrices cada vez más complejas (desde una única célula a un organismo entero). Por ello, estrategias como las separaciones multidimensionales o la combinación de diferentes plataformas se presentan como alternativas muy atractivas para ampliar la cobertura de los estudios actuales sobre el metaboloma. Sin embargo, estas nuevas metodologías llevan consigo un aumento en la complejidad y el tamaño de los datos a analizar, con lo que se necesita el uso de herramientas de análisis más potentes, como las basadas en la quimiometría.

Esta Tesis doctoral se centra principalmente en el desarrollo de protocolos analíticos basados en el uso de la cromatografía líquida bidimensional acoplada a espectrometría de masas (LC × LC-MS), así como en el desarrollo de estrategias quimiométricas que permitan su uso en aplicaciones medioambientales. Por un lado, se han optimizado métodos LC × LC-MS para análisis dirigidos y no dirigidos, implementando con éxito estrategias de modulación activa (en el caso de análisis de lípidos). Por otro lado, se ha validado la estrategia de regiones de interés (ROI) para comprimir y filtrar los datos obtenidos con LC × LC-MS. Asimismo, se ha preferido el uso de métodos de resolución multivariante de curvas mediante mínimos cuadrados alternados (MCR-ALS) para la resolución cualitativa y cuantitativa de muestras complejas en el caso de datos de LC × LC, debido a las deviaciones de la trilinealidad encontradas en dichos datos. Por otra parte, se han comparado diferentes estrategias cuantitativas aplicables a datos en LC × LC-MS, todas ellas basadas en el uso del método combinado de regiones de interés y resolución multivariante de curvas (ROIMCR). También se han comparado diversos métodos estadísticos multivariante basados en el análisis de varianza (ANOVA) y su aplicabilidad en estudios metabolómicos. Finalmente, se ha elegido una combinación de análisis estadístico efectuado con ANOVA-análisis de componentes simultáneos (ASCA) y un método de clasificación, análisis discriminante mínimos cuadrados parciales (PLS-DA) para seleccionar las variables (metabolitos) más relevantes. En resumen, se han validado diferentes flujos de trabajo para el estudio metabolómico del efecto de contaminantes emergentes en organismos modelo ambientales.

Acronyms

^{1}D	First-dimension
1D	One-dimensional
1DLC	One-dimensional liquid chromatography
^{2}D	Second-dimension
2D	Two-dimensional
2DLC	Two-dimensional liquid chromatography
3D	Three-dimensional
ACD	At-column dilution
Ag	Argentation
AIF	All ion fragmentation
ALS	Alternating least squares
AMOX	Amoxicillin
ANOVA	Analysis of variance
APCI	Atmospheric pressure chemical ionization
API	Atmospheric pressure ionization
APPI	Atmospheric pressure photoionization ionization
ASCA	Anova-simultaneous component analysis
ASM	Active solvent modulation
ATLD	Alternating trilinear decomposition
B	Magnetic sector
BD	Bligh and Dyer
BHT	Butylated hydroxytoluene
BPA	Bisphenol A
BPC	Base peak chromatograms
C	Control samples
CBZ	Carbamazepine
CE	Capillary electrophoresis
CE-MS	Capillary electrophoresis mass spectrometry
Cer	Ceramides
Ch4EO	Chemometrics for environmental omics group
CHEMAGEB	Chemometric and high-throughput omics analytical methods for assessment of global change effects on environmental and biological systems
COShift	Correlation-optimized shifting
COW	2D correlation optimized warping
CSS	Collision cross-section
CV	Coefficient of variance
DAD	Diode array detector
DAG	Diacylglycerol
DB	Database
DDA	Data dependendt acquistion
DESI	Desorption electrospray ionization mass spectrometry
DGDG	Digalactosyldiacylglycerol
DhCer	Dihydroceramides
DIA	Data independent acquisition

DNA	Deoxyribonucleic acid
DOE	Design of the experiment
dpf	Days post-fertilization
E	Electric sector
E2	17-β-estradiol
E2	Estradiol
EC	Emerging contaminants
ECHA	European chemical agency
EDC	Endocrine disrupting chemicals
EIC	Extracted ion chromatogram
ESI	Electrospray
FA	Fatty acyls
FC	Fold-changes
FDA	Food and drug administration
FDR	False discovery rate
FN	False negatives
FP	False postives
FSM	Fixed solvent modulation
FTICR	Fourier transform ion cylotron resonance
FT-IR	Fourier transform infrared spectroscopy
FT-OT	Fourier transform orbitrap
GASCA	Group-wise ANOVA simultaneous component analysis
GB	Gigabytes
GC	Gas chromatography
GC× GC	Comprehensive two-dimensional gas chromatography
GC-MS	Gas chromatography coupled to mass spectrometry
GL	Glycerolipids
GlucCer	Glucosylceramides
GNPS	Global natural product social molecular networking
GP	Glycerophospholipids
GUI	Graphical user interface
GWPCA	Group-wise principal component analysis
H	High dose
HCA	Hierarchical clustering analysis
HegG2	Human hepatocellular carcinoma cell line
HILIC	Hydrophilic interaction chromatography
HMDB	Human metabolome database
hpf	Hours post-fertilization
HPLC	High performance liquid chromatography
HRMS	High resolution mass spectrometry
IDAEA	Environmental assessment and water research Institute
IEX	Ion exchange
IM	Ion mobility
IM-MS	Ion mobility coupled to mass spectrometry
IP	Ion paring chromatograhy
IPA	Isopropanol

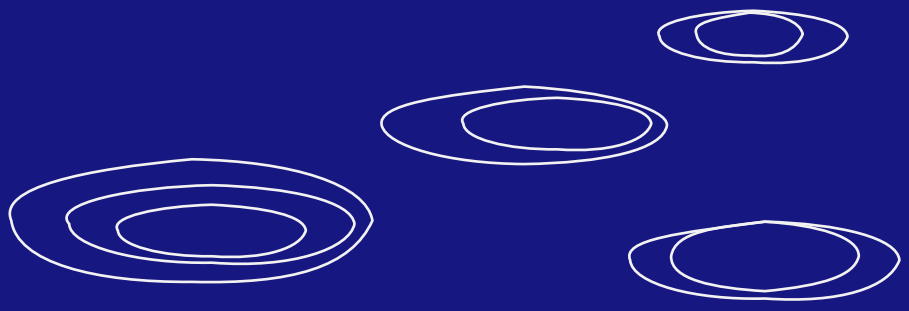
IS	Internal standard
IT	Ion trap
KEGG	Kyoto encyclopedia of genes and genome
L	Low dose
LC	Liquid chromatography
LC × LC	Comprehensive two-dimensional liquid chromatography
LC × LC-HRMS	Comprenhensive two-dimensional liquid chromatography coupled to high resolution mass spectrometry
LC × LC-MS	Comprenhensive two-dimensional liquid chromatography coupled to mass spectrometry
LC-HRMS	Liquid chromatography coupled to high resolution mass spectrometry
LC-LC	Multiple heart-cutting two-dimensional liquid chromatography
LC-MS	Liquid chromatography coupled to mass spectrometry
LD50	Median lethal dose
LLE	Liquid-liquid extraction
LOF	Lack of fit
LPC	Lysophosphatidylcholine
LPE	Lysophosphatidylethanolamine
LPG	Lysophosphatidylglycerol
LV	Latent variables
M	Medium dose
m/z	Mass-to-charge
MALDI	Matrix-assisted laser desorption ionization
MANOVA	Multivariate ANOVA
MCC	Matthews correlation coefficient
MCR	Multivariate curve resolution alternating least squares
MCR-ALS	Multivariate curve resolution
MeOH	Methanol
MG	Megabytes
MG	Monoacylglycerols
MGDG	Monogalactosyldiacylglycerol
mLC-LC	Heart-cutting two-dimensional liquid chromatography
MM	Mixed mode
MOM	Model organism metabolomes
MOREDISTRIBUTIONS	Multivariate and otherwise rapid and efficient determination and identification software for thorough representation and interpretation by unveiling traits informing on novel synthetics
MOREPEAKS	Multivariate optimization and refinement program for efficient analysis of key separations
MRM	Multiple reaction monitoring
MS	Mass spectrometry
MS/MS	MS fragmentation (MS2 level)
MS^n	Iterative mass fragmentation
MTBE	Methyl tert-buthyl ether
NAMs	New approach methodologies
NIMS	Nanostructure-imaging mass spectrometry
NMR	Nuclear magnetic resonance
NP	Normal phase
PA	Phosphophatidic acids

PARAFAC	Parallel factor analysis
PARAFAC2	Parallel factor analysis 2
PC	Phosphatidylcholine
PC	Phosphatidylcholine
PCA	Principal component analysis
PC-DFA	Principal component-discriminant function znalysis
PCR	Polymerase chain reaction
PDA	Photodiode array
PE	Phosphatidylethanolamine
PE	Phosphoethanolamine
PG	Phosphatidylglycerol
PG	Phosphoglycerols
PGP	Phosphatidylglycerophosphate
PI	Phosphoinositiols
PK	Polyketides
PLS	Partial least squares
PLS-DA	Partial least squares discriminant analysis
PQN	Probabilistic quotient normalization
PR	Prenol lipids
PS	Phosphoserines
Q	Quadrupole
QA	Quality assurances
QC	Quality controls
QqQ	Triple quadrupole detector
QTOF	Quadrupole-time of flight
RE	Relative errors
RF	Random forests
rMANOVA	Regularized manova
ROI	Regions of interest
ROIMCR	Regions of interest multivariate curve resolution
RP	Reversed phase
RT	Retention time
S/N	Signal-to-Noise ratio
SCA	Simultaneous component analysis
SEC	Size exclusion chromatography
SFC	Supercritical fluid chromatography
SIMPLISMA	Simple-to-use iterative self-modeling mixture analysis
SIMS	Secondary ion mass spectrometry
SL	Saccharolipids
sLC × LC	Selective comprehensive two-dimensional liquid chromatography
SM	Sphingomyelin
SML	Specific migration limit
SP	Sphingolipids
SQDG	Sulfoquinovosyldiacylglycerols
SRM	Selected reaction monitoring
ST	Sterol lipids

SVD	Singular value decomposition
SVM	Support vector machines
SWATH	Sequential window acquisition of all theoretical fragment-ion spectra
TAG	Triacylglycerol
TDI	Tolerable daily intake
TIC	Total ion chromatogram
TIMS-TOF-MS	Trapped Ion Mobility Spectrometry coupled to time of flight mass spectrometry
TN	True negatives
TOF	Time of flight
TP	True positives
TRZ	Trazodone
UB	Univeristy of Barcelona
UHPLC	Ultra high performance liquid chromatography
UV	Ultraviolet-Visible spectroscopy
VIP	Variable important in projection
w4m	Workflow4Metabolomics
WAX	Weak anion exchange
WHO	World health organization
WWTP	Wastewater treatment plants
YMDB	Yeast metabolome database
ZFIN	Zebrafish information network

Chapter one

Context, goals and structure of the Thesis



1.1 Context

Chemical exposure to emerging contaminants (ECs) and assessment of their effects on biological organisms are major concerns nowadays. ECs are a global environmental threat which require their in-depth characterization and evaluation of their occurrence and impact. There is an urgent need to unravel their effects and mode of action to evaluate their potential risk for both, the health of the different ecosystems and the human health. In this context, omic sciences have arisen as powerful tools to shed some light on the biological mechanisms affected by exposure to these chemicals. Particularly, metabolomics and lipidomics can provide a snapshot of what is actually happening at the molecular level, such as the understanding of what metabolic pathways are affected by the contaminants. New approach methodologies (NAMs), designed to protect humans and the environment, are being used to extract the sought information in complex biological matrices (from single cells to whole organisms). In this PhD Thesis different multidimensional separation methods have been proposed and developed to increase the metabolome coverage, such as the online comprehensive two-dimensional liquid chromatography coupled to mass spectrometry (LC×LC-MS) methodology. These novel approaches bring associated huge datasets, whose complexity challenges their analysis considerably. It is in this aspect that chemometric strategies are useful to get through this data processing bottleneck providing powerful tools to obtain the most from the data collected.

Hence, this PhD Thesis agrees with this environmental context briefly described above and with two of the research lines of the Chemometrics for Environmental Omics research group (Ch4EO) from the Environmental Assessment and Water Research Institute (IDAEA). These research lines are:

1) The development of chemometric strategies to analyze datasets from multiple analytical platforms,
2) The application of these chemometric methods to assess the effects of chemical pollutants on model organisms at molecular level (omics).

This PhD Thesis has beneficiated from the previous work of the research group which was awarded with the CHEMAGEB (CHEMometric and High-throughput

Omics Analytical Methods for Assessment of Global Change Effects on Environmental and Biological Systems) ERC Advanced Grant. The main goal of this project was the development of new analytical and chemometric methods to assess the effects of pollution and climate change in model organisms, representative of ecosystems. Several omic levels were studied, including genomics, transcriptomics and metabolomics (and lipidomics), in different model biosystems such as yeast, *Saccharomyces cerevisiae* [1,2]), zebrafish, *Danio rerio* [3,4], water flea, *Daphnia magna* [5,6], rice, *Oryza sativa* L. [7–9], and human cell lines [10,11]. The effect of environmental stressors such as chemicals (heavy metals [12,13], endocrine disrupting chemicals [4,14], pesticides [15,16] or pharmaceutical compounds [17,18]) and physical impacts (temperature [19,20], hydric stress [9] or UV light [10,21]) on the previously mentioned biosystems were investigated. The bioanalytical techniques employed throughout this project were genomic DNA-microarray chips, RNA-sequencing, nuclear magnetic resonance and liquid and gas chromatography in one and two dimensions coupled to high-resolution mass spectrometry. Several chemometric and multivariate data analysis tools were developed and tested in different omic studies and toxicological assessments.

Considering this previously acquired environmental omics experience, as well as the historical precedents and research lines of the group, the goals of this PhD Thesis are listed below.

1.2 Goals

Metabolomics is the analytical approach selected in this PhD Thesis for studying the effects of different pollutants in model biosystems. The major emphasis has been put on developing analytical protocols based on mass spectrometry, especially related to two-dimensional liquid chromatography mass spectrometry, and to data analysis strategies applicable to environmental metabolomic studies.

Consequently, the **main goal** of this PhD Thesis has been to **develop** and **optimize new analytical methodologies** employing **liquid chromatography** and

mass spectrometry, as well as new **chemometric** strategies to extract relevant environmental and metabolic information about the exposure of emerging pollutants in model biosystems.

This main goal can be divided into specific objectives classified according to their field of study:

Analytical goals

- Development of two-dimensional liquid chromatography coupled to high-resolution mass spectrometry (LC × LC-HRMS) methodology for untargeted lipidomics analysis.
- Development of LC × LC-MS methods for the analysis of small and polar metabolites.
- Assessment of new active modulation strategies for their use in LC × LC-MS metabolomics (and lipidomics) studies.
- Validation of different metabolomic workflows for assessing the impact of environmental stressors (i.e., emerging contaminants) in model biosystems.

Chemometric goals

- Assessment of the multiway data structure and multilinear behavior of LC × LC-UV-MS datasets and development of new data fusion strategies to combine the information provided by both detectors.
- Comparison of ANOVA-based multivariate approaches and their suitability in untargeted metabolomic studies.
- Validation of the spectral compression strategy based on the regions of interest (ROI) approach for LC × LC-MS datasets.
- Comparison of different quantification strategies based on the regions of interest and multivariate curve resolution (ROIMCR) for targeted LC × LC-MS analyses.

1.3 Structure

This PhD Thesis is divided into six Chapters, distributed in three main sections. The first section (Chapters one and two) is an introductory section related with the structure and general description of the different PhD Thesis topics. The second section (Chapters three to five) presents the main work of this PhD Thesis, classified according to pursued goals: chemometric evaluations and results (Chapter three), analytical developments and applications (Chapter four), and evaluation of metabolomic workflows (Chapter five). Finally, the third section (Chapter six) ends with a general conclusions section and final remarks. Additionally, there is one final Chapter (Chapter seven) that includes the Annexes of the PhD Thesis. The content of each of the main six Chapters is detailed below.

The **first Chapter** (current Chapter) summarizes the context and the previous work of the research group, the goals and structure of the PhD Thesis and the list of scientific publications that have resulted from that work.

The **second Chapter** is an introduction to the omics research field. This Chapter introduces the proposed environmental metabolomics workflow, including the studied model organisms and environmental stressors, and the analytical techniques and data analysis strategies applied throughout this PhD Thesis. There is an especial emphasis on the description of the development of the LC×LC-MS analytical methodology using active modulation strategies, and data processing workflows employing the regions of interest multivariate curve resolution (ROIMCR) method.

The **third Chapter** proposes different chemometric tools for specific bottlenecks commonly encountered in metabolomics data analysis and offers some examples of applications. The multiway structure and multilinear behavior of the LC×LC-MS datasets are studied, to propose the optimal approach for the proper resolution of this type of data sets. Information provided by two detectors, mass spectrometry (MS) and ultraviolet-visible spectroscopy (UV) is simultaneously analyzed using a new data fusion strategy. Three multivariate ANOVA-based strategies are evaluated and compared for the statistical analysis of metabolomic datasets. The most appropriate chemometric pipelines among the tested approaches are selected for the studies presented in the following Chapters.

The **fourth Chapter** focuses on the development and application of LC×LC-MS analytical methodology combined with a spectral compression based on the regions of interest (ROI). The optimization of the proposed LC×LC-MS method for the analysis of lipids and metabolites is presented. In addition, the untargeted LC×LC-HRMS method using active solvent modulation strategy (ASM) is developed and optimized for lipidomic studies. This untargeted approach is validated *via* a previously developed targeted approach performed on the analysis of the same samples. The proposed analytical methods and data processing approaches are applied in the assessment of the effects of two endocrine disrupting chemicals (EDCs): bisphenol A (BPA) and estradiol (E2) on the lipidome of zebrafish (*Danio rerio*) embryos. Finally, another proposed LC×LC-MS method has been also optimized for the targeted analysis of amino acids and the results were used to compare different quantification strategies using the ROIMCR approach.

The **fifth Chapter** describes different metabolomic workflows employed for the evaluation of emerging pollutants in model biosystems. Experimental studies in this Chapter were performed using metabolomic and lipidomic platforms that employed liquid chromatography coupled to mass spectrometry (LC-MS). The first study evaluates the arsenic uptake in rice (*Oryza sativa* L.) by comparing two different exposure routes (watering and soil) by means of untargeted metabolomic and lipidomic approaches. The second application is focused on the consequences of the exposure of pharmaceutical compounds on human hepatic cells (HepG2 cell line) at low doses, mimicking environmental concentration levels. A combination of targeted and pseudo-targeted methods has been used. The main advantages and limitations of the data analysis workflows employed in both studies are compared. Finally, the advantages and disadvantages of using targeted *versus* untargeted approaches are also briefly discussed.

The **sixth Chapter** gathers the most relevant conclusions of this PhD Thesis.

1.4 List of scientific publications of this PhD Thesis

The research carried out in this PhD Thesis has brought forth the following scientific publications:

I. SCIENTIFIC PUBLICATION I

Title: Two-Dimensional Liquid Chromatography in Metabolomics and Lipidomics

Authors: **Miriam Pérez-Cova**, Romà Tauler, Joaquim Jaumot

Citation reference: Wood P.L. (eds) Metabolomics. Neuromethods, vol 159.

DOI: 10.1007/978-1-0716-0864-7_3

II. SCIENTIFIC PUBLICATION II

Title: Untangling comprehensive two-dimensional liquid chromatography data sets using regions of interest and multivariate curve resolution approaches

Authors: **Miriam Pérez-Cova**, Joaquim Jaumot, Romà Tauler

Citation reference: Trends in Analytical Chemistry 137 (2021) 1162072

DOI: 10.1016/j.trac.2021.116207

III. SCIENTIFIC PUBLICATION III

Title: Chemometrics in comprehensive two-dimensional liquid chromatography: A study of the data structure and its multilinear behavior

Authors: **Miriam Pérez-Cova**, Romà Tauler, Joaquim Jaumot

Citation reference: Chemometrics and Intelligent Laboratory Systems 201 (2020) 104009

DOI: 10.1016/j.chemolab.2020.104009

IV. SCIENTIFIC PUBLICATION IV

Title: Comparison of multivariate ANOVA-based approaches for the determination of relevant variables in experimentally designed metabolomic studies

Authors: **Miriam Pérez-Cova**, Stefan Platikanov, Dwight R. Stoll, Romà Tauler, Joaquim Jaumot

Citation reference: Molecules 27 (2022), 3304

DOI: 10.3390/molecules27103304

V. SCIENTIFIC PUBLICATION V

Title: Untargeted lipidomics of zebrafish (*Danio rerio*) eleutheroembryos exposed to endocrine disrupting chemicals using comprehensive two-dimensional liquid chromatography and advanced chemometrics

Authors: **Miriam Pérez-Cova**, Laia Navarro-Martin, Gabriel Leme, Romà Tauler, Benjamin Piña, Joaquim Jaumot, Dwight R. Stoll

In preparation

VI. SCIENTIFIC PUBLICATION VI

Title: Quantification strategies for two-dimensional liquid chromatography datasets using regions of interest and multivariate curve resolution approaches

Authors: **Miriam Pérez-Cova**, Stefan Platikanov, Romà Tauler, Joaquim Jaumot

Citation reference: Talanta 247 (2022) 123586.

DOI: 10.1016/j.talanta.2022.123586

VII. SCIENTIFIC PUBLICATION VII

Title: Adverse Effects of Arsenic Uptake in Rice Metabolome and Lipidome Revealed by Untargeted Liquid Chromatography Coupled to Mass Spectrometry (LC-MS) and Regions of Interest Multivariate Curve Resolution

Authors: **Miriam Pérez-Cova**, Romà Tauler and Joaquim Jaumot

Citation reference: Separations 9 (2022) 79.

DOI: 10.3390/separations9030079

VIII. SCIENTIFIC PUBLICATION VIII

Title: Metabolomics and sphingolipidomics study of human hepatoma cells exposed to environmental concentrations of pharmaceutical compounds

Authors: **Miriam Pérez-Cova**, Carmen Bedia, Antonio Checa, Isabel Meister, Romà Tauler, Craig E Wheelock, Joaquim Jaumot

To be submitted - June 2022

References

[1] F. Puig-Castellví, I. Alfonso, B. Pinã, R. Tauler, 1H NMR metabolomic study of auxotrophic starvation in yeast using Multivariate Curve Resolution-Alternating Least Squares for Pathway Analysis, Scientific Reports. 6 (2016) 1–12. https://doi.org/10.1038/srep30982.

[2] F. Puig-Castellví, I. Alfonso, B. Piña, R. Tauler, A quantitative 1H NMR approach for evaluating the metabolic response of Saccharomyces cerevisiae to mild heat stress, Metabolomics. 11 (2015) 1612–1625. https://doi.org/10.1007/s11306-015-0812-9.

[3] E. Gorrochategui, J. Li, N.J. Fullwood, G.G. Ying, M. Tian, L. Cui, H. Shen, S. Lacorte, R. Tauler, F.L. Martin, Diet-sourced carbon-based nanoparticles induce lipid alterations in tissues of zebrafish (Danio rerio) with genomic hypermethylation changes in brain, Mutagenesis. 32 (2017) 91–103. https://doi.org/10.1093/mutage/gew050.

[4] E. Ortiz-Villanueva, L. Navarro-Martín, J. Jaumot, F. Benavente, V. Sanz-Nebot, B. Piña, R. Tauler, Metabolic disruption of zebrafish (Danio rerio) embryos by bisphenol A. An integrated metabolomic and transcriptomic approach, Environmental Pollution. 231 (2017) 22–36. https://doi.org/10.1016/J.ENVPOL.2017.07.095.

[5] R. Jordaõ, B. Campos, B. Piña, R. Tauler, A.M.V.M. Soares, C. Barata, Mechanisms of action of compounds that enhance storage lipid accumulation in daphnia magna, Environmental Science and Technology. 50 (2016) 13565–13573. https://doi.org/10.1021/ACS.EST.6B04768/ASSET/IMAGES/LARGE/ES-2016-04768S_0007.JPEG.

[6] B. Campos, D. Fletcher, B. Piña, R. Tauler, C. Barata, Differential gene transcription across the life cycle in Daphnia magna using a new all genome custom-made microarray, BMC Genomics. 19 (2018). https://doi.org/10.1186/S12864-018-4725-7.

[7] M. Navarro-Reig, J. Jaumot, R. Tauler, An untargeted lipidomic strategy combining comprehensive two-dimensional liquid chromatography and chemometric analysis, Journal of Chromatography A. 1568 (2018) 80–90. https://doi.org/10.1016/j.chroma.2018.07.017.

[8] M. Navarro-Reig, J. Jaumot, A. García-Reiriz, R. Tauler, Evaluation of changes induced in rice metabolome by Cd and Cu exposure using LC-MS with XCMS and MCR-ALS data analysis strategies, Analytical and Bioanalytical Chemistry. (2015). https://doi.org/10.1007/s00216-015-9042-2.

[9] M. Navarro-Reig, R. Tauler, G. Iriondo-Frias, J. Jaumot, Untargeted lipidomic evaluation of hydric and heat stresses on rice growth, Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences. 1104 (2019) 148–156. https://doi.org/10.1016/j.jchromb.2018.11.018.

[10] N. Dalmau, N. Andrieu-Abadie, R. Tauler, C. Bedia, Phenotypic and lipidomic characterization of primary human epidermal keratinocytes exposed to simulated solar UV radiation, Journal of Dermatological Science. 92 (2018) 97–105. https://doi.org/10.1016/j.jdermsci.2018.07.002.

[11] E. Gorrochategui, J. Casas, C. Porte, S. Lacorte, R. Tauler, Chemometric strategy for untargeted lipidomics: Biomarker detection and identification in stressed human placental cells, Analytica Chimica Acta. 854 (2015) 20–33. https://doi.org/10.1016/j.aca.2014.11.010.

[12] M. Farré, B. Piñ, R. Tauler, LC-MS based metabolomics and chemometrics study of the toxic effects of copper on Saccharomyces cerevisiae, | Metallomics. 8 (2016) 790. https://doi.org/10.1039/c6mt00021e.

[13] M. Navarro-Reig, J. Jaumot, B. Piña, E. Moyano, M.T. Galceran, R. Tauler, Metabolomic analysis of the effects of cadmium and copper treatment in: Oryza sativa L. using untargeted liquid chromatography coupled to high resolution mass spectrometry and all-ion fragmentation, Metallomics. 9 (2017) 660–675. https://doi.org/10.1039/c6mt00279j.

[14] E. Ortiz-Villanueva, J. Jaumot, R. Martínez, L. Navarro-Martín, B. Piña, R. Tauler, Assessment of endocrine disruptors effects on zebrafish (Danio rerio) embryos by untargeted LC-HRMS metabolomic analysis, Science of the Total Environment. 635 (2018) 156–166. https://doi.org/10.1016/j.scitotenv.2018.03.369.

[15] C. Gómez-Canela, D. Tornero-Cañadas, E. Prats, B. Piña, R. Tauler, D. Raldúa, Comprehensive characterization of neurochemicals in three zebrafish chemical models of human acute organophosphorus poisoning using liquid chromatography-tandem mass spectrometry, Analytical and Bioanalytical Chemistry. 410 (2018) 1735–1748. https://doi.org/10.1007/s00216-017-0827-3.

[16] V. Olmos, C. Bedia, R. Tauler, A. de Juan, Preprocessing Tools Applied to Improve the Assessment of Aldrin Effects on Prostate Cancer Cells Using Raman Spectroscopy, Applied Spectroscopy. 72 (2018) 489–500. https://doi.org/10.1177/0003702817746947.

[17] E. Garreta-Lara, A. Checa, D. Fuchs, R. Tauler, S. Lacorte, C.E. Wheelock, C. Barata, Effect of psychiatric drugs on Daphnia magna oxylipin profiles, (2018). https://doi.org/10.1016/j.scitotenv.2018.06.333.

[18] C. Gómez-Canela, T.H. Miller, N.R. Bury, R. Tauler, L.P. Barron, Targeted metabolomics of Gammarus pulex following controlled exposures to selected pharmaceuticals in water, Science of The Total Environment. 562 (2016) 777–788. https://doi.org/10.1016/J.SCITOTENV.2016.03.181.

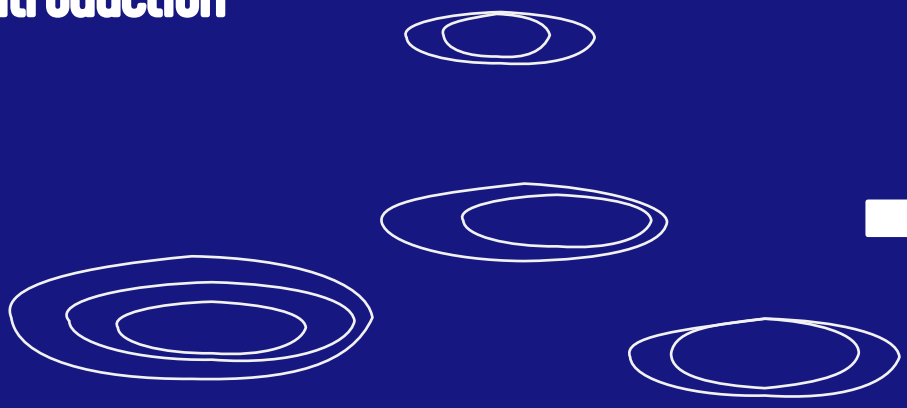
[19] F. Puig-Castellví, C. Bedia, I. Alfonso, B. Piña, R. Tauler, Deciphering the Underlying Metabolomic and Lipidomic Patterns Linked to Thermal Acclimation in Saccharomyces cerevisiae, Journal of Proteome Research. 17 (2018) 2034–2044. https://doi.org/10.1021/acs.jproteome.7b00921.

[20] E. Garreta-Lara, B. Campos, C. Barata, S. Lacorte, R. Tauler, Combined effects of salinity, temperature and hypoxia on Daphnia magna metabolism, Science of the Total Environment. 610–611 (2018) 602–612. https://doi.org/10.1016/j.scitotenv.2017.05.190.

[21] N. Dalmau, N. Andrieu-Abadie, R. Tauler, C. Bedia, Untargeted lipidomic analysis of primary human epidermal melanocytes acutely and chronically exposed to UV radiation, Molecular Omics. 14 (2018) 170–180. https://doi.org/10.1039/c8mo00060c.

Chapter two

Introduction



2.1 Metabolomics and environmental assessments

Omic sciences have attracted the attention of the scientific community in recent years, with many applications yet to explore. These applications vary from personalized medicine and new treatments for non-curable diseases to improvements in food quality or evaluation of ecological risks. This PhD Thesis aims to expand the horizons in the environmental field, especially in assessing the effects caused by emerging contaminants (ECs). Among omics, metabolomics has experienced an important growth from the point of view of cutting-edge methodology development and implementation. However, its full potential has not been reached yet; there is still a long way ahead. In this section, metabolomics is introduced in the omic context. Then, the usefulness of metabolomics in environmental assessments, as well as the environmental metabolomic workflow, are discussed. Finally, the model biosystems and emerging contaminants employed in this PhD Thesis are presented as study cases in environmental metabolomics.

2.1.1 Metabolomics at the heart of omics technologies

'Metabolite' is a broad term applied to a small molecule considered intermediate or end product of cellular regulatory processes. When referring to metabolites, it comprehends a heterogeneous group of compounds with diverse physicochemical properties and biological roles. These low molecular weight biochemicals (< 2000 Da) include peptides/aminoacids, lipids, carbohydrates, and nucleotides/nucleosides. The molecules from these main four classes are also known as primary metabolites. Moreover, metabolic intermediates, signaling molecules, (e.g., hormones), and secondary metabolites (e.g., flavonoids and alkaloids) can also be considered metabolites. The exact number of existing metabolites is still unknown. Estimations in humans can range from 2000 to 3000 or more, depending on the biofluid and/or part of the body selected for the analysis [1]. In contrast, more than 20000 are expected to be present in other organisms, such as plants, where many specific secondary metabolites are involved [2].

Metabolomics is the field that aims for the comprehensive detection and quantification of as many metabolites as possible in a biological system, i.e., the complete metabolomic profile or **metabolome**. It is an analogous term to genomics, transcriptomics, and proteomics, these focused on genes, transcripts, and proteins, respectively. All four are considered the main omics technologies and constitute the

so-called 'omic cascade'. As shown in **Figure 2.1**, metabolomics lies downstream and ultimately links genotype with phenotype.

According to the type of metabolites studied, there are different subclassifications among metabolomics, (e.g., lipidomics for lipids or glycomics for carbohydrates).

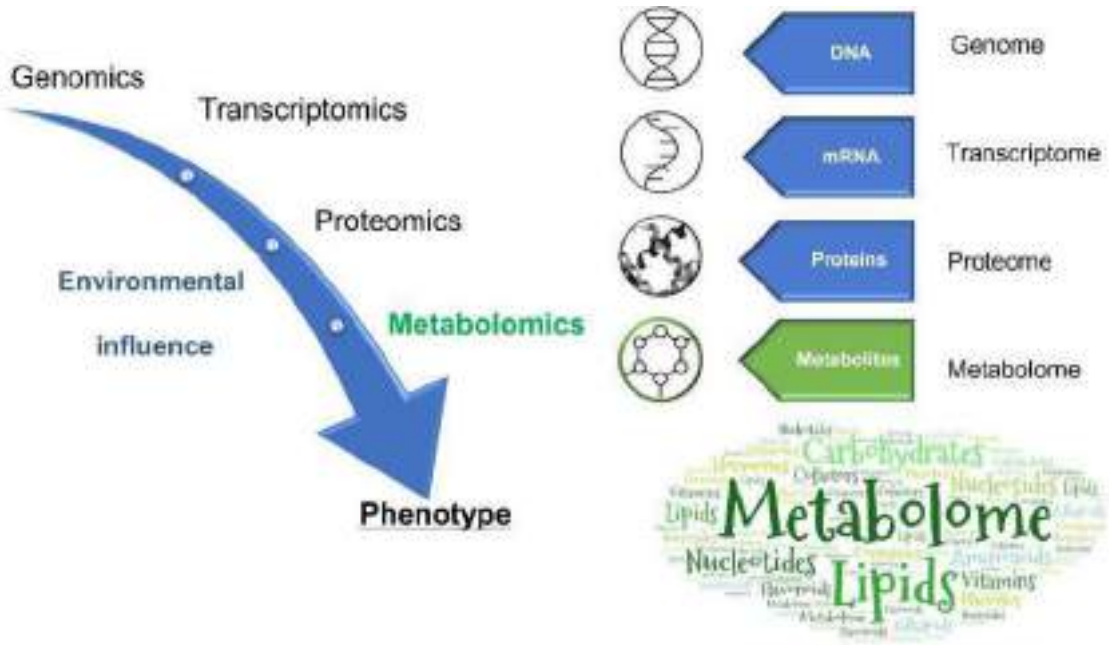


Figure 2.1. Scheme of the 'omic cascade', from genomics to metabolomics.

Compared to the other principal omics, metabolomics is the newest and presents its own characteristics. On the one hand, it is a dynamic approach that allows a snapshot of what is happening at cellular level. Therefore, it represents a blueprint of the molecular phenotype [3]. Much of the cellular activity occurs at the metabolic level, including energy storage and transfer, signaling, and cell to cell communication. Thus, it can be directly affected by nutrition, environment, and many more exogenous factors [1].

On the other hand, the metabolome is highly conserved across biology. Hence, if other organisms are used for the studies, the information about metabolic alterations can be translated to humans. Analytical methods can also be easily adapted from different biological systems, simplifying the experiment by reducing costs and optimization time [4]. In addition, metabolomics in clinical applications can be non-invasive (i.e., a sample volume of biofluids can be relatively low), and periodically sampling can be performed over time.

Since metabolomics is a rapid indicator of metabolic perturbations, its applications are broad and continuously augmenting nowadays. There is an important branch of metabolomics focused on health and pharmacology. Some examples are the discovery of biomarkers of diseased conditions for therapeutic purposes [5], personalized medicine [6], indicators of drug abuse or intoxication [7], or drug effectiveness, useful in drug discovery and toxicological assessments [8]. There is a whole new omic science called exposomics that considers all sources of exposure (especially both endogenous and exogenous chemicals) with the aim of linking them to adverse health outcomes. With the aid of metabolomics, it is possible to discover biomarkers related to environmental exposure and the apparition of certain diseases or medical conditions [9].

Metabolomics is the most sensitive omic to external stressors; therefore, also the quickest to show changes due to exposure. Thus, metabolomics is used to study the effects of environmental stressors on wildlife. Likewise, mimicking natural conditions in a laboratory instead (as in this case) can also provide valuable information of their mode of action [10]. **Environmental metabolomics** comprises all studies assessing organism-environment interactions to characterize molecular processes involved, in the context of evaluating environmental health [11]. From a global perspective, it includes the discovery of markers of natural or anthropogenic stressors in the environment [12]. More specifically, potential markers of emerging contaminants exposure can be discovered, and the mode of action of these compounds can be elucidated [9]. This PhD Thesis is framed in this last group.

2.1.2 Environmental metabolomic workflow

An appropriate experimental design is critical prior to any environmental metabolomic study. It includes four key aspects: model organism, type of external stressor, mode or route of exposure, and tissue/organ/biofluid of analysis [10]. **Figure 2.2** summarizes the usual metabolomic workflow, emphasizing the experimental design steps.

The first step of the metabolomic workflow is whether the **hypothesis** answers a specific question or obtains a general overview of what is happening at molecular level. The whole analytical approach employed will be conditioned by this decision. There are two main ways of action: pre-selecting the metabolites of interest in advance (**targeted approach**) or performing a screening of different families of metabolites at once without *a priori* assumptions (**untargeted approach**). The

differences between these approaches can be found in **Section 2.2.1 Targeted versus untargeted analysis**.

Steps one to three of the workflow, i.e., the initial **hypothesis**, the choice of **biosystem** and **external stressor**, are highly correlated. For instance, if the aim is to study the effect of hepatotoxic compounds on signaling molecules, the assessment can be performed by analyzing sphingolipids on hepatic cell lines or the whole liver. More complex models (e.g., whole organisms) will provide more biological information, but also will require more resources and time, and biological interpretation is more complex.

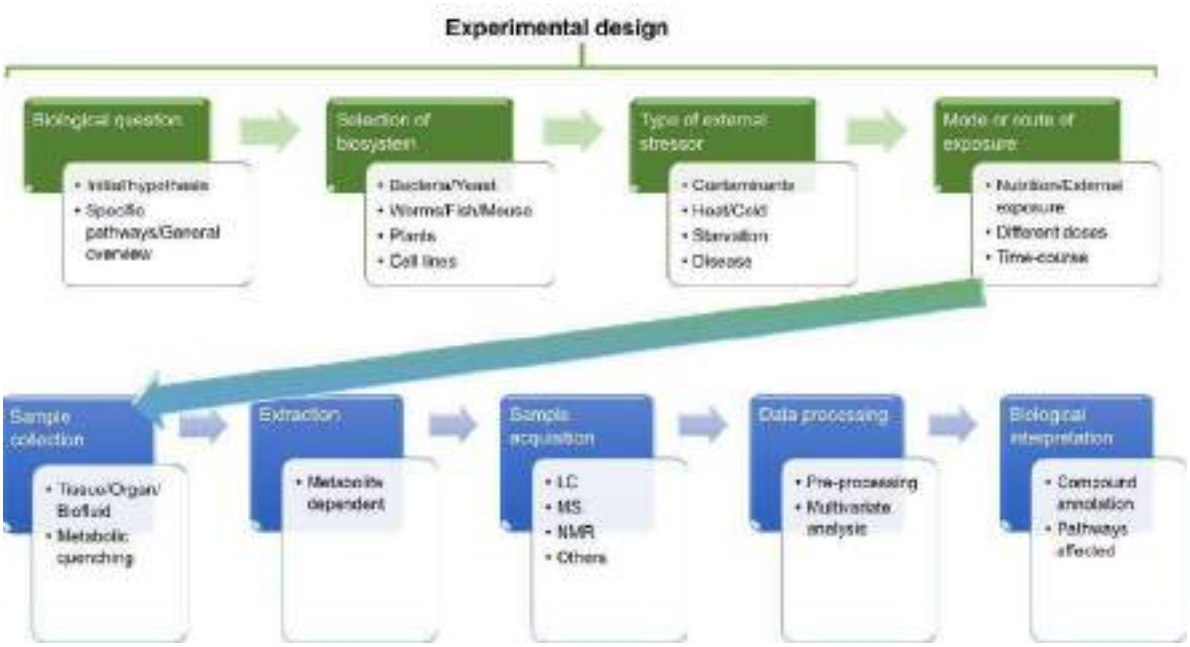


Figure 2.2. Common steps of the metabolomic workflow, highlighting the most important decisions when designing the metabolomic experiment. LC: Liquid Chromatography; MS: Mass Spectrometry; NMR: Nuclear Magnetic Resonance. *Adapted from* [10].

In metabolomic studies, there is usually a comparison between a minimum of two groups: control (or healthy) and exposed (or diseased). Then, the design can include different **routes and modes of exposure**, multiple doses (several concentration levels of exposure for the same compound), or time-course experiments (measurements of the same individual to evaluate effects over time).

After exposure, several steps are directly related to the analysis itself, starting with **sample collection.** The protocol will depend on the specific type of tissue, organism or biofluid (e.g., urine, blood, saliva). In this step, metabolic quenching is

critical. The goal is to suddenly stop metabolism and ensure fast and reliable metabolic inactivation. A standard procedure is introducing the sample in liquid nitrogen and, subsequently, store at -80 °C [13].

The **extraction** step will also be dependent on the type of metabolites, and liquid-liquid extraction (LLE) is widely employed. For polar analytes, a mixture of polar organic solvents (e.g., methanol) with water can be applied, whereas for more hydrophobic compounds, other organic solvents are recommended (e.g., chloroform, methyl tert-butyl ether) [10]. Internal standards are commonly added as surrogates, to correct from possible extraction losses, matrix effects and/or ionization suppression. These compounds are similar to the metabolites of interest but must be not present in the samples or interact with the analytes. For instance, isotopically labelled standards are an appropriate choice.

Most employed detectors for sample acquisition in metabolomic studies are nuclear magnetic resonance (NMR), Fourier transform infrared spectroscopy (FT-IR) and **mass spectrometry** (MS)[14–16]. In NMR, the sample is dissolved in a deuterated solvent and analyzed directly. Direct infusion can also be performed into the MS, but often ion suppression occurs due to complex sample matrices. Therefore, prior separation of the compounds of interest is highly recommended.

Standard separation techniques in metabolomics are gas or **liquid chromatography** (GC or LC, respectively), supercritical fluid chromatography (SFC), capillary electrophoresis (CE) or ion mobility (IM) [4,17,18]. Imaging techniques can also be applied, mainly mass-spectrometry based (e.g., matrix-assisted laser desorption ionization (MALDI), nanostructure-imaging mass spectrometry (NIMS), desorption electrospray ionization mass spectrometry (DESI) or secondary ion mass spectrometry (SIMS)) [19–21]. However, the main analytical techniques in metabolomics include the combination of gas or liquid chromatography mass spectrometry (MS) [10,22,23]. These two powerful couplings offer high sensitivity, selectivity, reproducibility, and versatility. Chromatography improves selectivity by separating the sample constituents before entering the detector and avoiding its saturation. Besides, the chromatographic dimension facilitates compound identification by providing information on the retention times for each compound in addition to mass spectra. This PhD Thesis focuses on the development of analytical methodologies employing one- or two-dimensional liquid chromatography coupled to mass spectrometry, as will be further explained in **Section 2.2**.

It is essential in metabolomics to reduce sources of variability not related to the biological process itself, mainly due to the analytical procedure. Different actions

pursue to reduce this variability. First, samples can be analyzed randomly to increase their repeatability and correct possible batch drifts. Second, a minimum of three biological replicates is highly recommended from a statistical point of view, although five replicates (or more) are desirable when possible. Third, quality controls (QCs) are commonly employed in metabolomic studies to correct possible batch effects, and to condition the system before injecting the real samples [24]. QCs are representative mixtures of aliquots composed by a pool of all types of samples which are run repeatedly along the sequence; hence, they can be used for inter- and intra-batch correction when required. More information about the use of QCs is detailed in **Section 2.3.3 normalization and data scaling**.

Blank correction can also be carried out to ensure the compounds detected are not coming from the extraction process or the analysis itself. Instrumental blanks are often run at the beginning of the analysis sequence, to check that no peaks are detected prior to the injection of real samples. Extraction blanks are commonly included too. Their purpose is to check possible contamination from the extraction process and proceed to correct it when needed.

The last step of the environmental metabolomic workflow is **data processing,** which is actually the main bottleneck. Recent developments in online bioinformatic tools have facilitated the implementation of metabolomics. These resources can be used to pre-process the metabolomic data (e.g., XCMS [25], Workflow4Metabolomics (w4m) [26], MZmine2 [27] or MS-DIAL [28,29]). Some of them also perform statistical/multivariate analysis, and pathway associations (e.g., MetaboAnalyst [30,31]). Nevertheless, there are still challenges. On the one hand, related to data pre-processing, such as compression, filtering, chromatographic alignment, or feature detection. On the other hand, linked to the annotation of unknown compounds, e.g., mass fragmentation interpretation or structure elucidation [32]. In this PhD Thesis, the stress is on the development of data pre-processing tools for metabolomic studies. Data analysis strategies developed during this PhD Thesis are included in **Section 2.3 Data analysis strategies in metabolomics**.

Once the metabolites of interest are identified (and quantified), a biological interpretation is pursued. The aim is to relate the individual metabolites with affected metabolic pathways, to assess the effects and mechanisms of action of environmental pollutants at biochemical levels of *in vitro* and *in vivo* models.

Thus, metabolomics (and lipidomics) arises as an emerging powerful tool in environmental toxicology and chemical risk analyses. It allows hazard identification and characterization, as well as exposure assessments [9].

2.1.3 Model biosystems for metabolomic studies

One of the decisive choices in the experimental design of environmental metabolomic studies is the appropriate choice of the model biosystem, as already stated when introducing the whole workflow. In the Directive 2010/63/EU, European legislation has promoted the use of non-animal models to replace traditional experimental testing [33]. This PhD Thesis employs three model organisms (*Saccharomyces cerevisiae, Danio rerio, Oryza sativa)* and a human hepatic cell line (HepG2). Since zebrafish embryos in their early stage of growth are not considered animals under the mentioned directive [34], the chosen biosystems are in total agreement with the legislation.

Model organisms have many advantages, i.e., easy reproduction, growth, and maintenance, similarities with other targeted organisms, availability, and accessibility. They have also been widely studied. Therefore, previous knowledge of the organism facilitates data processing and biological interpretation of the results. Thus, the more information at all omic levels, the better. Model organisms can be diverse (including bacteria, yeast, insects, worms, fish, rodents, and plants). Nevertheless, a broad part of the biochemical operating principles is preserved between kingdoms. It means that even non-mammalian can contribute to expand information on human biological processes. Besides, an in-depth characterization of other omic levels (e.g., genome and proteome) has also been conducted for these specific organisms [35].

Achieving a comprehensive metabolomic insight is, however, a major challenge. The reason is the vast amount and diversity of existing metabolites, especially secondary metabolites from plants, fungi, and microbes. With the aim of unifying the effort of the scientific metabolomic community, the Metabolomics Society created a task group focused on Model Organism Metabolomes (MOM) [36]. This group aims to identify and quantify all metabolites from several model organisms, assign compounds to the metabolic pathways where they are involved and compare the metabolic networks throughout evolution. In this context, a list of prioritized model organisms was presented, as shown in **Figure 2.3**. A more detailed description of each biosystem employed in this PhD Thesis is included below.

Yeast (*Saccharomyces cerevisiae*)

Yeast (*S. cerevisiae*) is one of the simplest eukaryotic organisms and belongs to the fungi kingdom. It is widely used in biochemical research mainly because it is cheap, grows quickly, and has the best-known genome among eukaryotes as its genome has already been completely sequenced. Yeast also shares many biological properties with more complex pluricellular organisms. This fact allows the understanding of metabolic regulation and gene expression in a simpler manner [37]. Hence, yeast presents one of the most well-characterized metabolisms. All current metabolic information of this versatile organism can be found in the Yeast Metabolome Database (YMDB) [38].

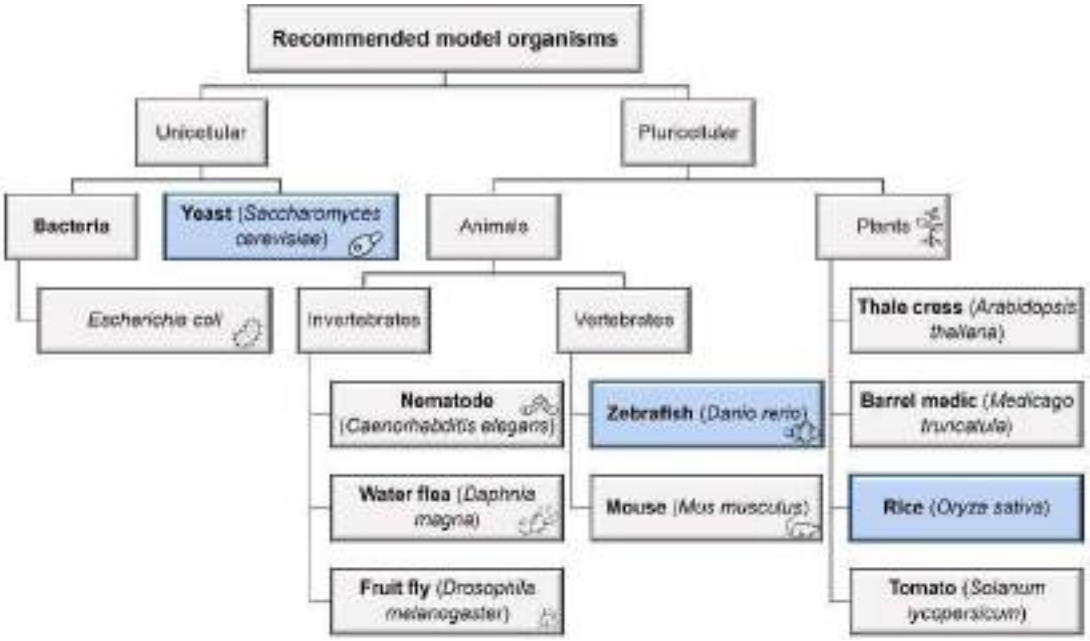


Figure 2.3. A scheme of the recommended model organism for in depth metabolome studies, selected by the model Organism Metabolome task group from the Metabolomic Society [36]. The organisms employed in this PhD Thesis are marked in blue.

S. cerevisiae is also often employed when aiming to validate analytical methodologies, mainly due to the reduced costs and simpler experimental designs [39,40].

Zebrafish (*Danio rerio*)

Zebrafish (*D. rerio*) is a small tropical fish native to Asia. Nowadays, it is commonly found in aquariums. In laboratories, normal living conditions are neutral pH and temperature of about 25 °C. Adults are usually between 3 and 5 cm long, and 1 cm

wide. They present several characteristic blue stripes along both sides of the body. Males are often slender with golden reflexes, whereas females are fatter and present silver reflexes. The life regeneration cycle from egg to egg can last around four months.

D. rerio has been widely selected as a model organism in research for several reasons. Firstly, it adapts easily to different environmental conditions and usually gets along well with other species. Second, its rearing is easy and cheap, fertilization is performed externally, and large offspring can be obtained (usually around 200 eggs per female). In addition, both eggs and larvae are transparent. Therefore, the whole embryonic process and the formation of the organs at early stages of growth can be followed [41]. Besides, zebrafish is a suitable alternative to mammal models for toxicology evaluations because it shares 70% of its genes with humans [42]. This fact promotes an extrapolation of results to humans.

Zebrafish embryos and larvae are widely employed for *in vivo* essays in pharmacology, toxicology, and ecotoxicology. They are used for testing new pharmaceutical compounds, rapid chemical toxicity screening and assessing sublethal effects of emerging contaminants [43–46]. In particular, there is an increasing interest in using zebrafish for studying lipid-related diseases. Lipids are involved in energy storage, signaling and embryonic development at the early stages in these animals [47]. Therefore, *D. rerio* is a promising choice for studying the effects of endocrine-disrupting chemicals (EDCs) [48,49], known for causing obesity, cardiovascular diseases, and diabetes, among others [50–52].

In this PhD Thesis, zebrafish eleutheroembryos up to six days have been employed to assess the effect of EDCs, the study shown in **scientific publication V**. The workflow followed (see **Figure 2.4**) starts with the eggs fertilization, which are kept separated from the adults with a mesh to avoid being eaten by their progenitors. Embryos remain in the chorion up to 48 hours post-fertilization (hpf). When hatching occurs, from 2 days post-fertilization (dpf), embryos are exposed to the pollutants of interest. Swim bladders begin to inflate around the 4 or 5th dpf. This period is called the eleutheroembryo stage, and the regulatory frameworks for animal experiments do not apply. This is because they do not need to be fed until the 6 or 7th dpf as nutrients come from the yolk sac, which is mainly a lipid reservoir [43]. Once the yolk sac is completely reabsorbed, an exogenous feeding starts, and they are considered as larvae [53].

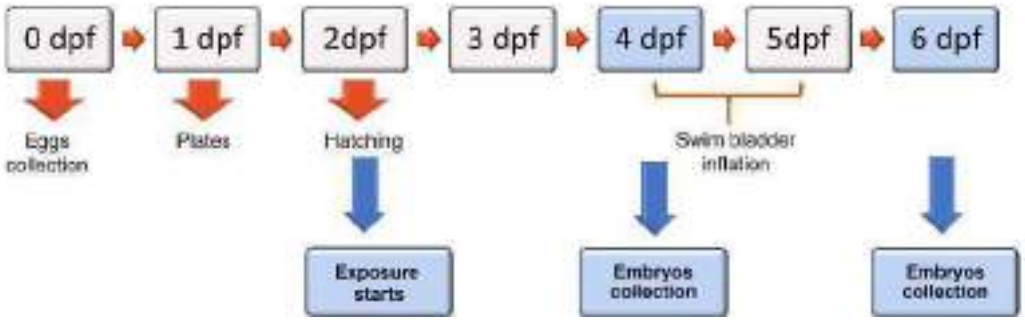


Figure 2.4. Growth and exposure workflow employed in the experiments with D. rerio in this PhD Thesis, from fertilization to eleutheroembryo collection. Dpf= days post-fertilization.

Information about zebrafish genetic, genomic, phenotypic and development data can be found in the Zebrafish Information Network (ZFIN) [54].

Rice (*Oryza sativa* L.)

Rice (*Oryza sativa* L.) is one of the most important crops worldwide, being the main staple food in Asia, Africa and South America. This cereal belongs to the *Poaceae* family, such as wheat (*Triticum aestivum* L.), maize (*Zea mays* L.) or barley (*Hordeum vulgare*). Among cereals, rice is the most suitable candidate for DNA sequencing due to its small and well-mapped genome. It is also the easiest to be genetically transformed.

Thus, *O. sativa* L. has arisen as the second-best plant model, after *Arabidopsis thaliana*. The main advantage faced to *A. thaliana* is that rice is a monocotyledonous plant (it has just one embryonic leaf in the seed instead of two). Therefore, rice has some important taxonomical, structural and phenotypical differences, and it is involved in specific processes only present in crop production (e.g., mycorrhization) [55].

In this context, the International Rice Genome Sequencing Project [56] was set. Several of its biotechnological applications for feeding purposes include improving rice production or enriching rice micronutrients content (e.g., golden rice to combat vitamin A deficiency) [57]. Other omics can also complement genomic information. An extensive list of current databases in rice genomics, transcriptomics, proteomics, and metabolomics can be found in a review by Hong et al. [58].

Metabolomics can clarify the function of unknown genes and improve the breeding of crops [59]. It is possible to identify metabolites related to rice quality, yield, and

nutrient content [60,61], or plant-pathogens interactions and defense mechanisms [62,63]. Environmental applications assess the effects of temperature or hydric stress [64,65], salt tolerance [66], microplastics [67], pesticides [68], heavy metals [69,70] or other inorganic pollutants [71]. This PhD Thesis is focused on the last of the cited applications, more specifically, arsenic exposure from the 11th day after germination until the 22nd day.

There are two main subspecies employed in rice cultivars: japonica and indica, both with broad genetic and metabolic diversity [72]. Oryza sativa ssp. japonica has been selected as the plant model in **scientific publication VII** in this PhD Thesis.

Human cell lines

Human cells and tissues are useful in metabolomics due to their stable phenotype and high availability. *In vitro* studies are easier to control, maintain and interpret, plus there is no specie extrapolation required. There are also fewer ethical considerations, and population factors (e.g., age, gender) are not applicable. Although *in vitro* studies lack the whole-organism integration, three-dimensional (3D) cell cultures are a very realistic and physiologically relevant approach. They simulate *in vivo* microenvironment, allowing cell-to-cell interaction in all directions and with the extracellular matrix [73,74].

More specifically, derived cell lines from the liver and kidney are often employed in toxicity tests. The aim is to unravel the toxic effects and mode of action of certain compounds directly on humans [75,76]. **Human hepatocellular carcinoma (HepG2) cell line** is the immortal and nontumorigenic cell line employed in this case. It is the immortalized cell line derived from the liver tissue of a Caucasian male of 15-year-old due to well-differentiated hepatocellular carcinoma. This cell line is commonly used in drug metabolism and hepatotoxicity studies because its phenotype and molecular expression have been deeply characterized, and it can express many specific functions of the liver [77]. HepG2 cells present high proliferation rates and express many liver-specific metabolic functions [78]. 3D liver spheroids present a hepatic cellular phenotype and are a robust liver model. They are used for *in vitro* liver toxicity screening assays due to their adaptability and viability for up to a month [79].

This PhD Thesis studies the effects produced in 3D liver spheroids from HepG2 cells by certain pharmaceutical compounds, known for their hepatotoxicity in **scientific publication VIII**.

2.1.4 Emerging contaminants

This PhD Thesis is focused on assessing the effects of different **emerging contaminants** (ECs). These compounds are considered chemical compounds released into the environment due to anthropogenic sources. They are a heterogeneous group of compounds that can be mainly clustered in six main classes: endocrine-disrupting chemicals; heavy metals and metalloids; agrochemicals, pesticides, and fertilizers; organo-halogen compounds; pharmaceuticals and personal care products; and micro or nano plastics [80]. These pollutants threaten human health and aquatic life, even at low concentrations. Their effects are diverse, but can include high toxicity, carcinogenicity, neurological/developmental/immune/memory/reproductive disruptions, cardiovascular and obesity disorders, apoptosis, or hypertension, among others [81]. Therefore, the characterization of their consequences at metabolic levels becomes crucial to regulate or forbidden their use.

The ECs employed in this PhD Thesis are listed below.

Metalloids

Heavy metals and metalloids are released into the biosphere during mining and manufacturing in industrial areas. Chemical and biological conditions of farmlands and aquatic reservoirs nearby are seriously affected, altering surrounding ecosystems [82,83]. As a result, deposits of metals and metalloids can be found even at 40 km from old mines in freshwaters and sediments [84]. Thus, these contaminants become a threat due to their high bioavailability, bioaccumulation, and persistence in nature [85].

In this PhD Thesis, a study has been conducted to unravel the effects of metalloids, more specifically, arsenic contamination.

Arsenic is a metalloid included in the top 10 chemicals of major public health concern by the World Health Organization or WHO [86]. Contamination by this metalloid has been related to natural sources like volcanism and geothermal activity, copper production, mining, burning of fossil fuels, thermal power plants and use of arsenical fungicides, herbicides, and insecticides in agriculture [87,88]. Chronic arsenic exposure can cause skin cancer, severe skin and liver disorders, asthmatic bronchitis, anemia, diabetes, increased blood pressure, and reproductive disorders [89,90].

High arsenic levels have been reported across the globe, mainly North and South America and southeast Asia [87,91]. Contaminated groundwaters are an important

source of arsenic [91], in which this metalloid has been reported up to 10 ppm [87]; its content in soils can be up to 40 ppm [92]. Nevertheless, an arsenic content higher than 20 ppb in drinking waters poses considerable risk of health hazards [93]. A safe standard of 10 ppb in drinking water has been established by World Health Organization (WHO) [86]. Besides, some countries, such as Denmark or The Netherlands have hardened the legislation tending to standards below this level between 5 and 1 ppb [94].

Inorganic forms of arsenic are more toxic than the organic compounds, and arsenate is more toxic than arsenite [92]. Inorganic arsenic is highly bioavailable because it uses the same transport system as other essential elements, such as silicon or phosphorous. Both arsenate and arsenite can be captured by plants through the roots and accumulate in the edible parts [95].

Rice is the most consumed food with high arsenic content, due to its presence in many daily products, according to the Food and Drug Administration (FDA) from the United States [92]. *Oryza sativa* L. planted in arsenic-contaminated soil is the main source of arsenic intake in a non-seafood diet [96]. Compared to other cereals, rice is more susceptible to arsenic because it grows under flooded conditions, and its content decreases from roots to leaves to grains [97]. Safe levels of this metalloid in polished rice grains have been set to 200 $\mu g \cdot kg^{-1}$ for adults and 100 $\mu g \cdot kg^{-1}$ for inorganic arsenic in rice-based products for children [98]. However, arsenic levels in polished rice have been reported to be up to 0.629 $mg \cdot kg^{-1}$ and 0.055 $mg \cdot kg^{-1}$ in the cases of total and inorganic arsenic content, respectively [99]. Strategies such as polishing or washing rice with abundant water have been suggested to reduce its arsenic content [98].

This PhD Thesis aims to unravel arsenic uptake mechanisms in rice and its metabolomic (and lipidomic) effects in **scientific publication VI**.

Endocrine-disrupting Chemicals

Endocrine disruptors englobe a group of compounds capable of disturbing the endocrine system and hormone regulation, causing reproduction and development alterations [100]. Sources can be natural or anthropogenic, but there is a special concern in Endocrine-disrupting Chemicals (EDCs). Diseases related to these chemicals are metabolic syndrome, obesity, type 2 diabetes, cardiovascular and pulmonary complications, miscarriages, endometriosis, liver lipid disorders or cancer, among others [101–103]. In this PhD Thesis, the emphasis is on understanding the obesogenic and estrogenic effect of bisphenol A (BPA) and 17-β-estradiol (E2).

Bisphenol A (BPA) is a monomer used in plastic and plastic derivatives, such as polycarbonates and epoxy resins. BPA can be found in food contact products, toys, medical devices, components for electronics, flame retardants, water pipes etc. Some of the unbonded monomers can be released into food and beverage containers or being in contact with the skin or eyes, becoming a threat to human health [100,104,105]. This EDC is also released into the environment. It can be found at non-negligible concentrations in effluent and surface waters, soils, sediments, and air, as summarized in **Table 1** [106].

The European Chemical Agency (ECHA) has forbidden and/or regulated its use in certain products (e.g., baby bottles: prohibited, food packing: <0.05 $mg \cdot kg^{-1}$, toys: <0.04 $mg \cdot kg^{-1}$, thermal paper <0.02%); the tolerable daily intake (TDI) for BPA has been set to 4 $\mu g \cdot kg^{-1}$ body weight per day and the specific migration limit (SML) to 0.05 mg of BPA per kg of food [107]. The median lethal dose (LD50) in rats via oral has been established as up to 3000 $mg \cdot kg^{-1}$ body weight [108], whereas malformations in zebrafish embryos have been reported from 25 μM of BPA [109].

Table 2.1. Summary of BPA concentration ranges in the environment, *extracted from* [106].

Source	BPA concentration range
Wastewater treatment plants (WWTP)	0-370 $\mu g \cdot L^{-1}$
Surface water	0-56 $\mu g \cdot L^{-1}$
Sea water	39-193 $\mu g \cdot L^{-1}$
Sewage sludge	10-10000 $\mu g \cdot kg^{-1}$
Soil	0.01-1000 $\mu g \cdot kg^{-1}$
Sediments	10-100 $\mu g \cdot kg^{-1}$
Air	100-50000 $ng \cdot m^3$

Hence, understanding the mode of action of this disrupting agent is crucial to continue legislating its applications. It is also important to compare its effect with other new substituents like Bisphenol S or F, to ensure a safe replacement, and current trends in toxicological studies are focused on comparisons between the new bisphenols [110].

17-β-estradiol (E2) is an estrogen steroid hormone. This endogenous estrogen is mainly involved in reproduction but also metabolism regulation, bodyfat distribution, energy consumption, lipogenesis and lipolysis [102]. Other EDCs usually mimic the way of action of gonadal hormones like E2, producing hormonal responses, blocking the receptor sites and substituting natural hormones, or modifying endocrine responses [103,111].

This steroid hormone is released into the environment, becoming a threat to aquatic life, soil, plants, water resources and humans [112]. Sources include sewage and wastewater treatment plants from hospitals, industries and domestic wastes, animal manure or effluent from livestock feedlots, and hospitals [113]. Thus, estrogen pollution is rising an international concern [114].

A comparison between the effects of BPA and E2 has been performed in this PhD Thesis, to shed a light on the changes these EDCs cause in the lipidome in **scientific publication V**. In this study, E2 has been used as an estrogenic control to characterize the non-estrogenic effect of BPA.

Pharmaceutical compounds

The increasing consumption of pharmaceutical compounds worldwide has led to consider these compounds as emerging contaminants. These drugs and their metabolites are released into wastewaters, not only from urban areas but also from farming and agriculture [115]. Then, an accumulation in sewage sludge, sediments and surface waters occurs because wastewater treatment plants cannot completely remove these compounds [116]. Other sources include hospital effluents or senior residences, which should be treated differently from domestic wastewaters [117,118]. The consequences are severe in the environment (e.g., producing addiction, bioaccumulation, and antibiotics resistance) and also pose a threat to human health [119].

In this PhD Thesis, the effects of commonly released hepatotoxic drugs are studied: an antiepileptic (**carbamazepine**), an antibiotic (**amoxicillin**) and an antidepressant (**trazodone**). They are present in relevant environmental concentrations in wastewaters, typically ranging from $\mu g \cdot L^{-1}$ to $ng \cdot L^{-1}$ [120–124]. **Table 2.2** exemplifies concentration ranges found for these drugs at wastewaters of different senior residences in Catalonia, according to the study by Gómez-Canela et al. [125].

Table 2.2. Concentration ranges of carbamazepine, amoxicillin and trazodone found in the wastewaters of senior residences in Catalonia.

Drug	Concentration range in senior residences ($\mu g \cdot L^{-1}$)
Carbamazepine	0-5.4
Amoxcillin	0-0.5
Trazodone	0-314

Carbamazepine (CBZ) is an anticonvulsive drug used to treat epileptic crisis, neuropathic pain, schizophrenia, and bipolar disorder. CBZ bioaccumulates, biotransforms, and biomagnifies through the trophic chain [120,126]. This drug also produces Endocrine-disrupting effects [127] and changes in the lipidome, especially in glycerophospholipids, triacylglycerides, cholesterol esthers, sphingolipids and oxylipins [128–130].

Amoxicillin (AMOX) is a generic β-Lactam antibiotic used to treat bacterial infections. AMOX is one of the most consumed pharmaceuticals [131]. Its presence in the environment poses a direct threat to human health by developing multi-resistant strains of bacteria and producing changes in the microbial community structure [132]. Thus, the removal of this penicillin-like drug from wastewaters becomes crucial [133].

Trazodone (TRA) is an antidepressant drug used for the treatment of depression, anxiety disorders or insomnia. It acts as a serotonin uptake inhibitor, and it is heavily metabolized and activated in the liver by the CYP3A4 enzyme [134] to the two precedent drugs. Therefore, TRA may cause drug-induced liver injury [135].

These three drugs are eliminated by urine, and their half-lives are varied (i.e., up to 36h for carbamazepine *versus* 7h for trazodone or 1h for amoxicillin) [124]. Carbamazepine is highly excreted unchanged (around 70%), whereas this percentage is lower in trazodone (around 20%) [134,136]. Amoxicillin and carbamazepine can form up to 7-9 metabolites, whereas trazodone forms less metabolites, only 4 [124]. All three are related to hepatic metabolism. Carbamazepine presents the highest hepatotoxicity, followed by amoxicillin and trazodone [137].

The toxicity of these drugs and the effects on the lipidome and metabolome have been studied through the exposure to HegG2 3D liver spheroids for a 24h period, in **scientific publication VIII.**

2.2 Analytical approaches for metabolomics

This PhD Thesis aims to develop an analytical methodology based on one- and two-dimensional liquid chromatography coupled to mass spectrometry (LC-MS or LC×LC-MS, respectively) for metabolomic applications. In the following sections, the analytical part of the metabolomic workflow is discussed, with especial emphasis in the techniques and approaches employed in this PhD Thesis.

2.2.1 Targeted *versus* untargeted analysis

From the analytical perspective, the first decision in the metabolomic workflow is whether the stress is on performing a general exploration of the pathways affected or rather quantifying a reduced number of metabolites from specific pathways. A more qualitative and global approach presents mainly identification purposes and relative abundances of *a priori* known and not known compounds. A more quantitative approach means absolute quantitation of a list of metabolites of interest, all known *a priori* [4]. These two analytical strategies are summarized in **Figure 2.5**.

A **targeted** approach is preferred to answer a precise and particular biochemical question, that is generally related to specific metabolic pathways. Hence, the main objective is to provide absolute concentrations of a reduced number of *a priori* known metabolites (usually less than fifty). Both extraction and chromatographic analyses are usually optimized for these compounds. Tandem mass spectrometry (MS/MS) concatenating two quadrupoles as analyzers (QqQ) is frequently employed. In this technique, one precursor ion is selected, isolated, and fragmented into one or several product ions, according to selected reaction monitoring (SRM) or multiple reaction monitoring (MRM) modes, respectively. Each precursor/product ion pair is compound-specific, called transition.

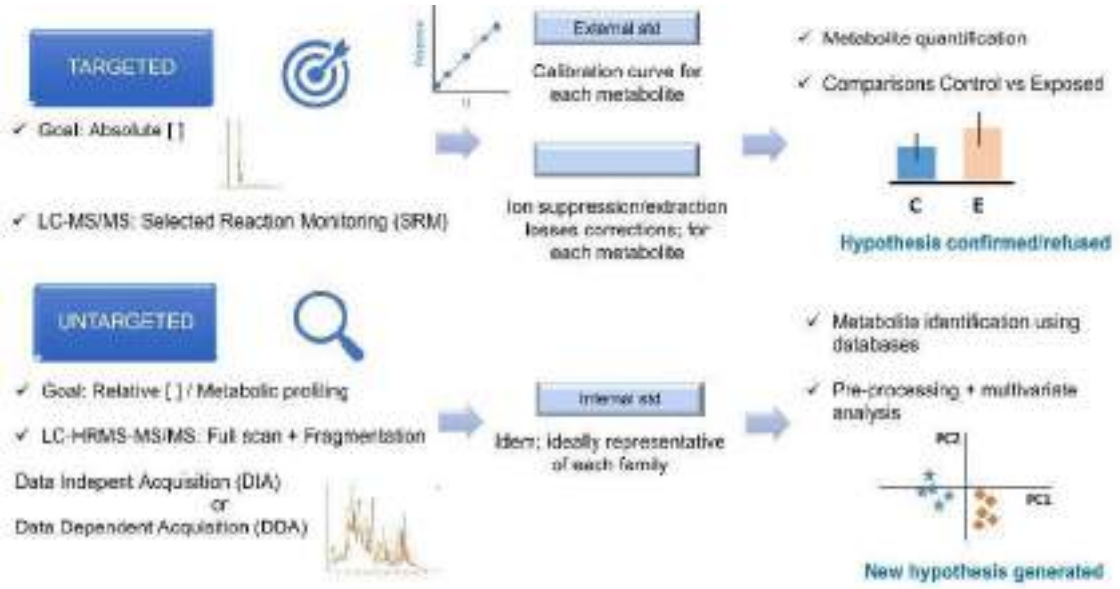


Figure 2.5. Comparison of targeted versus untargeted approaches in metabolomic workflows.

In targeted approaches, external standards are employed for building calibration curves comprising a wide range of concentration levels. Internal standards are used

to correct for ion suppression and/or matrix effects. Peak areas of the analytes are integrated and corrected with the areas of the internal standards. The amount of data generated is reduced. Therefore, statistical analysis is often straightforward and can be univariate. At the end of the data processing and interpretation, the hypothesis is validated or refused.

On the contrary, **untargeted** approaches aim to detect and identify the maximum number of metabolites possible, especially unknown compounds (commonly up to several hundreds). Relative abundances of the identified metabolites are determined instead. This strategy is suitable for obtaining a global overview of unexpected changes in the metabolic pathways and covering a broader extension of the metabolome. More general extraction and chromatographic separation procedures are preferred, in order to include multiple families of metabolites at once. High-resolution mass spectrometry (HRMS), which provides high mass accuracy in the measurements, is highly recommended for identifying the metabolites. MS/MS is also useful, but fragmentation is performed differently. There are two main MS/MS modes in untargeted analysis. If MS/MS data is acquired from all the detected precursors in specific, narrow, and pre-selected mass windows, then it is referred to as data independent acquisition (DIA). However, if only the most intense precursors are selected for further fragmentation in a second step, it is called data dependent acquisition (DDA). In this PhD Thesis, both modes have been employed. A more in-depth discussion of these two modes can be found in the **Tandem mass spectrometry subsection** from **Section 2.2.7 Mass spectrometry and metabolomics.**

There are some important differences between the two analytical approaches. On the one hand, in untargeted studies, it is not possible to have such a large number of internal standards to cover all possible compounds that can be analyzed. Thus, usually one representative internal standard per family of metabolites is added. In this case, relative abundances are obtained after integration and normalization, which allows for the calculating 'fold-changes' (FCs). FCs are a measure to describe how much a value has changed compared to an original (control) situation. In environmental metabolomics, the default situation is the control sample (not exposed), and the final value is obtained after exposure (exposed or treated). Indeed, instead of absolute concentrations as in targeted approaches, FC ratios commonly express the relative abundances in untargeted analyses.

On the other hand, the combination of HRMS with MS/MS in untargeted approaches produces a higher amount of data than targeted approaches. This

drawback challenges the data processing steps and appears as one important bottleneck. Pre-processing in untargeted analysis requires several steps (e.g., filtering and alignment) prior to integration, and multivariate analysis is usually necessary. That is why the development of data analysis strategies becomes crucial in untargeted approaches, and therefore, it is one of the main axes of this PhD Thesis (see **Section 2.3 Data analysis strategies in metabolomics** for more details). Although a few of the works in this PhD Thesis included also targeted studies (**scientific publications III, VI, VIII),** the major emphasis has been on developing analytical and data analysis strategies for untargeted approaches (**scientific publications I, II, IV, V, VII**).

In untargeted approaches, there is the additional necessary step of compound identification, which is frequently time-consuming and complex. Databases (DBs) are required for this purpose. Unfortunately, spectral DBs with fragmentation patterns depend on the instrument employed and the MS/MS parameters selected (e.g., collision energies). However, in general, spectral DBs are usually reproducible between different laboratories. Thus, recent efforts have been focused on creating free and user-friendly spectral libraries (e.g., Human Metabolome Database (HMDB) [138,139], METLIN [140], Massbank [141,142], LIPID MAPS [143,144], National Institute of Standards and Technology (NIST) [145]. On the other side, peak retention times can be used as an orthogonal and complementary property that can add more confidence to the identification. Nevertheless, these DBs depend on the chromatographic separations and usually change within different analytical strategies (e.g., different column stationary phase composition and/or batch, mobile phase mixture, gradient, dead volumes). Software programs are being developed at present that can predict peak retention times in these variable experimental conditions [146–148]. These predictions could reduce both time and costs in method development and identification steps in untargeted metabolomics.

At the end of the untargeted analysis, rather than confirming a previous hypothesis, usually a new one is generated. A subsequent targeted study of specific pathways affected should be performed to double-check the findings and quantify. Consequently, the untargeted approach provides a holistic view of the changes produced due to a certain exposure, and it is a very suitable starting point for collecting information about, for instance, new emerging contaminants.

There is an intermediate option, the **pseudo-targeted** approach. The goal is to provide approximate metabolite concentrations of a reduced number of compounds, but a considerably larger number than in the standard targeted approach. One single

calibration curve and a single internal standard are used for multiple metabolites with similar chemical structures. On the one hand, the amount of data produced is still easily manageable (unlike the data analysis strategies needed in untargeted approaches) and causes less time and economic costs than the standard targeted one (fewer calibration curves and fewer external and internal standards are needed). On the other hand, this strategy does not guarantee an accurate absolute quantitation, and it presents a bias towards certain metabolites, instead of exploring 'the big picture'. Besides, there is an important assumption that extraction, chromatographic separation and ionization conditions will be preserved among different metabolites from the same group, which may not be entirely true.

Sometimes, the term pseudo-targeted is employed in metabolomics when an untargeted method has been employed (extraction plus chromatography plus mass spectrometry conditions), but a reduced list of masses from metabolites of interest is selected for further integration and processing.

2.2.2 Sample preparation prior to metabolomic analysis

Extraction is a critical step in every analysis. A good recovery of the aimed analytes is a key step to obtain representative results of the sample. Besides, the choice of solvents in the case of liquid-liquid extraction (LLE), which is the most common extraction procedure in metabolomics, will determine the whole analysis. This mixture is analyte-dependent and will differ whether the study is a metabolomic (i.e., polar metabolite-based) or a lipidomic (i.e., lipid-based) approach.

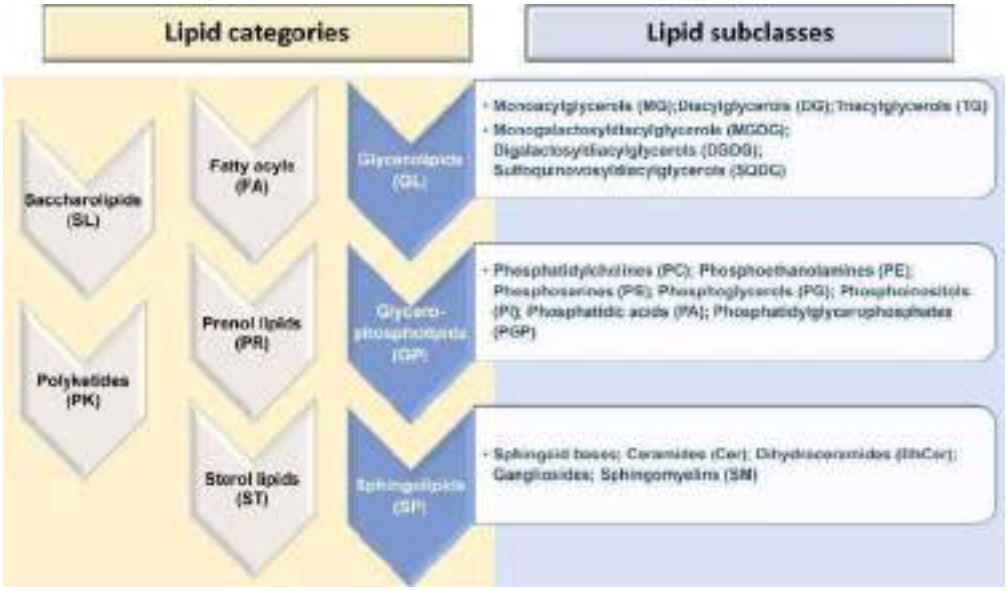


Figure 2.6. Summary of eight lipids categories according to LIPIDMAPS classification [143], with especial emphasis in the main lipid subclasses studied in this PhD Thesis (marked in blue).

Lipids are hydrophobic biomolecules, not soluble in water but soluble in organic solvents (e.g., chloroform ($CHCl_3$), isopropanol (IPA), methanol (MeOH)). These compounds have a crucial role in cell metabolism, mainly involved in energy storage and transfer, signaling and structural functions. The LIPID MAPS classification [143] includes eight lipid categories, as shown in **Figure 2.6**.

Different lipid extraction procedures are presented throughout this PhD Thesis. The extractions proposed by Folch et al. [149] and Bligh et al. (BD) [150] are the oldest and most typical extractions. These protocols mix methanol (MeOH) and chloroform ($CHCl_3$) in different ratios. A two-layer system is formed, the aqueous phase up and the organic phase at the bottom. The goal is to generically extract lipids, especially glycerolipids (GL), glycerophospholipids (GP) and sphingolipids (SP). Although these protocols were designed for animal tissues, they have been used in many other applications [151]. In the case of this PhD Thesis, the solvent proportion employed was $CHCl_3$:MeOH (2:1), and butylated hydroxytoluene (BHT) was added in the first step to prevent lipid oxidation [152].

A variant of this procedure was also selected in this PhD Thesis to study sphingolipids. The proportion of the solvents is inversed, $CHCl_3$:MeOH (1:2), and a saponification step is added [153]. Briefly, samples are incubated overnight, then potassium hydroxide (KOH) is added, and there is a further incubation step in an oven for 2 hours at 37 ºC. Alkaline hydrolysis occurs, where all ester bonds are hydrolyzed (i.e., from GP and GL) but not the amide bonds (i.e., from SP). Then, the excess base is neutralized with acid (e.g., acetic acid). Hence, the only lipids that remain belong to the sphingolipid family. As these compounds are usually present at lower concentrations compared to other lipid species (e.g., phosphatidylcholines (PC) or triacylglycerols (TG)), this extra step enhances their detectability.

Another generic and widespread lipid extraction that was used in this PhD Thesis involves methyl tert-butyl ether (MTBE) instead of $CHCl_3$ [154]. The main two advantages are that MTBE is less harmful, and that the organic layer is the upper one, which facilitates its recovery. This protocol has been proved to better extract unsaturated fatty acids, glycerophospholipids and ceramides [155].

If a simultaneous extraction of polar lipids and metabolites is desired, a protein precipitation protocol with methanol is recommended. This protocol is much simpler and shorter than the previously mentioned, as no layers are formed. Another benefit is that this precipitation also enhances the detectability of sphingolipids if a targeted chromatographic method is selected [156]. This procedure was applied to both, targeted lipidomics and untargeted metabolomic analysis.

Lastly, for generically extracting polar metabolites, a protocol combining methanol, water and chloroform ($MeOH:H_2O:CHCl_3$) was preferred in this PhD Thesis [157,158]. In this case, the layer kept for the analysis is the upper phase, as it will retain water-soluble compounds.

All extraction protocols employed throughout this PhD Thesis can be found in **Table 2.3** with reference to the study in which they were used.

Once the approach (targeted/untargeted/pseudo-targeted) and the extraction procedure for the metabolites of interest are decided, the next step is to select the analytical technique for the analysis conditions according to physicochemical properties of the compounds of interest. In the following sections, LC in one and two dimensions coupled to MS will be proposed as the separation technique and detector, respectively. The principles of LC-MS and LC×LC-MS are described and exemplified with applications in the metabolomic field.

Table 2.3. Summary of extraction protocols employed in this PhD Thesis. BD: Bligh and Dyer; EDCs: Endocrine-disrupting chemicals; MeOH: methanol; MTBE: Methyl tert-butyl ether.

Extraction type	Analytes	Study	Scientific publication
MTBE	Lipids	Arsenic in rice	**VII**
Folch-BD	Lipids	Yeast	**VI**
		EDCs in zebrafish	**V**
Folch-BD plus saponification step	Sphingolipids	Yeast	**VI**
		EDCs in zebrafish	**V**
Protein precipitation with MeOH	Sphingolipids Metabolites	Pharmaceutical compounds in cells	**VIII**
$MeOH:H_2O:CHCl_3$	Metabolites	Arsenic in rice	**VII**
		EDCs in zebrafish	**VI**

2.2.3 Liquid chromatography conditions in metabolomics

Liquid chromatography (LC) is a suitable choice for the analysis of non-volatile and thermolabile compounds. LC is a versatile and selective analytical technique due to its many adjustable parameters: the variety of separation modes and stationary phases, column length, inner diameter, and particle size; mobile phases solvent mixtures and modifiers; gradient program; temperature and pH etc.

Although direct infusion could be also performed in the mass spectrometer, adding a previous chromatographic dimension presents some major advantages: 1) separate

isobaric compounds, 2) reduce ion suppression when reaching the ionization source, and 3) decrease complexity of the measured mass spectra. Thus, LC-MS offers the broadest metabolites coverage of, and can be applied to both, targeted and untargeted studies.

There are many separation modes applicable to metabolomics. Some examples are reversed phase (RP), hydrophilic interaction (HILIC), ion exchange (IEX), ion pairing (IP), hydrophobic interaction (HIC), normal phase (NP), argentation (Ag), mixed mode (MM) or chiral. However, the most employed are RP and HILIC for non-polar and polar compounds, respectively. This PhD Thesis focuses on these two separation modes, separately and in combination (in the case of LC × LC) and will be further explained. A comparison between the two can be found in **Figure 2.7**.

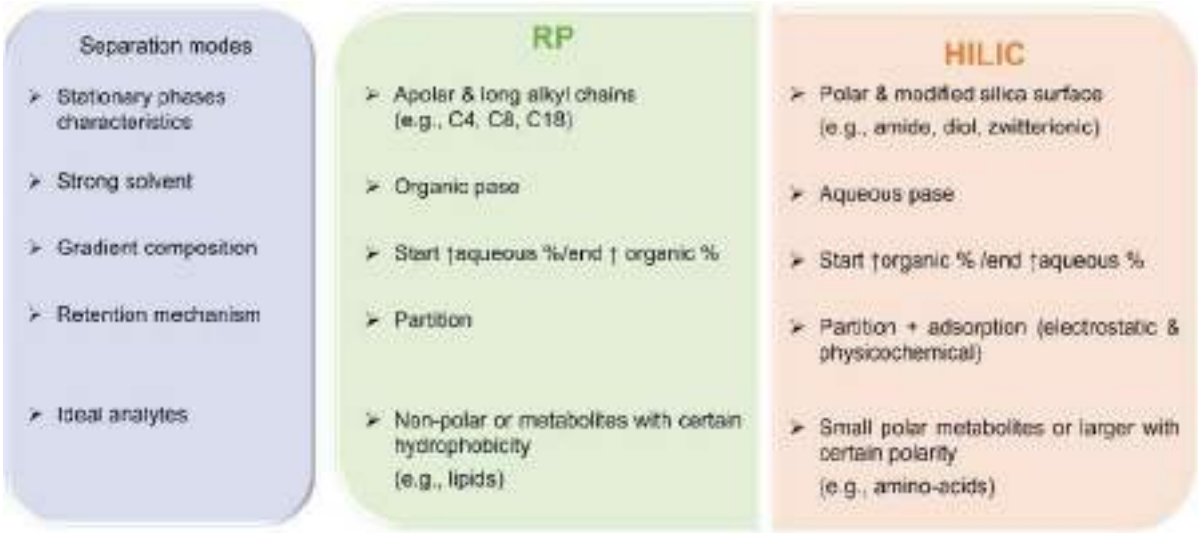


Figure 2.7. Comparison of the two main separation modes employed in metabolomics: reversed phase (RP) and hydrophilic interaction chromatography (HILIC).

Reversed phase chromatography (RP)

RP is the most classical retention mechanism due to its robustness and repeatability. Stationary phases are commonly composed of hydrophobic functional groups attached to silica particles. Typical stationary phases in this mode are C18, C8, C4, cyano, phenyl, amino, polyhydroxymethacrylate and graphic carbon. The first three refer to linear alkylsilane phases (e.g., C18 refers to 18 carbons bound to the silica). The longer the chain of the alkyl groups (e.g., C18), the higher the hydrophobicity and, consequently, the higher the retention of non-polar compounds. This means the analytes will interact more with the stationary phase, and the whole

separation will be wider and slower. Contrarily, shorter chains (e.g., C8) will result in shorter separations and sharper peaks. However, complex molecules and long-chain fatty acid chains may not be enough retained and can be eluted during the dead volume.

Some practical considerations when working with RP mode are:

- Less water-soluble compounds will be more retained than more water-soluble ones.
- Longer carbonated chains will present longer retention times compared to shorter chains. The same trend is found for non-branched isomers *versus* branch-chained, and saturated *versus* unsaturated compounds.
- Very polar and ionic compounds elute in the dead volume of the column.
- The elution order of different organic functional groups (from less retained to more retained) are carboxylic acids < alcohols, phenols < amines < esters, aldehydes, ketones < aliphatic chains.
- The weakest solvent is the aqueous phase, whereas the strongest solvent is the organic one. Gradients normally end at 100% organic solvent. This way, the most retained compounds are eluted, by increasing affinity with the mobile phase instead of with the stationary phase (i.e., partition mechanism).

RP is the preferred separation mode when analyzing lipids [151,159]. In this PhD Thesis, C18 or C8 have been employed in most studies (**scientific publications III, IV, V, VI, VII, VIII).** Lipids often present two well-differentiated parts: a hydrophobic one, composed of one to three fatty acyl chains; and a hydrophilic one, which allows discrimination between families (see **Figure 2.8**). Thus, reversed phase is not the only separation mode suitable for lipids. RP separates lipids according to the length of the fatty acyl changes and the number and positions of the unsaturations. On the other side, HILIC or NP can be employed to separate these compounds by the polar head groups [160]. This duality is very useful for achieving orthogonal separations in LC×LC, as will be further explained in **Section 2.2.4 2DLC applications in metabolomics**, especially in the **scientific publication I**.

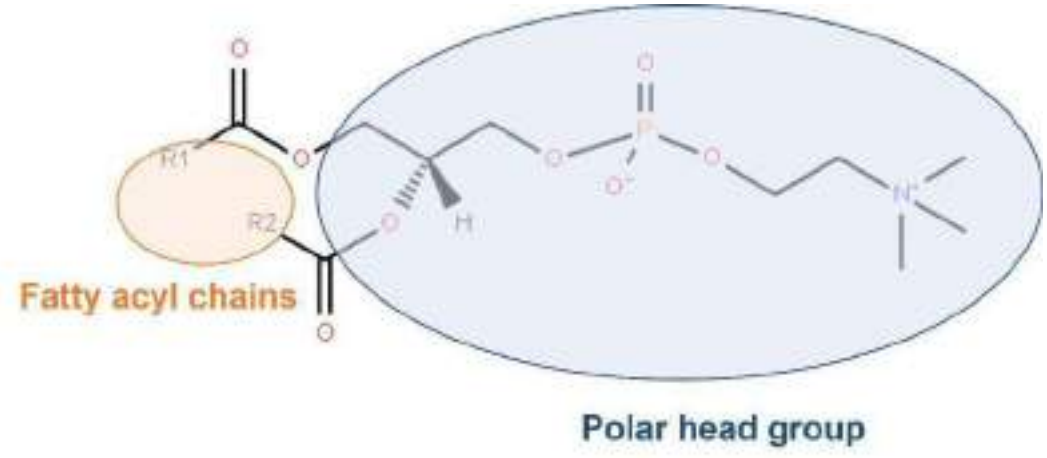


Figure 2.8. Duality of lipid molecules: hydrophobic and hydrophilic. Representation of a phosphatidylcholine. *Image adapted from* LIPID MAPS [143].

Hydrophilic interaction chromatography (HILIC)

HILIC has recently arisen as an alternative to RP because it is especially designed for polar and hydrophilic compounds, usually barely retained in RP conditions [161,162]. Contrary to NP, its major competitor, HILIC can employ semi-aqueous mobile phases (i.e., very similar to RP), increasing analyte solubility and MS compatibility [163]. The stationary phase in HILIC mode can be underivatized silica particles or bonded to hydrophilic functional groups. Typical HILIC stationary phases can be neutral (e.g., diol, cyano, amide), charged (e.g., amine, pure silica), or zwitterionic (e.g., sulfobetaine). A fixed water-enriched layer is also formed next to the stationary phase as a result of its high polarity, creating a hydrophilic environment that favors its affinity with the analytes. Retention mechanisms are more complex than in RP, because partition between this hydrophilic environment around stationary phases and mobile phases is not the only interaction. For instance, adsorption of the analytes onto the surface of the stationary phases occurs due to electrostatic (e.g., Van der Waals, hydrogen bonds, dipole-dipole) and physicochemical interactions.

Retention order is usually inversed compared to RP. Highly polar compounds are highly retained, whereas hydrophobic are eluted in the dead volume. In HILIC, the strongest solvent is water, whose content normally increases along the separation. However, a minimum water percentage is required from the beginning to form the water-enriched layer in the vicinity of the stationary phases.

Some practical considerations when working with HILIC mode are:

- Stationary phases are generally pH and ionic strength dependent. Therefore, mobile phases often require salts and/or buffers, especially when charged stationary phases are employed. However, a too high salt content will decrease MS sensitivity.
- Water content in the samples needs to be minimized, as it will lead to peak distortion and broadening.
- A longer re-equilibration time (compared to RP) is often necessary to reset conditions between chromatograms.
- As water is the strongest solvent and increases with the run. Therefore, column pressure will increase accordingly. It is recommended to check that pressure at the maximum water content during gradient is lower than column backpressure.

HILIC applications are varied, and its use has widely increased in recent years, especially in metabolomics [159,164]. Some examples of suitable compounds for this separation mode are amino acids, nucleosides, nucleotides, organic acids, lipids, vitamins, flavonoids, pharmaceutical compounds, and their metabolites [163,165,166].

In this PhD Thesis, HILIC mode has been employed for untargeted metabolomic analysis using LC-MS (**scientific publication IV, VII, VIII**). Besides, combinations with RP in LC×LC have also been selected for untargeted lipidomics (**scientific publication V**) and also targeted approaches, i.e., the analysis of pharmaceutical compounds (**scientific publication III**) and amino acids (**scientific publication VI**).

2.2.4 2DLC applications in metabolomics

One of the main limitations in untargeted metabolomics is the coverage that can handle available instrumentation [167]. Many metabolites are measured simultaneously from very different origins, i.e., endogenous, exogenous, environmental, plant, microbial or pharmaceutical sources, among others [168]. Besides, there are still many compounds to discover that may be related to specific exposures or diseases. In complex samples as is the case of biological matrices, overlapping, coeluting and non-resolved peaks are frequently encountered [169]. This problem often leads to unidentified peaks or difficulties in quantitation. To solve these drawbacks, current methodological trends are moving towards increasing peak capacity, while reducing ion suppression and diminishing interferences between

analytes [170,171]. Thus, there is an increasing tendency to combine different separation platforms or modes, with the aim of increasing resolution power [168].

Scientific publication I introduces two-dimensional liquid chromatography (2DLC) as a powerful option to cope with the previously described difficulties. Different 2DLC set-ups are discussed, especially the comprehensive mode (LC × LC), commonly employed in the untargeted analysis. Some common terms in the 2DLC world are also introduced (e.g., orthogonality, modulation, breakthrough). In addition, examples of applications in the metabolomics and lipidomics fields are listed, as well as some of the most recurrent combinations of the separation modes in both dimensions, with their pros and cons.

I. SCIENTIFIC PUBLICATION I

Title: Two-Dimensional Liquid Chromatography in Metabolomics and Lipidomics

Miriam Pérez-Cova, Romà Tauler, Joaquim Jaumot

Citation reference: Wood P.L. (eds) Metabolomics. Neuromethods, vol 159.

Publisher Name: Humana, New York, NY

Print ISBN:978-1-0716-0863-0

Online ISBN: 978-1-0716-0864-7

DOI:10.1007/978-1-0716-0864-7_3

Chapter 3

Two-Dimensional Liquid Chromatography in Metabolomics and Lipidomics

Miriam Pérez-Cova, Romà Tauler, and Joaquim Jaumot

Abstract

Multidimensional separation systems have arisen in the last years to overcome certain limitations of the classical one-dimensional separations. Multidimensional analytical approaches achieve a greater separation power, a crucial aspect when dealing with highly complex samples, as in metabolomics and lipidomics. Online comprehensive two-dimensional chromatography is a particularly interesting mode when pursuing untargeted analysis, which allows for separating, and consequently identifying a more extensive number of analytes. This chapter aims to summarize current applications of 2D-LC to the fields of metabolomics and lipidomics, and setups of different separation mechanism that are being employed, showing the most suitable combinations of chromatographic modes, depending on the target compounds.

Key words Two-dimensional liquid chromatography, LC×LC-MS, Comprehensive, Metabolomics, Lipidomics

1 Introduction

One-dimensional liquid chromatography (1D-LC) combined with mass spectrometry (MS) is the leading analytical technique employed in metabolomics and lipidomics studies. It is selective and versatile for nonvolatile analytes, due to the wide variety of different separation modes and adjustable parameters (e.g., mobile and stationary phases, pH, temperature, additives). However, the complexity of metabolomics (and lipidomics) samples with a large number of compounds and the presence of isobaric molecules difficult its identification and quantification. From a chromatographic viewpoint, a new hypothesis can be tested: "Extremely complex matrices require higher peak capacities to achieve a complete resolution." In the last years, different analytical techniques have been proposed to add another dimension to the separation, and, consequently, achieve a greater separation power (e.g., ion-mobility spectrometry, IMS [1]; supercritical fluid

Paul L. Wood (ed.), *Metabolomics*, Neuromethods, vol. 159, https://doi.org/10.1007/978-1-0716-0864-7_3,
© Springer Science+Business Media, LLC, part of Springer Nature 2021

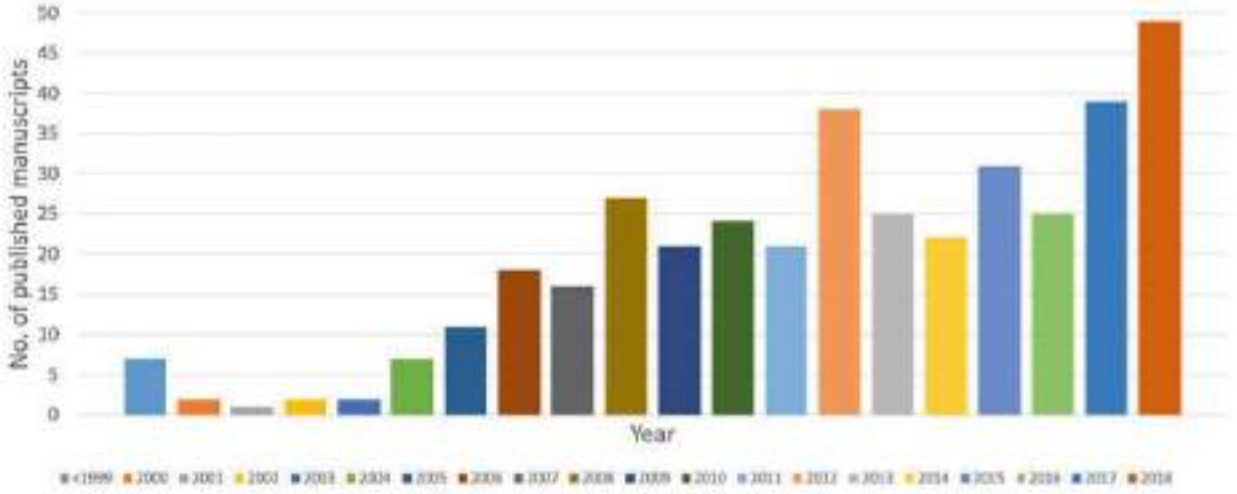


Fig. 1 Evolution of the number of published manuscripts related to 2D-LC (source Scopus, March 2019)

chromatography, SFC [2]; two-dimensional liquid chromatography, 2D-LC [3]).

Recently, 2D-LC has gained popularity for the analysis of metabolomics and lipidomics samples [4]. Two-dimensional chromatography is an attractive alternative for overcoming these limitations when characterizing these highly complex samples. The second column provides additional selectivity assuming that, ideally, both separations are orthogonal [5], which means that non-correlated retention time mechanisms are employed in the two dimensions. Consequently, the elution times from the two dimensions can be treated as statistically independent [6]. 2D-LC is commonly coupled to high-resolution mass spectrometry (HRMS) and/or tandem mass spectrometry (MS/MS), but other suitable detectors and analytical instruments include IMS-MS, ultraviolet–visible spectroscopy (UV-Vis), and fluorescence [7].

Figure 1 shows the continuous growth in the use of 2D-LC from the early 2000s. In the first years of the decade, only few manuscripts describing new instrumental developments allowing the evolution from offline to online analysis and from the UV-Vis to the MS detection (widely used nowadays) can be found. Also, pioneering works describing data analysis approaches for the processing of these data sets were published. During this period, most of the application of 2D-LC were related to proteomics and polymers analysis. In the second decade, the instrumental developments have allowed a major diversity of applications, in which metabolomics and lipidomics application stand out. However, these new examples cannot be explained without the recent technical progress that will be described below.

There are several ways of implementing 2D-LC that define the mainly used classification criteria.

The first classification approach dealing with the sample collection refers to offline, stop-flow, or online configurations.

Offline implies that fractions from the first-dimension (^{1}D) are collected, and afterward are transferred to the second-dimension (^{2}D). This setup is experimentally less complicated, and, as the two columns from the two dimensions are independent, greater peak capacities usually are achieved. The offline configuration also has the advantage of allowing for preconcentration, derivatization, and reconstitution into an appropriate mobile phase between the two separations [8].

In the stop-flow configuration, fractions are transferred directly to ^{2}D, but the flow is stopped in the ^{1}D when performing the analysis in the ^{2}D. This procedure is repeated for the whole ^{1}D separation [9]. In this approach, it is possible to increase the ^{1}D flow rate, as it will be continuously stopped before entering in the ^{2}D column. One of the most relevant drawbacks that this setup presents is the additional first dimension band broadening during stop-flow periods, especially when small molecules are analyzed. This is the main reason why stop-flow is less common than offline and online configurations [9]. However, it has been successfully employed for bigger molecules, such as peptides [10] or polymers [11].

In the online mode, the connection is made directly through an interface, that is, a high-pressure switching valve. ^{2}D separation needs to be very fast to analyze the whole fraction before the subsequent fraction is transferred. Therefore, short ^{2}D columns with large internal diameters and fast ^{2}D gradients are highly desirable, whereas long columns and low flow rates are commonly used in ^{1}D. Consequently, injection volume in ^{2}D is also reduced [12]. Chromatographic resolution is maintained from ^{1}D to ^{2}D because each peak is sampled 3–4 times, appearing in consecutive modulations (i.e., two successive injections into the second column [3]). This online arrangement allows for decreasing the total analysis time (an essential drawback in 2D-LC) thanks to full automatization, and also diminishes sample loss, cross-contamination, and degradation [8].

Another classification criterion is based on how fractions are collected and injected from first-dimension (^{1}D) to the second-dimension (^{2}D). When a single fraction from the ^{1}D effluent is reinjected in the ^{2}D column, it is a single heartcutting approach (LC–LC). Multiple heartcutting (mLC–LC) refers to the cases where several regions of the first separation are selected and introduced one at a time in the second separation. When all fractions are transferred at regular intervals and, therefore, are subjected to both separations is called comprehensive (LC × LC). Additionally, selective comprehensive (sLC × LC) implies that only specific and successive regions of the first separation are targeted for further separation, while the separation already obtained in the ^{1}D is maintained [3, 5].

This chapter is focused on online comprehensive 2D-LC. LC × LC is a mode particularly interesting when pursuing untargeted analysis, in which the main aim is to obtain the maximum information possible from the sample in one single run. This strategy allows separating and, consequently, identifying a more extensive number of analytes. Therefore, it is the main approach employed in screening and profiling, as it enables the separation of different classes of metabolites and their constituents [8]. The hyphenation of mass spectrometry analyzers and hybrid systems that permit fragmentation (e.g., Q-TOF, Orbitrap, Ion Trap, Triple Quadrupole) are also handy tools for achieving the identification of unknown compounds.

A generic scheme of a passive modulation interface is shown in Fig. 2. The concept of passive modulation implies that no modification takes place between the two dimensions. On the contrary, if some alterations have been made to the modulation process compared to the passive mode (i.e., presence of a bypass capillary, trapping columns, vacuum evaporation systems), it is called active modulation [4].

In the scheme, a setup including two six-port valves is detailed, as well as the two positions that valves acquire during subsequent modulations. Position 1 implies that ^{1}D effluent passes through

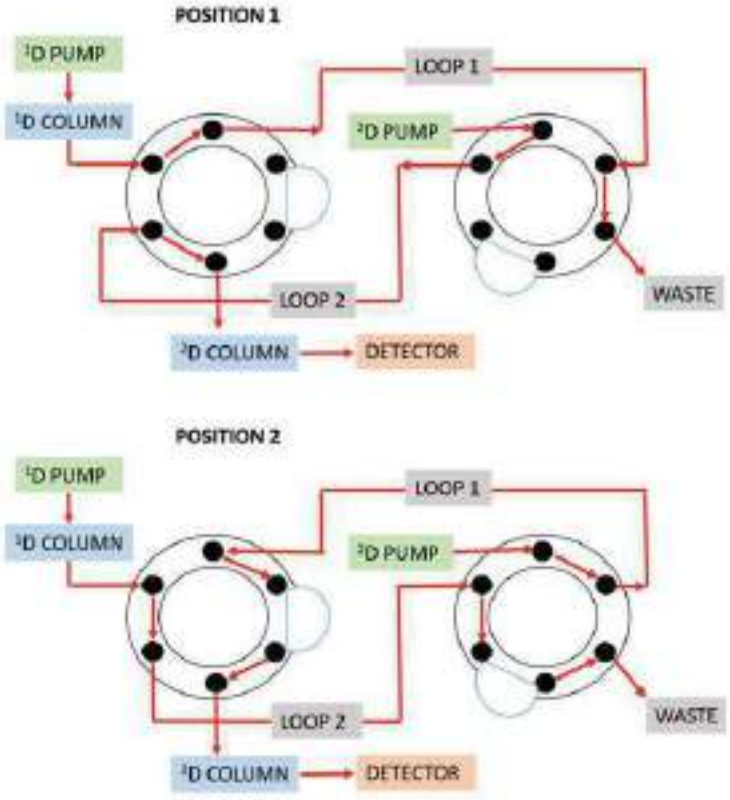


Fig. 2 Generic scheme of a passive-modulation interface in 2D-LC. Top—Position 1: Sampling loop 1— Injecting loop 2 in ^{2}D; Bottom—Position 2: Sampling loop 2—Injection ^{2}D

the ^{1}D column, and it is sampled afterward by loop 1, while ^{2}D effluent allows that content of loop 2 enters in the ^{2}D column; whereas in position 2, ^{1}D effluent passes through the ^{1}D column, but, in this case, it is sampled by loop 2, at the same time that ^{2}D effluent introduces the content of loop 2 in the ^{2}D column. Other common arrangements include the use of 8 or 10-port valves instead [4].

Although the predominant use of comprehensive 2D-LC is led by biopharmaceuticals, peptides, synthetic polymers, and traditional Chinese medicines (TCMs), metabolomics and lipidomics applications are currently widespread and are starting to gain importance. There is an increasing tendency in its use in both fields in the last few years [4]. Until 2010, the application of 2D-LC to metabolomics or lipidomics was negligible. A continuous increase in the number of published manuscripts within these omics fields can be observed (following a similar trend to the observed in Fig. 1 for the overall of 2D-LC references).

Setting up a 2D-LC is not straightforward. Many considerations are required to select and optimize different parameters (e.g., modulation interface, column dimensions, compatibility of mobile phases, isocratic or gradient elution in both dimensions, flow rates) [5].

Special attention is required in the choice of stationary phases and mobile phase composition. As there are two columns, the number of possible combinations is large, and uncorrelated separations (i.e., orthogonal mechanisms) are highly desirable. Depending on the application, some separation mode combinations could suit better than others, although there are some still unexplored. The preferred separation modes for 2D-LC metabolomics and lipidomics include reverse phase (RP), hydrophilic interaction liquid chromatography (HILIC), normal phase (NP), ion-exchange (IEC) and chiral chromatography [4]. Examples of different setups already employed will be described later in this chapter.

Optimization of comprehensive 2D-LC methods can be done through two different approaches: sample-independent (e.g., Pareto optimization [13]), or sample-dependent (e.g., the PIOTR program [14]). The first one aims to define two or more objective parameters, based on the impact that multiple method settings (e.g., gradient slope, column dimensions, flow rates) have through theoretical relations [13]. In contrast, the sample-dependent approach optimizes all chemical parameters that affect retention and selectivity, such as the mobile phase composition, temperature, pH, and buffer strength, for a specific sample. It is based on a very small number of experiments, modeling of the retention, and generalizing band-broadening behavior of individual sample components [14].

When MS detectors are coupled to LC × LC, some widespread experimental issues such as ion suppression effects (typical in LC-MS) are reduced. This hyphenation implies that quantitation

is also more reliable, as matrix effects are minimized. However, in LC × LC-MS, the splitting ^{2}D flow before entering the ion source is frequent, which may cause a decrease in sensitivity and a significant peak broadening [12]. Other main drawbacks associated with comprehensive 2D-LC are related to conceptual and instrumental complexity, long analysis time (normally up to one hour, instead of a few minutes analysis offered by 1D-LC), solvent compatibility, lower detection sensitivity and increased difficulty in data analysis [4]. Different strategies have started to arise to solve these downsides.

A common problem is solvent strength mismatch, which causes peak distortion. In the cases that the ^{1}D effluent is a relatively strong injection solvent when compared to ^{2}D eluent, analytes are not strongly retained by the stationary phase and, consequently, the retention mechanism employed may be affected. If the unretained peak elutes mostly in the dead volume instead of in its normal location, this phenomenon is known as breakthrough [4].

For instance, active modulation has emerged as an effective manner of handling solvent strength mismatch, reducing breakthrough and possibly enhancing detection sensitivity. Several active modulation approaches have been recently proposed (e.g., active solvent modulation (ASM) [15], stationary-phase-assisted modulation (SPAM) [16], vacuum-evaporation modulation (VEM) [17]). More details about these approaches can be found in a review article by Pirok et al. [4].

The need for data analysis software is another bottleneck in 2D-LC, which is also currently being faced. Examples of commercially available software packages are Chromsquare from Shimadzu and GC Image LC × LC Edition Software from GC Image™ (including LC Image, LC Project, and Image Investigator). Other used approaches employ more sophisticated chemometric tools for the deconvolution and resolution of overlapping peaks. Examples of these methods include curve resolution algorithms such as the Multivariate Curve-Resolution by Alternating Least Squares (MCR-ALS) or Parallel Factor Analysis (PARAFAC) [18]. Both methods had been successfully used in the analysis of comprehensive gas chromatography data (GC × GC). MCR-ALS decomposes the experimental data following a bilinear model in the spectral and chromatographic contributions. In contrast, PARAFAC employs a trilinear model which needs higher data quality (i.e., high reproducibility in the observed retention times and peak shapes). Consequently, the use of these methods for the analysis of LC × LC data has also been proposed. However, the inherent properties of LC × LC data prevent from this totally trilinear behavior (i.e., minor changes in the retention time between modulations and different peak shapes) and, therefore, it seems that the use of MCR-ALS is more advisable (in particular, when dealing with LC × LC-MS, as in the case of metabolomics and lipidomics studies).

However, due to the huge size of the generated datasets, it is also highly advisable a preliminary step of data reduction. This compression can be applied either in the spectral mode, to reduce the total number of m/z values leaving only these that can be considered as more interesting (i.e., using, for instance, the Regions of Interest strategy), or in the time direction by compacting all the measured retention times (e.g., applying one-dimensional wavelet analysis strategy) [19].

Another challenge that is being faced is trying to reduce analysis time as much as possible, achieving in some cases, chromatographic runs of less than one hour [20–22].

All these recent instrumental and data analysis developments have allowed the application of this two-dimensional chromatography in new research fields. Examples in the fields of metabolomics and lipidomics are described more in detail throughout the chapter, with special emphasis in stationary phases chemistry and most employed ways to combine them, to obtain acceptable selectivities. It is important to notice that mobile phase composition is also a decisive parameter that needs to be taken into account when developing an LC × LC method. Although lipidomics is a branch of metabolomics, lipids are analyzed separately from the more water-soluble metabolites due to their generally hydrophobic character. The distinction between lipidomics and metabolomics 2D-LC methods is required though, pursuing a more specific approach for each group of compounds.

Thus, the following sections will be focused on comprehensive 2D-LC in lipidomics and metabolomics, respectively. There will be a special emphasis on strategies to set up methods for each field separately, according to most suitable combinations that have been applied.

2 Lipidomics

The main LC chromatographic modes employed in lipid analysis include RP, HILIC, NP, and silver-ion (Ag), which can be considered a particular type of NP [23]. Consequently, it is not surprising that the most common comprehensive 2D-LC combinations in lipidomic studies have been RP × HILIC, HILIC × RP, NP × RP, and Ag × RP. In addition, some studies using strong anion exchange (SAX) – RP, although not fully comprehensive, are also found in the literature [24]. All these combinations are considered orthogonal separations, which basically means that their selectivities are complementary enough and allows achieving enhanced peak capacities. A more detailed explanation of the main characteristics of each separation mode and their combinations is presented below.

RP is the most selected chromatographic mode for lipid analysis, and in general, in most of LC × LC applications. RP is a flexible and versatile retention mechanism, and therefore, it is suitable for any of the two chromatographic dimensions. In the particular lipidomics case, RP separates lipids according to the degree of hydrophobicity of fatty acids alkyl chains and the number and position of the double bonds present in these chains [23]. RP presents a series of experimental advantages as can be considered robust, fast, uses MS-compatible solvents, needs low reequilibration times, handles high temperatures and pressures, and provides high-resolution separations [5]. Hence, RP is widely employed as the ^{2}D.

NP and HILIC chromatographic mechanisms can be employed to separate lipids according to the different polarity of their head groups and, as a consequence, allows distinguishing between the different lipid classes [5]. Thus, the combination of the previously described RP separation with either of these two modes presents several benefits, due to the amphiphilic nature of lipids. Therefore, in the NP × RP and HILIC × RP, separation of lipids takes place firstly depending on their classes and polarities, and subsequently, according to the hydrophobicity of fatty acid alkyl chains and the number and positions of the double bonds.

NP × RP coupling is much more complicated, as the miscibility of the two mobile phases from the two chromatographic dimensions presents a high degree of incompatibility (i.e., the eluting strength of the ^{1}D eluent in NP mode is stronger than that of the mobile phase on the head of the ^{2}D RP column [17]). In an attempt to overcome these difficulties, some approaches with increased experimental drawbacks can be applied, that is, thermal and vacuum-evaporation modulations [25]. This NP × RP coupling also presents significant solvent-immiscibility problems that require complex interfaces [26]. For these reasons, RP × NP is usually discarded due to its slow separations (extremely long analysis times), slow column reequilibration times of the NP column and minor solvent-MS compatibility.

HILIC mode presents several improvements when compared to NP for 2D-LC. For instance, HILIC shows better repeatability, shorter equilibration times, and longer column lifetimes [26, 27]). Thereby, HILIC is becoming a popular alternative to NP in the lipidomics (and, also metabolomics) field, as shown in recent publications which are often based in combinations between RP and HILIC [18, 20–23, 26]. Nevertheless, in the case of HILIC × RP, some experimental difficulties should also be considered, that is, solvent strength mismatch. However, recent work has been focused on solving this drawback and the application of active modulation strategies (see above) has already been proposed.

Another less common setup is RP × HILIC. The main advantage of this coupling is that it does not require active modulation strategies due to good solvent compatibility and lower risk of

solvent strength mismatch. However, the use of HILIC as the ^{2}D mechanism is less frequent because it may suffer from injection effects [28]. An alternative to overcome this issue is using a narrow gradient span [26]. Particularly in lipidomics, RP × HILIC could be a good option despite these limitations. HILIC in ^{2}D allows for partial separation of lipid species within individual classes [23], because more variations in the nonpolar side of the structure are found than due to the polar head-groups. Consequently, RP as ^{1}D would perform a longer separation and therefore, a higher separation power can be obtained [23].

Differences between these HILIC × RP and RP × HILIC couplings are more easily understood when comparing the obtained 2D-LC chromatograms. An example of each type is presented below (Fig. 3).

In HILIC × RP mode (Fig. 3a), the ^{1}D (HILIC) separates several lipid classes according to their polarity. For instance, it is possible to differentiate lysophosphatidylethanolamines (LPE) and lysophosphatidylcholines (LPC) according to their different head groups. Then, the ^{2}D (RP) separates lipids on the basis of the length of the fatty acyl chains and the number and position of the double bonds. This is the reason why carotenoids, cholesteryl esters (CEs), and triacylglycerides (TGs) are separated in the ^{2}D, despite they elute in the same fraction of the ^{1}D.

In RP × HILIC (Fig. 3b), an opposite separation is observed. For example, it becomes clearer how shorter fatty acids alkyl chains of smaller phosphatidylcholines (PC) elute before longer ones, as

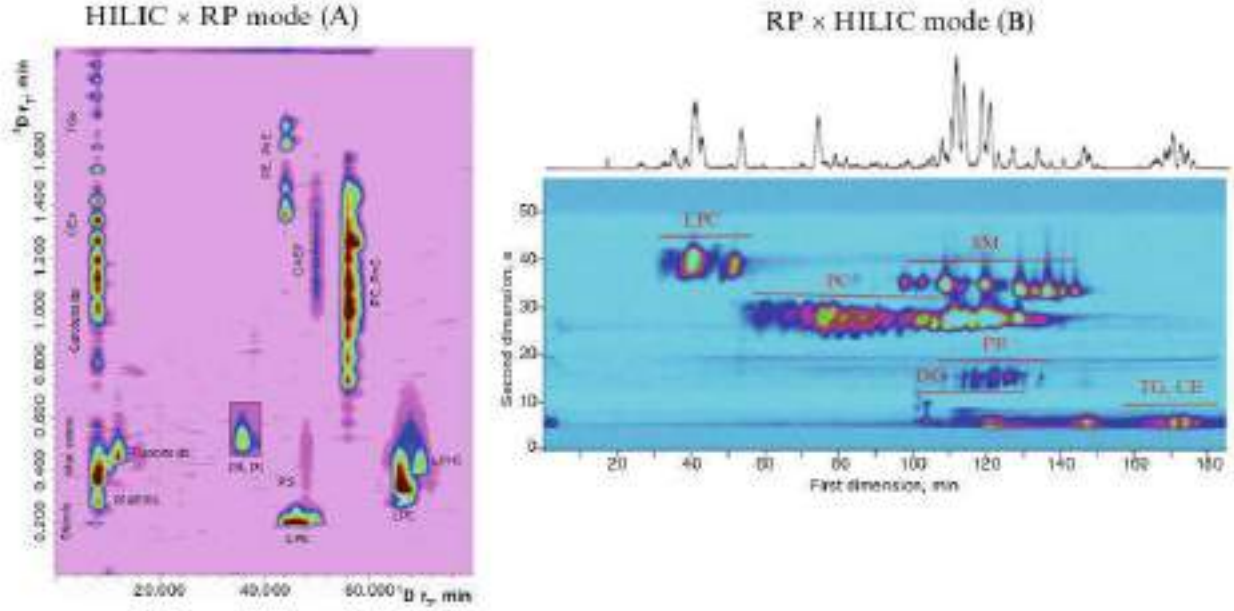


Fig. 3 Comparison of the application of HILIC × RP (**a**) and RP × HILIC (**b**) modes to lipidomic studies, and how lipids are though separated depending on their polarity and afterward by the length of the fatty acid alkyl chains, or vice versa. Reprinted with permission from [20] (**a**) and [26] (**b**)

they are less hydrophobic, and consequently, less retained by the RP stationary phase. On the contrary, the ^{2}D (HILIC) allows differentiation between phosphatidylcholines (PC) and sphingomyelins (SM), that otherwise will be overlapped, if only the ^{1}D (RP) was employed.

Finally, two chromatographic modes are less common in the literature but present some specific advantages for lipidomics studies. For instance, the use of Ag × RP is especially interesting when triacylglycerides (TGs) are the target analytes [29], as the ones with a different number of double bonds can be separated by the ^{1}D column. In constrast, those with different partition numbers are well resolved by the ^{2}D column. SAX resin chromatography allows for separating anionic lipids from neutral lipids. SAX × RP combination benefits from easier modulation strategies (i.e., relatively lack of difficulty in developing active modulation methods), although this coupling is slightly less orthogonal than other combinations [24].

Table 1 outlines the benefits and downsides of the chromatographic mode combinations already described, emphasizing its suitability in lipidomics analysis.

In contrast, the main aim of Table 2 is to illustrate examples of applications of 2D-LC to lipidomics, with special emphasis in comprehensive setups, including some details of the chromatographic conditions employed and some insights on the type of samples and target analytes.

Table 2 shows that the main separation mechanisms that have dominated lipidomic in 2D-LC include combinations of HILIC, RP, and NP. The recent improvements in solvent compatibility issues due to active modulation strategies have been especially relevant in the development of HILIC × RP and NP × RP, reducing solvent immiscibility issues, solvent strength mismatch, and breakthrough.

These applications are mainly focused on clinical studies dealing, for instance, with human plasma [23, 24, 26, 30] and in different animal tissues (e.g., mouse brain regions) [21], to study lipidome changes due to illnesses such as anxiety disorder [21], lacunar infarction [25], or atherosclerosis [32]). These studies allowed for the identification of some potential lipid biomarkers from a variety of families. There are also some studies focused on environmental research, which targeted some bioindicator organisms [18] or model organisms exposed to pollutants [20].

Comprehensive 2D-LC is a powerful approach in untargeted studies, such as lipid profiling, which is the goal of most of the publications shown in Table 2. The number of lipid species identified in these studies varies between 100 and 500, and from up to 17 lipid classes. Some examples where a specific group of compounds are targeted are also found (i.e., phospholipid and sphingomyelin [21], TGs [29]).

Table 1
Advantages and disadvantages of different combinations of separations in lipidomics

Advantages	Drawbacks
HILIC × RP	
• Separation based on polar head-groups of lipids in ^{1}D, and on hydrophobic interactions with nonpolar parts of lipids in ^{2}D • Good solvent compatibility • Analytes are preconcentrated in the interface when using traps (gain in sensitivity) • Possibility to do lipid profiling	• Requires active modulation • RP as ^{2}D cannot provide a highly efficient separation in a short period of time in the case of lipids
RP × HILIC	
• Separation of lipids based on hydrophobic interactions with the length of the fatty acyl chains and the number and positions of double bonds in ^{1}D, and on polar head-groups in ^{2}D, providing partial separation of lipid species within individual classes • Excellent solvent compatibility • Possibility to do lipid profiling	• HILIC as ^{2}D may suffer from injection effects
NP × RP	
• Separation of lipids based on the polar head groups (^{1}D) and on the different fatty-acyl chains (^{2}D) • Possibility to do lipid profiling • Analytes are enriched in the interface	• Requires thermal and/or vacuum-evaporation modulations • Competes with HILIC mode in ^{1}D. NP presents longer equilibration times, lower run-to-run repeatability, rapid column deterioration • Lower solvent compatibility
SAX × RP	
• Separation according to electrical propensities SAX resin, and molecular separation in RP • Capability of detecting low abundant species • Possibility to do lipid profiling	• Lower peak capacity than UPLC due to the use of nanoLC • Lower orthogonality
Ag × RP	
• Good separation of TGs, in ^{1}D according to different double bonds (critical pairs) and in ^{2}D with different partition number	• Applied in target analysis of a specific family (e.g., TGs)

3 Metabolomics

In 1D-LC metabolomics, RP is often used for the analysis of hydrophobic compounds, and the profiling of medium to nonpolar metabolites. However, for hydrophilic and neutral compounds, HILIC or NP modes are more suitable. Although NP presents

Table 2
Summary of applications of online 2D-LC in lipidomics

Year	Separation	Columns	Mobile phase solvents	Gradient	Detection	Sample	Ref.
2018	HILIC × RP	^{1}D: Ascentis Express HILIC (150 × 2.1 mm i.d., 2.7 μm d.p) ^{2}D: Titan C18 (50 × 4.6 mm i.d., 1.9 μm d.p)	^{1}D: (A) ACN:10 mM $HCOONH_4$ (98:2, v/v); (B) ACN:MeOH:10 mM $HCOONH_4$ (55:35:10, v/v/v) ^{2}D: (A) ACN:MeOH:10 mM $HCOONH_4$ (55:35:10, v/v/v); (B) IPA + 0.1% HCOOH	Both	HRMS; MS/MS	Mediterranean mussels; lipid profiling	[20]
2018	HILIC × RP	^{1}D: Kinetex HILIC (150 × 2.1 mm i.d., 2.6 μm, d.p) ^{2}D: Acquity UPLC BEH C18 (50 × 2.1 mm i.d., 1.7 μm d.p)	^{1}D: (A) 10 mM $HCOONH_4$; (B) Acetone ^{2}D: (A) H_2O:ACN:5 mM $HCOONH_4$ (50:50 v/v); (B) H_2O:ACN:5 mM $HCOONH_4$ (5:95 v/v)	Both	HRMS	Mouse brain; phospholipid and sphingomyelin profiling	[21]
2018	RP × HILIC	^{1}D: RP ZORBAX Eclipse XDB-C18 (150 × 2.1 mm i.d..; 5 μm d.p) ^{2}D: Kinetex HILIC (30 × 3 mm i.d.; 2.6 μm d.p)	^{1}D: (A) ACN:IPA (1:2, v/v) + 0.1% HCOOH; (B) H_2O + 0.1% HCOOH ^{2}D: (A) H_2O:5 mM CH_3COONH_4 pH = 5.5 (CH_3COOH); (B) ACN	Only ^{1}D	MS/MS	Rice exposed to As; lipid profiling	[18]
2017	NP − RP	^{1}D: Rx-SIL silica (150 × 2.1 mm i.d., 5 μm d.p) ^{2}D: Poroshell 120 EC C8 (50 × 2.1 mm i.d.., 2.7 μm d.p)	^{1}D: (A) Hexane; (B) IPA (2% H_2O + 5 mM $HCOONH_4$); (C) MeOH (2% H_2O + 5 mM $HCOONH_4$) ^{2}D: (A) MeOH:5 mM $HCOONH_4$; (B) H_2O:5 mM $HCOONH_4$	Both	HRMS; MS/MS	Human plasma (patients with lacunar infarction); lipid profiling	[25][a]

2017	RP × HILIC	^{1}D: Acquity UPLC (BEH) C18 (150 × 2.1 mm i.d., 1.7 μm d.p) ^{2}D: Acquity UPLC (BEH) HILIC (50 × 2.1 mm i.d., 1.7 μm d.p)	^{1}D: (A) ACN:H_2O:10 mM $HCOONH_4$ (60:40, v/v); (B) IPA:ACN:10 mM $HCOONH_4$ (90:10, v/v) ^{2}D: (A) H_2O:10 mM $HCOONH_4$; (B) ACN	Both	HRMS	lipid standards; pooled human plasma purchased	[26]
2015	HILIC – RP	^{1}D: Ascentis Express HILIC (150 × 2.1 mm i.d., 2.7 μm d.p) ^{2}D: Ascentis Express C18 (75 × 2.1 mm i.d., 2.7 μm d.p)	^{1}D: (A) ACN; (B) H_2O:10 mM $HCOONH_4$; pH = 3 (HCOOH) ^{2}D: (A) THF:ACN:IPA: H_2O:10 mM $HCOONH_4$; pH = 3 (HCOOH); 45 °C	Only ^{1}D	ozonolysis; HRMS	Egg yolk phospholipids extract	[22][b]
2015	NP – RP	^{1}D: Rx-SIL silica (150 × 2.1 mm i.d., 5 μm d.p) ^{2}D: Poroshell 120 EC C8 (50 × 2.1 mm i.d.., 2.7 μm d.p)	^{1}D: (A) Hexane; (B) IPA (2% H_2O + 5 mM $HCOONH_4$); (C) MeOH (2% H_2O + 5 mM $HCOONH_4$) ^{2}D: (A) MeOH:5 mM $HCOONH_4$; (B) H_2O:5 mM $HCOONH_4$	Both	HRMS; MS/MS	Human plasma (patients with benign breast tumor and breast cancer); lipid profiling	[30][a]
2015	RP × HILIC	^{1}D: Acquity UPLC (BEH) C18 (150 × 1 mm i.d., 1.7 μm d.p) ^{2}D: Core–shell silica Cortecs HILIC (50 × 3 mm i.d., 2.7 μm d.p)	^{1}D: (A) H_2O:5 mM $HCOONH_4$; (B) ACN:IPA (1:2, v/v) + 0.5% $HCOONH_4$ ^{2}D: (A) H_2O:5 mM $HCOONH_4$ pH = 5.5; (B) ACN	Both	HRMS; MS/MS	Human plasma; porcine brain; lipid profiling	[23]
2014	HILIC – RP	^{1}D: Ascentis Express HILIC (150 × 2.1 mm i.d., 2.7 μm d.p) ^{2}D: Ascentis Express C18 (75 × 2.1 mm i.d., 2.7 μm d.p)	^{1}D: (A) ACN; (B) H_2O:10 mM $HCOONH_4$; pH = 3 (HCOOH) ^{2}D: (A) THF:ACN:IPA: H_2O:10 mM $HCOONH_4$; pH = 3 (HCOOH); 45 °C	Only ^{1}D	ozonolysis; HRMS	Rat liver; elucidation of phosphatidylcholine isomers	[31][b]

(continued)

Table 2
(continued)

Year	Separation	Columns	Mobile phase solvents	Gradient	Detection	Sample	Ref.
2014	NP – RP	^{1}D: Rx-SIL silica (150 × 2.1 mm i.d., 5 μm d.p) ^{2}D: Poroshell 120 EC C8 (50 × 2.1 mm i.d., 2.7 μm d.p)	^{1}D: (A) Hexane; (B) IPA (2% H_2O + 5 mM $HCOONH_4$); (C) MeOH (2% H_2O + 5 mM $HCOONH_4$) ^{2}D: (A) MeOH:5 mM $HCOONH_4$; (B) H_2O:5 mM $HCOONH_4$	Both	HRMS; MS/MS	Human plasma from atherosclerosis patients; lipid profiling	[32][a]
2013	NP – RP	^{1}D: Rx-SIL silica (150 × 2.1 mm i.d., 5 μm d.p) ^{2}D: Poroshell 120 EC C8 (50 × 2.1 mm i.d., 2.7 μm d.p)	^{1}D: (A) Hexane:IPA:H_2O (30:70:2, v/v/v) + 5 mM $HCOONH_4$; (B) MeOH:H_2O (100:2, v/v) + 5 mM $HCOONH_4$ ^{2}D: (A) MeOH:H_2O:5 mM $HCOONH_4$ (50:50, v/v); (B) MeOH:5 mM $HCOONH_4$	Both	HRMS; MS/MS	Peritoneal dialysis patients; lipid profiling	[33][a]
2013	SAX – RP	^{1}D: lab made ^{2}D: lab made	^{1}D: From 10 mM to 1 M CH_3COONH_4 ^{2}D: (A) 0.05% NH_4OH; (B) MeOH/CH_3CN/IPA + 0.05% NH_4OH; and 5 mM $HCOONH_4$	Both	MS/MS	Healthy human plasma; lipid profiling	[24][b]
2012	Ag × RP	^{1}D: Microbore TSK gel SP-2SW column (150 × 1 mm i.d., 5 μm d.p) ^{2}D: Platinum™ EPS C18 (33 × 7 mm, i.d., 1.5 μm)	^{1}D: (A) Hexane +0.3% ACN; (B) Hexane +1.3% ACN ^{2}D: (C) ACN; (D) IPA–Hexane (2:1, v/v)	Only ^{1}D	HRMS; MS/MS	Peanut oil and mouse tissue; TGs analysis	[29]

[a]Non-stop-flow NP/RP 2D-LC system with vacuum evaporation interface
[b]Not fully comprehensive setups

some solvent incompatibilities issues when coupled to MS detectors and, consequently, other options such as aqueous normal phase chromatography (ANP) are preferred [8, 34]. Besides, for separating hydrophilic and charged small molecules, capillary electrophoresis seems a good option to be considered [34]. Chiral columns are also very useful in specific applications, for instance, in the pharmaceutical industry [35].

2D-LC combinations of separation modes more used in metabolomics include HILIC × RP, RP × RP, and Chiral × Chiral. The main advantages of HILIC × RP, as already described in the lipidomics section, are good orthogonality, good MS and solvent compatibility, and high applicability (i.e., for the simultaneous separation of polar and nonpolar metabolites). However, active modulation strategies are highly recommended when employing this combination. Below, a more detailed explanation of RP × RP and Chiral × Chiral is included, as they are not described in the previous lipidomics section.

In the case of highly polar metabolites, one of the most common combinations is HILIC × RP [19, 36–39], although NP × RP combinations have also been employed in other fields, such as in food analysis [40]. However, when dealing with hydrophobic compounds such as anabolic steroids, HILIC may not be the best option, as it is unlikely to provide sufficient separation [41].

RP × RP seems a good alternative in these cases. To overcome orthogonality problems that arise from the fact that both dimensions employed the same retention mechanism, some differences (i.e., different pH values, temperature, organic modifiers, or varying gradient composition between subsequent ^{2}D runs) are highly desirable from one chromatographic dimension to the other one [12, 28]. Thus, a better chromatographic resolution is achieved. Other advantages of RP × RP are an excellent robustness, repeatability, well-known separation mechanisms and adequate selectivity [12]. However, as well as HILIC × RP, the RP × RP coupling presents some solvent-compatibility issues, that can also be solved using active modulation strategies already described.

Chiral × Chiral has been used for enantioselective amino acid analysis of peptide and protein hydrolysates. Two chiral stationary phases are combined with similar chemoselectivity but orthogonal stereochemistry (i.e., opposite chiral recognition due to inverted configurations in C8 and C9 of the recognition site [35]). Therefore, this methodology is useful for stereoconfiguration profiling as, for instance, the analysis of chiral amino acids in peptide therapeutics [35].

Another potential strategy that has not been widely employed yet in metabolomics is HILIC × HILIC. This fact is mainly due to the shallow degree of orthogonality between both dimensions depending on the target analytes and the used stationary phases, For instance, D'Attoma et al. found low orthogonality in peptide

analysis using a variety of chromatographic columns, organic modifiers, and temperatures [28]. However, this drawback was not found by Wang et al. separating saponins using a TSK gel Amide-80 and a polyhydroxyethyl A columns as the ^{1}D y ^{2}D separations respectively, employing acetonitrile and aqueous mobile phases, at acid pH [42]. Nevertheless, further improvements are needed in terms of chromatography efficiency [5].

When considering the metabolomics applications found in the literature, the most frequent combinations are HILIC × RP and RP × RP, depending on the type of metabolites targeted.

Table 3 includes a series of metabolomics examples of applications of online comprehensive 2D-LC, specifying chromatographic conditions employed, the type of samples and target analytes.

The main difference with the lipidomics field is that applications in metabolomics are usually more specific, as they target some groups of compounds such as polyphenols [12, 38], flavones [36, 48], or steroids [41, 43]. Matrices, where these compounds are found, are quite diverse, as it includes, for instance, green cocoa beans [36], bovine urine [41], apples [38], or wine [12]. Food analysis is the principal field of current applications, with some exceptions in environmental studies [19], microbial metabolites [45] or biological fluids [41, 43].

Although there are less frequent than in lipidomic analysis, there are also some examples of metabolic profiling in licorice [39, 44] or rice [19]. In these untargeted studies, the total number of metabolites detected ranges between 80 and 150, while it was possible to identify between 40 and 140.

4 Conclusions

The inherent complexity of metabolomics and lipidomics samples in which hundreds (or thousands) of compounds can coexist, has pushed to the limits the traditional one-dimensional separation techniques. The quest for a better description of these samples has led to the development of new technologies merging different separation strategies. Therefore, the combination of different chromatographic modes (i.e., GC × GC or LC × LC) or its combination with another separation technique such as capillary electrophoresis or ion mobility, has allowed for improved sample characterization.

Among all these new technologies, comprehensive 2D-LC seems a promising technique for the analysis of complex samples, with especial interest in untargeted lipidomics and metabolomics. Recent technical developments have increased their applicability overcoming experimental issues such as the solvent immiscibility and enhancing detection sensitivity. Also, the development of a new generation of chromatographic stationary phases increases the

Table 3
Summary of applications of online comprehensive 2D-LC in metabolomics

Year	Separation	Columns	Mobile phase solvents	Gradient	Detection	Sample and analyte	Ref.
2018	HILIC × RP	^{1}D: LiChrospher DIOL 5 (150 × 1.0 mm i.d., 5 μm d.p) ^{2}D: Ascentis Express C18 (30 × 4.6 mm, 2.7 μm d.p)	^{1}D: (A) ACN (2% CH_3COOH); (B) MeOH:H_2O:CH_3COOH (95:3:2 v/v) ^{2}D: (A) H_2O (0.05% TFA); (B) ACN (0.05% TFA)	Both	UV-Vis; MS/MS	Green cocoa beans; flavan-3-ol monomers and procyanidin oligomers	[36]
2018	HILIC × RP	^{1}D: Develosil 100 Diol column (250 × 1.0 mm i.d., 5 μm d.p) ^{2}D: Kinetex C18 (50 × 4.6 mm i.d., 2.6 μm d.p)	^{1}D: (A) ACN (99:1, v/v); (B) MeOH: H_2O:CH_3COOH (94.05:4.95:1, v/v/v) ^{2}D: (A) H_2O (0.1% HCOOH); (B) ACN (0.1% HCOOH)	Both	UV-Vis	Cocoa extract; procyanidins isomers	[37]
2018	RP × RP	^{1}D: Zorbax Eclipse Plus C18 (50 × 2.1 mm i.d., 1.8 μm d.p) ^{2}D: Chromolith Performance NH_2 (10 × 4.6 mm)	^{1}D: (A) H_2O (0.1% HCOOH); (B) MeOH ^{2}D: (A) 20 mM $HCOONH_4$ (pH = 4.3); (B) ACN	Both	UV-Vis	standards; phenolic and flavone antioxidants	[12]
2018	RP × RP	^{1}D: Waters HSS Cyano (150 × 1 mm i.d., 1.8 μm d.p) ^{2}D: Waters Phenyl column (50 × 1 mm i.d., 1.7 μm d.p)	^{1}D: (A) H_2O:ACN (90:10 v/v) (0.1% HCOOH); (B) H_2O:ACN (10:90 v/v) (0.1% HCOOH) ^{2}D: (A) H_2O:ACN (90:10 v/v) (0.1% HCOOH); (B) H_2O:ACN (10:90 v/v) (0.1% HCOOH)	Both	MS	Bovine urine; residues of sulfonamides, beta-agonists, and steroids	[43]
2018	RP × RP	^{1}D: Ascentis Cyano column (150 × 1.0 mm i.d., 2.7 μm d.p) ^{2}D: Ascentis Express C18 column (50 × 2.1 mm i.d., 2.7 μm d.p)	^{1}D: (A) H_2O (0.1% CH_3COOH); (B) ACN (0.1% CH_3COOH) ^{2}D: (A) H_2O (0.1% CH_3COOH); (B) ACN (0.1% CH_3COOH)	Both	UV-Vis; MS/MS	Liquorice; metabolic profiling	[44]
2018	RP × RP	^{1}D: Ascentis Express C8 column (150 × 2.1 mm i.d., 2.7 μm d.p) ^{2}D: Cyano column Zorbax SB-CN (30 × 4.6 mm i.d., 1.8 μm d.p)	^{1}D: (A) H_2O (0.1% HCOOH); (B) MeOH (0.1% HCOOH) ^{2}D: (A) H_2O (0.1% HCOOH); (B) ACN (0.1% HCOOH) 1-butanol	Both	HRMS	Bovine urine; hydrophobic compounds; anabolic-steroid residues	[41]

(continued)

Table 3
(continued)

Year	Separation	Columns	Mobile phase solvents	Gradient	Detection	Sample and analyte	Ref.
2018	Chiral × Chiral	lab made	^{1}D: MeOH:CH_3COOH:$HCOONH_4$ (98:2:0.5, v/v/w) ^{1}D: MeOH:CH_3COOH:$HCOONH_4$ (98:2:0.5, v/v/w)	None	UV-Vis	Peptide antibiotic drugs gramicidin and bacitracin; amino acids	[35]
2017	HILIC × RP	^{1}D: HILIC TSK gel amide-80 column (250 mm × 2.0 mm i.d., 5 μm d.p) ^{2}D: Kinetex C18 (50 mm × 2.1 mm i.d., 1.7 μm d.p)	^{1}D: (A) ACN; (B) 5 mM CH_3COONH_4 pH = 5.5 (CH_3COOH) ^{2}D: (A) H_2O (0.1% HCOOH); (B) ACN (0.1% HCOOH)	Both	HRMS	Rice; metabolite profiling	[19]
2017	HILIC × RP	^{1}D: Luna HILIC (150 mm × 2.0 mm i.d., 3.0 μm d.p) ^{2}D: Titan™C18 (50 mm × 3.0 mm i.d., 1.9 μm d.p)	^{1}D: (A) H_2O:ACN (80:20 v/v) (0.1% CH_3COOH); (B) ACN (0.1% CH_3COOH) ^{2}D: (A) H_2O (0.1% CH_3COOH); (B) ACN (0.1% CH_3COOH)	Both	MS/MS	Typical Italian apple variety; polyphenols	[38]
2016	HILIC × RP	^{1}D: SeQuant ZIC HILIC (150 × 1.0 mm i.d., 3.5 μm d.p) ^{2}D: Ascentis Express C18 (50 × 4.6 mm i.d., 2.7 μm d.p)	^{1}D: (A) ACN; (B) H_2O:10 mM $HCOONH_4$ (pH = 5.0) ^{2}D: (A) H_2O (0.1% HCOOH); (B) ACN	Both	MS/MS	Roots licorice; metabolic profiling	[39]

2016	RP × RP	^{1}D: COSMOSIL silica-based reversed phase (250 mm × 4.6 mm i.d., 5 μm d.p) ^{2}D: COSMOSIL C18-MS-II UPLC (50 mm × 30 mm i.d., 2.6 μm d.p)	^{1}D: (A) H_2O; (B) MeOH ^{2}D: (A) H_2O; (B) ACN	Both	HRMS	Rice medium of *C. globosum* SNSH1-5; microbial metabolites	[45]
2016	RP × RP	^{1}D: monolithic zwitterionic polymethacrylate 0.53 mm i.d. BIGDMA–MEDSA microcolumn ^{2}D: Five short commercial core-shell or silica-based monolithic columns/three commercial silica-based monolithic columns with octadecyl stationary phases	^{1}D: (A) H_2O:10 mM $HCOONH_4$ (pH = 3.1 HCOOH); (B) ACN:10 mM $HCOONH_4$ (pH = 3.1 HCOOH) ^{2}D: (A) H_2O:10 mM $HCOONH_4$ (pH = 3.1 HCOOH); (B) ACN:10 mM $HCOONH_4$ (pH = 3.1 HCOOH)	Both	UV-Vis	Standards; flavones and related polyphenolic compounds	[46]
2016	RP × RP	^{1}D: Brownlee Choice C18 (50 mm × 2.1 mm i.d., 5 μm d.p) ^{2}D: Xbridge Shield C18 (50 mm × 2.1 mm i.d., 5 μm d.p)	^{1}D: H_2O:MeOH (87:13, v/v) (0.1% CH_3COOH) ^{2}D: (A) H_2O:MeOH (90:10, v/v) (0.1% HCOOH)	None	UV-Vis	Standards; formic acid, benzyl alcohol, catechol, vanillin, and guaiacol	[47]

number of potential setup combinations, allowing for an expanded number of applications either in a targeted or an untargeted manner. Finally, it is worth to mention the complementarity of the information provided by the same stationary phases depending on the order used in the analysis. A clear example in lipidomics is the possibility of obtaining the primary separation on the lipid class or the chain length/hydrogen bonds depending on the first chromatographic mode (HILIC or RP, respectively) which can be a significant help for the compound identification.

However, there are still some weaknesses that should be improved to achieve a major popularization in omics communities. On the one hand, a simplification in the required instrumentation will ease its use for new research groups, which could be promoted by vendor companies. New developments in data mining tools could help the automation of the entire pipeline, which could also favor the use of these techniques in routine analysis. On the other hand, more work has to be done in order to improve experimental factors such as the analysis time, because it can limit the use of 2D-LC in high-throughput studies. If these limitations can be overcome, a successful future can be expected for the 2D-LC.

Acknowledgments

The research leading to these results has received funding from the Spanish Ministry of Science and Innovation (MCI, Grant CTQ2017-82598-P). MPC acknowledges a predoctoral FPU 16/02640 scholarship from Spanish Ministry of Education and Vocational Training (MEFP).

References

1. Zhang H, Jiang JM, Zheng D et al (2019) A multidimensional analytical approach based on time-decoupled online comprehensive two-dimensional liquid chromatography coupled with ion mobility quadrupole time-of-flight mass spectrometry for the analysis of ginsenosides from white and red ginsengs. J Pharm Biomed Anal 163:24–33. https://doi.org/10.1016/j.jpba.2018.09.036
2. Sarrut M, Corgier A, Crétier G et al (2015) Potential and limitations of on-line comprehensive reversed phase liquid chromatography×supercritical fluid chromatography for the separation of neutral compounds: An approach to separate an aqueous extract of bio-oil. J Chromatogr A 1402:124–133. https://doi.org/10.1016/j.chroma.2015.05.005
3. Stoll DR, Carr PW (2017) Two-dimensional liquid chromatography: a state of the art tutorial. Anal Chem 89(1):519–531. https://doi.org/10.1021/acs.analchem.6b03506
4. Pirok BWJ, Stoll DR, Schoenmakers PJ (2019) Recent developments in two-dimensional liquid chromatography – fundamental improvements for practical applications. Anal Chem 91(1):240–263. https://doi.org/10.1021/acs.analchem.8b04841
5. Pirok BWJ, Gargano AFG, Schoenmakers PJ (2018) Optimizing separations in online comprehensive two-dimensional liquid chromatography. J Sep Sci 41:68–98. https://doi.org/10.1002/jssc.201700863
6. Marriott PJ, Wu Z-Y, Schoenmakers P (2012) Nomenclature and conventions in

comprehensive multidimensional chromatography – an update. Chromatogr Online 25 (5):266–275

7. Cheng C, Liao CF (2018) Novel dual two-dimensional liquid chromatography online coupled to ultraviolet detector, fluorescence detector, ion-trap mass spectrometer for short peptide amino acid sequence determination with bottom-up strategy. J Chin Chem Soc 65:714–725. https://doi.org/10.1002/jccs.201700380
8. Pandohee J, Stevenson P, Zhou X-R et al (2015) Multi-dimensional liquid chromatography and metabolomics, are two dimensions better than one? Curr Metabolomics 3:10–20. https://doi.org/10.2174/2213235X03666150403231202
9. Kalili KM, De Villiers A (2013) Systematic optimisation and evaluation of on-line, off-line and stop-flow comprehensive hydrophilic interaction chromatography × reversed phase liquid chromatographic analysis of procyanidins, Part I: theoretical considerations. J Chromatogr A 1289:58–68. https://doi.org/10.1016/j.chroma.2013.03.008
10. Bedani F, Kok WT, Janssen HG (2006) A theoretical basis for parameter selection and instrument design in comprehensive size-exclusion chromatography × liquid chromatography. J Chromatogr A 1133:126–134. https://doi.org/10.1016/j.chroma.2006.08.048
11. Striegel AM (2001) Longitudinal diffusion in size-exclusion chromatography: a stop-flow size-exclusion chromatography study. J Chromatogr A 932:21–31. https://doi.org/10.1016/S0021-9673(01)01214-6
12. Donato P, Rigano F, Cacciola F et al (2016) Comprehensive two-dimensional liquid chromatography–tandem mass spectrometry for the simultaneous determination of wine polyphenols and target contaminants. J Chromatogr A 1458:54–62. https://doi.org/10.1016/j.chroma.2016.06.042
13. Vivo G, Van Der Wal S, Schoenmakers PJ (2010) Comprehensive study on the optimization of online two-dimensional liquid chromatographic systems considering losses in theoretical peak capacity in first- and second-dimensions: a pareto-optimality approach. Analysis Anal Chem 82(20):3090–3100. https://doi.org/10.1021/ac101420f
14. Pirok BWJ, Pous-Torres S, Ortiz-Bolsico C et al (2016) Program for the interpretive optimization of two-dimensional resolution. J Chromatogr A 1450:29–37. https://doi.org/10.1016/j.chroma.2016.04.061
15. Stoll DR, Shoykhet K, Petersson P, Buckenmaier S (2017) Active solvent modulation: a valve-based approach to improve separation compatibility in two-dimensional liquid chromatography. Anal Chem 89:9260–9267. https://doi.org/10.1021/acs.analchem.7b02046
16. Vonk RJ, Gargano AFG, Davydova E et al (2015) Comprehensive two-dimensional liquid chromatography with stationary-phase-assisted modulation coupled to high-resolution mass spectrometry applied to proteome analysis of saccharomyces cerevisiae. Anal Chem 87:5387–5394. https://doi.org/10.1021/acs.analchem.5b00708
17. Tian H, Xu J, Xu Y, Guan Y (2006) Multidimensional liquid chromatography system with an innovative solvent evaporation interface. J Chromatogr A 1137:42–48. https://doi.org/10.1016/j.chroma.2006.10.005
18. Navarro-Reig M, Jaumot J, Tauler R (2018) An untargeted lipidomic strategy combining comprehensive two-dimensional liquid chromatography and chemometric analysis. J Chromatogr A 1568:80–90. https://doi.org/10.1016/j.chroma.2018.07.017
19. Navarro-Reig M, Jaumot J, Baglai A et al (2017) Untargeted comprehensive two-dimensional liquid chromatography coupled with high-resolution mass spectrometry analysis of rice metabolome using multivariate curve resolution. Anal Chem 89:7675–7683. https://doi.org/10.1021/acs.analchem.7b01648
20. Donato P, Micalizzi G, Oteri M et al (2018) Comprehensive lipid profiling in the Mediterranean mussel (Mytilus galloprovincialis) using hyphenated and multidimensional chromatography techniques coupled to mass spectrometry detection. Anal Bioanal Chem 410:3297–3313. https://doi.org/10.1007/s00216-018-1045-3
21. Berkecz R, Tömösi F, Körmöczi T et al (2018) Comprehensive phospholipid and sphingomyelin profiling of different brain regions in mouse model of anxiety disorder using online two-dimensional (HILIC/RP)-LC/MS method. J Pharm Biomed Anal 149:308–317. https://doi.org/10.1016/j.jpba.2017.10.043
22. Sun C, Zhao YY, Curtis JM (2015) Characterization of phospholipids by two-dimensional liquid chromatography coupled to in-line ozonolysis-mass spectrometry. J Agric Food Chem 63:1442–1451. https://doi.org/10.1021/jf5049595
23. Holčapek M, Ovčačíková M, Lísa M et al (2015) Continuous comprehensive two-dimensional liquid chromatography-

electrospray ionization mass spectrometry of complex lipidomic samples. Anal Bioanal Chem 407:5033–5043. https://doi.org/10.1007/s00216-015-8528-2

24. Bang DY, Moon MH (2013) On-line two-dimensional capillary strong anion exchange/reversed phase liquid chromatography-tandem mass spectrometry for comprehensive lipid analysis. J Chromatogr A 1310:82–90. https://doi.org/10.1016/j.chroma.2013.08.069
25. Yang L, Lv P, Ai W et al (2017) Lipidomic analysis of plasma in patients with lacunar infarction using normal-phase/reversed-phase two-dimensional liquid chromatography–quadrupole time-of-flight mass spectrometry. Anal Bioanal Chem 409:3211–3222. https://doi.org/10.1007/s00216-017-0261-6
26. Baglai A, Gargano AFG, Jordens J et al (2017) Comprehensive lipidomic analysis of human plasma using multidimensional liquid- and gas-phase separations: two-dimensional liquid chromatography–mass spectrometry vs. liquid chromatography–trapped-ion-mobility–mass spectrometry. J Chromatogr A 1530:90–103. https://doi.org/10.1016/j.chroma.2017.11.014
27. Brouwers JF (2011) Liquid chromatographic-mass spectrometric analysis of phospholipids. Chromatography, ionization and quantification. Biochim Biophys Acta Mol Cell Biol Lipids 1811:763–775. https://doi.org/10.1016/j.bbalip.2011.08.001
28. D'Attoma A, Grivel C, Heinisch S (2012) On-line comprehensive two-dimensional separations of charged compounds using reversed-phase high performance liquid chromatography and hydrophilic interaction chromatography. Part I: orthogonality and practical peak capacity consideratio. J Chromatogr A 1262:148–159. https://doi.org/10.1016/j.chroma.2012.09.028
29. Yang Q, Shi X, Gu Q et al (2012) On-line two dimensional liquid chromatography/mass spectrometry for the analysis of triacylglycerides in peanut oil and mouse tissue. J Chromatogr B Anal Technol Biomed Life Sci 895–896:48–55. https://doi.org/10.1016/j.jchromb.2012.03.013
30. Yang L, Cui X, Zhang N et al (2015) Comprehensive lipid profiling of plasma in patients with benign breast tumor and breast cancer reveals novel biomarkers. Anal Bioanal Chem 407:5065–5077. https://doi.org/10.1007/s00216-015-8484-x
31. Sun C, Zhao YY, Curtis JM (2014) Elucidation of phosphatidylcholine isomers using two dimensional liquid chromatography coupled in-line with ozonolysis mass spectrometry. J Chromatogr A 1351:37–45. https://doi.org/10.1016/j.chroma.2014.04.069
32. Li M, Tong X, Lv P et al (2014) A not-stop-flow online normal-/reversed-phase two-dimensional liquid chromatography-quadrupole time-of-flight mass spectrometry method for comprehensive lipid profiling of human plasma from atherosclerosis patients. J Chromatogr A 1372:110–119. https://doi.org/10.1016/j.chroma.2014.10.094
33. Li M, Feng B, Liang Y et al (2013) Lipid profiling of human plasma from peritoneal dialysis patients using an improved 2D (NP/RP) LC-QToF MS method. Anal Bioanal Chem 405:6629–6638. https://doi.org/10.1007/s00216-013-7109-5
34. Krastanov A (2010) Metabolomics - the state of art. Biotechnol Biotechnol Equip 24:1537–1543. https://doi.org/10.2478/V10133-010-0001-y
35. Woiwode U, Reischl RJ, Buckenmaier S et al (2018) Imaging peptide and protein chirality via amino acid analysis by chiral × chiral two-dimensional correlation liquid chromatography. Anal Chem 90:7963–7971. https://doi.org/10.1021/acs.analchem.8b00676
36. Toro-Uribe S, Montero L, López-Giraldo L et al (2018) Characterization of secondary metabolites from green cocoa beans using focusing-modulated comprehensive two-dimensional liquid chromatography coupled to tandem mass spectrometry. Anal Chim Acta 1036:204–213. https://doi.org/10.1016/j.aca.2018.06.068
37. Muller M, Tredoux AGJ, de Villiers A (2018) Predictive kinetic optimisation of hydrophilic interaction chromatography × reversed phase liquid chromatography separations: Experimental verification and application to phenolic analysis. J Chromatogr A 1571:107–120. https://doi.org/10.1016/j.chroma.2018.08.004
38. Sommella E, Ismail OH, Pagano F et al (2017) Development of an improved online comprehensive hydrophilic interaction chromatography × reversed-phase ultra-high-pressure liquid chromatography platform for complex multiclass polyphenolic sample analysis. J Sep Sci 40:2188–2197. https://doi.org/10.1002/jssc.201700134
39. Montero L, Ibáñez E, Russo M et al (2016) Metabolite profiling of licorice (Glycyrrhiza glabra) from different locations using comprehensive two-dimensional liquid chromatography coupled to diode array and tandem mass spectrometry detection. Anal Chim Acta

913:145–159. https://doi.org/10.1016/j.aca.2016.01.040

40. Dugo P, Herrero M, Kumm T et al (2008) Comprehensive normal-phase × reversed-phase liquid chromatography coupled to photodiode array and mass spectrometry detection for the analysis of free carotenoids and carotenoid esters from mandarin. J Chromatogr A 1189:196–206. https://doi.org/10.1016/j.chroma.2007.11.116
41. Baglai A, Blokland MH, Mol HGJ et al (2018) Enhancing detectability of anabolic-steroid residues in bovine urine by actively modulated online comprehensive two-dimensional liquid chromatography – high-resolution mass spectrometry. Anal Chim Acta 1013:87–97. https://doi.org/10.1016/j.aca.2017.12.043
42. Wang Y, Lu X, Xu G (2008) Development of a comprehensive two-dimensional hydrophilic interaction chromatography/quadrupole time-of-flight mass spectrometry system and its application in separation and identification of saponins from Quillaja saponaria. J Chromatogr A 1181:51–59. https://doi.org/10.1016/j.chroma.2007.12.034
43. Blokland MH, Zoontjes PW, Van Ginkel LA et al (2018) Multiclass screening in urine by comprehensive two-dimensional liquid chromatography time of flight mass spectrometry for residues of sulphonamides, beta-agonists and steroids. Food Addit Contam Part A Chem Anal Control Expo Risk Assess 35:1703–1715. https://doi.org/10.1080/19440049.2018.1506160
44. Wong YF, Cacciola F, Fermas S et al (2018) Untargeted profiling of Glycyrrhiza glabra extract with comprehensive two-dimensional liquid chromatography-mass spectrometry using multi-segmented shift gradients in the second dimension: Expanding the metabolic coverage. Electrophoresis 39:1993–2000. https://doi.org/10.1002/elps.201700469
45. Yan X, Wang LJ, Wu Z et al (2016) New on-line separation workflow of microbial metabolites via hyphenation of analytical and preparative comprehensive two-dimensional liquid chromatography. J Chromatogr B Anal Technol Biomed Life Sci 1033–1034:1–8. https://doi.org/10.1016/j.jchromb.2016.07.053
46. Hájek T, Jandera P, Staňková M, Česla P (2016) Automated dual two-dimensional liquid chromatography approach for fast acquisition of three-dimensional data using combinations of zwitterionic polymethacrylate and silica-based monolithic columns. J Chromatogr A 1446:91–102. https://doi.org/10.1016/j.chroma.2016.04.007
47. Corgier A, Sarrut M, Crétier G, Heinisch S (2016) Potential of online comprehensive two-dimensional liquid chromatography for micro-preparative separations of simple samples. Chromatographia 79:255–260. https://doi.org/10.1007/s10337-015-3012-x
48. Hájek T, Jandera P, Staňková M, Česla P (2016) Automated dual two-dimensional liquid chromatography approach for fast acquisition of three-dimensional data using combinations of zwitterionic polymethacrylate and silica-based monolithic columns. J Chromatogr A 1446:91–102. https://doi.org/10.1016/j.chroma.2016.04.007

2.2.5 Recent advances in 2DLC applied to metabolomics

This section aims to update current knowledge about online comprehensive two-dimensional liquid chromatography (LC × LC) after **scientific publication I,** specifically those published in the last years (from 2019 until present).

The use of multidimensional liquid chromatography in metabolomics and lipidomics is increasing vertiginously in recent years, as shown in the reviews from 2019 by Lv et al. [170] and Brandao et al. [172]. Nevertheless, it is worth noticing that many 2021 publications in the metabolomic field are still focused on GC × GC [173–177]. Some efforts have also been directed to the combination of different multidimensional chromatographic separations (e.g., liquid and gas chromatography) [178,179], and the use of multiple detectors (e.g., mass spectrometry and nuclear magnetic resonance) [179,180]. The reason why 2DLC seems to be less implemented still in laboratories compared to GC × GC may be due to the complexity that still represents in terms of its set-up (e.g., the mismatch between two dimensions; need for active modulation interface) and to data processing. For instance, peak alignment is not often required in GC × GC, whereas it can be critical in 2DLC, as it will be explained in **Section 2.3 Data analysis strategies in metabolomics**. However, 2DLC applications are enormous and will continue increasing in the next decade, incorporating the new technological advances.

One major 2DLC aspect subject to new improvements is the modulation interface. The term **modulation** is referred to each of the fractions in which the first dimension (^{1}D) effluent is divided after going through the ^{1}D column. These fractions pass through the modulator and are stored in loops, until further separated in the second dimension (^{2}D) column. The modulator is the interface between the two dimensions, and it is usually composed of a high-pressure valve (6, 8 or 10 ports) with a minimum of two positions. When fractions are collected and transferred from the ^{1}D to the ^{2}D without modifications, it is called **passive modulation**. On the contrary, if the ^{1}D effluent is diluted before reaching the ^{2}D column, loop traps or an evaporative system are employed in between dimensions, then the term **active modulation** is preferred.

Apart from the generic review from 2019 by Pirok et al. [181] mentioned in **scientific publication I**, a review by Chen et al. specifically focused on modulation in 2DLC has been recently published [182]. This review details all possible modulations, with single or multiple valves, and with and without assistant technology. A summary of the challenges in modulation for 2DLC applications, classifying retention mechanisms according to their compatibility, can also be found.

More information on separation combinations for 2DLC can be found in a previous review from 2017 by Pirok et al. [183].

A novel modulation system has also been recently proposed by Chen et al. [184]. It is known as at-column dilution (ACD) and the main difference with previous strategies is an adjustable and optimized dilution factor which does not require splitting the ^{1}D effluent. Thus, solvent compatibility between dimensions, orthogonality and sensitivity are improved. This modulator was tested for the analysis of butterfly bush, using RP × HILIC [185].

Among recent applications of 2DLC in metabolomics, there is an RP × RP approach for the analysis of modified nucleosides in different biological matrices [186]. Another RP × RP configuration coupled to a photodiode array (PDA) and a mass spectrometer was employed for the analysis of metabolites in brown mustard [187]. ^{2}D presented a segmented-in-fraction gradient composed of three different full-in-fraction steps. The advantage of this system is that it adequates ^{2}D mobile phase composition to metabolites retention along with the separation, i.e., lower slope in co-eluting areas and steeper gradient for more retained compounds. A third RP × RP method was set up employing a pentafluorophenyl (PFP) column coupled to a C18 for the analysis of cannabinoids and phenolic compounds [188]. The ^{2}D separation employed a shifted gradient, and it was coupled to a double detection system, first diode array detection (DAD), and second, an HRMS detector.

New applications of 2DLC, specifically in lipidomics, include the set-up of an RP × HILIC method for analyzing the lipidome of zebrafish embryos [189]. In this publication, a comparison between 1D and different 2DLC approaches (C18 × HILIC, HILIC × C18, HILIC × PFP) is performed. The authors concluded that the most suitable method is the C18 × HILIC combination, due to the higher lipid separation. The ^{1}D separation can separate by the hydrophobic part of the molecules, differentiating the lipids by the length of the chains and the number of the double bonds, whereas the ^{2}D separation provides a quick screening between the lipid classes. Similar results were obtained in a previous comparison of Holčapek et al. [190], and a similar set-up was proposed in **scientific publication V.** In addition, a Chiral × RP set-up was also developed for the analysis of conjugated polyunsaturated fatty acid isomers and structurally related compounds [191]. The ^{1}D column separated lipid isomers, whereas the ^{2}D column separated according to the number of double bonds and degree of oxidation. A combined DAD-HRMS detection with the possibility of obtaining additional MS/MS information, allowed a thorough characterization of the compounds and the structural annotation of unknowns.

Parallel analysis of metabolome and lipidome is also frequent. It is the case of the study of potential biomarkers of esophageal squamous cell carcinoma [192]. A column C8 was employed in the ^{1}D to pre-separate metabolome from lipidome. Fractions were transferred online, and ^{2}D were analyzed in parallel (C18 for the metabolome and C8 again for the lipidome). Another parallel column-based 2DLC pseudo-targeted method from the same research group was also developed for the analysis of a mixture of different biofluids and tissues [193]. The combinations were the same as the previously mentioned publication: C8 × C18 for metabolomics, and C8 × C8 for lipidomics. Parallel 2DLC was also used in untargeted metabolomics of rat livers [194]. A HILIC × RP dual system was employed, with two RP columns, for measuring positive and negative ionization modes in parallel.

The combination of GC × GC and LC × LC also seems promising, as shown in the untargeted study of polar metabolites involved in colorectal cancer [168]. Again, a dual column system that allowed parallel analysis of positive and negative ionization modes was employed in the 2DLC-MS set-up. RP and HILIC columns were selected as retention mechanisms.

2.2.6 Practical considerations about LC × LC

2DLC is an analytical technique that holds a great potential for growth in many different applications. From the different existing set-ups, online and comprehensive two-dimensional liquid chromatography (LC × LC) presents some important advantages, as shown in **Figure 2.9**. First, higher peak capacity, and peak-production rates compared to, for instance, heart-cutting approaches. This means that separation power is considerably higher. Second, the selectivity is increased due to the combination of two complementary retention mechanisms (i.e., **orthogonal**). Thus, analytes can be grouped by classes thanks to one of the separations, and further differentiated within the class thanks to the other. Third, most of the LC × LC combinations are MS compatible. In addition, the use of MS/MS can improve analyte identification, especially relevant in untargeted approaches, as previously mentioned. LC × LC is also useful in targeted separations of structurally similar compounds.

Unfortunately, the use of LC × LC is not exempt from some big challenges (see **Figure 2.9**). The intrinsic and primary difficulty is related to the conceptual and instrumental complexity. Manual method optimization is time-consuming, rather expensive and requires expert personnel. In the review by Bos et. al [195], different *in silico* strategies to reduce costs and simplify process optimization are summarized.

A recent more recent publication, not included in the review, about computer-assisted modelling from Makey et al. [196] is also remarkable.

One of the main goals of this PhD Thesis is to propose solutions and improvements for LC × LC metabolomics analysis. The three principal faced challenges are: 1) to increase solvent compatibility between mobile phases from ^{1}D and ^{2}D, 2) to enhance the analytical sensitivity, and 3) to develop data analysis and visualization strategies to cope with the huge and complex datasets produced by this technique. All in all, the ultimate goal of this PhD Thesis is to facilitate and encourage the implementation of LC × LC in routine work in the analytical laboratories.

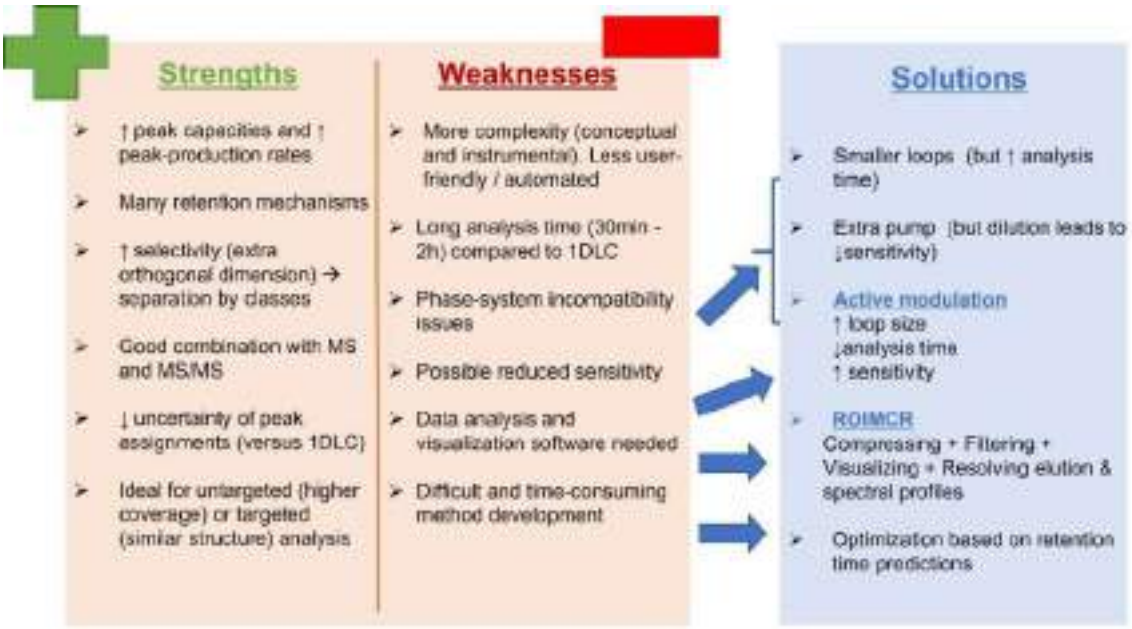


Figure 2.9. Compilation of current strengths and weaknesses of LC × LC, and proposed solutions. The main developments of this PhD Thesis are marked in blue. *Adapted from* [181].

How to increase solvent compatibility while enhancing sensitivity

In this PhD Thesis, online and comprehensive two-dimensional liquid chromatography methods have been developed for different applications, as will be further discussed in the following Chapters. For untargeted lipidomics (**scientific publication V**) and targeted analysis of pharmaceutical compounds (**scientific publication III**), RP × HILIC has been selected. For targeted analysis of amino acids (**scientific publication VI**), the inverse combination, HILIC × RP, has been preferred. However, in both set-ups, there is a mismatch challenge when joining ^{1}D and ^{2}D effluents, as exemplified in **Figure 2.10.**

When two different peaks are observed for the same compound, one in the dead volume and one at its usual retention time, this phenomenon is known as **breakthrough** or solvent-plug peak [197]. The cause is the presence of too much strong solvent at the beginning, which hinders proper retention of the analyte, and only a fraction of it will be able to stick to the stationary phase. The rest will be eluted in the solvent front, unretained. This excess of strong solvent can come, for example, from the initial mobile phase composition. Another source can be the sample. For instance, if a too big sample volume is employed or the sample is too concentrated. The immediate consequence is that quantification will be compromised, because the retained peak area will not represent the analyte content in the sample.

Breakthrough is a critical issue to consider when dealing with LC × LC [181,183,198,199]. **Figure 2.10.A** shows the possible case of breakthrough RP × HILIC. Firstly, it is necessary to consider that although RP and HILIC are usually almost orthogonal mechanisms (not directly correlated), the strongest solvent is the opposite. In RP, gradients increase organic percentage along separation, whereas, in HILIC, the aqueous phase content is augmented accordingly. This means that at the beginning of the RP × HILIC separation (if no modification is performed between the separations, e.g., no active modulation strategies are used), a non-negligible amount of water (and other polar solvents, e.g., isopropanol in the case of lipid separations) will access the ^{2}D column. The result may lead to polar compounds not completely retained, and a fraction of them eluting with the front. Contrarily, in the case of HILIC × RP, as summarized in **Figure 2.10.B**, the initial organic solvent coming from the ^{1}D mobile phase composition poses a threat to the ^{2}D separation. This is because hydrophobic compounds may not be well attached to the stationary phase and important losses of these analytes may occur due to peak splitting.

Besides breakthrough, other phenomena can lead to a poor chromatographic separation due to incompatibilities between the two mobile phases. For instance, peaks can be severely distorted, plus both resolution and sensitivity can be put at risk in the ^{2}D separation.

The solvent mismatch between ^{1}D and ^{2}D can be approached from different perspectives, as stated in the work by Chapel et al. [200]. Here only a few solutions will be detailed, mainly related to this PhD Thesis work.

A common attempt is to add an extra third pump to dilute the ^{1}D before reaching the ^{2}D column. It is the solution employed in the work by Navarro-Reig et al. [201] for the untargeted analysis of rice metabolome employing a HILIC × RP method. A 10-port two-position valve was the interface of both dimensions. The breakthrough was avoided by adding water at a constant flow rate to the sample solvent. A stainless-steel piece with a T form joined the water flow and the ^{2}D pump flow before the ^{2}D column.

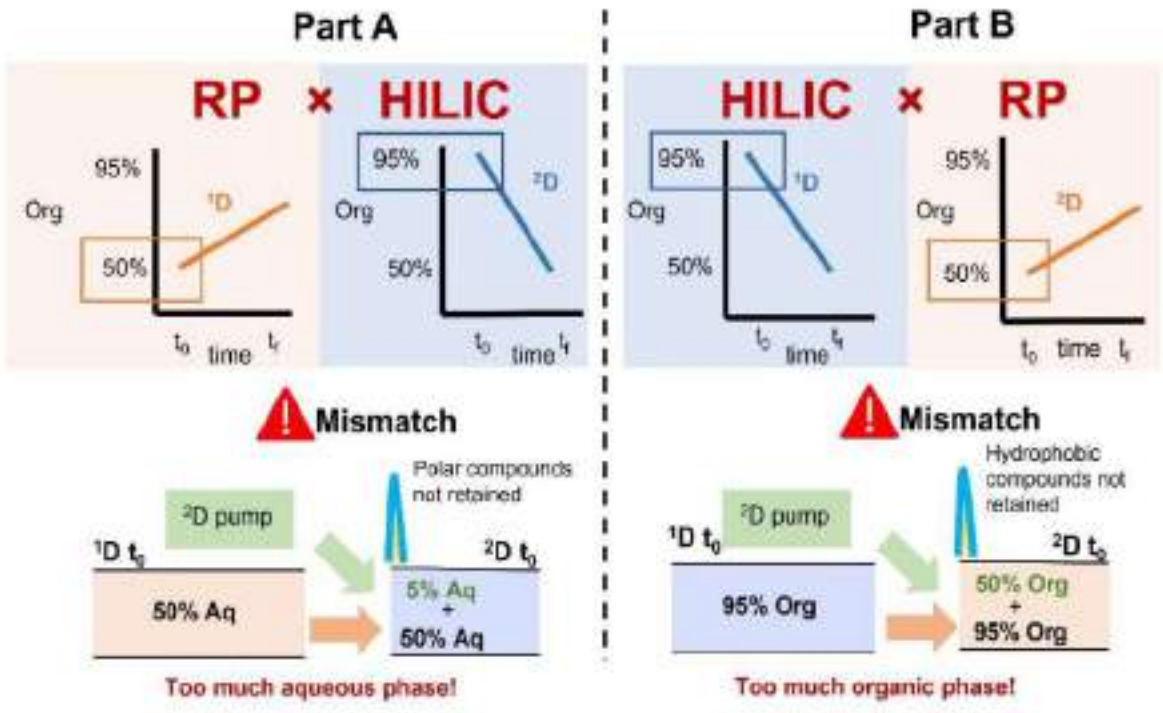


Figure 2.10. Exemplification of potential breakthrough for the combinations A) RP × HILIC and B) HILIC × RP.

Another alternative to deal with solvent strength mismatch is related to **sample loops.** Sample fractions from the ^{1}D are normally collected and stored in two sample loops until further analysis in the ^{2}D column. Interface valves used to join both columns usually present two positions. In each turn, one of the loops is fulfilled with a fraction from the ^{1}D column, containing the ^{1}D mobile phase plus the correspondent sample fraction. Meanwhile, the content of the other sample loop is discharged into the ^{2}D column, carried by the ^{2}D mobile phase from the ^{2}D pump. Then, valve switches and positions and roles are inversed. A clearer visualization of both loop functions is depicted in **Figure 2.11**.

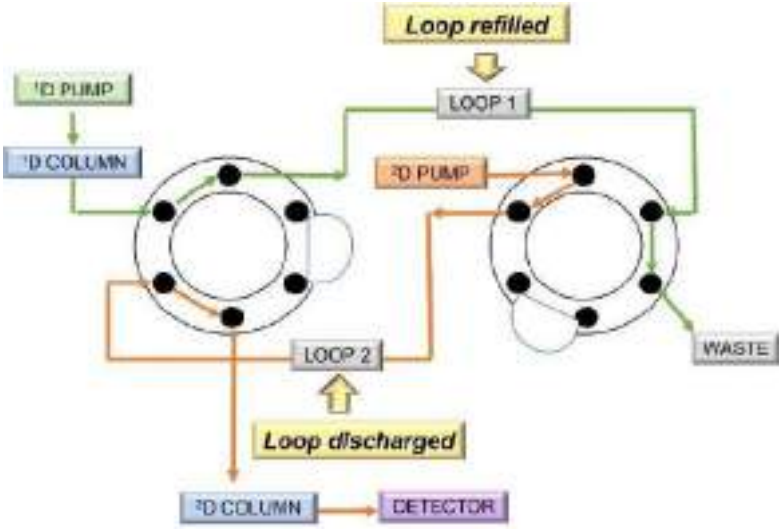


Figure 2.11. Scheme of a LC × LC separation with two 6-port valves with 2 positions each, acting as interface. *Adapted from* **scientific publication I.**

Loop sizes can vary depending on the method, normally ranging from 20 to 100 μL. One solution to increase solvent compatibility will be selecting smaller loops. Therefore, a lower amount of the strongest solvent will reach the ^{2}D column, and separation will be less jeopardized. This strategy was employed in **scientific publication VI** of this PhD Thesis. However, the main disadvantage of this solution is that the total analysis time will considerably increase. The reason is the direct proportion between the size loop and the ^{1}D flow, according to **Equation 1.** Consequently, loop size will condition the whole duration of the ^{1}D separation, i.e., lower flows mean a slower gradient and, therefore, a longer ^{1}D run.

Equation 1

$$^{1}D\ Flow = \frac{Loop\ size}{^{2}D\ modulation\ time}$$

It is important to notice that if loop size dimensions are in μL, resulting flow will be expressed as μL·min^{-1}, instead of mL·min^{-1} as usual.

The last approach considered in this PhD Thesis to solve solvent incompatibilities is the use of **active modulation**. Current trends involve these strategies since many new modulation systems have been implemented in the last decade [182]. In this PhD Thesis, Active Solvent Modulation (ASM) was employed in **scientific publication V**. ASM is a valve-based approach recently developed by Stoll et al. [202] that uses an 8-port interface with a 4-position design, modified with a bypass capillary. When the bypass path is isolated (**Positions A and C of Figure 2.12**), the valve acts as a normal 8- or 10-port valve with 2 positions. This means that one loop is refilled with ^{1}D effluent, while the other loop is discharged into the ^{2}D column with

^{2}D mobile phase. However, when the bypass is on (**Positions B and D of Figure 2.12**), the ^{1}D effluent from the loop is displaced and diluted with ^{2}D initial mobile phase composition.

This dilution step (also called the ASM step) depends on the flow rate and loop size and takes place at the beginning of each modulation. The dilution is performed according to split ratios (i.e., ¼ means 1 part through the loop and 4 parts through bypass).

In recent years, the number of ASM applications has increased. For instance, monoclonal antibodies were separated in a HILIC × RP-HRMS system [203]. This study contributed to enlightening the benefits of ASM, e.g., the is avoided and sensitivity is enhanced. An RP × RP system for separating complex peptide mixtures (e.g., characterization of therapeutic antibodies) was also set [203]. Peak capacities were considerably increased thanks to this modulation strategy. In addition, ASM allowed the use of longer and more efficient ^{1}D columns and longer injection volumes in the ^{2}D column, reducing the total analysis time.

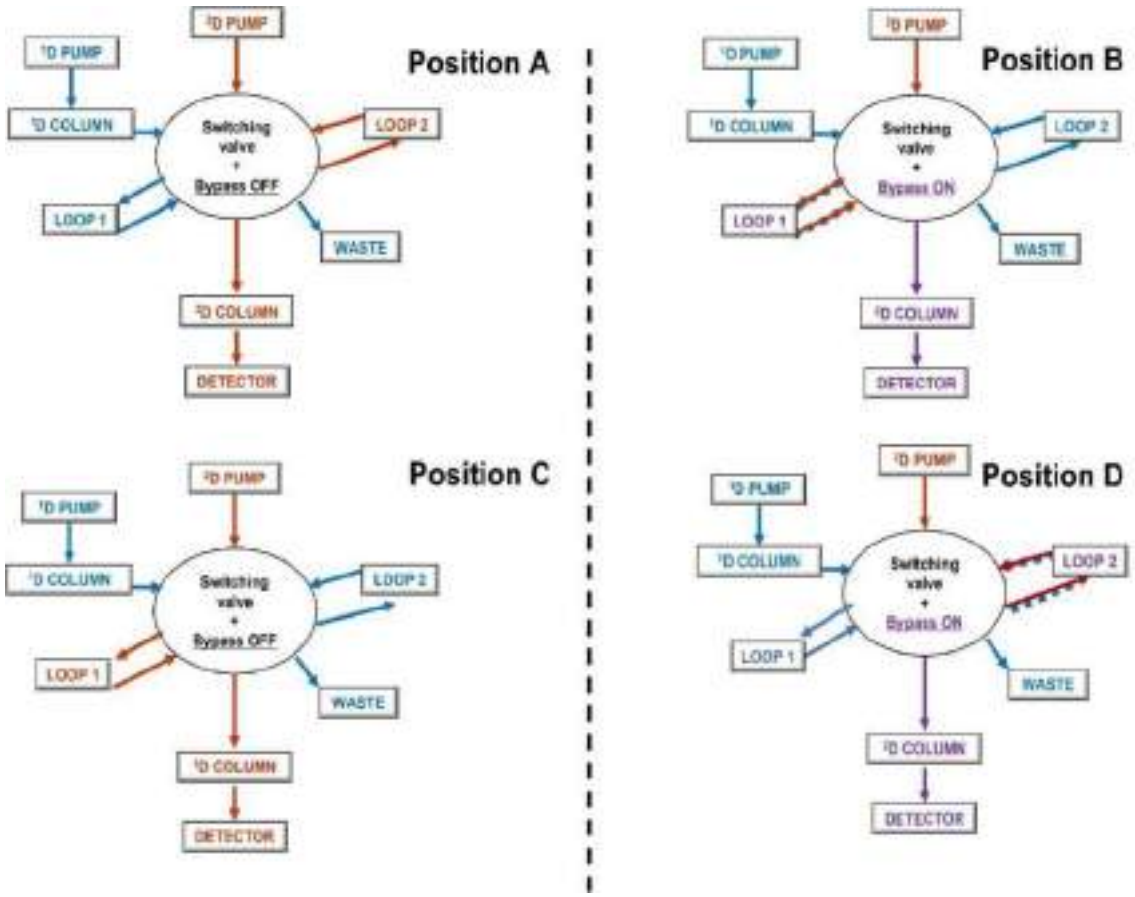


Figure 2.12. Scheme of Active Solvent Modulation (ASM). **Positions A and C** represent the ten-port valve when the bypass is off. **Positions D and D** show how dilution is performed when bypass is on.

The advantages of ASM also in the targeted analysis were proven with a size exclusion chromatography (SEC) combined with RP method applied to assess polymer blends and determine impurities in polymeric matrices [204]. Both heart-cutting and comprehensive modes were tested. Complementary, a pseudo-comprehensive SEC/RP method was also employed in polymer characterization [205]. The main achievement is that this type of separation will be no longer limited to water-soluble polymers, but a wider range of them because the solvent mismatch issue was resolved.

Other 2DLC set-ups (e.g., selective comprehensive or multiple heart-cutting) have also benefited from ASM. A selective comprehensive HILIC × RPLC method was developed for the simultaneous analysis of water- and fat-soluble vitamins [206]. In this case, ASM helped to prevent peak distortion and broadening. Hence, an optimal resolution was achieved. On the other side, a multiple heart-cutting method using a mixed-mode reversed phase/weak anion-exchange (RP/WAX) as ^{1}D and pure RP in the ^{2}D was developed for the analysis of synthetic oligonucleotides [207]. The main reason for employing ASM was to reduce the incompatibility with the mass spectrometer caused by the high non-volatile buffer components and ion-pair agent's content from the ^{1}D.

The last applications of ASM include its use in the simulation of elution profiles [208]. As previously stated, *in silico* optimization is a key issue in LC × LC to save both time and money. Moreover, due to the increasing parameters needed to be considered related to the ASM step (e.g., dilution factor, loop size conditions), this modulation strategy complicates method optimization. Thus, great efforts have been performed to predict elution profiles in LC × LC employing ASM, with the aim of increasing simulation capabilities for *in silico* optimization.

How to visualize two-dimensional liquid chromatograms

LC × LC acquisition is composed of multiple ^{2}D chromatograms (i.e., modulations) in which ^{1}D separation is split at equal time intervals. Usually, shorter sampling times lead to a better resolution, although the whole run becomes longer because the ^{1}D separation needs to be cut into smaller fractions. Hence, a single peak can be fractured into different subsequent modulations for further analysis in the ^{2}D column, as shown in **Figure 2.13.** Visualizing the LC × LC data requires that the single ^{1}D chromatogram is transformed into several stack ^{2}D chromatograms with multiple peaks [209]. Then, the data could be rearranged into a contour plot (see **Figure 2.13**). Each area where there is a compound is expressed as a spot, and the color scale

represents the peak intensity. A 10-20 contour amplitude value is commonly selected from minimum to maximum [210]. 3D plots can also be used, although these representations are often harder to interpret.

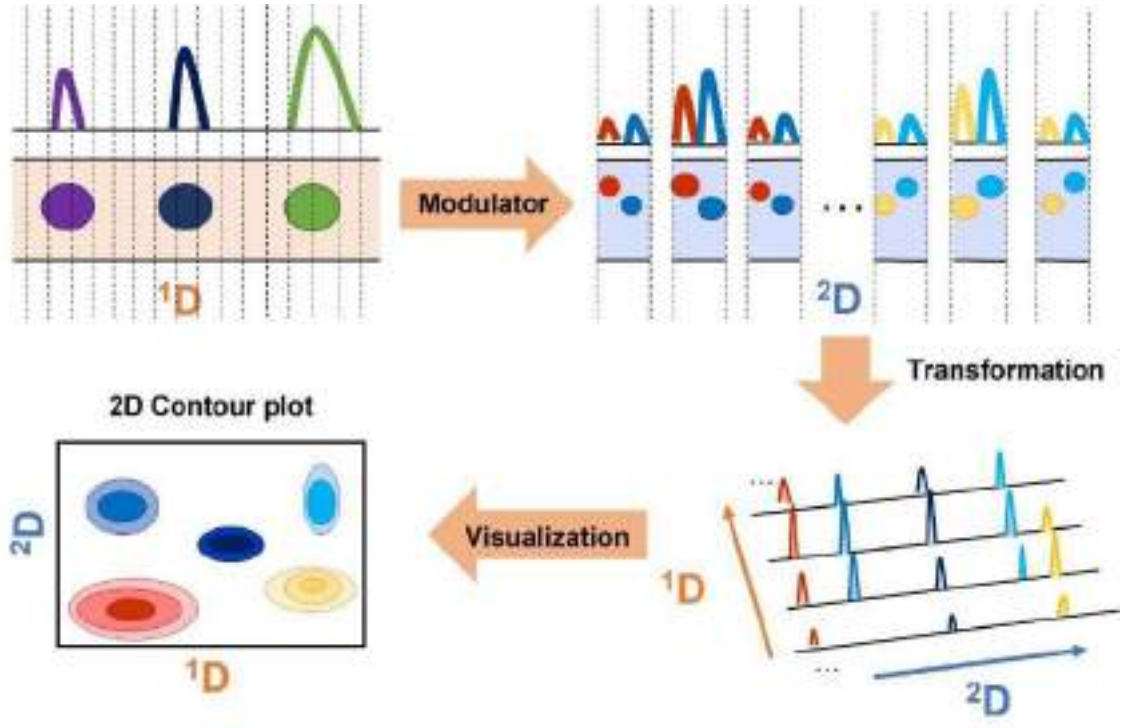


Figure 2.13. Transformation of ^{2}D chromatograms to provide a LC × LC visualization based on contour plots. *Adapted from* [209].

2.2.7 Mass spectrometry and metabolomics

Mass spectrometry (MS) allows for identifying and quantifying metabolites through their mass-to-charge ratios (*m/z*). Besides, high-resolution mass spectrometry can provide information about the metabolites chemical structure (e.g., accurate mass, isotope distribution patterns, and fragmentation patterns for structural elucidation). Thus, MS capabilities in biomarker discovery and in unraveling metabolic pathways are vast and unceasingly growing nowadays [211,212].

A mass spectrometer is composed of three main sections: ion source, analyzer, and ion detector. The first ionizes the molecules, the second separates the ions according to their *m/z* and may also fragment them, and the third counts the ions from every *m/z*. Once the sample accesses the ion source, the whole process undergoes in extreme vacuum conditions.

The choice of the ion source is directly dependent on the targeted molecules. For instance, volatile compounds analyzed after a gas chromatographic separation will employ more energetic ion sources (e.g., electron or chemical ionizations) to fragment molecules into smaller pieces. In the case of mass spectrometry imaging, a source able to desorb analytes from a solid matrix is preferred (e.g., Matrix-Assisted Laser Desorption Ionization (MALDI)). However, in LC-MS coupling, softer ionization sources are employed, commonly preserving the molecular ion, M^+ (i.e., the molecule with an electron less and though positively charged). Frequent ion sources include electrospray (ESI), atmospheric pressure ionization (API), Atmospheric Pressure Chemical Ionization (APCI), or atmospheric pressure photoionization ionization (APPI).

The selection of the mass analyzer is more related to the type of analytical approach. Hence, full scan mode is applied for untargeted studies, which means that all molecules are separated according to their *m/z*. On the contrary, if a targeted analysis is pursued, then the mass analyzer can act as a filter and deflect only specific ions to the detector. Examples of analyzers are electric sector (E), magnetic sector (B), quadrupole (Q), ion trap (IT), time of flight (TOF), Fourier transform ion cyclotron resonance (FTICR), and Fourier transform Orbitrap (FT-OT). Current instruments are usually hybrids, including more than a single analyzer (e.g., Quadrupole-time of flight (QTOF) or Quadrupole-Orbitrap, known as QExactive). In tandem mass spectrometry (MS/MS) mode, ions can be separated and then fragmented in a second mass analyzer, which can be the same type of analyzer (e.g., triple quadrupole (QqQ)) or a different type (e.g., hybrid mass analyzers). If ions are subsequently separated and fragment in an iterative process, it is called MS^n.

Once the ions hit the detector, a cascade of electrons is produced, and then amplified to increase sensitivity. Recurrent detectors are electron multipliers or microchannel plates, but other examples are the faraday cup or the photomultiplier.

Integrated software on the computers analyzes the data coming from the detector (e.g., relations between *m/z* values and their relative abundances). In combination with the information provided by the previous separation (e.g., the retention time for each *m/z* value), it is possible to compare with existing databases and identify the compounds detected.

In this PhD Thesis, all works presented employed ESI as the ion source, although different analyzers were selected according to each application. A brief explanation and comparison of the mass spectrometer components used during this PhD Thesis is now presented.

Ionization sources: Electrospray

In electrospray ionization (ESI), a strong electric field is applied under atmospheric pressure to the tip of a capillary tube from which a combination of sample and mobile phase is passing through at a slow flow. First, charged droplets are formed. Then the rest of solvent is removed when the droplets traverse a heated inert gas curtain (e.g., nitrogen) [213]. ESI allows positive and negative ionization modes, where adducts are formed accordingly (e.g., molecular ion plus a proton $[M+H]^+$, molecular ion plus ammonium $[M+NH_4]^+$ or molecular ion minus a proton $[M-H]^-$, etc). This ionization source is highly sensitive, and its coupling to LC is straightforward. One of the main advantages is that thermally labile compounds can be ionized with ESI. In addition, the mass range of the analytes can range from small metabolites to proteins or other multiple charged molecules (e.g., polymers). Some limitations are potential ion suppression if too complex matrix or concentrated samples are analyzed, and its high sensitivity to salts and detergents. In conclusion, ESI is the most widely employed ionization source in both metabolomics and lipidomics [151,214,215].

Mass analyzers

The used mass analyzers of this PhD Thesis are time of flight (TOF), triple quadrupole (QqQ), quadrupole – time of flight (QTOF) and Q – Orbitrap, commonly known by the commercial brand QExactive. **Table 2.4** summarizes the main characteristics, advantages, and limitations of each of them.

Time of flight is the fastest mass analyzer and presents the highest practical mass range [213]. The TOF principle of separation is velocity or flight time. Ions are separated according to their *m/z* ratio based on their flight time, which is the time they take to travel to a tube of known length until they reach the detector. Smaller molecules (lower *m/z* values) will travel fast and arrive firstly at the detector, whereas bigger molecules (higher *m/z* values) will be slower and arrive later. The separation of *m/z* values occurs in the space, which means that a very wide range of *m/z* values can be acquired in parallel and that is the reason why scan speed is very fast.

The main limitation of TOF alone is the inability to fragment unless it is coupled to a second TOF. The first TOF will provide the ions with the energy they require to fragment, and the second TOF will separate the fragments according to their *m/z* values. If MS/MS is not acquired, then false positives in compound identification can be obtained due to isobaric interferences [213]. TOF has been employed in this PhD

Thesis for the untargeted analysis of rice lipidome after arsenic exposure (**scientific publication VII**).

The **triple quadrupole** is the most used mass analyzer in targeted analysis because it presents a wide dynamic range, high sensitivity, and good scan speed. Besides, sample preparation is often minimal as selected or multiple reaction monitoring are employed, which increases selectivity. QqQ is a good choice for quantitation due to its high throughput, but it cannot be employed in qualitative analysis or compound discovery due to its low resolution.

This analyzer is composed of a quadrupole, followed by a collision cell, and lastly, another quadrupole [213]. This configuration allows MS/MS. In the first quadrupole, there is a first *m/z* selection of specific *m/z* values of interest. Thus, the principle of separation is the *m/z*, based on its trajectory stability. The collision cell is where the reaction takes place and, therefore, fragments are formed by colliding with an inert gas (e.g., Ar, He, or N_2). In the third quadrupole, another selection is performed, but this time only of the targeted fragments. Several precursor/product ion pairs can be measured simultaneously.

In this PhD Thesis, QqQ has been employed in the targeted analysis of sphingolipids in an LC-MS platform (**scientific publication VIII**), in the analysis of pharmaceutical compounds in an LC × LC-DAD-MS coupling, where DAD is a diode array detector (**scientific publication III**), and also in the analysis of pharmaceutical compounds and amino acids in LC × LC-MS (**scientific publications III and VII**).

Apart from TOF-TOF, the most common set-up for TOF in MS/MS is a **quadrupole – time of flight,** in the so-called QTOF. This hybrid instrument combines the high efficiency of the quadrupole in compound fragmentation, i.e., two quadrupoles, the first for separating *m/z* and the second as collision cell, plus a time-of-flight analyzer for separating the fragments afterwards. The advantage of the TOF compared to a third quadrupole is the high scan speed and high mass resolution. Contrarily to a single TOF, false positives in compound identification are avoided thanks to fragmentation. QTOF presents good sensitivity and high mass accuracy in both precursor and product ions, which allows compound discovery, structure elucidation and, therefore, simultaneous qualitative and quantitative analysis. The main limitations are the lower dynamic range compared to QqQ and the impossibility of performing MS^n. Another important drawback is that switching polarity is not straightforward, and the whole sequence needs to be acquired twice, in positive and negative modes separately.

Table 2.4. Main characteristics of the mass analyzers employed in this PhD Thesis.

Name	TOF	QqQ	QTOF	QExactive
Analyzers	Time of flight	Quadrupole - Colission cell - Quadrupole	Quadrupole - Time of flight	Quadrupole - Orbitrap
Tandem MS	--------	In space	In space	In time
Mass acuracy	< 20 ppm	0.05 Da	< 5 ppm	< 5 ppm
Cost	Medium	Cheapest	High	Most expensive
Advantages	-High resolution mass spectrometry at a reduced price	-Minimal sample preparation -High throughput for quantitation -High sensitivity -Wide dynamic range -Good scan speed	-Qualitative and quantitative analysis -Structural elucidation -Wider mass range than QExactive -Accurate mass measurements both precursors and product ions	-Qualtitative and quantitative analysis -Structural elucidation -Higher resolution than QTOF -Possible to perform multiple fragmentation steps, MS^n -Fast polarity switching
Limitations	-MS^2 required to obtain complementary information and avoid false positives and isobaric interferences	-Not suitable for qualitative analysis or compound discovery -Low resolution	-Lower sensitivity than QExactive -Lower dynamic range than QqQ	-Higher costs and lower analysis speed than QTOF (currently being improved) -Lower dynamic range than QqQ -Only accurate mass measurements for precursor ions

The applications of QTOF in metabolomics, but also in toxicology or environmental screening of emerging contaminants, have continuously increased in recent years [216–220]. Here, QTOF has been chosen for the untargeted lipidomic study of zebrafish embryos exposed to endocrine disruptor chemicals (**scientific publication V**), and for the untargeted metabolomic analysis of cells exposed to pharmaceutical compounds (**scientific publication VIII**).

The last analyzer used in this PhD Thesis is **Quadrupole** – **Orbitrap,** commonly known as QExactive. The principle of separation in this instrument is the resonance frequency of the *m/z*. Ions are trapped and stored in a potential dwell and turn around the central electrode. The *m/z* value is related to the ion oscillation frequency, which is measured. Fourier transforms are required to measure the frequency of the time-domain signal [213].

This hybrid instrument presents the highest sensitivity and resolution among the previously explained ones. Both qualitative and quantitative analyses can be performed. MS^n is allowed because *m/z* separation is performed in time (not in space) which allows subsequent fragmentations of the ions. Therefore, the outputs are both accurate mass and fragmentation patterns. Thus, structure elucidation is also possible. Besides, both polarity modes (i.e., positive and negative) can be measured simultaneously, because polarity switch is easy. QExactive used to be considerably more expensive than a QTOF, but recently, prices are becoming more competitive.

The untargeted metabolomic analysis of rice metabolome after arsenic exposure was performed on a QExactive (**scientific publication VII**).

Both QTOF and Orbitrap are widely employed, especially in untargeted analysis. Nevertheless, an important concern about the results provided with both instruments has arisen, as a recent interlaboratory experiment has proven [222]. Due to the adducts, fragments, charge states, and clusters generated by different mass spectrometers, the detected features are not the same. This issue can lead to errors in the annotation, because not all compounds found in spectral libraries are often acquired with both instruments. Hence, MS/MS fragmentation patterns at certain conditions may not be available for comparison with experimental data. Further research is required to ensure measurements with several instruments are comparable, like the study of Szabó et al. about collision energies in proteomics [223], and to increase information from spectral libraries in all possible conditions.

Tandem mass spectrometry

Metabolomic studies, especially untargeted approaches, usually employ one of these three acquisition modes: full scan, data dependent acquisition or DDA, and data independent acquisition or DIA. All three modes were employed throughout this PhD Thesis. Therefore, a brief comparison of them is presented below, and **Figure 2.14** summarizes the main advantages and limitations of each of the modes.

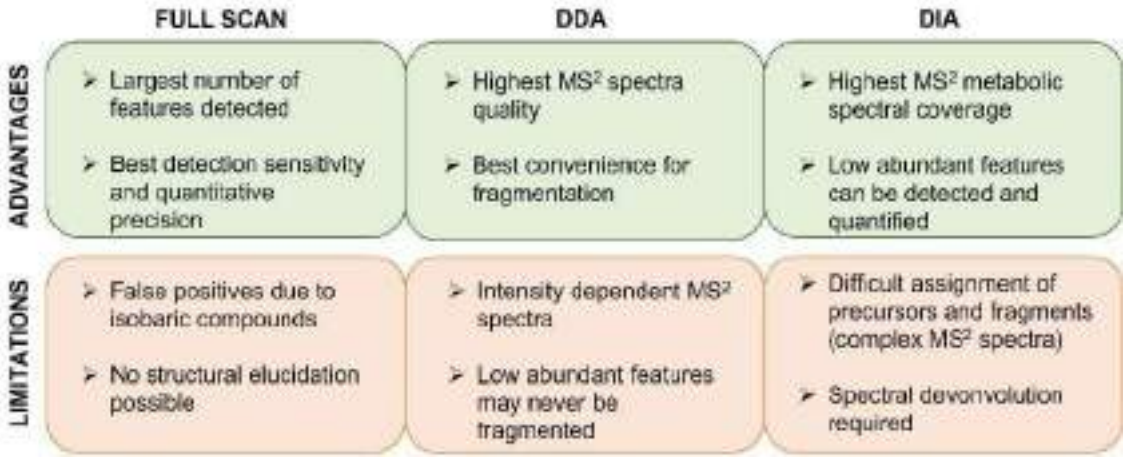


Figure 2.14. Summary of advantages and limitations of the three acquisition modes selected in this PhD Thesis: full scan, data dependent acquisition and data independent acquisition. *Based on the publication by Guo et al.* [224].

The choice of the acquisition mode is of great importance because it is directly related to the quality of the metabolomic results. A comparison of these three modes in metabolomic studies was recently published by Guo et al. [224]. Several parameters were evaluated, including metabolic coverage (i.e., number of features detected), quantitative precision, MS^2 spectral quality and spectral coverage, and convenience (e.g., practicality in untargeted analysis).

Unsurprisingly, the largest number of detected features was obtained with full scan mode. All the ions are captured in a single run and the best sensitivity detection and quantitative precision is achieved, because the instrument focuses only on the MS^1 level. However, only accurate mass is obtained, and structure elucidation or discrimination between isobaric compounds is impossible. To ensure maximum sensitivity, some examples in the literature propose the acquisition of all samples in full scan mode and then, only fragment the QCs, which will be representative pools of all the samples [225].

In DDA, the instrument selects a list of precursor ions above an intensity threshold and follows user-guided criteria for further fragmentation at the MS^2 level. Clean MS/MS spectra for each of the precursors are collected sequentially. Hence, DDA provides the highest quality of MS^2 spectra, and the association between the precursors and their fragments is straightforward. Therefore, metabolite annotation is significantly improved [226]. The main drawback of DDA is that low abundant features may not be selected due to the prior selection of the instrument and, consequently, ignored leading to reduced coverage of the metabolome due to instrumental factors. DDA method was used in this PhD Thesis for the untargeted lipidomic analysis of zebrafish embryos exposed to Endocrine-disrupting chemicals (**scientific publication V**).

Table 2.5. Summary of all studies that appear in this PhD Thesis.

Model biosystem	**Yeast**	**Zebrafish eleutheroembryos**		**Rice**		**HegG2 cells**	
Emerging pollutant exposure	Not applicable; method development	Endocrine disruptor chemicals (*bipshenol A and estradiol*)		Metalloids (a*rsenic*)		Pharmaceutical compounds (*carbamazepine, amoxicillin, trazodone*)	
Omic study	Lipidomics	Lipidomics	Metabolomics	Lipidomics	Metabolomics	Lipidomics	Metabolomics
Approach	Untargeted	Untargeted	Untargeted	Untargeted	Untargeted	Targeted	Untargeted
Analytical technique	LC-HRMS	LC × LC-HRMS	LC-HRMS	LC-HRMS	LC-HRMS	LC-MS	LC-HRMS
MS acquisition mode	Full scan	Full scan + MS/MS (DDA)	Full scan + MS/MS (DDA)	Full scan	Full scan + MS/MS (DIA)	SRM	Full scan + MS/MS (DIA)
MS analyzer	TOF	QTOF	QTOF	TOF	Orbitrap	QqQ	QTOF

In contrast, DIA can also detect and quantify low abundant ions because it generates MS^2 spectra for all precursors without any discrimination, increasesing reproducibility between experiments. Thus, the MS^2 metabolic spectral coverage is the highest of the three modes. Besides, no undersampling can occur due to fast acquisition rates [226]. All ion fragmentation (AIF) and sequential window acquisition of all theoretical fragment-ion spectra (SWATH) are common DIA modes. The main downside of DIA is that spectral deconvolution is required to match precursor ions with their fragments because of the complexity of MS^2 spectra. However, a

retrospective analysis of DIA data can be performed if, for instance, a new deconvolution algorithm is available.

For instance, in this PhD Thesis, CorrDec deconvolution algorithm [227] and the original MSDIAL deconvolution method (MS2Dec) [28] were employed in the untargeted metabolomic analysis of culture cells exposed to pharmaceutical compounds (**scientific publication VIII**). The principle of CorrDec is that the intensities of the precursors and fragments correlate across samples. In this case, MS/MS spectra from different sample types are required for deconvoluting, which means that fragmentation should not be held only on the quality control samples.

To sum up, **Table 2.5** summarizes all studies included in this PhD Thesis, emphasizing the analytical techniques and the addressed environmental issue. MS acquisition mode and MS analyzer are also specified.

2.2.8 Metabolite annotation

Once *m/z* values and MS/MS fragmentation patterns are obtained for the compounds of interest, they need to be associated with the name of a metabolite or lipid. The metabolite annotation step is one of the main bottlenecks often found in untargeted metabolomics, because of the huge number of metabolites detected and the presence of isobar compounds with similar fragmentation patterns. Besides, identifying potential markers of environmental exposure is crucial, and a prior step to identify the pathways affected and draw biological conclusions of the effects of certain pollutants.

Metabolite identification can be organized in four levels of confidence [228], as schematized in **Figure 2.15**. Level 1 refers to fully identified metabolites. Usually, information on two orthogonal properties of a tentative metabolite is confirmed with its commercial standard under the same experimental conditions (e.g., same RT and *m/z*). In levels 2 and 3, putative identification is achieved. The most common scenario is that the metabolite is annotated based on RT, *m/z* and MS/MS fragmentation patterns (Level 2). If no MS/MS information was obtained, the metabolite family or class could be indicated based on RT and *m/z* (Level 3). The last level, number 4, refers to metabolites whose identity remains unknown.

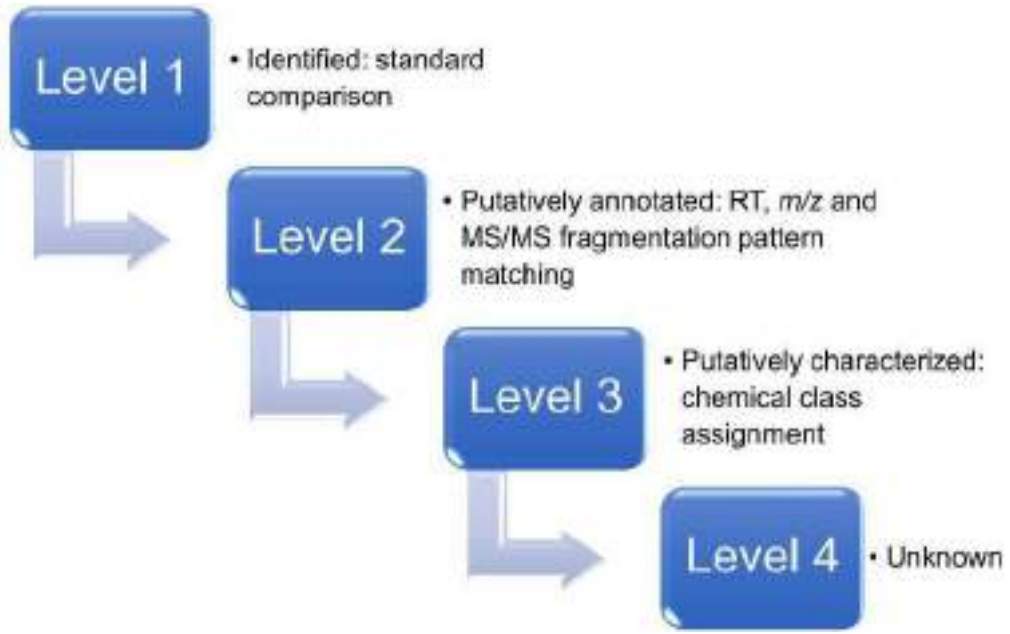


Figure 2.15. Confidence levels in metabolites identification for untargeted metabolomics studies.

In this PhD Thesis, most of the metabolites have been annotated in confidence levels 2 and 3, except for **scientific publication VIII,** in which sphingolipids were fully identified (level 1) thanks to the match with their internal standards. In putative identification, the *m/z* value obtained is compared to the theoretical values of different adducts for the exact mass of the potential metabolite. MS/MS fragmentation patterns are also verified with the theoretical ones. Databases provide both exact mass and MS/MS information with spectral libraries. The databases employed during this PhD Thesis are: HMDB [138,139], YMDB [38], METLIN [140], Massbank [141,142], LIPID MAPS [143,144], NIST [145] and PlantCyc [229]. On the other hand, the Kyoto Encyclopedia of Genes and Genomes (KEGG) [230] has been employed specifically to identify the affected metabolic pathways.

2.3 Data analysis strategies in metabolomics

Metabolomic datasets are usually complex and overwhelming in size, especially in untargeted studies where high-resolution mass spectrometry has been employed. Hence, obtaining the biological information sought is not always straightforward. There are many aspects to consider, such as peak alignment, data compression, feature detection, chromatogram resolution, isotope recognition, exclusion of false-

positive peaks, and which type of multivariate analysis is more suitable in every case [32].

Every year, new online tools, databases and resources appear with the aim of facilitating this difficult task and automate the whole process. A recent review from Misra [231] summarizes latter tools and resources currently available for metabolomic data analysis, a field of great expansion nowadays.

Chemometrics is the science that aims to extract information from chemical systems. This discipline relies on mathematical and statistical methods to exhaustively analyze and ease the interpretation of complex datasets. From experimental design and optimization steps to the biological interpretation of the metabolites affected by the exposure, chemometrics is a powerful bridge that provides the tools needed. Therefore, cutting-edge chemometric methods seem to be the perfect match to deal with current limitations due to the huge size of metabolomic datasets [232].

In this section, the chemometric tools selected for the analysis throughout the PhD Thesis will be described. An especial emphasis has been dedicated to the steps of data compression and resolution, in a combined strategy known as ROIMCR, coupling the regions of interest (ROI) and the multivariate curve resolution (MCR) approaches.

2.3.1 Data analysis workflow for metabolomic datasets

The standard metabolomics data analysis workflow contains three main steps: pre-processing, resolution and post-processing, shown in **Figure 2.16.A**. In the first step, raw chromatograms are submitted to spectral or time compression (e.g., binning or time-windowing, respectively), background correction (e.g., denoising and smoothing or baseline corrections), and/or retention time alignment. This group of strategies aims to improve reproducibility between all chromatograms before peak detection, as well as filter and convert the datasets into more manageable matrices. Besides, overlapping signals are commonly encountered in metabolomic datasets due to the high complexity of the biological matrices. Thus, the second step aims to enhance resolution and ease peak detection. Other peak-tracking strategies are also of great utility, above all when LC × LC is employed [233]. The third step is related to all statistical procedures that can be applied to the data matrix containing the areas of the most relevant features. The choice of the analysis type will depend on the goal of the metabolomic study. For instance, exploratory analysis is a useful approximation for an overview of the dataset in untargeted analysis. On the other

hand, univariate analysis such as ANOVAs or t-tests can be a good option for targeted analysis with a reduced number of features.

A recent review by Bos et al. collects chemometric strategies that can be applied to one and two-dimensional liquid chromatography data for each of these main steps, plus a section dedicated to chromatographic optimization [195]. More specifically into the metabolomic field, the review from Paul et al. includes chemometric tools commonly applied in metabolomic studies [234].

Figure 2.16.B exemplifies how these three steps previously mentioned are applied to an LC-MS metabolomic dataset throughout this PhD Thesis. In LC-MS, experimental data are arranged into a matrix with the *m/z* values in the columns and the retention times in the rows. A single LC-MS run is considered two-way data (i.e., a two-dimensional matrix), whereas a whole LC-MS dataset with multiple samples may be considered three-way data (i.e., a three-dimensional data set or data cube), formed by the set of LC-MS individual matrices from the different samples when they are arranged together and analyzed simultaneously.

In this PhD Thesis, the **regions of interest** or ROI strategy is proposed [235,236] for the mass spectral data noise filtering and compression, together with the **multivariate curve resolution alternating least squares** or MCR-ALS [237–239], for the full resolution of the chromatograms and mass spectra of the constituents (metabolites, lipids) of the analyzed samples. These two data analysis strategies have been combined in the ROIMCR method [236,240,241]. The output of the ROI step is a column-wise augmented data matrix where the different samples are concatenated vertically, one below the other. This step is crucial in noise filtering and data size reduction. The output data matrix will only contain the most important *m/z* values, i.e., the ones above a certain intensity threshold established *a priori* while still conserving the full instrumental mass accuracy. Results of MCR-ALS are on one side the elution profiles, and on the other side the spectra profiles. Thus, quantitative and qualitative information can be obtained from MCR-ALS resolved elution profiles of all the components in the different samples and their peak areas are readily available. These peak areas matrix is used afterwards for performing all multivariate analyses (e.g., exploratory or classification studies). In the following sections, both ROIMCR and the different multivariate analysis methods for metabolomics data analysis studies are discussed in more detail.

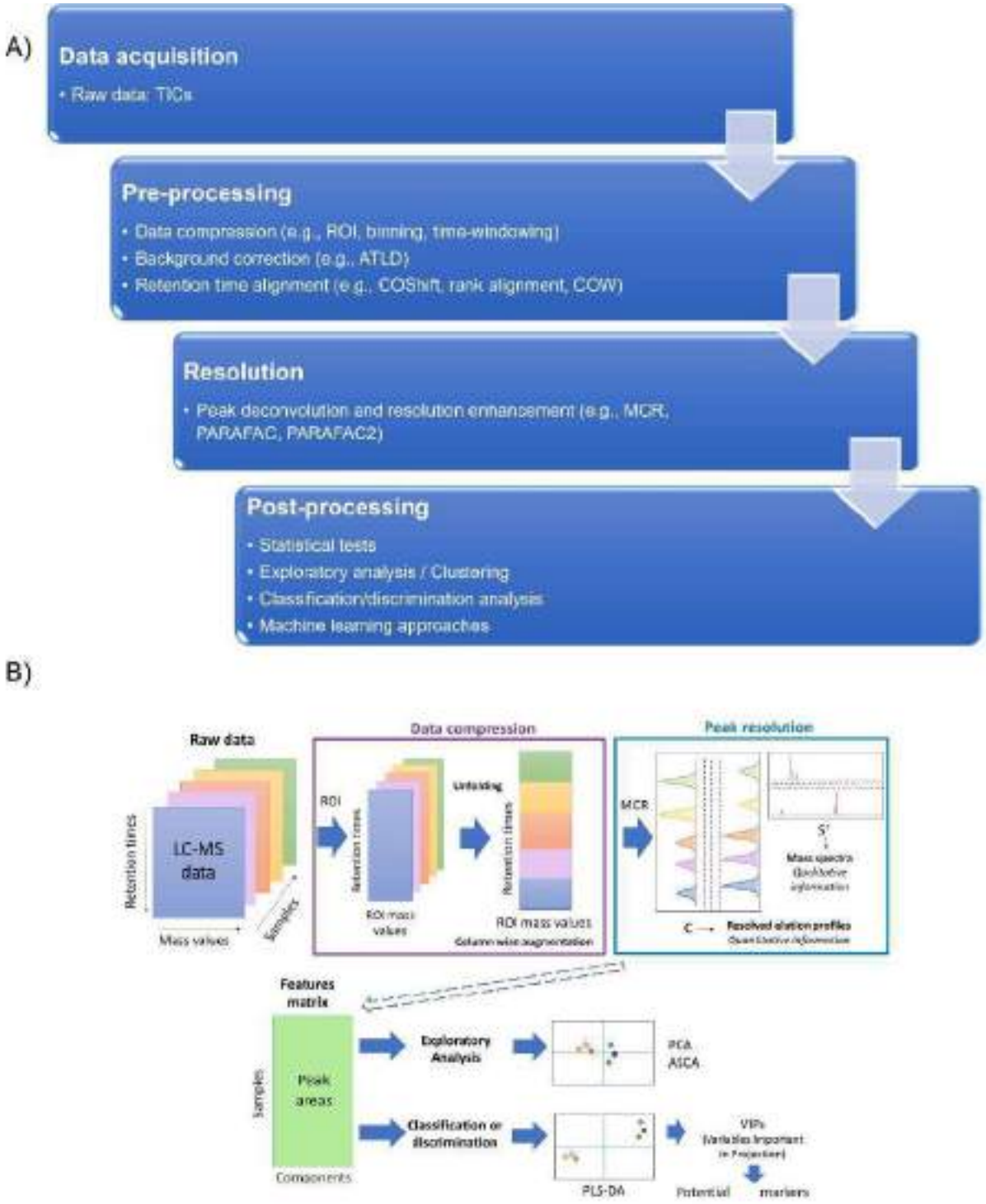


Figure 2.16. **A**) Generic workflow for metabolomics data analysis. **B**) Example of the pipeline employed in this PhD Thesis for the analysis of untargeted LC-MS metabolomics data. TIC: total ion chromatogram; ROI: regions of interest; ATLD: alternating trilinear decomposition; COShift: correlation-optimized shifting; COW: 2D correlation optimized warping; MCR: multivariate curve resolution; PARAFAC: parallel factor analysis; PCA: principal component analysis; ASCA: anova-simultaneous component analysis; PLS-DA: partial least squares discriminant analysis.

2.3.2 ROIMCR

The ROIMCR method is composed of the coupling of the Regions of Interest (ROI) and the Multivariate Curve Resolution Alternating Least Squares (MCR-ALS) methods. It has proven to be a powerful strategy for data analysis, especially in metabolomics studies using MS data [236]. This approach has been used in untargeted metabolomic and lipidomic studies with several analytical techniques and many different applications, e.g., LC-MS [157,240,242–246], LC×LC-MS [154], or MSI [247]. Apart from metabolomics, ROIMCR has also been employed in other applications, such as the analysis of proteins [241], contaminants of emerging concern [248] both in samples from wastewater treatment plants, as well as in environmental studies [248–251].

Compared to other compression approaches, like binning [252], ROI filters the data by searching only for the relevant features without any loss of spectral accuracy. Besides, one of the major advantages of applying MCR-ALS is that it does not require any prior peak alignment nor modelling step. This simplifies the whole data treatment workflow, especially in 2DLC, where large chromatographic shifts are commonly encountered. The combination of ROI and MCR-ALS strategies allow the resolution of the chemical constituents of the analyzed samples, their concentration (elution) and spectra profiles from which relative quantitative information and qualitative information associated with metabolite identification can be acquired.

Regions of interest

The ROI method was initially described by Stolt et al. [253,254], and afterwards introduced into the centWave algorithm of the XCMS platform for analyzing metabolomic data [25,255]. ROI was later adapted to the MATLAB environment [236]. A user-friendly interface has been recently released for its use in the analysis of LC-MS GC-MS, LC×LC-MS, GC-GC-MS and MSI data [235], and also capillary electrophoresis mass spectrometry (CE-MS) [256] or ion mobility mass spectrometry (IM-MS) datasets.

The ROI strategy is based on spectral data filtering and compression by determining the so-called regions of interest. These are *m/z* data regions where the MS signals are considerably more intense than the *a priori* set signal-to-noise ratio threshold. Therefore, these are the data spots where the most relevant analytes are located and will be preserved for further analysis. In contrast, the rest of very low intensity values (below the noise threshold) are discarded. The ROI selection produces a filtered compressed ROI data matrix in the spectral dimension, as already shown in **Figure 2.16**.

There are three main parameters to consider in the ROI selection approach, which are visualized in **Figure 2.17.** The first is the signal-to-noise ratio (S/N) threshold, basically an intensity filter placed just above the noise baseline level. The aim is to clean the mass spectra by keeping the relevant features and discarding the low intensity signals that are noise-related. The second parameter is the *m/z* error interval or mass accuracy that can be considered acceptable and will depend on the mass spectrometer spectral resolution. Mass accuracy provided by low-resolution instruments will be lower than high-resolution ones, and consequently, the acceptable *m/z* error will be higher. The third parameter is the minimum number of occurrences, or the minimum number of consecutive points in the time dimension needed to properly define the peak. This will depend on the number of readings needed to determine a chromatographic peak and on the chromatographic flow and mass detector speed.

ROIs are searched at every mass spectrum and retention time, and the MS signals at the *m/z* values in common within the *m/z* error established are jointed. For each ROI, the final *m/z* value assigned is the mean (or the median) of all points grouped in that specific ROI. This spectral compression can be applied simultaneously to multiple samples. If a certain peak is not present in some samples (or its intensity value is below the S/N threshold), a zero value (or better, a very low random value) is automatically dispensed. A matrix with the areas of the most relevant features can also be obtained from the ROI procedure.

A more detailed description of the ROI methodology can be found in **scientific publication II** (included below in this Section).

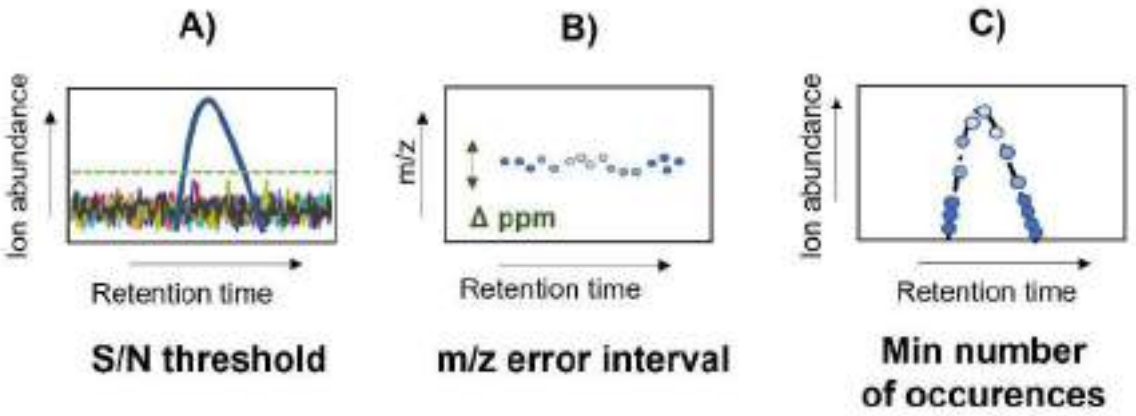


Figure 2.17. Three main parameters need to be optimized in the ROI procedure: A) signal-to-noise ratio (S/N) threshold; B) *m/z* error interval; and C) minimum occurrences to define a peak. *Adapted from* [236].

Multivariate curve resolution alternating least squares

MCR-ALS is a well-known chemometric tool employed in the past thirty years to resolve complex chemical mixtures. Applications are wide and include datasets from many different analytical techniques (e.g., chromatography coupled to mass spectrometry, hyperspectral imaging, NMR, spectroscopic or electrochemical analysis), as recently reviewed by de Juan et al. [237].

MCR-ALS is based on a bilinear decomposition model that corresponds with **Equation 2:**

Equation 2 $$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathbf{T}} + \mathbf{E}$$

D refers to the measured experimental data matrix. The two main outputs of the MCR-ALS procedure are the column factor matrix, C, which is the concentration matrix that contains the elution profiles, and the row factor matrix, $\mathbf{S}^{\mathbf{T}}$, which is the spectra matrix or response matrix including the measured variables. **E** is the residuals matrix which includes the variance not explained by the MCR model. The number of rows of **D** and **C** is the same, whereas the number of columns of **D** and $\mathbf{S}^{\mathbf{T}}$ matches as well. **C** and $\mathbf{S}^{\mathbf{T}}$ contain the information of elution or spectra profiles, respectively, for each of the resolved components. Ideally, each component should represent a relevant chemical compound. For instance, when applying MCR-ALS to an LC-MS dataset, each component refers to the different MS signals at the *m/z* values associated with the same elution profile. This means that the MS signals at the *m/z* values for the different adducts of the same chemical compound appear in the same MCR-ALS component.

The decomposition of matrix **D** follows three main steps, summarized in **Figure 2.18**. First, the number of components required for explaining the chemical variance (not the experimental noise) should be determined. Singular value decomposition (SVD) [257] is a common method employed for this purpose. If too many components are chosen, then it is possible that noise is being introduced in the model (overfitting), which should be avoided., Some chemical constituents may not be resolved if too few components are chosen. When complex datasets are analyzed, as in MS metabolomics, the choice of the exact number of components may be not straightforward. In these cases, several models can be compared and the one that best describes the data variance with the minimum number of components and with chemical meaningful elution and spectra profiles, will be selected.

An initial estimation, either of spectra ($\mathbf{S}^{\mathbf{T}}$) or elution profiles (**C**), is needed to start the ALS optimization. A widely employed method for detecting pure variables in

any of the two data directions or modes (elution or spectra) is an adaptation of the Simple-to-use Iterative Self-modeling Mixture Analysis (SIMPLISMA) algorithm [258].

During the alternating least squares (ALS) optimization, matrices **C** and $\mathbf{S}^T$ are re-estimated iteratively under constraints to be chemically meaningful, not only from the mathematical perspective. Constraints help to reduce the ambiguity associated with the bilinear MCR model [259,260]. Frequently used constraints are non-negativity (both elution and spectra profiles can only be positive), unimodality (there is only one chromatographic peak for each elution profile), and peak areas (concentrations) correlation (used to build calibration curves).

The optimization process finishes when convergence is achieved, meaning that the relative difference of standard deviations from residuals between experimental values and the adjusted with ALS in consecutive iterations are below an *a priori* established value (typically 0.1%) [261].

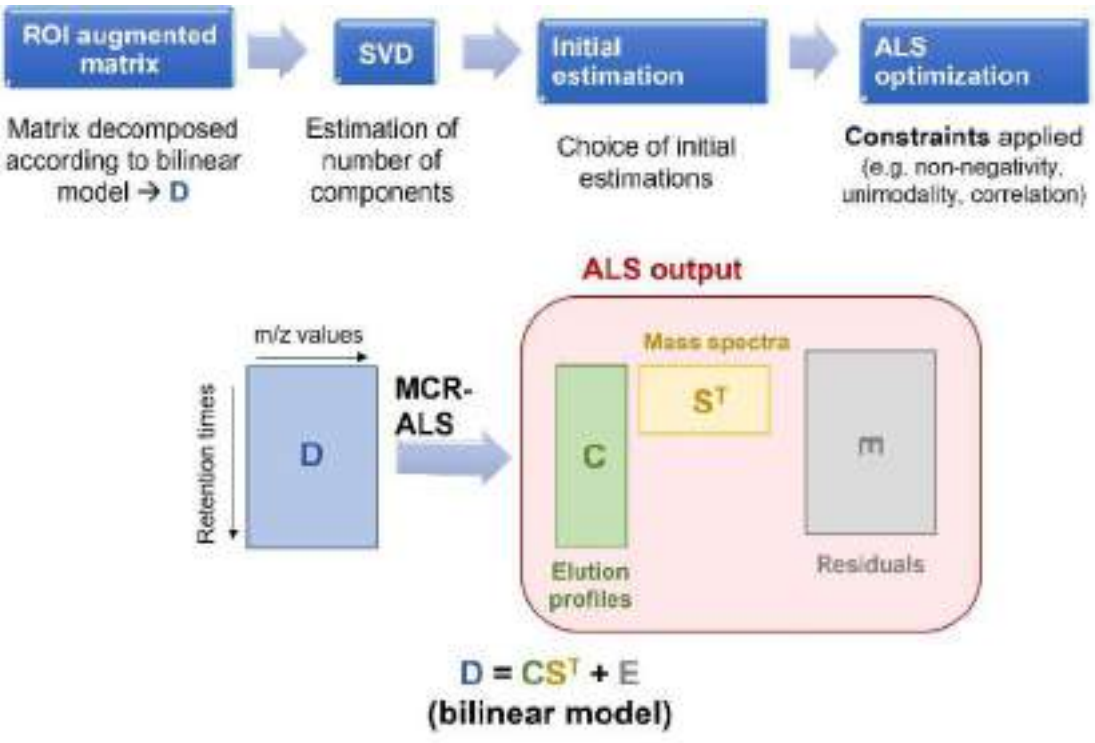


Figure 2.18. MCR-ALS implementation to a LC-MS dataset step by step. SVD: Singular value decomposition; ALS: alternating least squares.

From the peak areas (or the maximum intensity of the peaks) from the MCR-ALS resolved elution profiles in the **C matrix**, it is possible to know the relative abundances of each metabolite from which quantitative information can be derived. Likewise, from the spectral profiles in the $\mathbf{S}^T$ matrix, it is possible to extract qualitative information and annotate the metabolites present in the samples.

The MCR methodology used in this PhD Thesis is explained more in-depth in **scientific publication II** (included below in this section), especially describing how matrices multiple samples can be simultaneously analyzed in column-wise matrix augmented (1DLC) and superaugmented (2DLC) data matrices.

ROIMCR for 2DLC datasets

When dealing with 2DLC datasets, a compression step becomes even more urgent due to their huge data size, up to 13 Gb in the analysis of a single sample (a single data file) is coupled to high-resolution mass spectrometry [235]. Previous works in our group attempted different combinations of spectral compression plus data segmentation in different time-windows in order to reduce data dimensionality and speed up further calculations [201,262]. Nevertheless, in fully comprehensive and untargeted studies, fractioning the chromatograms in different windows or regions makes the analysis more complicated and slows down the whole analysis.

In **scientific publication II**, the strategy ROIMCR is proposed for dealing with 2DLC datasets. First, the LC×LC data structure is discussed (i.e., multiway datasets), and different possible bilinear and multilinear models are discussed for their analysis. Then, an extensive description of the whole ROIMCR procedure is presented, as well as other suitable strategies for pre-processing 2DLC data. A list of post-processing approaches that can be used to analyze 2DLC datasets is additionally given, which can also be applied to most LC-MS datasets indistinctly. Finally, different examples of applications of the ROIMCR strategy to 2DLC datasets are included.

II. SCIENTIFIC PUBLICATION II

Title: Untangling comprehensive two-dimensional liquid chromatography data sets using regions of interest and multivariate curve resolution approaches

Authors: Miriam Pérez-Cova, Joaquim Jaumot, Romà Tauler

Citation reference: Trends in Analytical Chemistry 137 (2021) 1162072

DOI:10.1016/j.trac.2021.116207

Trends in Analytical Chemistry 133 (2021) 116297

Contents lists available at ScienceDirect

Trends in Analytical Chemistry

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

Untangling comprehensive two-dimensional liquid chromatography data sets using regions of interest and multivariate curve resolution approaches

Miriam Pérez-Cova [a, b], Joaquim Jaumot [a], Romà Tauler [a, *]

[a] Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, 08034 Barcelona, Spain
[b] Department of Chemical Engineering and Analytical Chemistry, University of Barcelona, Diagonal 647, 08028, Barcelona, Spain

ARTICLE INFO

Article history:
Available online 28 January 2021

Keywords:
LC × LC
MCR-ALS
ROIMCR
Chemometrics
Data analysis

ABSTRACT

Data analysis remains a major challenge in the global application of comprehensive two-dimensional liquid chromatography (LC × LC). Advanced chemometric tools have been proposed to reduce the complexity of LC × LC datasets. In this work, key aspects of LC × LC are summarized from a chemometrics perspective. In particular, the recently developed ROIMCR method is proposed and adapted for LC × LC data analysis. First, this strategy consists of selecting of the Regions of Interest (ROI), in which data are filtered and compressed. Second, the resolution of the elution profiles of the sample constituents using the Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS) method. A detailed overview of this recently developed tool and examples of its application in LC × LC are given, as well as pre-processing and post-processing tools to facilitate and complement the analysis of LC × LC data and the optimal interpretation of results.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

Multidimensional separations have gained popularity in the last decades as a response to the need for increasing separation power [1,2] in the analysis of complex natural samples with matrix effects. The combination of different analytical platforms in the same study overcomes the limitations of one-dimensional approaches, i.e., resolution of overlapping peaks. Coelution of multiple analytes in one-dimensional liquid chromatography (1D-LC) might difficult its resolution, whereas the aid of an extra separation dimension can potentially separate, identify and quantify them. For instance, in comprehensive two-dimensional liquid chromatography (LC × LC), the effluent coming out from the first dimension (^{1}D) column is collected and divided into fractions at fixed periods (so-called modulations), which are then further separated in the second dimension (^{2}D) column. However, the large number of detected signals in these experiments, especially when coupled to high resolution mass spectrometry (HRMS), makes the manual inspection of chromatograms not feasible. In addition, beyond the analytical challenges, a significant increase in the data size and complexity is produced when considering multidimensional chromatography [2,3].

Thus, there are two main challenges of LC × LC data analysis compared to 1D-LC. The first is the higher complexity of the obtained data structures, e.g., multiple chromatographic peaks are obtained for a single analyte in the second column [2]. The second is the need for data analysis software. Some examples of commercial software available are GC Image LC × LC Edition Software from GC Image™, AnalyzerPro® XD from SpectralWorks, and Chromsquare from Shimadzu. However, they present some limitations, as only basic pre-processing steps are commonly included in their workflows. Besides, some of them present vendor-specificity, which difficulties general strategies applicable independently of how data have been acquired [2]. Consequently, the resolution of these complex mixtures and the extraction of the maximum amount of information from them is still a major challenge nowadays [2,4]. Hence, the lack of well stablished data-analysis protocols and software [3,5] emerges as the main drawback of multidimensional separations, and specifically of LC × LC.

Chemometrics is the discipline that applies mathematical and statistical methods to chemical systems with two main goals [6,7]. On the one hand, to improve the measurement process, in order to obtain optimal procedures and experiments. On the other hand, to extract maximum relevant qualitative and quantitative information from the chemical measurements. Consequently, chemometrics

* Corresponding author.
E-mail address: roma.tauler@idaea.csic.es (R. Tauler).

https://doi.org/10.1016/j.trac.2021.116297
0165-9936/© 2021 Elsevier B.V. All rights reserved.

List of acronyms

^{1}D	first dimension
^{2}D	second dimension
2DALC	2D assisted LC
ACD	at-column dilution
ANOVA	analysis of variance
ANOVA TP	ANOVA by target projection
ANOVA-PCA	ANOVA-principal component analysis
APTLD	alternating penalty trilinear decomposition
ASCA	ANOVA-simultaneous component analysis
ASM	active solvent modulation
ATLD	alternating trilinear decomposition
CDShift	correlation-optimized shifting
COW	correlation optimized warping
DAD	diode array detector
DOE	experimental design
DWT	dynamic time warping
GASCA	group-wise ANOVA simultaneous component analysis
GC × GC	comprehensive 2D GC
IKSFA	iterative key set factor analysis
IOPA	iterative orthogonal projection approach
LC × LC	comprehensive 2D LC
LC × LC-HRMS	comprehensive 2D LC-high resolution MS
LC-LC	heart-cutting 2D LC
MANOVA	multivariate analysis of variance
MCR-ALS	multivariate curve resolution-alternating least squares
mLC-LC	multiple heart-cutting 2D LC
NEB	normal-exponential bernoulli
NG	normal-gamma
NGB	normal-gamma-bernouilli
NN	neural networks
OPLS-DA	orthogonal partial least squares-discriminant analysis
PAFFT	peak alignment fast fourier transform
PARAFAC	Parallel Factor Analysis
PARAFAC2	Parallel Factor Analysis 2
PCA	principal component analysis
PFP	pentafluorophenyl phase
PLS-DA	partial least squares-discriminant analysis
RF	random forest
ROI	regions of interest
ROIMCR	regions of interest-multivariate curve resolution
SCA	simultaneous component analysis
SIMCA	soft independent modeling of class analogies
sLC × LC	selective comprehensive 2D LC
SOM	self-organizing maps
SPAM	stationary-phase-assisted modulation
SVD	singular value decomposition
SVM	support vector machines
SWALTD	self-weighted alternating trilinear decomposition
TAGs	triacylglycerides
VEM	vacuum-evaporation modulation
VIPs	variables important in projection

includes, for instance, experimental design, multivariate calibration, pattern recognition and classification. Advanced chemometric methods go a step further from basic data analysis tools, also embracing methods for more sophisticated data pre-processing and model improvement, variable selection or resolving very complex mixtures among others. Thus, cutting-edge chemometrics has arisen as powerful tools able to shed some light into this issue and get through this bottleneck, offering solutions that can facilitate data compression, resolution and eventually interpretation of results. These goals are often achieved by combining different strategies. First, there are methods which aim to pre-process the experimental data to improve their quality (i.e., methods for data compression, baseline correction, elimination of background signals, alignment of chromatographic peaks within modulations of the same or among multiple samples). Second, other methods pursue to extract the sought analytical information from the data, including the discovery, resolution and quantitation of the components present in the analyzed samples. Hence, the usual output of this step is a table or matrix of the peak areas or concentrations of the resolved components. Third, there are additional post-processing steps, which include multiple types of multivariate data analysis methods. Their objective is to explore the patterns present in these tables or matrices and classify them into different groups (in the case where several types of samples are compared). Statistical analysis can also be performed to assess the effects of the experimental factors in designed experiments or to discover the most important variables (i.e., biomarkers) defining the investigated processes, often from a multivariate point of view. Although the third step can be applied directly on the pre-processed chromatograms from step one, only major differences between samples will be observed. It might be useful for a quick and visual overview of patterns in the data. For a more in-depth analysis, the second step is highly recommended, as it provides quantitative information. Besides, if the aim is, for instance, to identify variables responsible for the differences between the samples, performing this second step is necessary.

The aims of the present work are, on one side, to briefly describe the state of the art of LC × LC, and on another side, to go in-depth into the chemometric point of view, focusing on the structure of the datasets (i.e., possible multilinear behavior), and on the best strategies to analyze these huge data sets. In this sense, we propose the use of the spectral compression strategy based on the searching of the Regions of interests (ROI) [8], and its combination with the Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS) [8,9] in the ROIMCR method for the optimal analysis of LC × LC data sets. In addition, some pre-processing and post-processing strategies of multivariate analysis of LC × LC data are also discussed, and some recent data examples of application are described to illustrate these data analysis strategies.

2. 2D-LC state of the art and recent developments

Two-dimensional chromatography (2D-LC) aims to increase peak capacity and peak production rates compared to 1D-LC when analyzing complex mixtures of compounds, often with strong peak coelutions and matrix effects. The second column or dimension, ^{2}D, adds extra selectivity to the separation, thanks to the possibility of combining two different retention mechanisms, ideally non-correlated (i.e., orthogonal) [1].

Here, only online chromatographic analysis is considered, where both dimensions are connected through a high-pressure valve that acts as the interface. However, other strategies, such as offline and stop-flow modes, have been also proposed [10]. Online setup allows a full automatization of the separation. There

are multiple types of online 2D-LC according to the number of fractions from the first column, ^{1}D, transferred to the second column, ^{2}D. Heart-cutting (LC-LC) implies that only one fraction from ^{1}D is collected and analyzed in the ^{2}D. In multiple heart-cutting (mLC-LC), several fractions of the ^{1}D separation are selected and stored in multiple loops before their further separation in the ^{2}D; whereas in selective comprehensive (sLC × LC), certain specific and successive regions of the ^{1}D separation are subsequent analyzed in the ^{2}D separation. Finally, comprehensive mode, LC × LC, implies the complete analysis of the effluent from the ^{1}D column also in the ^{2}D column [11]. In all cases, the separation already obtained in the ^{1}D needs to be maintained. Otherwise, undersampling will occur, which means that separated peaks are merged in the ^{2}D [12].

LC × LC is particularly powerful in untargeted analysis, where the comprehensive screening and profiling of all the constituents of a sample or set of samples are possible in one single chromatographic run. Non-destructive detectors (UV, Fluorescence) can be added in the middle of the two separations if needed, although it is not frequent. Nevertheless, destructive detectors (e.g., MS), can only be used at the end of both separations. Most of the examples of LC × LC applications are coupled to UV–Vis [13,14], MS [15,16] or both [17–19]. An extra dimension can be added if, for instance, LC × LC is coupled to ion mobility, as some other recent examples have proven [20–22].

One of the technical aspects that have suffered major improvements nowadays in LC × LC is the modulation strategy, i.e., how the interface between the two chromatographic separations is performed. A considerable effort has been spent in increasing solvent compatibility between dimensions. In passive modulation mode, there is no modification in the fractions from the ^{1}D prior to its further analysis in the ^{2}D column. Therefore, the ^{2}D separation may be highly affected by the ^{1}D effluent. Solvent strength mismatch should be avoided since it can lead to peak distortion and breakthrough, especially when peaks are not enough retained in the stationary phase due to the injection of a strong solvent coming from the previous separation. As a consequence, some of the sample constituents are eluted in the dead volume, instead of at their usual retention times [1]. As breakthrough issues arise as a major concern in LC × LC, special emphasis in its understanding has been made lately, to avoid or minimize it as much as possible. More information can be found, for instance, in the publications by Moussa et al. [23], about breakthrough from sampling loops, and van der Ven et al. [24], about how to improve the analysis of water-soluble biopolymers.

Active modulation strategies have been proposed to deal with solvent strength mismatch, breakthrough, and decreased detection sensitivity, caused by the dilution suffered by the sample in the LC × LC system. The main approaches developed to solve this problem are Active Solvent Modulation (ASM) [25], Stationary-Phase-Assisted Modulation (SPAM) [26], Vacuum-Evaporation Modulation (VEM) and at-column dilution (ACD) [27]. In ASM, the effluent from the ^{1}D, already stored in the sample loop, is diluted with weak solvent before this fraction exits the valve and enters the ^{2}D column. SPAM replaces standard storage loops by low-volume trapping columns, known as "traps", whose stationary phase composition is similar to the ^{2}D column. VEM applies heat under vacuum conditions in the ^{1}D effluent to deposit the analytes in the loop, which are re-dissolved prior to their introduction in the ^{2}D column. A more detailed description of these three approaches can be found in Ref. [1]. ACD is the most recent approach and aims to automatically regulate dilution factor by adjusting flows from a transfer pump and from the ^{2}D gradient, which are joint in a mixer before reaching the ^{2}D column. The result is modulation based on a dilution at-column [28].

Optimization of the separation conditions in LC × LC is not straightforward. There are many parameters and considerations to be taken into account. The first step is the choice of the retention mechanisms for both dimensions, and therefore the selection of the appropriate stationary phases. This selection would require chromatographic expertise and previous knowledge about the chemical properties of the sample and targeted analytes. Then, mobile phase solvents, composition and modifiers need to be selected. In order to calculate retention parameters, statistical screening tests are useful as an input for the design of the experiments, optimal retention and chemometric modeling. Preliminary predictions of the retention times and simulation of elution profiles of the different analytes will reduce experimental effort and speed up the whole optimization process [29], especially due to the usually long LC × LC runs. Besides, automated multicolumn LC × LC workflows can accelerate method development, as they allow fast identification of the best combinations of columns and mobile phases compositions, for both targeted and untargeted analysis [30,31].

Choosing what quality descriptors (i.e., orthogonality, resolution) are required is a relevant step prior to objectively evaluate the quality of the separation under multiple possible scenarios. Another parameter to consider is how the gradient is performed in the ^{2}D, as different gradients can be applied. The main three are full, shifted and parallel gradients. A recent comparison of the three options can be found in Ref. [32]. In short, full gradients are the most common. They widely vary mobile phase composition in a very short period of time, and need re-equilibration time at the end, which limits the space in the 2D separation, as the modulation time is not fully employed in the separation itself. In contrast, in shifted gradients, the mobile phase composition range in every modulation is narrower and changes in agreement with the retention of the compounds in the ^{1}D column. However, they require specific software and hardware. Parallel gradients seem an attractive alternative in the case that retention mechanism of ^{1}D and ^{2}D are correlated, as in the case of employing reverse phase in both, RP × RP. In this option, the ^{2}D gradient is practically isocratic. The main advantage is that peak capacity increases, as the space available for the separation is higher than in the case of full gradient, for instance. Orthogonality is also improved when parallel gradients are used in both dimensions.

Evolutionary algorithms have been employed for optimizing gradient separations in LC × LC. These algorithms can be applied to method development and retention modeling, but also to molecular design or molecular modeling [33,34]. A comparison between genetic algorithms, non-adaptive evolution strategies and the covariance matrix adaptation evolution strategy has recently been published [35].

More detailed information about the use of LC × LC methods, practical aspects and data analysis can be found in previous reviews. For instance, the evaluation of the combination of different retention mechanisms in LC × LC was studied by Pirok et al. in Ref. [36]. In addition, recent developments in LC × LC and their new applications can be found in another publication by Pirok et al. [1]. A summary on the peak detection and profiling strategies for multidimensional chromatography was published by Navarro-Reig et al. in Ref. [4]. More information about retention modeling is reviewed in an article by den Uijl et al. [37]. Lastly, the chemometric analysis in one and two-dimensional chromatography was reviewed by Bos et al. in Ref. [2].

3. Multiway structure of LC × LC data

One important consideration to keep in mind when analyzing LC × LC data is how the data structure complexity increases when moving from one dimensional to multidimensional separations. For

instance, in one-dimensional liquid chromatography coupled to mass spectrometry (LC-MS), data are usually arranged in matrices with retention times in the rows and mass-to-charge ratios values (*m/z* values) in the columns. In chemometrics, this means that a single sample provides a two-way data set. Other common multi-wavelength detectors (e.g., LC-UV-Vis, using diode array detector or DAD), provide equivalent data structures. Analogously, when several samples are analyzed, three-way data sets are obtained, as shown in part **A** of Fig. 1. In multidimensional chromatography, more complex multiway data sets are obtained. For instance, in two-dimensional separations, a single sample gives a three-way data set, whose ways (modes/directions) are: ^{1}D retention times, ^{2}D retention times, and the multivariate response of the detector. In LC × LC-MS, this third mode is composed of the *m/z* values of the measured mass spectra. Therefore, the analysis of multiple samples in this situation will lead to a four-way dataset, exemplified in part **B** of Fig. 1. When combining different detectors at the same time, (i.e., LC × LC-DAD-MS), a data fusion step is required in order to horizontally concatenate the spectral data from each detector.

Another key aspect that should be considered despite data structure is what type of mathematical model suits the most for the analysis of the multiway data structure. According to the data set complexity, several multilinear models can be employed. In some cases, the choice can be straightforward, e.g., linear models for one-way data, bilinear models for two-way data, trilinear models for three-way data, and so on. However, this is not always necessarily true. For instance, there are some cases where the application of a trilinear model is not adequate for three-way data. Then, a lower complexity multilinear model, as the bilinear model, could be a better option.

Hence, the choice of the multivariate data analysis method depends on the multilinear behavior of the measured data. The fulfillment of trilinear models can be assessed when every chemical constituent can be expressed with a unique profile in every one of the three data ways, e.g., along the ^{1}D, ^{2}D and spectral dimensions for all samples analyzed, and they are invariant through the experimental conditions of the study [38–41].

Trilinearity loss can be caused by retention time shifts or changes in the peak shapes. This means that the same peak in different chromatographic runs may not appear in the same position and/or the same elution profile, which can be caused, for example, by fluctuations in pressure, flow, mobile phase composition and/or temperature [42,43]. These temporal misalignments in the data are often found in chromatography. For instance, although data from two-dimensional gas chromatography coupled to mass spectrometry (GC × GC-MS) have been analyzed under the

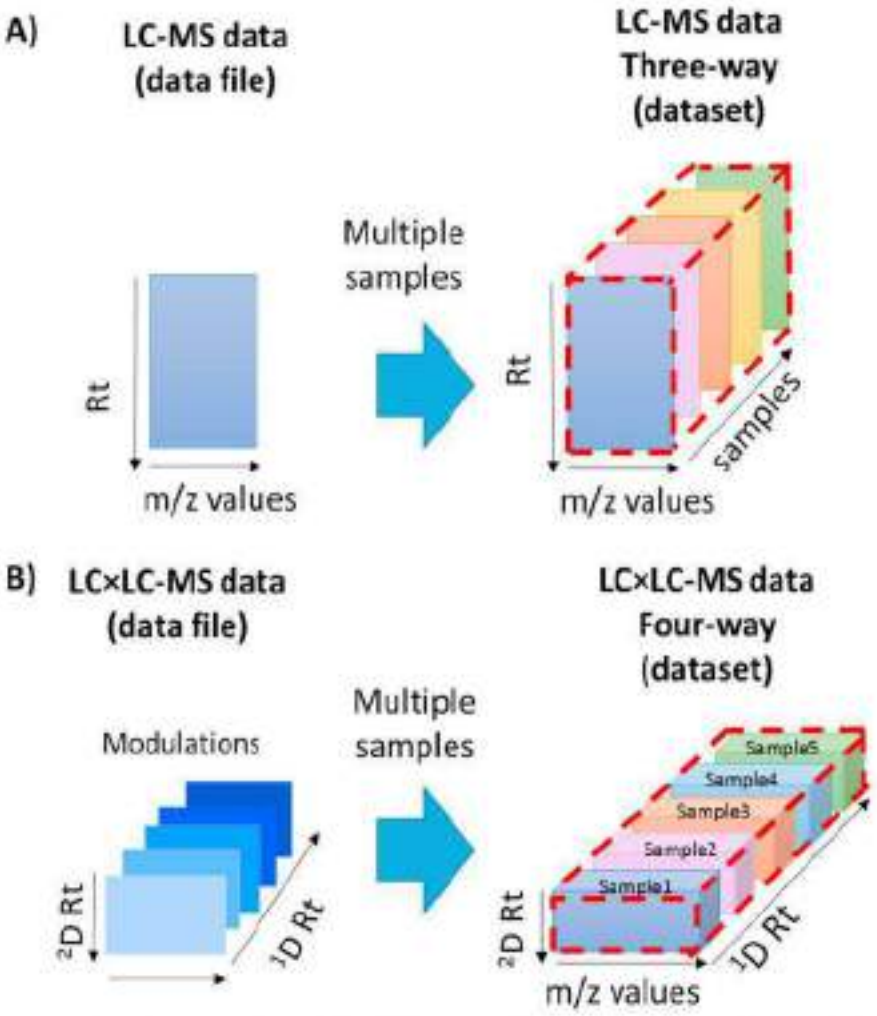


Fig. 1. Comparison of data arrangements in LC-MS (part A) and LC × LC-MS (part B). **A)** Results of LC-MS analysis of a single sample give a data matrix, whereas simultaneous LC-MS analysis of multiple samples provides a three-way data set. **B)** A single LC × LC-MS run is composed by the different modulations in which the ^{1}D separation is fractioned giving a three-way data set, and multiple LC × LC-MS runs provide a four-way data set.

assumption of the trilinear model [44,45], small deviations of this model have been encountered [46,47]. This issue is clearer in the case of liquid chromatography, even for 1D-LC, and it becomes critical in 2D-LC. Previous studies performed with LC × LC-MS and LC × LC-DAD, confirmed that deviations from the trilinear behavior were present [5,48,49]. The reason for this is because in LC × LC the total reproducibility of the elution profiles of the same component is not fulfilled, not only between the different chromatographic runs, but also within modulations. This variation is caused by both time shifts and changes in peak shapes. Therefore, in LC × LC, it is especially important to check if there are retention time shifts between modulations, and also between different samples, a problem which is commonly encountered in practice.

There are two manners of dealing with non-trilinear chromatographic data. The first implies the data pre-processing to achieve a proper alignment of peaks (see below, at the end of Section 4). Consequently, trilinearity is restored before employing methods based on trilinear models, such as Parallel Factor Analysis (PARAFAC). However, some difficulties may arise (e.g., increased complexity, the apparition of unexpected constituents, changes on chromatographic peak shapes when they are strongly coeluted, etc). A potential tool able to solve some of these difficulties is the Tchebichef moments' approach. This image moments method has demonstrated some advantages as pre-processing tools in the analysis of multiway chromatographic data with overlapping peaks and peak drifts [50]. However, this will be not possible when changes in the shapes of the chromatographic profiles of the same component in the different chromatographic runs are also produced.

The second involves the use of flexible algorithms which allows having different profiles in the samples mode for the same component, such as MCR-ALS or Parallel Factor Analysis 2 (an extension of PARAFAC that allows small deviations in its multilinear behavior). PARAFAC2 allows small time shift departures between the elution peaks of the same component in the different runs and modulations (a problem commonly encountered, especially when using parallel gradients). But when coelution occurs, this approach also fails, especially because of changes in the shapes of peak profiles.

A comparison of the effects in peak shapes and shifts in the trilinear structure of the data has been shown elsewhere [51,52]. These works discuss the limitations of PARAFAC2 and the recommendation of the use of MCR-ALS in the analysis of complex samples where the fulfillment of the trilinear model is not achieved. Thus, when trilinearity is not accomplished, the application of trilinear or higher multilinear models is not recommended, and simpler bilinear models should be applied instead [49,52–54].

In the case of LC × LC data, multiple strategies have been proposed to evaluate the adequacy of the application of the trilinear model and its extension in the analysis of a particular data set. For instance, one possibility is the comparison of the singular value decomposition (SVD) of the concatenated data matrices in their different augmentation modes (column- or row-wisely), obtained when the LC × LC data three-way array is unfolded [55]. Other options are the evaluation of the core consistency diagnostic of the PARAFAC decomposition [56], and the comparison of the data fitting obtained when bilinear and trilinear models are applied in multivariate curve resolution [49,57]. In the latter case, if the data behave following a trilinear model, data fitting results with the two models should be similar, apart from the effects of lower of degrees of freedom and some overfitting in the case of the softer bilinear model. These three strategies, together with the examination of the reliability of the resolved profiles in each case, help to investigate whether the trilinear model is adequate for the analysis of LC × LC-MS and LC × LC-DAD datasets, or if on the contrary, the use of trilinear and higher multilinear models should be avoided [49].

4. Pre-processing, data compression and feature selection

Multiple pre-processing methods can be proposed for the analysis of experimental chromatographic data to enhance their quality, e.g., eliminate noise and baseline contributions, signal smoothing, peak alignment or modeling. Most of them represent the adaptation from those approaches used in one-dimensional liquid chromatography. A brief summary of different pre-processing strategies is included below, with special emphasis to the Regions of Interest, ROI, approach.

The first type of pre-processing tools is focused on data compression. The size of LC × LC-MS data files is considerably large, especially if HRMS is employed, as intensity signals at thousands of m/z values can be acquired in a single scan. Besides, long analysis runs are also common in these multidimensional separations (i.e., more than one hour per chromatogram). This means that compression and filtering are crucial steps to reduce data dimensionality, filter noisy signals and make them more manageable, especially when multiple datasets (samples) are analyzed. Simultaneously, data compression and filtering can be performed in the spectral dimension (columns direction) or the chromatographic dimension (rows direction). The first type of strategy focuses on searching for the most relevant and significant m/z values, above an intensity threshold in the spectral direction, i.e., above noise-related instrumental signals. In contrast, the compression in the chromatographic direction can be especially useful when the focus of the study is on just one specific region of the 2D plot, i.e., on a specific range of retention times in the first or second dimensions.

There are several manners to reduce the size of LC × LC data sets. For instance, a classical approach is binning, which converts raw mass spectra into a matrix representation. The m/z axis is divided into equidistant pieces according to a specific bin size, related to the MS resolution [58]. However, some of its disadvantages include how difficult it is to choose the bin size in each data set, to avoid, for example, joining several peaks into the same bin, or the contrary, peak splitting into different bins. In contrast, wavelet compression reduces the data chromatographic dimension size without significant loss of information, while the effects of noise are minimized. Wavelets decompose the chromatographic data according to their frequency into a new reduced scale, but preserving the spatial location of the chromatographic peaks, their intensities and shapes [59]. It is also useful as a denoising tool, as wavelets are automatically adapted to remove noise-dependent high frequencies of a signal as well as to preserve low-frequency components [60,61]. Another strategy, known as time-windowing, aims to accelerate calculations while reducing storage is to divide the chromatogram into time sections or windows, which can be afterwards analyzed individually [62]. It is also worth to mention the approach proposed by Sisanian et al. consisting of a series of feature extraction processes (i.e., using resolution methods as described below) from low to high spectral resolution. This method allows evaluating the entire dataset in a fast and straightforward manner. However, there is a risk of missing low-concentration compounds (i.e., low explained variance) in the first steps of the analysis due to the large bins employed in the initial binning [63].

The Regions of Interest (ROI) strategy has been proposed to handle the large data sets obtained by HRMS analysis. The application of the ROI procedure allows a compression in the spectral dimension (column direction), filtering and pre-processing MS data in a straightforward manner in a single step, searching for the more relevant data regions and signals [64]. The main advantage of ROI compared to other procedures used for data compression and

matrix transformation, (i.e., binning or bucketing, [58]), is that its application is performed without any loss of the spectral accuracy of the raw measured MS data [64], an aspect which is fundamental for identification purposes and structure elucidation. The ROI algorithm proposed in this work is similar to the one implemented in the centWave of the popular XCMS metabolomic platform [58,65]. More information about practical aspects on how to process ROI in MATLAB or R can be encountered in Refs. [64,66].

In LC × LC, prior to analysis, some additional preliminary data rearrangements are required. In the case of considering a single sample, the data matrices from subsequent modulations should be concatenated vertically, one below the other, creating a column-wise augmented data matrix, where m/z values are in the columns. When several samples are analyzed at a time, a column-wise super-augmented data matrix is generated. This matrix includes the different modulations concatenated vertically, one below the other, repeatedly for all the samples, which are displayed analogously, sample two under sample one, etc. Fig. 2 depicts these two possible scenarios and how the column-wise augmented data matrices are built.

Therefore, a column-wise ROI super-augmented data matrix is generated with the common m/z values selected in its columns and filtered according to three main parameters: the MS signal intensity threshold, the mass accuracy (deviation error), and the minimum number of occurrences. The first parameter is an intensity threshold value, which should be above the noise baseline level. It is crucial not to select a threshold value too high, otherwise relevant low intensity signals from low concentration compounds can be lost. On the contrary, if the threshold is set too low, too much background noise will be included into the data analysis. In practice, this threshold value is usually established between 0.1% and 1% of the maximum intensity of the measured signal. It is also possible to combine this threshold value with a multiplication factor (so-called minmax factor), which allows setting a low threshold, but at the same time only consider these ROIs whose intensities are above the product of the threshold and the factor (e.g., intensity threshold x 3). The second parameter is the mass error uncertainty (mass accuracy) associated with the MS instrument, and should be proportional to its spectral resolution. HRMS measures have a very low mass error (high mass accuracy), but the specificities of every employed instrument are needed to be taken into account. This mass accuracy can be defined in terms of absolute mass units (daltons) or alternatively in relative ppm units, and it should be set at a multiple number

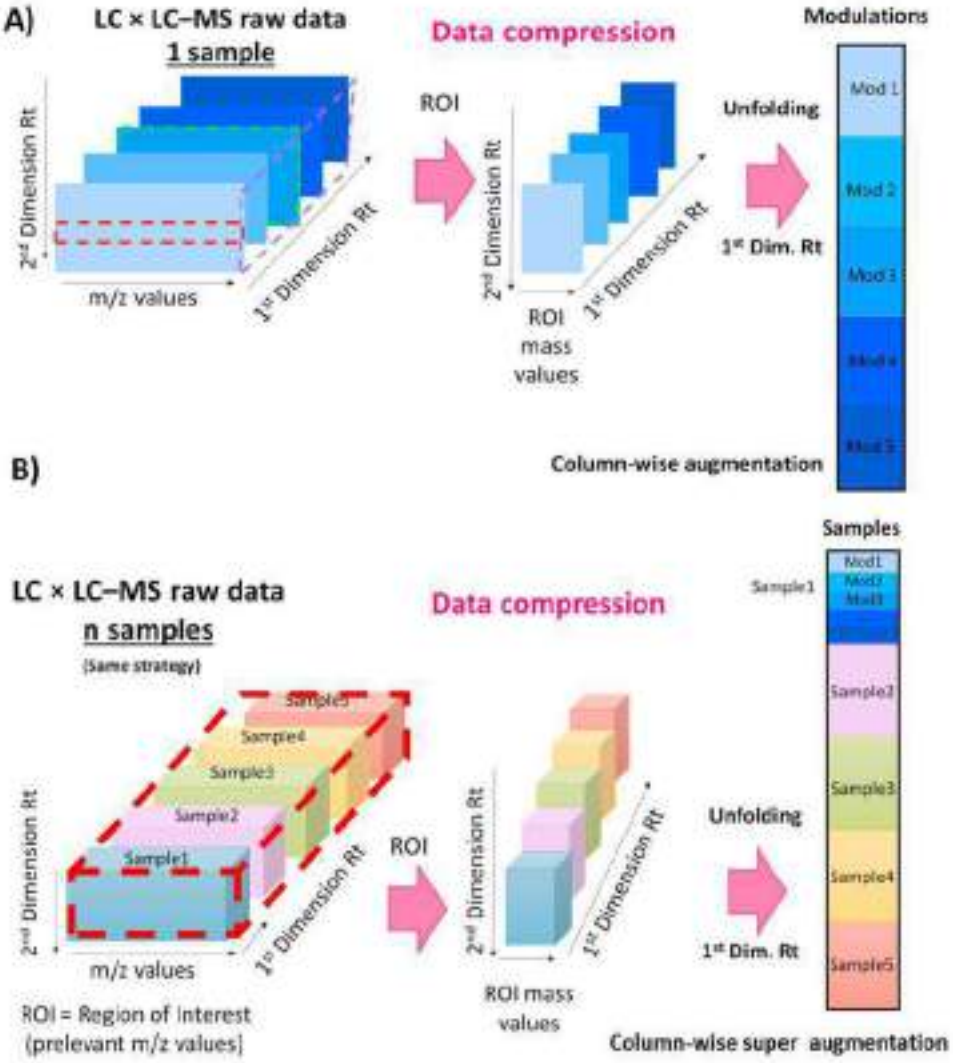


Fig. 2. Arrangement and ROI data compression of LC × LC-MS data (one or multiple samples) in a column-wise augmented ROI data matrix. **A)** All modulations from the LC × LC-MS analysis of the same sample are concatenated vertically, one below the other in a column-wise augmented data matrix. **B)** LC × LC-MS column-wise data matrices from different samples are further concatenated vertically in a new single column-wise super-augmented data matrix.

(i.e., 3 or 5 times) of the expected mass spectral resolution of the considered instrument. Consequently, features with very close m/z values within the selected mass accuracy are joined in one single m/z ROI, or they kept separately, if they differ more than the tolerated mass accuracy. The third parameter in the ROI evaluation is the minimum number of occurrences that a particular ROI signal should occur to be reliable. In chromatography, this value is related to the minimum number of retention times that are necessary to depict an elution peak correctly and to the acquisition frequency of the equipment. In the particular case of LC × LC, this value needs to take into consideration the width of retention times associated with the different peaks that belong to the same compound, which can be present in the subsequent modulations.

Although the ROI method was initially designed for the untargeted analysis of all significative masses of a particular sample [58], the ROI method can be adapted to the study of targeted ions or particular spectral regions, where only certain m/z values corresponding to *a priori* known compounds are selected and analyzed. This option also allows setting specific ranges of retention times of interest for its further analysis. Advantageously, the evaluation of the ROIs obtained provides information about the distribution of m/z values present in the same ROI, about the peak shape of the elution profiles assigned to the ROI in the samples, and about the relative peak areas of the same compound eluted in the subsequent modulations. The individual 2D plot representation of each LC × LC ROI is also useful and informative.

The ROI procedure gives two main outputs, as shown in Fig. 3. On the one hand, a vector including the list of ROI m/z values considered relevant according to the previously selected parameters. On the other hand, a data matrix containing the MS signal intensities at the selected ROIs, for all retention times and all the simultaneously analyzed samples. As detailed in Fig. 3, this data matrix will be arranged in a column-wise super-augmented data matrix where the intensities of the m/z values for each modulation, (individual mass spectra scans), are concatenated one below the other, for all first column retention times and all samples. The intensities at all retention times from both dimensions (rows) for the selected ROI m/z values (columns) are gathered. Then, this ROI augmented data matrix can be analyzed by the MCR-ALS method, as discussed in the following section.

Other pre-processing methods are focused on eliminating background noise, perform instrumental drift correction and signal smoothing. Despite the noise filtering step provided by the wavelets approach, modeling baseline and background contributions is recommended, using one or more additional components in the bilinear factor decomposition type of models, like in MCR methods [46] (see below). Other strategies that have been employed in GC × GC include removing the background contributions from the chromatographic 2D signals estimated using their structural and statistical properties [67], or by curve fitting and linear interpolation techniques [46,68]. Specifically, in LC × LC, the trilinear model decomposition based on the alternating trilinear decomposition (ATLD) method has been proposed to eliminate the background signal drifts without the need of running a blank sample or having previous knowledge about the sample composition [69]. Other methods, like PARAFAC, self-weighted alternating trilinear decomposition (SWALTD), or alternating penalty trilinear decomposition (APTLD) have also been proposed for the same purposes [70,71].

Alignment of chromatographic peaks is also a critical step in many methods analyzing LC × LC data [72,73]. In this case, retention time shifts within a sample are corrected among the subsequent ^{2}D modulations where the same sample constituent elutes. It is also important to remember that, in LC × LC, shifts in the chromatographic peaks between different samples are even more likely

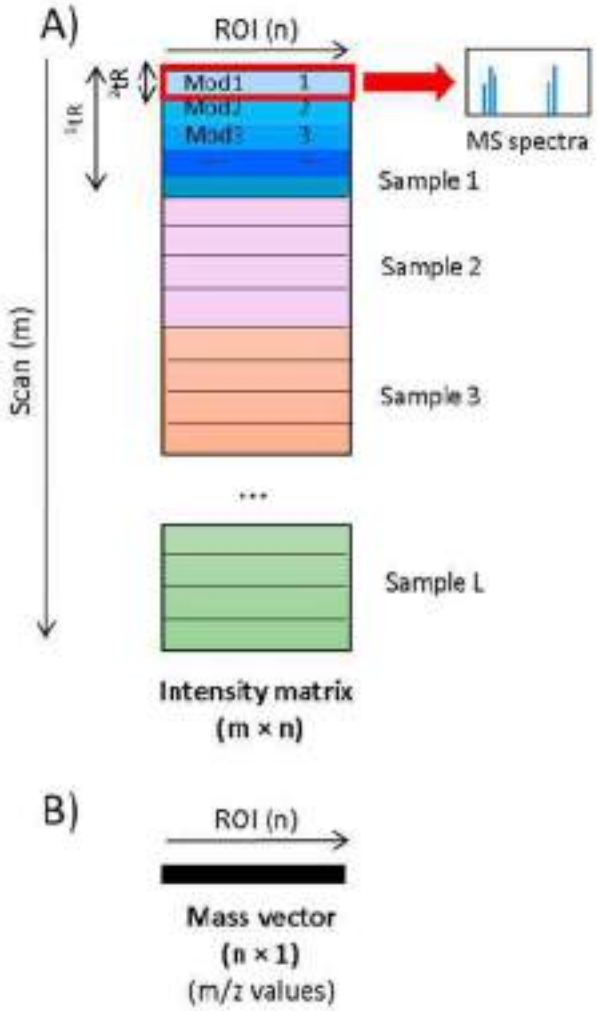


Fig. 3. Description of the two main outputs obtained by the ROI procedure: **A)** Data matrix with the MS signal intensities at the selected m/z ROIs for all modulations and samples analyzed; **B)** Data vector with the list of the m/z values selected by the ROI searching and filtering procedure.

than in LC, due to the long time of the ^{1}D, in contrast to the continuous and short times of analysis in the ^{2}D. Different methods have been proposed to perform this correction in multidimensional separations, such as rank alignment [74], correlation-optimized shifting (COShift) [75], 2D correlation optimized warping (COW) [76], dynamic time warping (DTW) [77], or peak alignment fast fourier transform (PAFFT) method [78]. However, in general, other related phenomena, e.g., peak swapping or absent peaks, are not fixed with peak alignment strategies. It is worth to emphasize here that in the case of the MCR based methods, chromatographic peak alignment correction is not needed since the only requirement for the data analysis is their alignment in the spectral direction, which is easily achieved, in contrast to the chromatographic peak alignment [46,79,80].

Finally, peak detection methods provide information about the different sample consituents resolved during the chromatographic separation in both dimensions. Usually, these methods include peak modeling and signal smoothing tools as a part of the detection process. Recently developed methods for peak detection in multidimensional chromatography include the Normal-Exponential Bernoulli (NEB) algorithm and mixture probability models [81], both procedures included in the *msPeak* R package [81]. In this case, baseline correction, background subtraction,

recognition of potential peak regions, peak picking, peak areas integration, and final peak detection by mass spectral similarity are part of the same software package. Similar adaptations of this method include Normal-Gamma (NG) [82], and Normal-Gamma-Bernouilli (NGB) [83]. Other peak detection alternatives have been also proposed [84]. In this case, initially, the method only considers the first chromatographic dimension, but then in a second step, peaks from the same compound are jointed in a two-dimensional peak, using a merging algorithm based on Bayesian inference [84]. More detailed descriptions about pre-processing methods for LC × LC data can be found elsewhere [2,85]. Apart from the pre-processing and feature selection methods briefly summarized above, multivariate curve resolution methods can be applied in order to recover the elution and spectra profiles of the different constituents present in the analyzed samples. In the following section, the MCR-ALS method will be discussed in more detail.

5. Multivariate Curve Resolution Alternating Least Squares (MCR-ALS)

Multivariate Curve Resolution methods resolve the constituents of unknown mixtures component by component using a bilinear model. It is important to notice that, despite its similarities, resolution and deconvolution are not synonyms. Whereas the deconvolution term is commonly used in the context of univariate signals (like in total ion or in single wavelength UV chromatograms), the term resolution is more appropriate in the context of multivariate signals, like in full scan LC-MS or multiwavelength LC-UV. In the first case, the deconvolution techniques used are usually model-based curve-fitting approaches (e.g., gaussian elution peak shapes). However, in the second case, the multivariate resolution techniques are model free, i.e., they do not need the postulation of any model to describe the shape of the profiles. They are derived directly from the bilinear multivariate data model. Thus, unlike deconvolution algorithms used with single channel univariate detection methods, Multivariate Curve Resolution methods deal with the whole response of multivariate detection methods (e.g., at multiple wavelengths, m/z, ...) and do not require the postulation of a signal shape type of model.

When MCR methods are applied to spectroscopically (multichannel, multivariate) hyphenated chromatography, they allow the direct mathematical resolution of the chromatographic overlapped peaks of the coeluted sample constituents as well as the resolution of their pure multichannel responses (spectra) of the analyzed constituents of the sample. In the case that MS is employed as a multichannel detector, the resolved mass spectra contain rich qualitative information useful for identifying of the different constituents of the analyzed mixtures. At the same time, from the resolved elution profiles, it is possible to obtain quantitative information, from their peak areas or their peak heights.

Mathematically, MCR methods decompose the experimental datasets according to a bilinear additive factor decomposition model which naturally corresponds to the generalization of Lambert-Beer's law of molecular absorption. This bilinear model can be described using linear algebra data matrix notation as:

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathbf{T}} + \mathbf{E} \quad (1)$$

where **D** is the data matrix with the experimental measures, which in the case of a chromatographic run monitored spectroscopically, have the spectra at the different retention times in the rows ($i = 1, \ldots, I$), and the chromatograms at several spectral channels/variables (wavelengths, m/z, ..., etc.) in the columns ($j = 1, \ldots, J$). The MCR bilinear decomposition provides the concentration/elution profiles of the different components ($n = 1, \ldots, N$) of the analyzed sample/mixture in the columns of the matrix **C**, and the spectra of these components in the rows of the matrix $\mathbf{S}^{\mathbf{T}}$. Matrix **E** in this Equation (1) represents the obtained residuals and accounts for the variance not explained by the bilinear model. When multiple hyphenated chromatographic runs are simultaneously analyzed, this bilinear model can be easily extended according to Equation (2):

$$\mathbf{D}_{\mathbf{aug}} = \begin{bmatrix} \mathbf{D}_1 \\ \ldots \\ \mathbf{D}_L \end{bmatrix} = \begin{bmatrix} \mathbf{C}_1 \\ \ldots \\ \mathbf{C}_L \end{bmatrix} \mathbf{S}^{\mathbf{T}} + \begin{bmatrix} \mathbf{E}_1 \\ \ldots \\ \mathbf{E}_L \end{bmatrix} = \mathbf{C}_{\mathbf{aug}}\mathbf{S}^{\mathbf{T}} + \mathbf{E}_{\mathbf{aug}} \quad (2)$$

where now the multiple chromatographic runs of different samples ($l = 1, \ldots, L$) analyzed with the same spectroscopic detector (UV, MS, etc.), each one of them giving an individual data matrix ($\mathbf{D}_1, \ldots \mathbf{D}_L$), are vertically concatenated in the column-wise augmented data matrix $\mathbf{D}_{\mathbf{aug}}$, with $I \times L$ rows and (J) columns. $\mathbf{C}_{\mathbf{aug}}$ has the concentration/elution profiles of the (N) components present in the (L) analyzed samples and $\mathbf{S}^{\mathbf{T}}$ the spectra of these components. Observe that, in this case, the bilinear model implies that the concentration/elution profiles of the same components in the simultaneously analyzed runs/samples, i.e., in $\mathbf{C}_1, \ldots \mathbf{C}_L$ can be different, whereas their spectra are the same in $\mathbf{S}^{\mathbf{T}}$. The non-explained variances are now in the $\mathbf{E}_{\mathbf{aug}}$ residual matrix.

A similar extension of the MCR bilinear model can be performed in the case of 2D chromatography. The data from one run/sample analyzed by 2D can also be arranged in an augmented data matrix $\mathbf{D}_{\mathbf{aug}}$ with $I \times K$ rows and J columns, where (I) refers to the number of retention times in the second column, and (K) refers to the number of modulations, as shown in Equation (3).

$$\mathbf{D}_{\mathbf{aug}} = \begin{bmatrix} \mathbf{D}_1 \\ \ldots \\ \mathbf{D}_K \end{bmatrix} = \begin{bmatrix} \mathbf{C}_1 \\ \ldots \\ \mathbf{C}_K \end{bmatrix} \mathbf{S}^{\mathbf{T}} + \begin{bmatrix} \mathbf{E}_1 \\ \ldots \\ \mathbf{E}_K \end{bmatrix} = \mathbf{C}_{\mathbf{aug}}\mathbf{S}^{\mathbf{T}} + \mathbf{E}_{\mathbf{aug}} \quad (3)$$

When this strategy is further extended to the simultaneous analysis of several LC × LC chromatographic runs (different second column modulations from multiple samples), the bilinear model applied to the column-wise super-augmented data matrices can be written as:

$$\mathbf{D}_{\mathbf{saug}} = \begin{bmatrix} \mathbf{D}_{\mathbf{aug}1} \\ \ldots \\ \mathbf{D}_{\mathbf{aug}L} \end{bmatrix} = \begin{bmatrix} \mathbf{C}_{\mathbf{aug}1} \\ \ldots \\ \mathbf{C}_{\mathbf{aug}L} \end{bmatrix} = \mathbf{C}_{\mathbf{saug}}\mathbf{S}^{\mathbf{T}} + \mathbf{E}_{\mathbf{saug}} \quad (4)$$

As also detailed in Fig. 4, $\mathbf{D}_{\mathbf{saug}}$ (size of $I \times K \times L$ rows, J columns) is the LC × LC column-wise super-augmented data matrix, where I corresponds to the number of retention times in the (K) modulations of the second chromatographic dimension, for the (L) samples, and (J) corresponds to the number of spectral channels (wavelengths in DAD and m/z values in MS). $\mathbf{C}_{\mathbf{saug}}$ ($I \times K \times L$, N) is the matrix containing the resolved 2D chromatographic profiles in the (L) samples, in the (K) modulations, for the (N) components, and $\mathbf{S}^{\mathbf{T}}$ (N, J) are their corresponding spectra. Finally, $\mathbf{E}_{\mathbf{saug}}$ ($I \times K \times L$, J) contains the residuals not explained by the bilinear model using this number of components.

The Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) is a popular version of MCR method which uses an iterative alternating linear least squares (ALS) optimization to resolve the component profiles. ALS is a matrix factorization

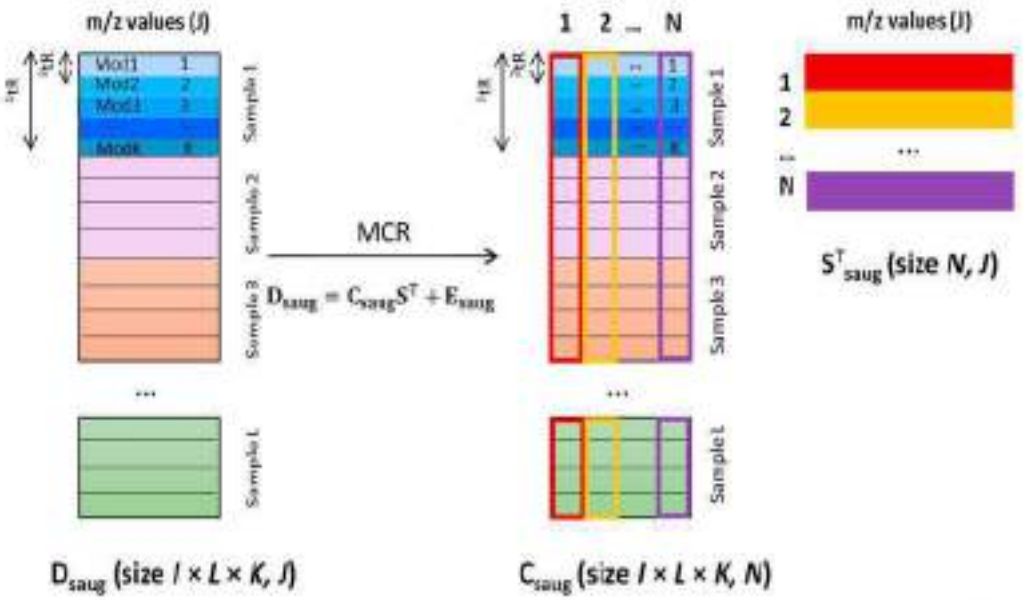


Fig. 4. MCR resolution of a single LC × LC data set. The column-wise augmented data matrix, $\mathbf{D}_{saug}$, is decomposed into two factor matrices: $\mathbf{C}_{saug}$ and $\mathbf{S}^T$, with the elution profiles and pure mass spectra of the resolved components.

algorithm. It includes regression steps to find solutions to local linear models where alternatively $\mathbf{C}$ or $\mathbf{S}^T$ are fixed (Equation (1)). The MCR-ALS method has been described in detail elsewhere [9,86], and it can be applied independently of the detector employed (e.g., MS, DAD, Fluorescence). The first step of the MCR-ALS procedure is selecting of the number of components, N, which is initially roughly estimated from the number of singular values [87] larger than experimental noise. In addition, other criteria that can be used in this selection are the evaluation of the amount of variance explained by each component, and the visual assessment of the reliability of the resolved chromatographic and spectral profiles associated with these components. The goal is that this number of components explains enough data variance, while avoiding the possibility of data overfitting with too many components explaining just experimental noise. In the case of MS, this step is more difficult, and may require a more detailed analysis and selection, specially to detect minor components contributing very little to the total measured signal. However, also in this case, the high selectivity of the MS detector facilitates this selection, specifically for the components contributing significantly (i.e., more than the noise) to the measured signal.

Next step is to get initial estimates (concentration profiles or spectra) for the selected number of components. Pure variable detection methods are usually a good choice in the LC × LC case since they can provide an initial estimation of the elution or spectral profiles for the selected number of components in a single step. These methods find the most dissimilar chromatographic elutions or mass spectra directly from the measured chromatographic raw data. When the overlapping in the chromatographic direction is lower than in the spectral direction, it is recommended to select the purest elution times and work with the initial estimates of the spectra measured at these selected times [9]. This is the case, for instance, of DAD detection. However, in the case of MS detection, the situation is different because the spectral direction is more selective. Therefore, the search for the elution profiles at the purest m/z channels is suggested.

With these initial estimations, the ALS iterative optimization starts. The ALS iterative process ends when the convergence criterion is achieved, i.e., when the relative change of the standard deviation of the residuals between consecutive iterations is lower than a preselected threshold value [9], or the maximum number of iterations selected is reached.

Since solving the bilinear MCR model does not assure unique solutions, different MCR solutions with same data fit but with distinct concentration, and spectral profiles are feasible. This number of feasible solutions, generally known as ambiguity, is one of the main concerns when working with MCR-based methods. There are different types of ambiguities but the more relevant are rotational ambiguities (due to the existence of rotation matrices that can affect the shape and intensity of the resolved profiles without changing data fitting and fulfilling the applied constraints). Application of inherent data properties can diminish the impact of these rotational ambiguities in the obtained solutions. For instance, sparse data (like in MS-generated datasets) show a minor effect of rotational ambiguities in the resolved elution profiles due to the high selectivity of MS signals. In addition, different data augmentation strategies have been proposed to minimize (and eventually eliminate) the effect of these rotational ambiguities. Finally, some methods have been recommended to quantify this rotational ambiguities impact [88,89].

During the ALS optimization, chemical, physical or mathematical information in the form of constraints are applied to reduce the ambiguity in the MCR bilinear solutions. Among the constraints imposed in MCR-ALS, the most commonly used in the case of analyzing LC × LC data is the non-negativity constraint, which is applied to both elution and spectral profiles in the respective factor matrices, $\mathbf{C}_{saug}$ and $\mathbf{S}^T$ (see Equation (4)) [9]. Due to the selectivity of the chromatographic separation and specially of the MS detection, the use of non-negativity constraints together with the simultaneous analysis of multiple chromatographic runs and samples with common constituents usually provide feasible solutions with very little or no ambiguity associated, giving unimodal elution profiles. Thus, other natural constraints such as unimodality are not needed when MS is employed, in contrast, to its common use in LC × LC-UV [90–93]. Alternatives to classical ALS optimization have also been considered. For instance, Cook et al. proposed an additional sparse regression step using an elastic net (MCR-ENALS) [94] to increase the algorithm's performance when analyzing MS data. This

approach takes advantage of the inherent sparseness of MS data to detect low intensity signals.

One major advantage of the application of MCR-ALS to 2D chromatographic systems is that it allows the resolution of the elution profiles of the constituents in the analyzed mixtures without the application of any peak alignment nor peak modeling pre-processing step. This is caused by the intrinsic flexibility provided by the bilinear modeling (see Equation (4) and Fig. 4 above) of the chromatographic peaks in both chromatographic dimensions (in $\mathbf{C}_{\mathbf{aug}}$ matrix of Equation (4)). Matching of elution times among chromatographic peaks of the same component in the different ^{1}D and ^{2}D columns is not required. Data alignment is only needed in the spectral (column) direction, in the columns (wavelengths or *m/z* values) of $\mathbf{S}^{\mathbf{T}}$. Hence, chromatographic time shifts within and between samples do not affect the MCR-ALS final results. Consequently, the spectral and chromatographic elutions of different components in both dimensions can be resolved independently.

As stated above, MCR-ALS resolved elution and spectra profiles can be used to get qualitative and quantitative information, respectively. From the resolved spectra (either MS, UV, fluorescence, etc), it is possible to retrieve qualitative information about the sample constituents and perform their identification. On the other side, from the peak areas or heights of the elution profiles, relative quantitative information of the different sample components can be obtained. From this relative quantitative information, pattern recognition and classification of the analyzed samples (differences and similarities) are possible, as it is shown in Fig. 5.

In the case of LC × LC analysis, from the resolved second-dimension elution profiles, it is also possible to perform the relative quantitation of the chemical constituents of the analyzed samples. Since the first-dimension elution profile (^{1}D peak) is split into the consecutive injections that are further separated into the ^{2}D column, it is possible to obtain the total peak contribution for a specific resolved component by summing all the peak areas of the same component in the second dimension. From the comparison of the peak areas of the components in the different samples, relative quantitative information can be obtained. Various integration strategies applied to LC × LC are described elsewhere [48,92]. Manual integration is recommended in the cases where signal to noise ratios are small, due to low concentrations of analytes in the samples, as residual background can have a significant effect on the areas when peak integrations are performed by an automatic summation.

It has also been proved that MCR-ALS can build good calibration models for LC analysis to quantify the concentrations of the different compounds present in the sample, if absolute reference concentrations are available or multiple standards are used [95]. When validation samples are employed, the quantitative results in the samples not used during the calibration process can also be calculated as well as their relative errors [46]. In the case of LC × LC, it is possible to calibrate and quantify employing the resolved ^{1}D chromatograms. Compared to other conventional calibration procedures, MCR-ALS calibrations can efficiently remove the background and interference contributions from the analyte signals, while resolving baseline contributions and compound coelutions separately. This has been demonstrated in previous works, like in GC × GC analysis of polyaromatic hydrocarbons in oil samples [96].

Recently, a new restriction called area correlation constraint has been proposed for second order quantitative calibration with MCR-ALS [97,98], to estimate the absolute concentrations of the constituents of a set of samples simultaneously analyzed by a chromatographic method in the presence of unknown interferences. This constraint is applied at each iteration step of the ALS process by regressing the peak areas or heights of the concentration profiles of the analytes in every analyzed sample against their known concentrations. In practical situations, the known and unknown samples are simultaneous analyzed, and this correlation constraint is applied to build the calibration curve with the known "calibration" samples, whereas the rest are used for validation and prediction of the unknown samples. This approach can also be employed using different calibration strategies, i.e., using external or internal standards and the standard addition method [41,95].

Besides, from the peak areas of the resolved elution profiles of the different sample constituents, other post-processing methods can be applied, from which it is possible to evaluate differences among samples and identification of potential markers, as it will be explained below in the following section.

6. Post-processing data analysis methods

Apart from quantitation, the table (matrix) of the peak areas from the resolved elution profiles of the different constituents of the analyzed samples (for instance obtained by MCR-ALS, or as by other feature detection methods), can be post-processed using several types of multivariate data analysis methods to gain additional information about the analyzed systems. As shown in Fig. 5, five main types of analysis can be performed with the output results obtained after application of MCR methods: exploratory, clustering, statistical, classification and machine learning-based analysis. A brief comment is given here on some examples of application of these multivariate analysis methods to extract additional information from the results of the analysis of LC × LC datasets, especially when different types of samples are analyzed simultaneously.

Among exploratory data analysis methods, the most common one is Principal Component Analysis (PCA) [99]. PCA is a non-supervised data analysis method that provides an overview of the number and nature data variance sources. In PCA, a reduced dimensional space (whose variables are known as principal components) explains the most relevant information about the major sources of data variance. On one side, sample scores plots allow the unsupervised visualization of the analyzed samples in the principal components vector space, where the different groups or trends in these samples can be distinguished in plots of reduced dimensions. On the other side, from the loadings plot, it is possible to detect the most relevant variables in the definition of each principal component, i.e., giving details of the nature of the major sources of data variance.

One example of the application of PCA in 2D-LC is, for instance, the analysis of the triacylglycerides (TAGs) composition of several oils [100,101]. The PCA scores plot enabled unsupervised differentiation among various types of oil samples. Several kinds of oils with dissimilar compositions were distinguished, whereas samples with similar TAG compositions were clustered together. In addition, from the loadings plot, it was possible to distinguish and identify the most significant TAG present in each sample type.

The PCA study of the selectivity of different 1D-LC systems allowed the classification and comparison of several types of LC columns [102,103]. Thus, another use of PCA, beyond the exploratory analysis of the samples, has been the determination of orthogonal column combinations for LC × LC, a key aspect to be taken into account in multidimensional chromatography. For instance, Græsbøll et al. developed a method to help in the selection of LC columns for its application in LC × LC. This work was based on the hydrophobic subtraction model and different approaches were tested to assess the orthogonality of the columns [104].

Clustering methods, either hierarchical or non-hierarchical, group objects and samples according to their similarities and

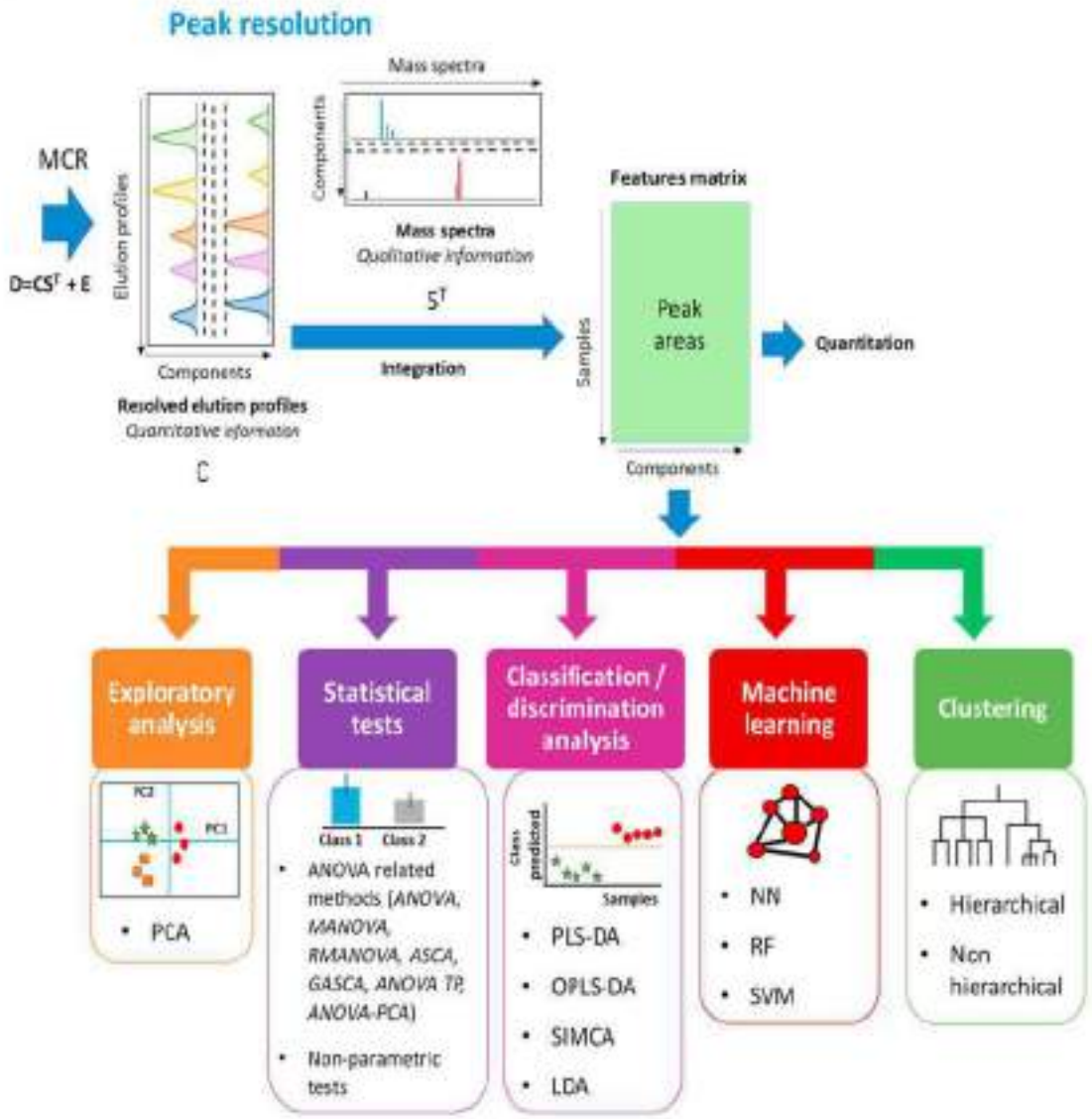


Fig. 5. Post-processing of the table of peak areas of the MCR resolved compounds. Five different types of analysis can be performed: exploratory, classification, statistical, machine learning-based and clustering.

differences within and between groups [105]. These methods are widely used in omics sciences, where some examples of applications in GC × GC [106], and in 2D-LC [101,107], have been already reported in combination with exploratory and classification methods. Multivariate peak area tables can be further analyzed using multiple statistical approaches. In the case of the simultaneous analysis of multiple variables per sample (i.e., when the tables of peak areas for several analytes and samples), different extensions of the ANOVA method have been proposed like the multivariate analysis of variance (MANOVA) [108], the ANOVA simultaneous component analysis (ASCA) [109], the group-wise ANOVA simultaneous component analysis (GASCA) [110], the ANOVA-PCA [111], the regularized MANOVA [112], or the ANOVA by target projection (ANOVA TP) [113]. In addition, linear mixed models have been generalized and adapted as an alternative to ANOVA based methods covering all possible experimental designs involving fixed factors [114,115]. For instance, in ASCA, which is based in the combination of ANOVA and Simultaneous Component Analysis (SCA), a method similar to PCA, individual factors as well as two or three-ways interactions between different factors can be analyzed, to determine whether they are significant or not [109]. ASCA has been applied to peak area tables obtained after MCR-ALS to statistically assess the significance of the different factors used in the design of the LC × LC-HRMS experiments [62].

Supervised classification analysis includes a group of methods that allow the class modeling and discrimination of samples in previously established groups before the analysis. The Soft Independent Modeling of Class Analogies (SIMCA) [116] method allows the assignment of a particular sample/observation to a well-defined class. This classification is performed through the size of the residuals of the disjoint PCA (a PCA model is generated for every class present in the dataset). In contrast, Partial Least Squares-Discriminant Analysis (PLS-DA) [117], has been proven to provide equal or even better performance in discrimination tasks (two-class type of problems). In PLS-DA, class distinction is established *a priori*, according to previous information known from the experimental design (DOE) about the different samples, which is then introduced in the model building. From the class assignment, contingency tables can be obtained where the quality of the model can be assessed, taking into account figures of merit

such as selectivity and sensitivity. In addition, PLS-DA results also allow identifying the features responsible for the differences between classes, for instance, through examining the changes in the size of the variables between classes, which in the case of MS data, can be easily associated with m/z values. The Variables Important in Projection (VIPs) method give the highest score values for the most significant variables (m/z values) explaining the differences between the classes. PLS-DA has also been employed to look for biomarkers through the VIPs, comparing healthy samples (or control samples) with samples from patients with a certain disease [118]. PLS-DA can also be useful in the analysis of 2D-LC data sets [62,101,107,119]. Other variants of PLS-DA methods are widely employed, such as the OPLS-DA (orthogonal PLS-DA) method [120], with classification models are often found easier to interpret.

Besides SIMCA and PLS-DA, other machine learning based methods can also be employed to explore and classify LC × LC data sets, although they have barely been explored until now. In this context, different variants of Neural Networks (NN), Random forests (RF) and Support vector machines (SVM) should be highlighted [121,122]. Examples of recent applications of machine learning methods in LC-MS can be found, especially in metabolomics [123–125], but also in method optimization, retention time predictions or in silico quantifications [126–128].

7. Examples of application of the MCR-ALS and ROIMCR methods to LC × LC data

In this section, some examples of the application of the MCR-ALS and the ROIMCR methods in the analysis of LC × LC data, taken from previous studies are succinctly described.

Several applications of MCR-ALS to LC × LC- UV–Vis (i.e., using DAD) data, have been described in detail in the works of Rutan et al., focused on two data analysis challenges, the evaluation of the structure of the LC × LC data, and the quantitative aspects of this type of analysis. For instance, Bailey et al. analyzed urine samples with the so-called IKFSA-ALS-ssel procedure [48]. Initial guesses for MCR-ALS analysis were obtained using the iterative key set factor analysis (IKSFA) method, in a similar way as in the application of other purest variable detection methods [105,129]. Non-negativity and selectivity constraints were also applied during the optimization. Relative concentrations of the sample constituents were estimated from the sum of the second-dimension peak areas that correspond to the same compound, after their resolution with the IKFSA/MCR-ALS procedure and manual baseline correction. Better integration results (lower % RSD) were obtained when this manual baseline correction of the MCR resolved elution peaks was applied, compared to the direct raw peak data integration using commercial software.

In another study about the LC × LC analysis of maize seedling samples, the PARAFAC-ALS and the IKSFA/MCR-ALS methods were compared. The authors concluded that retention time shifts were responsible for the worse results of PARAFAC-ALS (lower precision), whereas the IKSFA/MCR-ALS method gave better results and was recommended. Also, the IKFSA/MCR-ALS-ssel method coupled with a manual baseline subtraction quantification method was used in the analysis of wine samples using LC × LC-DAD [130]. In this work, different chemometric strategies (such as the Fisher Ratio or the Similarity Index methods) were evaluated as rapid screening methods in the LC × LC analysis of complex samples. In all these strategies, the focus was especially on determining of those sample constituents whose chromatographic peak areas (concentrations) differed among the investigated samples, rather than on the determination of all the sample constituents detected in the chromatograms. Thus, the time of the analysis was drastically reduced.

MCR-ALS has also been proven to be a useful tool combined with 2D assisted liquid chromatography (2DALC), where two DAD detectors were placed after each column. This set up allowed obtaining simultaneously ^{1}D and 2D chromatograms and improved the quantitation of targeted analytes [90]. 2DALC, in combination with the MCR-ALS method, presented some advantages over the traditional LC × LC-DAD method, since it could perform a proper background correction due to dilution in LC × LC, as well as avoiding, at least partially, the loss of resolution caused by the undersampling of the ^{1}D elution profile. This strategy takes advantage of the pure spectra of the components resolved in the ^{2}D by MCR-ALS, which can be used then for the MCR-ALS analysis of the ^{1}D chromatographic data. This method allowed the calibration and quantification of the different sample constituents resolved in the ^{1}D chromatograms.

Other strategies combining MCR-ALS and LC × LC data (including 2DALC, and its adapted algorithm, called combined-2DALC) have been proposed for the quantitative analysis of furanocoumarins in apiaceous vegetable samples [92], which were separated using an RP × RP-DAD chromatographic method (both column dimensions employed reverse phase as the retention mechanism, a pentafluorophenyl phase (PFP) and a C18 phase as ^{1}D and ^{2}D columns, respectively). In this example, initial estimates of the spectral profiles were obtained employing the iterative orthogonal projection approach (IOPA), in which the number of spectra extracted was preselected by the user from his previous knowledge of the sample complexity. Non-negativity and local rank constraints were applied in the chromatographic and spectral directions. Local rank constraints were applied in this case to force certain spectral regions of the resolved components to zero-intensity (i.e., intensities associated with wavelengths out of the range of absorption of the target analytes). Two integration strategies were also compared, one based on a manual selection of peak boundaries (including background subtraction) and another based on an automatic summation. The first strategy was especially recommended for samples at low concentrations of the analytes showing low signal-to-background ratios.

Other recent applications of MCR-ALS to LC × LC-MS have also been reported. For instance, mixtures of highly similar TAG structural isomers present in vegetable oil samples were separated with an IEX × RP-MS (ion exchange as ^{1}D and reverse phase as ^{2}D) and completely resolved by MCR-ALS, as shown in Fig. 6 [54]. Navarro-Reig et al. performed the comparison between MCR-ALS bilinear, MCR-ALS trilinear, MCR-ALS trilinear allowing time shifting, PARAFAC and PARAFAC2 methods [54]. It was demonstrated that the LC × LC data did not fulfill the trilinearity model requirements due to retention time shifts and to changes in the shapes of the elution peak profiles. Therefore, the MCR-ALS bilinear method was the most suitable choice for the analysis of this type of data. Non-negativity constrains were applied to both elution and spectra profiles, as well as spectral normalization (either of peak heights or peak areas). Regions where triacylglycerides (TAGs) isomers co-eluted in both chromatographic dimensions were specifically studied to check the limits of the resolution potential of the proposed method. In this work, MCR-ALS was also confirmed to resolve and identify separately the different positional and chain TAG isomers despite their very strong coelution (with embedded peaks) in both dimensions.

Next, some other examples applying the combination of the ROI method with the MCR-ALS method in the recently proposed ROIMCR method to LC × LC data are briefly described. Examples of the application of different post-processing multivariate analysis of the peak areas or heights of the MCR-ALS resolved elution profiles of the analyzed sample constituents are given.

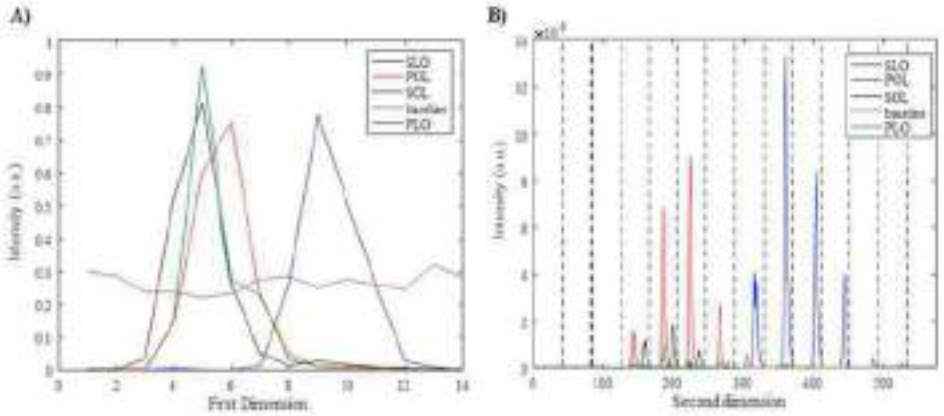


Fig. 6. ^{1}D and ^{2}D column elution profiles of a chromatographic region with highly overlapped triacylglycerol structural isomers (TAGs) resolved by MCR-ALS "Reprinted from Talanta, 160, M. Navarro-Reig et al., Chemometric analysis of comprehensive LC × LC-MS data: Resolution of triacylglycerol structural isomers in corn oil, 624–635, Copyright (2016), with permission from Elsevier."

In a recent publication, the ROIMCR method was applied to LC × LC-MS, LC × LC-UV and LC × LC-UV-MS data from the analysis of a mixture of 31 pharmaceutical compounds [49] using a combined RP × HILIC (an RP, reverse phase, column as ^{1}D, and a HILIC, hydrophilic interaction column, as ^{2}D). In this work, the selected ROI parameters employed for the analysis of the LC × LC-MS data were a signal threshold value of 0.1% of the maximum signal intensity, mass accuracy of 0.5 Da (a low resolution MS detector was employed), and a minimum number of *m/z* occurrences of 25. Non-negativity constraints were applied in the chromatographic direction (elution profiles of the two columns) and, also, in the spectral direction (mass spectra). Three different approaches assessed the trilinear behavior of the data. The first includes the comparison of the singular values of the data augmented matrices obtained by concatenation of consecutive ^{2}D modulations in their different directions (vertically, horizontally and slice by slice). The second implied the calculation of the core consistency in the PARAFAC decomposition of the data three-way array. The third compared the data fitting results obtained assuming either bilinear or trilinear decomposition models in MCR-ALS. Significant differences in the SVD decomposition of the augmented data matrices in their different directions or modes, poor trilinear core consistency values, and large data fitting deviations between bilinear and trilinear models were found for both types of data (MS or DAD). Consequently, the MCR-ALS bilinear model approach was preferred and recommended for the analysis of this LC × LC data set. In addition, the combined analysis of the data from the two detectors (DAD and MS) building up a "fused multidetector" multiset data structure, allowed resolving more components than from individual detector analysis.

ROIMCR has also been used in other LC × LC-MS metabolomics studies, such as in the untargeted analysis of the metabolites present in a particular sample. For instance, the changes in rice metabolism caused by environmental factors such as watering and harvesting time, were studied by LC × LC-MS [62,119] using a combination of HILIC × RP-HRMS separation methods in the analysis of the polar metabolites of the rice samples, using the hydrophilic interaction column as ^{1}D and the reverse phase column as ^{2}D. The strategy proposed in this work included two compression steps: a first spectral compression using the ROI strategy, and a second compression in the time direction, using wavelets. An additional windowing approach was used to divide the chromatograms into three different chromatographic windows. In this way, the size of the analyzed data sets and the global time of analysis were feasible and significantly reduced. Possible instrumental drifts were corrected using an internal normalization based on dividing all the metabolite peak areas by the peak area of a standard compound added to each sample. ASCA investigated the effects of the watering and the harvesting time factors. Peak areas of all MCR-ALS resolved metabolites were arranged in a data table/matrix, and the effects of both factors were evaluated statistically, as well as the interaction between them. In addition, the application of PLS-DA allowed discerning what metabolites were able to separate the effects of watered and non-watered samples through the calculation of VIPs as shown in Fig. 7. MCR-ALS was further applied in a post-processing step to study the evolution of the metabolic concentration profiles (peak areas) over time, and to investigate in this way the effects of the harvesting time factor, for both water and non-watered samples. In this extensive study, the proposed ROIMCR chemometrics strategy allowed the simultaneous untargeted direct resolution of 150 metabolites from 15 different families, with 134 of them identified by their HRMS spectra.

In another recent study, RP × HILIC-MS and MS/MS untargeted throughout analysis of the lipids present in rice samples under arsenic (As) exposure were presented. In this case, only the spectral compression was necessary to reduce the LC × LC-MS data size, which was also analyzed using the ROIMCR strategy. ROI parameters employed were similar to those mentioned before for the analysis of pharmaceuticals explained above, since the same settings of the instrumental equipment were used. Arsenic effects were investigated on two plant tissues: aerial parts and roots. MCR-ALS was able to resolve the elution profiles and mass spectra of the lipids present in the rice samples and allowed their identification. Complementary, ASCA analysis revealed that As exposure affected the rice lipidome significantly. PCA analysis of the changes on the peak areas of the MCR-ALS resolved elution profiles of the lipids detected in positive ionization mode (see Fig. 8) showed that the rice samples were well separated at the two As concentrations levels in aerial rice samples. In contrast, the separation from control and the lowest concentration level was not obvious in root rice samples. In the lipidomic analysis using negative ionization mode, aerial and root rice samples exposed at the two As concentrations were not discriminated from control rice samples. PLS-DA was applied to identify the potential lipid markers of As exposure effects at two concentration levels. Identification of these markers was a challenge, especially due to the low sensitivity of the detector and to the lack of theoretical MS/MS lipid spectra databases. Overall, this study confirmed that LC × LC combined with the ROIMCR data analysis is an excellent tool in omics studies. Other applications can

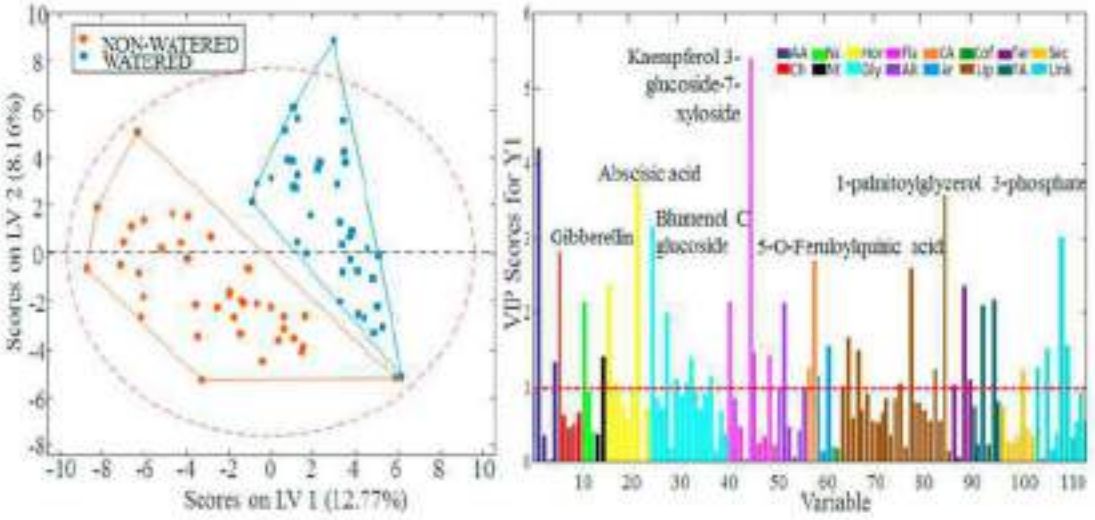


Fig. 7. PLS-DA scores and loadings plots for watered and non-watered rice samples analyzed by LC × LC-HRMS. "Reprinted from Analytical Chemistry, 89, M. Navarro-Reig et al., Untargeted Comprehensive Two-Dimensional Liquid Chromatography Coupled with High-Resolution Mass Spectrometry Analysis of Rice Metabolome Using Multivariate Curve Resolution, 7675–7683, Copyright (2017), with permission from American Chemical Society."

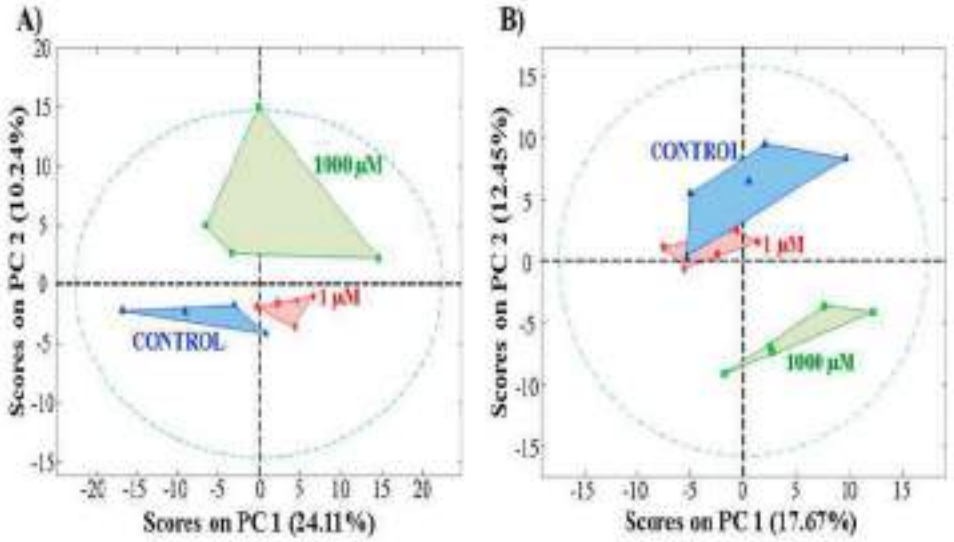


Fig. 8. Scores plot obtained for the lipids analyzed in aerial parts and roots, respectively, of rice samples treated with As. "Reprinted with permission from Journal of Chromatography A, 1568, M. Navarro-Reig et al., An untargeted lipidomic strategy combining comprehensive two-dimensional liquid chromatography and chemometric analysis, 80–90, Copyright (2018) Elsevier".

be envisioned, like in untargeted analysis of environmental monitoring samples, where huge data sets are also generated.

8. Conclusions

Although one of the main drawbacks of LC × LC coupled to multichannel detection (mainly MS and DAD) is the difficulty of performing a thorough analysis of the large data sets generated, recent chemometric advances provide a good option to untangle the complexity of this type of data and facilitate the extraction of the sought information in different types of studies such as in metabonomics. Since data from LC and LC × LC analysis usually present significative deviations from the ideal trilinear behavior, bilinear model-based methods are preferred in general for their analysis.

Among the recently proposed chemometric strategies, the ROIMCR method is a powerful tool to analyze LC × LC data, thanks to the combination of compression and resolution steps in both targeted and untargeted analysis. The possibility of filtering, compressing and arranging large datasets without loss of spectral accuracy is the main achievement of the ROI approach. With the application of this strategy, resolution of the LC × LC elution profiles of all the sample constituents of the analyzed samples in the two dimensions as well as their mass spectra can be achieved, providing simultaneously quantitative and qualitative (identification) information. The application of MCR-ALS to the ROI data adds further advantages such as the flexibility in modeling the complex structure of the data (bilinear or trilinear) which allows the proper analysis of the different types of datasets encountered in practice. MCR-ALS does not require the modeling or alignment of the chromatographic peak signals resolved in different multiple chromatographic runs in their simultaneous analysis. MCR-ALS can also be used as a preliminary pre-processing step in the application of other multivariate data analysis methods. Thus, in conclusion, the

ROIMCR method is confirmed to be a promising approach for integrating pre-processing and exhaustive data analysis of complex LC × LC-HRMS datasets.

CRediT author statement

Miriam Pérez-Cova: Conceptualization, Writing-Original Draft preparation, Writing-Reviewing and Editing. **Joaquim Jaumot**: Conceptualization, Writing-Reviewing, and Editing, Funding Acquisition. **Romà Tauler**: Conceptualization, Writing-Reviewing and Editing, Funding Acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The research leading to these results has received funding from the Spanish Ministry of Science and Innovation (MCI, Grants CTQ2017-82598-P and PID2019-105732GB-C21). The authors also want to grant support from the Catalan Agency for Management of University and Research Grants (AGAUR, Grant 2017SGR753) and the Spanish MCI (Severo Ochoa Project CEX2018-000794-S). MPC acknowledges a predoctoral FPU 16/02640 scholarship from the Spanish Ministry of Education and Vocational Training (MEFP).

References

[1] E.W.J. Pirok, D.R. Stoll, P.J. Schoenmakers, Recent developments in two-dimensional liquid chromatography – Fundamental improvements for practical applications, Anal. Chem. 91 (2018). https://doi.org/10.1021/acs.analchem.8b04841. acs.analchem.8b04841.

[2] T.S. Bos, W.C. Knol, S.R.A. Molenaar, L.E. Niezen, P.J. Schoenmakers, G.W. Somsen, B.W.J. Pirok, Recent applications of chemometrics in one- and two-dimensional chromatography, J. Sep. Sci. (2020) 1–50. https://doi.org/10.1002/jssc.202000011.

[3] M. Navarro-Reig, J. Jaumot, A. Baglai, C. Vivó-Truyols, P.J. Schoenmakers, R. Tauler, Untargeted comprehensive two-dimensional liquid chromatography coupled with high-resolution mass spectrometry analysis of rice Metabolome using multivariate curve resolution, Anal. Chem. (2017). https://doi.org/10.1021/acs.analchem.7b01648.

[4] M. Navarro-Reig, C. Bedia, R. Tauler, J. Jaumot, Chemometric strategies for peak detection and profiling from multidimensional chromatography, Proteomics 18 (2018) 1–12. https://doi.org/10.1002/pmic.201700327.

[5] M. Navarro-Reig, J. Jaumot, T.A. van Beek, C. Vivó-Truyols, R. Tauler, Chemometric analysis of comprehensive LC×LC-MS data: resolution of triacylglycerol structural isomers in corn oil, Talanta 160 (2016) 624–635. https://doi.org/10.1016/j.talanta.2016.08.005.

[6] B.R. Kowalski, Chemometrics, Anal. Lett. 11 (1978) 11–13. https://doi.org/10.1080/00032717808069728.

[7] H.T.L. dos Santos A.M. de Oliveira, P.G. de Melo, W. Freitas, A.P.R. de Freitas, Chemometrics: theory and application, in: Multivar. Anal. Manag. Eng. Sci., 2013. https://doi.org/10.1772/53166.

[8] E. Gorrochategui J. Jaumot, R. Tauler, ROIMCR: a powerful analysis strategy for LC-MS metabolomic datasets, BMC Bioinf. 20 (2019) 1–17. https://doi.org/10.1186/s12859-019-2848-8.

[9] A. De Juan, J. Jaumot, R. Tauler, Multivariate Curve Resolution (MCR). Solving the mixture analysis problem, Anal. Methods. 6 (2014) 4964–4976. https://doi.org/10.1039/c4ay00571f.

[10] E.M. Kalili, A. De Villiers, Systematic optimisation and evaluation of on-line, off-line and stop-flow comprehensive hydrophilic interaction chromatography × reversed phase liquid chromatographic analysis of procyanidins, Part I: theoretical considerations, J. Chromatogr. A 1289 (2013) 58–68. https://doi.org/10.1016/j.chroma.2013.03.008.

[11] D.R. Stoll, P.W. Carr, Two-dimensional liquid chromatography: a state of the art tutorial, Anal. Chem. (2017). https://doi.org/10.1021/acs.analchem.6b03506.

[12] R.E. Murphy, M.R. Schure, J.P. Foley, Effect of sampling rate on resolution in comprehensive two-dimensional liquid chromatography, Anal. Chem. 70 (1998) 1585–1594. https://doi.org/10.1021/ac971184b.

[13] M. Muller, A.G.J. Tredoux, A. de Villiers, Predictive kinetic optimisation of hydrophilic interaction chromatography × reversed phase liquid chromatography separations: experimental verification and application to phenolic analysis, J. Chromatogr. A 1571 (2018) 107–120. https://doi.org/10.1016/j.chroma.2018.08.004.

[14] J. Woiwode, R.J. Reischl, S. Buckenmaier, W. Lindner, M. Lämmerhofer, Imaging peptide and protein chirality via amino acid analysis by chiral × chiral two-dimensional correlation liquid chromatography, Anal. Chem. 90 (2018) 7963–7971. https://doi.org/10.1021/acs.analchem.8b00676.

[15] M.H. Blokland, P.W. Zoontjes, L.A. Van Ginkel, M.G.M. Van De Schans, S.S. Sterk, T.F.H. Bovee, Multiclass screening in urine by comprehensive two-dimensional liquid chromatography time of flight mass spectrometry for residues of sulphonamides, beta-agonists and steroids, Food Addit. Contam. Part A Chem. Anal. Control. Expo. Risk Assess. 35 (2018) 1703–1715. https://doi.org/10.1080/19440049.2018.1505168.

[16] M. Xu, J. Legradi, P. Leonards, Evaluation of LC-MS and LC×LC-MS in analysis of zebrafish embryo samples for comprehenive lipid profiling, Anal. Bioanal. Chem. 412 (2020) 4313–4325. https://doi.org/10.1007/s00216-020-02661-1.

[17] F. Cacciola, D. Mangraviti, F. Rigano, P. Donato, P. Dugo, L. Mondello, H.J. Cortes, Novel comprehensive multidimensional liquid chromatography approach for elucidation of the microbiosphere of shikimate-producing Escherichia coli SP1.1/pKD15.071 strain, Anal. Bioanal. Chem. 410 (2018) 3473–3482. https://doi.org/10.1007/s00216-017-0744-1.

[18] L. Montero, V. Sáez, D. von Baer, A. Cifuentes, M. Herrero, Profiling of Vitis vinifera L. canes (polyphenolic compounds using comprehensive two-dimensional liquid chromatography, J. Chromatogr. A 1536 (2018) 205–215. https://doi.org/10.1016/j.chroma.2017.06.013.

[19] R. Kirongo, T. Begani, D.R. Stoll, M. Lämmerhofer, A selective comprehensive reversed-phase×reversed-phase 2D-liquid chromatography approach with multiple complementary detectors as advanced generic method for the quality control of synthetic and therapeutic peptides, J. Chromatogr. A 1627 (2020) 461480. https://doi.org/10.1016/j.chroma.2020.461480.

[20] P. Venter, M. Muller, J. Vestner, M.A. Stander, A.G.J. Tredoux, H. Pasch, A. De Villiers, Comprehensive three-dimensional LC × LC × ion mobility spectrometry separation combined with high-resolution MS for the analysis of complex samples, Anal. Chem. 90 (2018) 11643–11650. https://doi.org/10.1021/acs.analchem.8b03234.

[21] S. Stephan, J. Hippler, T. Köhler, D. Brecht, O.J. Schmitz, A powerful four-dimensional separation method for complex samples, J. Anal. Test. 1 (2017) 1–9. https://doi.org/10.1007/s41664-017-0004-x.

[22] H. Zhang, J.M. Jiang, D. Zheng, M. Yuan, Z.Y. Wang, H.M. Zhang, C.W. Zheng, L.B. Xiao, H.X. Xu, A multidimensional analytical approach based on time-decoupled online comprehensive two-dimensional liquid chromatography coupled with ion mobility quadrupole time-of-flight mass spectrometry for the analysis of ginsenosides from white and red ginsengs, J. Pharm. Biomed. Anal. 163 (2019) 24–33. https://doi.org/10.1016/j.jpba.2018.09.036.

[23] A. Moussa, T. Lauer, D. Stoll, G. Desmet, K. Broeckhoven, Numerical and experimental investigation of analyte breakthrough from sampling loops used for multi-dimensional liquid chromatography, J. Chromatogr. A 1626 (2020) 461283. https://doi.org/10.1016/j.chroma.2020.461283.

[24] H.C. van de Ven, J. Rumova, G. Groeneveld, T.S. Bos, A.F.G. Gargano, S. van der Wal, Y. Mengerink, P.J. Schoenmakers, Living with breakthrough: two-dimensional liquid-chromatography separations of a water-soluble synthetically grafted bio-polymer, Separations 7 (2020) 1–16. https://doi.org/10.3390/separations7030041.

[25] D.R. Stoll, K. Shoykhet, P. Petersson, S. Buckenmaier, Active solvent modulation: a valve-based approach to improve separation compatibility in two-dimensional liquid chromatography, Anal. Chem. 89 (2017) 9058–9067. https://doi.org/10.1021/acs.analchem.7b02046.

[26] R.J. Vonk, A.F.G. Gargano, E. Davydova, H.L. Dekker, S. Eeltink, L.J. De Koning, P.J. Schoenmakers, Comprehensive two-dimensional liquid chromatography with stationary-phase-assisted modulation coupled to high-resolution mass spectrometry applied to proteome analysis of saccharomyces cerevisiae, Anal. Chem. 87 (2015) 5387–5394. https://doi.org/10.1021/acs.analchem.5b00308.

[27] H. Tian, J. Xu, Y. Xu, Y. Guan, Multidimensional liquid chromatography system with an innovative solvent evaporation interface, J. Chromatogr. A 1137 (2006) 42–48. https://doi.org/10.1016/j.chroma.2006.10.005.

[28] Y. Chen, J. Li, O.J. Schmitz, Development of an at-column dilution modulator for flexible and precise control of dilution factors to overcome mobile phase incompatibility in comprehensive two-dimensional liquid chromatography, Anal. Chem. 91 (2019) 10251–10257. https://doi.org/10.1021/acs.analchem.9b02291.

[29] S.L. Weatherbee T. Brau D.R. Stoll, S.C. Rutan, M.M. Collinson, Simulation of elution profiles in liquid chromatography – IV: experimental characterization and modeling of solute injection profiles from a modulator valve used in two-dimensional liquid chromatography, J. Chromatogr. A 1626 (2020) 1–10. https://doi.org/10.1016/j.chroma.2020.461371.

[30] H. Wang, H.R. Ihatka, R. Bennett, M. Potapenko, C.J. Pickens, B.E. Mann, I.A. Haidar Ahmad, E.L. Regalado, Introducing online multicolumn two-dimensional liquid chromatography screening for facile selection of stationary and mobile phase conditions in both dimensions, J. Chromatogr. A 1622 (2020) 461091. https://doi.org/10.1016/j.chroma.2020.461091.

[31] C.J. Pickens, I.A. Haidar Ahmad, A.A. Makarov, R. Bennett, B.E. Mann, E.L. Regalado, Comprehensive online multicolumn two-dimensional liquid chromatography-diode array detection-mass spectrometry workflow as a framework for chromatographic screening and analysis of new drug

substances, Anal. Bioanal. Chem. 412 (2020) 2655–2663. https://doi.org/10.1007/s00216-020-02498-8.
[32] A.A. Aly, M. Muller, A. de Villiers, B.W.J. Pirok, T. Górecki, Parallel gradients in comprehensive multidimensional liquid chromatography enhance utilization of the separation space and the degree of orthogonality when the separation mechanisms are correlated, J. Chromatogr. A 1628 (2020) 461452. https://doi.org/10.1016/j.chroma.2020.461452.
[33] A. Niazi, R. Leardi, Genetic algorithms in chemometrics, J. Chemom. 26 (2012) 345–351. https://doi.org/10.1002/cem.2426.
[34] T. Alvarez-Segura, S. López-Ureña, J.R. Torres-Lapasió, M.C. García-Alvarez-Coque, Multi-scale optimisation vs. genetic algorithms in the gradient separation of diuretics by reversed-phase liquid chromatography, J. Chromatogr. A 1609 (2020). https://doi.org/10.1016/j.chroma.2019.460427.
[35] B. Huygens, K. Efthymiadis, A. Nowé, G. Desmet, Application of evolutionary algorithms to optimise one- and two-dimensional gradient chromatographic separations, J. Chromatogr. A 1628 (2020) 461435. https://doi.org/10.1016/j.chroma.2020.461435.
[36] B.W.J. Pirok, A.F.G. Gargano, P.J. Schoenmakers, Optimizing separations in online comprehensive two-dimensional liquid chromatography, J. Sep. Sci. 41 (2018) 68–98. https://doi.org/10.1002/jssc.201700863.
[37] M.J. den Uijl, P.J. Schoenmakers, B.W.J. Pirok, M.R. Bonvin, Recent applications of retention modelling in liquid chromatography, J. Sep. Sci. (2020) 1–27. https://doi.org/10.1002/jssc.202000905.
[38] R. Tauler, Multivariate curve resolution of multiway data using the multilinearity constraint, J. Chemom. (2020) 1–24. https://doi.org/10.1002/cem.3279.
[39] R. Bro, PARAFAC. Tutorial and applications, Chemometr. Intell. Lab. Syst. 38 (1997) 149–171. https://doi.org/10.1016/S0169-7439(97)00032-4.
[40] J.D. Carroll, J.J. Chang, Analysis of individual differences in multidimensional scaling via an n-way generalization of 'Eckart-Young' decomposition, Psychometrika 35 (1970) 283–319. https://doi.org/10.1007/BF02310791.
[41] A.C. Olivieri, G.M. Escandar, A.M. de la Peña, Second-order and higher-order multivariate calibration methods applied to non-multilinear data using different algorithms, TrAC - Trends Anal. Chem. 30 (2011) 607–617. https://doi.org/10.1016/j.trac.2010.11.018.
[42] S.A. Bortolato, J.A. Arancibia, G.M. Escandar, A.C. Olivieri, Time-alignment of bidimensional chromatograms in the presence of uncalibrated interferences using parallel factor analysis. Application to multi-component determinations using liquid-chromatography with spectrofluorimetric detection, Chemometr. Intell. Lab. Syst. 101 (2010) 30–37. https://doi.org/10.1016/j.chemolab.2009.12.001.
[43] K.M. Pierce, B. Kehimkar, L.C. Marney, J.C. Hoggard, R.E. Synovec, Review of chemometric analysis techniques for comprehensive two dimensional separations data, J. Chromatogr. A (2012). https://doi.org/10.1016/j.chroma.2012.05.050.
[44] S.E. Prebihalo, D.K. Pinkerton, R.E. Synovec, Impact of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry experimental design on data trilinearity and parallel factor analysis deconvolution, J. Chromatogr. A 1605 (2019) 460368. https://doi.org/10.1016/j.chroma.2019.460368.
[45] R.E. Mohler, K.M. Dombek, J.C. Hoggard, E.T. Young, R.E. Synovec, Comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry analysis of metabolites in fermenting and respiring yeast cells, Anal. Chem. 78 (2006) 2700–2709. https://doi.org/10.1021/ac052100o.
[46] H. Parastar, J.R. Radović, J.M. Bayona, R. Tauler, Solving chromatographic challenges in comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry using multivariate curve resolution-alternating least squares ABC Highlights: authored by Rising Stars and Top Experts, Anal. Bioanal. Chem. 405 (2013) 6235–6249. https://doi.org/10.1007/s00216-013-7067-y.
[47] Y. Izadmanesh, E. Garreta-Lara, J.B. Ghasemi, S. Lacorte, V. Matamoros, R. Tauler, Chemometric analysis of comprehensive two dimensional gas chromatography–mass spectrometry metabolomics data, J. Chromatogr. A 1488 (2017) 113–125. https://doi.org/10.1016/j.chroma.2017.01.052.
[48] H.P. Bailey, S.E. Ratan, Chemometric resolution and quantification of four-way data arising from comprehensive 2D-LC-DAD analysis of human urine, Chemometr. Intell. Lab. Syst. 106 (2011) 131–141. https://doi.org/10.1016/j.chemolab.2010.07.004.
[49] M. Pérez-Cova, R. Tauler, J. Jaumot, Chemometrics in comprehensive two-dimensional liquid chromatography: a study of the data structure and its multilinear behavior, Chemometr. Intell. Lab. Syst. 201 (2020). https://doi.org/10.1016/j.chemolab.2020.104009.
[50] B.Q. Li, J. Chen, J.J. Li, X. Wang, H.L. Zhai, The application of a Tchebichef moment method to the quantitative analysis of multiple compounds based on three-dimensional HPLC fingerprint spectra, Analyst 140 (2015) 630–636. https://doi.org/10.1039/c4an01736f.
[51] M.B. Amardi, J.A. Arancibia, A.C. Olivieri, Interpretation of matrix chromatographic-spectral data modeling with parallel factor analysis 2 and multivariate curve resolution, J. Chromatogr. A 1604 (2019). https://doi.org/10.1016/j.chroma.2019.460502.
[52] S.A. Bortolato, A.C. Olivieri, Chemometric processing of second-order liquid chromatographic data with UV-vis and fluorescence detection. A comparison of multivariate curve resolution and parallel factor analysis 2, Anal. Chim. Acta 842 (2014) 11–19. https://doi.org/10.1016/j.aca.2014.07.007.
[53] G.M. Escandar, A.C. Olivieri, A road map for multi-way calibration models, Analyst 142 (2017) 2862–2873. https://doi.org/10.1039/c7an00822h.
[54] M. Navarro-Reig, J. Jaumot, T.A. van Beek, G. Vivó-Truyols, R. Tauler, Chemometric analysis of comprehensive LC×LC-MS data: resolution of triacylglycerol structural isomers in corn oil, Talanta (2016). https://doi.org/10.1016/j.talanta.2016.08.005.
[55] R. Tauler, I. Marqués, E. Casassas, Multivariate curve resolution applied to three-way trilinear data: study of a spectrofluorimetric acid-base titration of salicylic acid at three excitation wavelengths, J. Chemom. 12 (1998) 55–75. https://doi.org/10.1002/(sici)1099-128x(199801/02)12:1<55::aid-cem501>3.0.co;2-323.
[56] R. Bro, H.A.L. Kiers, A new efficient method for determining the number of components in PARAFAC models, J. Chemom. 17 (2003) 274–286. https://doi.org/10.1002/cem.801.
[57] A. Malik, R. Tauler, Performance and validation of MCR-ALS with quadrilinear constraint in the analysis of noisy datasets, Chemometr. Intell. Lab. Syst. 135 (2014) 223–234. https://doi.org/10.1016/j.chemolab.2014.04.002.
[58] R. Tautenhahn, C. Bottcher, S. Neumann, Highly sensitive feature detection for high resolution LC/MS, BMC Bioinf. 9 (2008) 1–16. https://doi.org/10.1186/1471-2105-9-504.
[59] A.W. Dowsey, J.A. English, F. Lisacek, J.S. Morris, G.Z. Yang, M.J. Dunn, Image analysis tools and emerging algorithms for expression proteomics, Proteomics 10 (2010) 4226–4257. https://doi.org/10.1002/pmic.200900835.
[60] C.R. Mittermayr, S.G. Nikolov, H. Hutter, M. Grasserbauer, Wavelet denoising of Gaussian peaks: a comparative study, Chemometr. Intell. Lab. Syst. 34 (1996) 187–202. https://doi.org/10.1016/0169-7439(96)00026-3.
[61] M. Daszykowski, B. Walczak, Use and abuse of chemometrics in chromatography, TrAC - Trends Anal. Chem. 25 (2006) 1081–1096. https://doi.org/10.1016/j.trac.2006.09.003.
[62] M. Navarro-Reig, J. Jaumot, A. Baglai, G. Vivó-Truyols, P.J. Schoenmakers, R. Tauler, Untargeted comprehensive two-dimensional liquid chromatography coupled with high-resolution mass spectrometry analysis of rice Metabolome using multivariate curve resolution, Anal. Chem. 89 (2017) 7675–7683. https://doi.org/10.1021/acs.analchem.7b01648.
[63] M.M. Sinanian, D.W. Cook, S.C. Rutan, D.S. Wijesinghe, Multivariate curve resolution-alternating least squares analysis of high-resolution liquid chromatography-mass spectrometry data, Anal. Chem. 88 (2016) 11092–11099. https://doi.org/10.1021/acs.analchem.6b03130.
[64] E. Gorrochategui, J. Jaumot, R. Tauler, ROMICS: a powerful analysis strategy for LC-MS metabolomic datasets, BMC Bioinf. 20 (2019) 1–17. https://doi.org/10.1186/s12859-019-2848-8.
[65] C.A. Smith, E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification, Anal. Chem. 78 (2006) 779–787. https://doi.org/10.1021/ac051437y.
[66] R. Tauler, E. Gorrochategui, J. Jaumot, R. Tauler, A protocol for LC-MS metabolomic data processing using chemometric tools, Protoc. Exch. (2015). https://doi.org/10.1038/protex.2015.102.
[67] S.E. Reichenbach, M. Ni, D. Zhang, E.B. Ledford, Image background removal in comprehensive two-dimensional gas chromatography, J. Chromatogr. A 985 (2003) 47–56. https://doi.org/10.1016/S0021-9673(02)01498-X.
[68] Z. Da Zeng, S.T. Chin, H.M. Hegel, P.J. Marriott, Simultaneous deconvolution and re-construction of primary and secondary overlapping peak clusters in comprehensive two-dimensional gas chromatography, J. Chromatogr. A 1218 (2011) 2301–2310. https://doi.org/10.1016/j.chroma.2011.02.025.
[69] Y. Zhang, H.L. Wu, A.L. Xia, L.H. Hu, H.F. Zou, R.Q. Yu, Trilinear decomposition method applied to removal of three-dimensional background drift in comprehensive two-dimensional separation data, J. Chromatogr. A 1167 (2007) 178–183. https://doi.org/10.1016/j.chroma.2007.08.055.
[70] Z.P. Chen, H.L. Wu, J.H. Jiang, Y. Li, R.Q. Yu, A novel trilinear decomposition algorithm for second-order linear calibration, Chemometr. Intell. Lab. Syst. 52 (2000) 75–86. https://doi.org/10.1016/S0169-7439(00)00081-2.
[71] A.L. Xia, H.L. Wu, D.N. Fang, Y.J. Ding, L.Q. Hu, R.Q. Yu, Alternating penalty trilinear decomposition algorithm for second-order calibration with application to interference-free analysis of excitation-emission matrix fluorescence data, J. Chemom. 19 (2005) 65–76. https://doi.org/10.1002/cem.911.
[72] J.M. Amigo, T. Skov, R. Bro, ChroMATHography: solving chromatographic issues with mathematical models and intuitive graphics, Chem. Rev. 110 (2010) 4582–4605. https://doi.org/10.1021/cr900394n.
[73] J.T.V. Matos, R.M.B.O. Duarte, A.C. Duarte, Trends in data processing of comprehensive two-dimensional chromatography: state of the art, J. Chromatogr. B Anal. Technol. Biomed. Life Sci. 910 (2012) 31–45. https://doi.org/10.1016/j.jchromb.2012.06.039.
[74] C.G. Fraga, B.J. Prazen, R.E. Synovec, Objective data alignment and chemometric analysis of comprehensive two-dimensional separations with run-to-run peak shifting on both dimensions, Anal. Chem. 73 (2001) 5833–5840. https://doi.org/10.1021/ad010656q.
[75] H. Parastar, M. Jalali-Heravi, R. Tauler, Comprehensive two-dimensional gas chromatography (GC×GC) retention time shift correction and modeling using bilinear peak alignment, correlation optimized shifting and multivariate curve resolution, Chemometr. Intell. Lab. Syst. 117 (2012) 80–91. https://doi.org/10.1016/j.chemolab.2012.02.003.
[76] D. Zhang, X. Huang, F.E. Regnier, M. Zhang, Two-dimensional correlation optimized warping algorithm for aligning GCxGC-MS data, Anal. Chem. 80 (2008) 2664–2671. https://doi.org/10.1021/ac702017.

[77] J. Vial, H. Noçairi, P. Sassiat, S. Mallipatu, C. Cognon, D. Thiébaut, B. Teillet, D.N. Rutledge, Combination of dynamic time warping and multivariate analysis for the comparison of comprehensive two-dimensional gas chromatograms. Application to plant extracts, J. Chromatogr. A 1216 (2009) 2866–2872. https://doi.org/10.1016/j.chroma.2008.09.027.

[78] S. Parbo, A.B. Hansen, T. Skov, J.H. Christensen, Pixel-based analysis of comprehensive two-dimensional gas chromatograms (color plots) of petroleum: a tutorial, Anal. Chem. 86 (2014) 7100–7170. https://doi.org/10.1021/ac4036504.

[79] R. Tauler, A. de Juan, Chapter 5 - multivariate curve resolution for quantitative analysis, in: A.M. de la Peña, H.C. Goicoechea, G.M. Escandar, A.C. Olivieri (Editors), Fundam. Anal. Appl. Multiw. Calibration, Elsevier, 2015, pp. 247–292.

[80] A. de Juan, R. Tauler, Factor analysis of hyphenated chromatographic data Exploration, resolution and quantification of multicomponent systems, J. Chromatogr. A 1158 (2007) 184–195. https://doi.org/10.1016/j.chroma.2007.05.045.

[81] S. Kim, M. Ouyang, J. Jeong, C. Shen, X. Zhang, A new method of peak detection for analysis of comprehensive two-dimensional gas chromatography mass spectrometry data, Ann. Appl. Stat. 8 (2014) 1209–1231. https://doi.org/10.1214/14-AOAS731.

[82] S. Plancade, Y. Rozenholc, E. Lund, Generalization of the normal-exponential model: exploration of a more accurate parametrisation for the signal distribution on Illumina BeadArrays, BMC Bioinf. 13 (2012) 1–16. https://doi.org/10.1186/1471-2105-13-329.

[83] S. Kim, H. Jang, I. Koo, J. Lee, X. Zhang, Normal–Gamma–Bernoulli peak detection for analysis of comprehensive two-dimensional gas chromatography mass spectrometry data, Comput. Stat. Data Anal. 105 (2017) 96–111. https://doi.org/10.1016/j.csda.2016.07.015.

[84] G. Vivó-Truyols, Bayesian approach for peak detection in two-dimensional chromatography, Anal. Chem. 84 (2012) 2622–2630. https://doi.org/10.1021/ac202124t.

[85] N. Navarro-Reig, C. Bedia, R. Tauler, J. Jaumot, Chemometric strategies for peak detection and profiling from multidimensional chromatography, Proteomics 18 (2018) 1–12. https://doi.org/10.1002/pmic.201700327.

[86] C. Ruckebusch, L. Blanchet, Multivariate curve resolution: a review of advanced and tailored applications and challenges, Anal. Chim. Acta 765 (2013) 28–36. https://doi.org/10.1016/j.aca.2012.12.028.

[87] G.H. (Gene H. Golub, C.F. Van Loan, Matrix Computations, Johns Hopkins University Press, 1989.

[88] A. Golshan, H. Abdollahi, S. Beyramysoltan, M. Maeder, K. Neymeyr, R. Rajkó, N. Sawall, R. Tauler, A review of recent methods for the determination of ranges of feasible solutions resulting from soft modelling analyses of multivariate data, Anal. Chim. Acta 911 (2016) 1–13. https://doi.org/10.1016/j.aca.2016.01.011.

[89] R. Tauler, Calculation of maximum and minimum band boundaries of feasible solutions for species profiles obtained by multivariate curve resolution, J. Chemom. 15 (2001) 627–646. https://doi.org/10.1002/cem.654.

[90] D.W. Cook, S.C. Rutan, D.R. Stoll, P.W. Carr, Two dimensional assisted liquid chromatography – a chemometric approach to improve accuracy and precision of quantitation in liquid chromatography using 2D separation, dual detectors, and multivariate curve resolution, Anal. Chim. Acta 859 (2015) 87–95. https://doi.org/10.1016/j.aca.2014.12.009.

[91] C. Tistaert, H.P. Bailey, R.C. Allen, Y. Vander Heyden, S.C. Rutan, Resolution of spectrally rank-deficient multivariate curve resolution: alternating least squares components in comprehensive two-dimensional liquid chromatographic analysis, J. Chemom. 26 (2012) 474–486. https://doi.org/10.1002/cem.2434.

[92] D.W. Cook, M.L. Burnham, D.C. Harmes, D.R. Stoll, S.C. Rutan, Comparison of multivariate curve resolution strategies in quantitative LC×LC: application to the quantification of furanocoumarins in apiaceous vegetables, Anal. Chim. Acta (2017). https://doi.org/10.1016/j.aca.2017.01.047.

[93] H.P. Bailey, S.C. Rutan, P.W. Carr, Factors that affect quantification of diode array data in comprehensive two-dimensional liquid chromatography using chemometric data analysis, J. Chromatogr. A 1218 (2011) 8411–8422. https://doi.org/10.1016/j.chroma.2011.09.057.

[94] D.W. Cook, S.C. Rutan, Analysis of liquid chromatography-mass spectrometry data with an elastic net multivariate curve resolution strategy for sparse spectral recovery, Anal. Chem. 89 (2017) 8405–8412. https://doi.org/10.1021/acs.analchem.7b01932.

[95] E. Peré-Trepat, S. Lacorte, R. Tauler, Alternative calibration approaches for LC-MS quantitative determination of coeluted compounds in complex environmental mixtures using multivariate curve resolution, Anal. Chim. Acta 595 (2007) 228–237. https://doi.org/10.1016/j.aca.2007.04.011.

[96] H. Parastar, J.R. Radović, M. Jalali-Heravi, S. Diez, J.M. Bayona, R. Tauler, Resolution and quantification of complex mixtures of polycyclic aromatic hydrocarbons in heavy fuel oil sample by means of GC × GC-TOFMS combined to multivariate curve resolution, Anal. Chem. 83 (2011) 9289–9297. https://doi.org/10.1021/ac201799r.

[97] N. Bayot, M. Marín-García, J.B. Ghasemi, R. Tauler, Application of the area correlation constraint in the MCR-ALS quantitative analysis of complex mixture samples, Anal. Chim. Acta 1113 (2020) 52–65. https://doi.org/10.1016/j.aca.2020.03.057.

[98] A.C. de O. Neves, R. Tauler, K.M.G. de Lima, Area correlation constraint for the MCR–ALS quantification of cholesterol using EEM fluorescence data: a new approach, Anal. Chim. Acta 937 (2016) 21–28. https://doi.org/10.1016/j.aca.2016.08.011.

[99] I.T. Jolliffe, B. Morgan, Principal component analysis and exploratory factor analysis, Stat. Methods Med. Res. 1 (1992) 69–95. https://doi.org/10.1177/096228029200100105.

[100] F. Wei, S.X. Ji, N. Hu, X. Lv, X.Y. Dong, Y.Q. Feng, H. Chen, Online profiling of triacylglycerols in plant oils by two-dimensional liquid chromatography using a single column coupled with atmospheric pressure chemical ionization mass spectrometry, J. Chromatogr. A 1312 (2013) 69–79. https://doi.org/10.1016/j.chroma.2013.08.005.

[101] X.Y. Dong, J. Zhong, F. Wei, X. Lv, L. Wu, Y. Lei, B.S. Liao, S.Y. Quek, H. Chen, Triacylglycerol composition profiling and comparison of high-oleic and normal peanut oils, JAOCS, J. Am. Oil Chem. Soc. 92 (2015) 233–242. https://doi.org/10.1007/s11746-014-2580-5.

[102] F. Andrić, K. Héberger, How to compare separation selectivity of high-performance liquid chromatographic columns properly? J. Chromatogr. A 1488 (2017) 45–56. https://doi.org/10.1016/j.chroma.2017.01.066.

[103] M.R. Euerby, P. Petersson, W. Campbell, W. Roe, Chromatographic classification and comparison of commercially available reversed-phase liquid chromatographic columns containing phenyl moieties using principal component analysis, J. Chromatogr. A 1154 (2007) 138–151. https://doi.org/10.1016/j.chroma.2007.03.119.

[104] R. Grasshoff, N.J. Nielsen, J.H. Christensen, Using the hydrophobic subtraction model to choose orthogonal columns for online comprehensive two-dimensional liquid chromatography, J. Chromatogr. A 1326 (2014) 39–46. https://doi.org/10.1016/j.chroma.2013.12.034.

[105] S. Brown, R. Tauler, Comprehensive Chemometrics: Chemical and Biochemical Data Analysis, Elsevier Science, 2015. https://books.google.es/books?id=heXEDwAAQBAJ.

[106] D. Bressanello, E. Liberto, M. Collino, F. Chiazza, R. Mastrocola, S.E. Reichenbach, C. Bicchi, C. Cordero, Combined untargeted and targeted fingerprinting by comprehensive two-dimensional gas chromatography: revealing fructose-induced changes in mice urinary metabolic signatures, Anal. Bioanal. Chem. 410 (2018) 2723–2737. https://doi.org/10.1007/s00216-018-0958-9.

[107] X. Sun, Y. Lv, J. Wang, H.Q. Cheng, J. Huang, Y. Du, J. Dong, Differential protein expression profiling by iTRAQ-2D-LC-MS/MS of rats treated with oxaliplatin, J. Cell. Biochem. 120 (2019) 18128–18141. https://doi.org/10.1002/jcb.29116.

[108] J.G. Schreiner, S.M. Mamwa, multiple response variables and multispecies interactions, Design and Analysis of Ecological Experiments, in: C. Press (Editor), Des. Anal. Ecol. Exp., 1993, pp. 94–112.

[109] A.K. Smilde, J.J. Jansen, H.C.J. Hoefsloot, R.J.A.N. Lamers, J. van der Greef, M.E. Timmerman, ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data, Bioinformatics 21 (2005) 3043–3048. https://doi.org/10.1093/bioinformatics/bti476.

[110] E. Saccenti, A.K. Smilde, J. Camacho, Group-wise ANOVA simultaneous component analysis for designed omics experiments, Metabolomics 14 (2018) 1–18. https://doi.org/10.1007/s11306-018-1369-1.

[111] P.D.B. Harrington, N.E. Vieira, J. Espinoza, J.K. Nien, R. Romero, A.L. Yergey, Analysis of variance–principal component analysis: a soft tool for proteomic discovery, Anal. Chim. Acta 544 (2005) 118–127. https://doi.org/10.1016/j.aca.2005.02.042.

[112] J. Engel, L. Blanchet, B. Bloemen, L.P. Van den Heuvel, U.H.F. Engelke, R.A. Wevers, L.M.C. Buydens, Regularized MANOVA (rMANOVA) in untargeted metabolomics, Anal. Chim. Acta 899 (2015) 1–12. https://doi.org/10.1016/j.aca.2015.06.042.

[113] F. Marini, D. de Beer, E. Joubert, B. Walczak, Analysis of variance of designed chromatographic data sets: the analysis of variance-target projection approach, J. Chromatogr. A 1405 (2015) 94–102. https://doi.org/10.1016/j.chroma.2015.05.060.

[114] B. Govaerts, B. Francq, R. Marion, M. Martin, M. Thiel, The Essentials on Linear Regression, ANOVA, General Linear and Linear Mixed Models for the Chemist, second ed., Elsevier, 2020. https://doi.org/10.1016/b978-0-12-409547-2.14579-2.

[115] M. Martin, B. Govaerts, LiMM-PCA: combining ASCA+ and linear mixed models to analyse high-dimensional designed data, J. Chemom. 34 (2020) 1–20. https://doi.org/10.1002/cem.3232.

[116] A. Rácz, A. Gere, D. Bajusz, K. Héberger, Is soft independent modeling of class analogies a reasonable choice for supervised pattern recognition? RSC Adv. 8 (2018) 10–21. https://doi.org/10.1039/c7ra08901e.

[117] L. Stahle, Cross-validation for the two-class Problem: a monte CarlO study, Analysis 1 (1987) 185–196.

[118] Y. Sato, T. Nakamura, K. Aoshima, Y. Oda, Quantitative and wide-ranging profiling of phospholipids in human plasma by two-dimensional liquid chromatography,mass spectrometry, Anal. Chem. 82 (2010) 9858–9864. https://doi.org/10.1021/ac102211c.

[119] M. Navarro-Reig, J. Jaumot, R. Tauler, An untargeted lipidomic strategy combining comprehensive two-dimensional liquid chromatography and chemometric analysis, J. Chromatogr. A 1568 (2018) 80–90. https://doi.org/10.1016/j.chroma.2018.07.017.

[120] M. Bylesjö, M. Rantalainen, O. Cloarec, J.K. Nicholson, E. Holmes, J. Trygg, OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification, J. Chemom. 20 (2006) 341–351. https://doi.org/10.1002/cem.1006.

[121] E.D. Kantz, S. Tiwari, J.D. Watrous, S. Cheng, M. Jain, Deep neural networks for classification of LC-MS spectral peaks, Anal. Chem. 91 (2019) 12407–12413. https://doi.org/10.1021/acs.analchem.9b02983.

[122] Y. Xu, S. Zomer, R.G. Brereton, Support vector machines: a recent method for classification in chemometrics, Crit. Rev. Anal. Chem. 36 (2006) 177–188. https://doi.org/10.1080/10408340600969486.

[123] A. Checa, C. Bedia, J. Jaumot, Lipidomic data analysis: tutorial, practical guidelines and applications, Anal. Chim. Acta 885 (2015) 1–16. https://doi.org/10.1016/j.aca.2015.02.068.

[124] U.W. Liebal, A.N.T. Phan, M. Sudhakar, K. Raman, L.M. Blank, Machine learning applications for mass spectrometry-based metabolomics, Metabolites 10 (2020) 1–23. https://doi.org/10.3390/metabo10060243.

[125] M. Marcinkiewicz-Siemion, M. Kaminski, M. Ciborowski, K. Ptaszynska-Kopczynska, A. Szpakowicz, A. Lisowska, M. Jasiewicz, E. Tarasiuk, A. Krenowski, B. Sobkowicz, K.A. Kaminski, Machine-learning facilitates selection of a novel diagnostic panel of metabolites for the detection of heart failure, Sci. Rep. 10 (2020) 1–11. https://doi.org/10.1038/s41598-019-56889-8.

[126] S.I. Abba, A.C. Usman, S. Isik, Simulation for response surface in the HPLC optimization method development using artificial intelligence models: a data-driven approach, Chemometr. Intell. Lab. Syst. 201 (2020). https://doi.org/10.1016/j.chemolab.2020.104007.

[127] P. Bonini, T. Kind, H. Tsugawa, D.K. Barupal, O. Fiehn, Retip: retention time prediction for compound annotation in untargeted metabolomics, Anal. Chem. 92 (2020) 7515–7522. https://doi.org/10.1021/acs.analchem.9b05765.

[128] D. Panagopoulos Abrahamsson, J.S. Park, R.R. Singh, M. Sirota, T.J. Woodruff, Applications of machine learning to in silico quantification of chemicals without analytical standards, J. Chem. Inf. Model. 60 (2020) 2718–2727. https://doi.org/10.1021/acs.jcim.9b01096.

[129] W. Windig, J. Guilment, Interactive self-modeling mixture analysis, Anal. Chem. 63 (1991) 1425–1432. https://doi.org/10.1021/ac00014a016.

[130] H.P. Bailey, S.C. Rutan, Comparison of chemometric methods for the screening of comprehensive two-dimensional liquid chromatographic analysis of wine, Anal. Chim. Acta 770 (2013) 18–28. https://doi.org/10.1016/j.aca.2013.01.052.

Obtaining quantitative information in 2DLC by ROIMCR

Quantifying analytes in 2DLC is more complex than in one-dimension liquid chromatography (1DLC). The same chemical compound is commonly fractioned into different peaks in the subsequent modulations. Therefore, obtaining the global peak area for each analyte is more difficult.

Several strategies can be performed, such as the summation of the ^{2}D chromatographic peaks, the determination of the area in the ^{2}D plot or estimating the peak height or volume [263]. Most 2DLC quantifications have been carried out through vendor software or by specific data processing software designed for 2DLC data (e.g., GC Image LC×LC Edition Software from GC Image™, AnalyzerPro® XD from SpectralWorks, and ChromSquare from Shimadzu). These tools present two main disadvantages. The first is that they are usually costly, especially for the laboratories that start using 2DLC and have not implemented routine methods yet. The second disadvantage is the limitation of the proposed pre- and post-processing approaches. Although an important effort has been put into implementing user-friendly software to fully analyze 2DLC data, the reality is that more than one software is usually required.

In this PhD Thesis, three different strategies have been used for 2DLC quantification. All of them present the main benefit that can be run in the MATLAB environment through specific toolboxes. The comparison between the three strategies is shown in **scientific publication VII** (**Chapter 4**). The first strategy uses the ROI procedure to obtain the peak areas, summing the individual areas of each ^{2}D peak corresponding to the same compound and associated with a specific region of interest (i.e., *m/z* value). This process can be performed automatically by the MSroi app [235], or manually, in the case that isobaric compounds with different retention times in the ^{1}D have been associated with the same ROI. In the latter case, a specific retention time range can be selected in the software application to obtain the peak area of that region only. Once the areas are obtained for each analyte, a calibration curve can be built.

The second and third strategies consider the peaks of the elution profiles resolved by MCR-ALS focusing on the quantitative analysis [264]. In the specific case of 2DLC data, several studies proved its usefulness when a diode array detector (DAD) was used as detector [263,265–267]. In this PhD Thesis, the emphasis is

placed on applying MCR-ALS for quantification of LC × LC-MS datasets, considering the advantages of this flexible bilinear model (e.g., no peak alignment requirements and the ability to handle overlapping peaks). Both peak areas and peak heights can be used to build calibration curves with the MCR-ALS resolved peak areas [268]. In this work, LC × LC-MS quantitation is performed using the peak areas of the elution profiles resolved by MCR-ALS.

Recently, the area correlation constraint has been proposed for cases like those analyzed in this PhD Thesis using the MCR-ALS method in its bilinear model variant [269,270]. Two approaches have been compared to assess whether the area correlation constraint would improve quantification results in the case of LC × LC-MS datasets. Thus, in the first strategy, calibration curves are built from the areas of the elution profiles finally obtained in the results of the application of the MCR-ALS method with only applying non-negativity constraint. Since the concentration profiles obtained are in arbitrary units and only have relative quantitation information, they should be calibrated using the known concentrations of the analytes in some of the simultaneously analyzed samples, for instance, by linear regression (see more details in [240]). The other strategy tested in this PhD Thesis, applies bot, non-negativity and the area correlation constraint during the ALS optimization of the bilinear MCR model. This last strategy performs the calibration (linear regression) of the peak areas of the elution profiles at every ALS step until convergence, which allows the application of the quantification of the analytes of interest in the presence of unknown interferences, even if these interferences are not present in the calibration samples [259,269,270].

When the area correlation constraint is applied (during the ALS optimization), the area of an elution profile from the calibration samples can be correlated to a concentration value known *a priori*. Then, a local correlation model is created. The intensity of the elution profile from the calibration samples is adjusted with this known value. Then, predictions and adjustments on the elution profiles of the component concerned from unknown samples can be performed. The concentration profiles from the calibration samples are updated with the ones from the unknown samples and the optimization process continues until the convergence is achieved. This iterative process is summarized in **Figure 2.19**. The peak areas (concentration values) are rescaled to the real concentration units during the iterative optimization, by the application of this constraint [271].

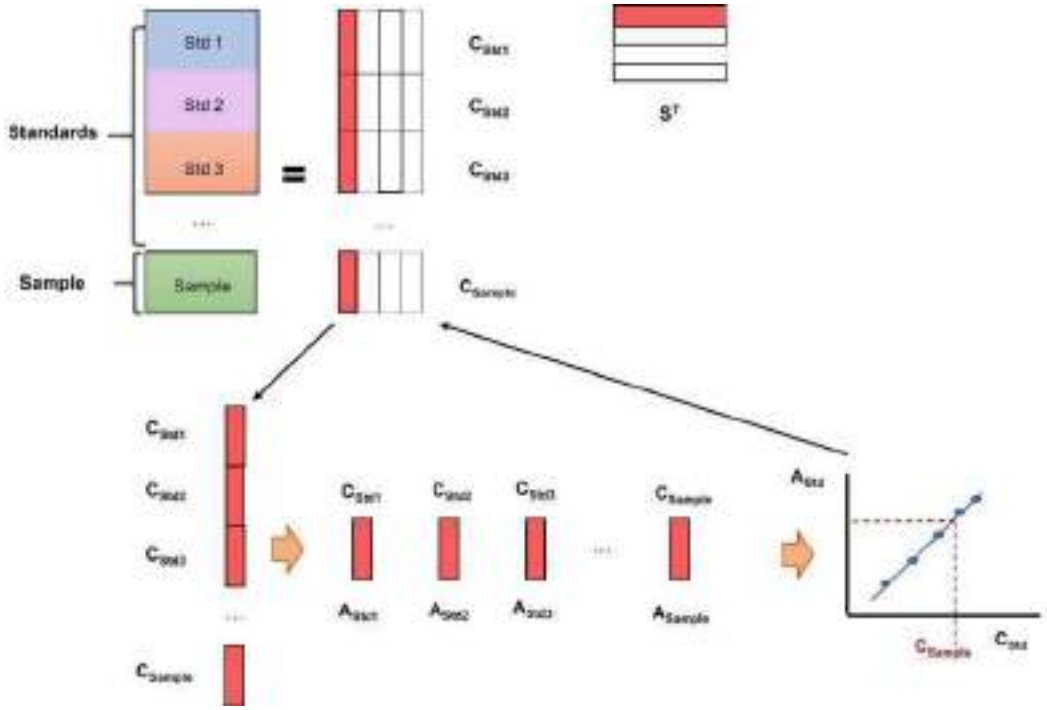


Figure 2.19. Graphical representation of the implementation of the area correlation constraint in MCR-ALS.

2.3.3 Normalization and data scaling

As stated in the previous section, the areas of the compounds can be obtained directly from the ROI procedure or after applying MCR-ALS. In both cases, the output is a data table or matrix with the samples in the rows and the features or compounds in the columns, as already shown when describing the ROIMCR workflow in **Figure 2.16.B**. Then, post-processing steps, such as multivariate analysis, are performed on this matrix. However, depending on the purpose of the study, a preliminary step of normalization and data scaling is required.

Metabolomic analysis needs a minimum of reporting standards to ensure data quality, as proposed by Goodacre et al. [272]. The discovery of biomarkers will directly depend on data quality. Therefore, different metrics were proposed to guarantee data quality, repeatability, and reproducibility of results (e.g., coefficient of variation, missing values, retention time shifts) [273].

A major issue that can cause a lack of reproducibility in metabolomic studies is the systematic variation between samples (e.g., instrumental drifts along the sequence, extraction efficiency), which should be reduced. There are different

strategies to minimize this systematic variation, but they can be classified into two main groups. On the one side, normalization is performed in the rows because the aim is to correct by sample (e.g., normalize by the quantity of tissue or protein). On the other side, data scaling approaches are performed in the columns because they affect the variables (e.g., centering or autoscaling).

Normalization

Normalization can be performed chemically and/or mathematically. In this PhD Thesis, the first step carried out PhD Thesis was a correction considering the amount of biological material of each biological replicate. For instance, in the analysis of rice, zebrafish embryos, or cell cultures, normalization can be applied according to the weight of tissue (root or leave), the exact number of embryos, the amount of protein or by the number of cells per replicate, respectively.

In a second step, surrogates and internal standards have been used in this PhD Thesis for chemical normalization. The purpose of the surrogates (added just at the beginning of the extraction) is to correct possible extraction losses. In contrast, the aim of the internal standards (added just before starting the analysis) is to correct instrumental drifts along the analysis itself or by ion suppression. In the case that no internal standards have been added, another possibility is to correct according to signals present in the background or associated with known metabolites (**scientific publication VI**).

There are two main quality management procedures in metabolomics: quality assurance (QA) and quality control (QC) [274]. QA ensures that quality requirements are fulfilled before sample collection. QC guarantees that data quality meets specific requirements after data acquisition. In this PhD Thesis, QC is referred to as a type of sample composed of a pool of all kinds of different sample types, which is measured repeatedly along the whole analytical sequence each 5-10 samples, depending on the whole length of the batch. QC samples have been measured in all the metabolomic/lipidomic analyses presented in this PhD Thesis.

In the first step, the variance in the control samples is calculated and should be lower than 20%. Moreover, all QCs should cluster together when samples are plotted (e.g., in a principal component analysis scores plot [275]). Quality controls are a good indicator of batch effects [276,277]. They provide a good overview of the possible baseline and signal intensity variations throughout the whole analytical sequence. QCs can prove intra and inter batch repeatability and reproducibility, or help to correct batch effects when they occur, which unfortunately happens frequently [24,277]. QC recommendations have been followed in this PhD Thesis to guarantee

the quality of the analyzed metabolomics datasets, as well as for correcting batch effects when necessary (**scientific publication VIII**).

Among mathematical normalizations, the Probabilistic Quotient Normalization (PQN) [278] has been selected in this PhD Thesis in the untargeted analysis of rice exposed to arsenic (**scientific publication VII**), due to its better performance compared to other normalization methods for untargeted metabolomic analysis [279]. PQN states that changes throughout the concentration of a sample influence the whole spectrum, but changes in the concentration of a single analyte will only affect a specific part of the spectra. Hence, a normalization factor is calculated using the signals of a reference spectrum. The reference spectrum can be the average metabolite abundance of all samples (blanks not included).

Data scaling

Data scaling allows the comparison among variables in different samples. The concentration range of metabolites may significantly change, and their simultaneous analysis may be biased towards those variables showing larger variances. There are different data scaling pre-treatments that reduce this variation in the scale to enhance biological differences, independently of the actual concentration value of the metabolites. A list of methods for this purpose can be found elsewhere [272]. The most commonly employed are mean-centering, scaling and autoscaling (which includes mean-centering). Mean-centering removes the differences in scale magnitude because the mean value of every variable is subtracted from each individual value. Consequently, the mean value is zero. However, differences in scale amplitude (e.g., derived from standard deviation) are still present. In this PhD Thesis, autoscaling has been employed because it eliminates both scale amplitude and scale magnitude. The mean value is subtracted from each individual value and divided by the variable standard deviation. Therefore, the mean value is zero and the standard deviation of each variable equals one.

2.3.4 Other multivariate resolution methods

The simultaneous resolution of datasets forming three-way data structures can be performed with methods based on the trilinear model like parallel factor analysis (PARAFAC) or PARAFAC2. Most of the applications of these approaches in metabolomics have been related to GC × GC-MS datasets [280–285], but there are also some applications in LC × LC-MS [286]. PARAFAC has also been applied to LC-DAD datasets for metabolomic fingerprinting [287,288], and LC-Fluorescence studies of plant roots exudes [289] or colorectal cancer [290]. In this PhD Thesis,

trilinear methods have been compared to trilinear and bilinear ones obtained by MCR-ALS in the analysis of LC × LC-MS.

PARAFAC is a factor decomposition method of a data cube according to the trilinear model [291] of **Equation 3** and exemplified for LC × LC-MS in **Figure 2.20:**

Equation 3

$$\underline{\mathbf{B}} = \sum_{i=1}^{N} \boldsymbol{f_i} \otimes \boldsymbol{g_i} \otimes \boldsymbol{h_i} + \underline{\mathbf{E}}$$

In the case of LC × LC-MS data of a single samples (file), **B** refers to the three-way data (data cube), which is decomposed into the product of three contribution factors with *N* as the number of components; f_i describes the elution profiles of the ^{1}D; g_i represents ^{2}D elution profiles and h_i the pure spectra profiles. Analogously to MCR-ALS, **E** refers to non-explained variance. The different factors are linked through an external product, represented by the symbol ⊗. **F**, **G** and **H** from **Figure 2.20** are the joint f_i, g_i, and h_i profiles.

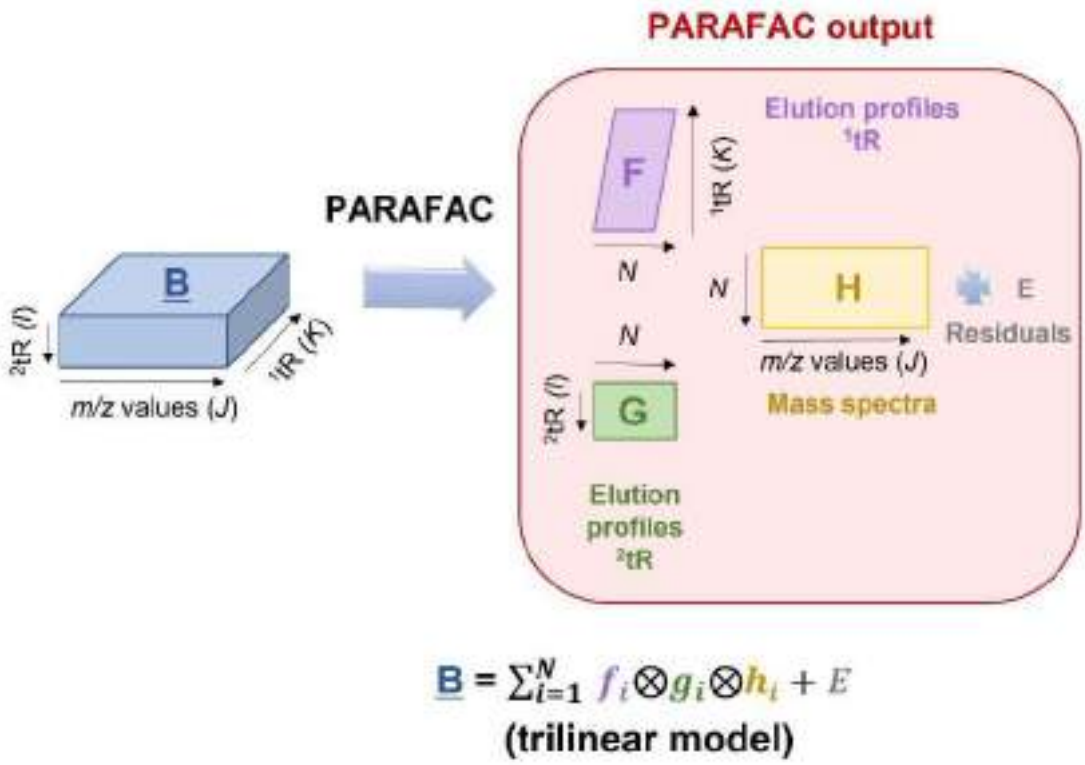


Figure 2.20. Graphical representation of PARAFAC resolution for a data cube.

PARAFAC reduces the dimensionality of a dataset. This method assumes three-way data of dimensions I × J × K. Thus, the instrumental response combines pure responses in each dimension [292]. In the case of multidimensional chromatographic

separations coupled to mass spectrometry, the three dimensions are the chromatographic profiles in the two dimensions, and the mass spectra of each compound.

The main advantage of PARAFAC and trilinear MCR compared to bilinear MCR is that rotational ambiguities are avoided. The application of the trilinear model provides unique solutions and theresolution is often more robust and easier to interpret if the trilinear model is fullfilled [291]. However, PARAFAC does not allow small deviations from the trilinear behavior [293].

Therefore, the first question to consider is whether three-way LC × LC-MS data are trilinear or not. This issue has been addressed in **scientific publication II** (included in **Section 2.3.2. ROIMCR**, previously in this introductory Chapter), but also is discussed in further detail in **scientific publication III in Chapter 3**. Briefly, in multidimensional chromatographic separations, deviations from trilinearity are commonly encountered (e.g., retention time shifts or peak shape changes) [294–296]. Hence, the two options are a) trying to restore trilinearity before applying PARAFAC (e.g., with Tchebichef moments' approach [297], or b) using more flexible algorithms.

PARAFAC2 arises as an alternative to PARAFAC thanks to the fewer restrictions, which allow the handling of small time shifts. In the PARAFAC2 model, elution profiles are no longer considered parallel and proportional, but only the inner-product structure is preserved across different samples. The result is that the cross-products of the matrix that contains the elution profiles in its columns are constant [293]. Consequently, in the case of LC × LC-MS data, elution profiles can present some differences between modulations but still be considered the same compound. Nevertheless, although PARAFAC2 is able to cope with small swings in retention times, there are still deviations of trilinearity caused by changes in the peak shapes due to, for instance, coelution of multiple compounds. When these deviations are encountered, PARAFAC2 model also fails.

The MCR-ALS method can also be applied to three- and higher-order data sets and adapted to the implementation of constraints to fulfill trilinear and multilinear models [298,299]. In fact, the application of such constraints in MCR-ALS is very flexible and allows the simultaneous implementation of mixed bilinear and multilinear models for the same dataset. These constraints also can deal with deviations of the trilinear model caused by peak shifting. The comparison with these variants of the MCR-ALS method is not shown in this PhD Thesis, where the main point was to check whether LC × LC-MS data are trilinear and could be analyzed by PARAFAC and

PARAFAC2 methods, which are the two more currently used trilinear model-based methods in this field. On the other hand, MCR-ALS bilinear does not require the fulfillment of the trilinear model nor prior chromatographic alignment or peak modelling steps, which simplifies the analysis of multi-way datasets.

Previous comparisons between PARAFAC, PARAFAC2 and MCR-ALS bilinear and trilinear from the literature have been reported on both GC × GC-MS [300] and LC × LC-MS [301]. However, stronger deviations from trilinear behavior are very common and serious in 2DLC data. Therefore, in these cases, especially in 2DLC untargeted analysis, bilinear MCR-ALS is highly recommended [292,296]. That is why MCR-ALS, based only in the fulfillment of the bilinear model (without trilinearity constraints), has been preferred in the LC × LC-MS studies in this PhD Thesis.

Lastly, a novel tool for supervised discovery-based experimentation has been recently proposed [302]. The study compares the purified mass spectrum obtained with more classic chemometric approaches (e.g., MCR-ALS, PARAFAC, PARAFAC2).

More information about the comparison between MCR-ALS and PARAFAC methods for the analysis of 2DLC data, and the fulfillment of the trilinear model requirements by er 2DLC datasets can be found in **scientific publication II** (at the end of **Section 2.3.2. ROIMCR** of this introduction Chapter) and in **scientific publication III (Chapter 3).**

2.3.5 Post-processing strategies

Once the data table or matrix with the areas of the metabolites analyzed is obtained by, for instance, the application of the MCR-ALS method, different post-processing strategies can be applied to establish the existing relations between the samples and the experimental factors. There are five main groups of post-processing methods, summarized in **Figure 2.21**: statistical tests, exploratory analysis, classification or discrimination analysis, machine learning and clustering. Strategies from all groups, except machine learning, were used throughout this PhD Thesis to extract the sought information and will be briefly introduced below.

More information on other methods for metabolomics and lipidomics analysis is included in **scientific publication II (at the end of Section 2.3.2 ROIMCR of this introduction Chapter**), and also in the review by Paul et al. [234], Feizi et al. [303], Yi et al. [304], and Checa et al. [305].

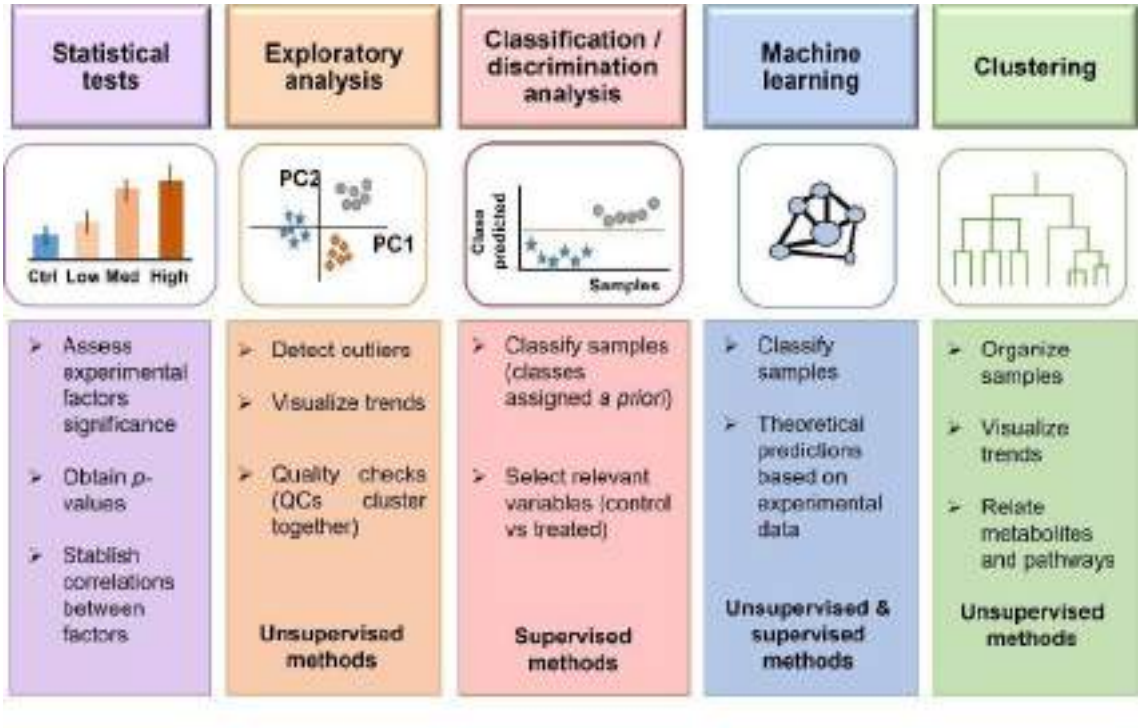


Figure 2.21. The five groups of post-processing methods that can be applied to the data matrix with the metabolite areas.

Exploratory analysis

Principal component analysis (PCA) is the most popular non-supervised exploratory data analysis method used in this PhD Thesis. The main advantage of PCA is that allows visualizing the data in a space of reduced dimensions. New axes are defined, called principal components, which keep the most relevant information about differences and similarities between the samples and eliminate experimental noise-related information. Classes or sample types are not established *a priori*, and in general, no information about the samples is provided to the model. That is why it is considered a non-supervised method. The new PCA axes are orthogonal, and they are built up by the linear combinations of the original variables that more efficiently describe the data variation. A reduced number of principal components is usually needed (typically two or three), to explain a large amount of the data variance. Principal components are ordered according to their relative importance, i.e., the first principal component explains the maximum amount of data variance, etc.

PCA is based on a bilinear decomposition model, as MCR, that follows **Equation 4:**

Equation 4 $$\boldsymbol{X = TP^T + E}$$

The original data matrix is **X**; the orthogonal scores matrix is **T** and describes the samples in the principal components space. $\mathbf{P}^T$ is the orthonormal loadings matrix, which accounts for the linear combination of the original variables in the new principal components space.

Some practical applications of the scores plot provided by PCA include the detection of outliers and the visualization of possible clusters and trends in the data (e.g., similar samples coming from the same class are together, samples from different classes cluster separately, etc). PCA also provides a quick data quality check. For instance, if all QCs are clustered together in the PCA space, this means that batch effects are negligible. On the contrary, if there is an instrumental drift along the sequence or significant differences among different extraction batches, these differences will be probably pointed out in the scores plot. **Figure 2.22** shows an example of a scores plot where different sample classes (one control, four different treatments and QCs) are clustered separately.

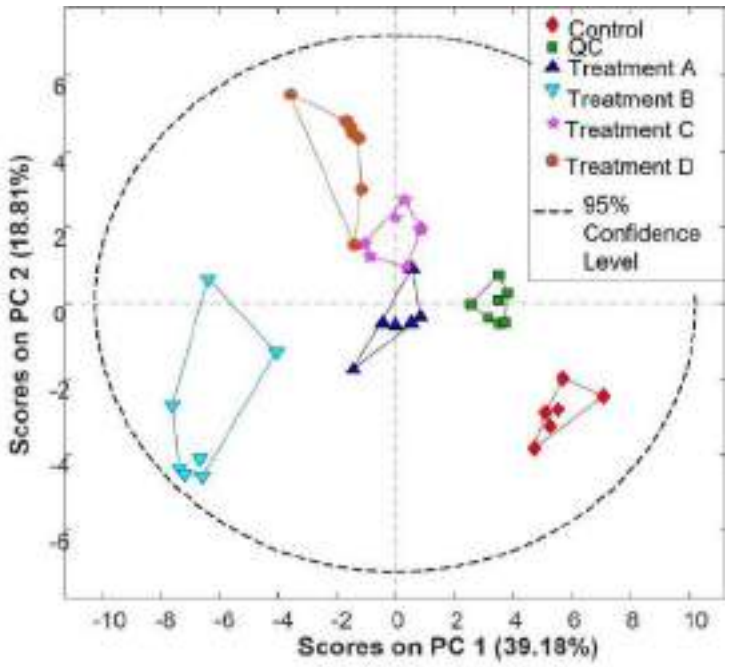


Figure 2.22. Example of score plot of PCA with multiple sample types clustered together, including QCs.

On the other side, from the loadings plot, it is possible to identify the most relevant variables (*m/z* values in MS analysis) that cause differences between the samples. Correlations between experimental variables can also be established. In the case of metabolomic studies, where direct analysis of MS data is performed, the

number of variables (*m/z* values) measured is much higher than the number of samples, normally from hundreds to thousands of *m/z* values. Hence, extracting information from loadings plot is much more complex than if a reduced number of variables was analyzed. However, after the application of the ROI approach (or even better after application of the combination of ROI and MCR-ALS), PCA analysis of peak areas can produce also useful information in the loadings plot.

Classification or discrimination analysis

Metabolomics is frequently based on comparing a minimum of two groups of samples, i.e., control *versus* treated or exposed. These groups of samples are already established from the experimental design and may differentiate the tested effects on the investigated samples (e.g., according to the treatment, or the time of exposure, etc). Supervised methods incorporate this information on different sample classes into the model and classify or discriminate between classes.

Partial Least Squares Discriminant Analysis (PLS-DA) [306] is the discriminant method used in this PhD Thesis. In PLS-DA, two inputs are required. The first, **X**, is the original data matrix (e.g., peak areas data matrix from MCR). The other, **Y**, is a vector or a matrix containing the information about the groups or sample types in the experiment. Although multiple class comparison can be performed simultaneously, in this PhD Thesis, all comparisons were by pairs (0 or 1 assigned, accordingly), incorporating only two classes by model (e.g., control and treated).

The PLS regression model correlates both inputs, **X** and **Y**, with a bilinear model using as fewer components or relevant factors as possible (i.e., latent variables or LV). **Figure 2.23** presents an example of the scores plot on the LVs, and the predictions of a PLS-DA model with two groups (control/treated) of three samples each.

The PLS-DA latent variables are calculated in order to include the maximum covariance between the two inputs, and the weight matrix (**W**) is generated during regression. The weight vectors ($\mathbf{w}_k$) reveal which are the most relevant variables in the prediction or projection model, which can derive in the variables important in projection or VIP scores [307]. Hence, the most relevant variables have a higher VIP value. In this PhD Thesis, significant variables are associated with VIP values higher than 1. In PLS-DA, these variables with higher VIP values are the ones responsible for the differences between the groups and, therefore, provide potential exposure markers. Consequently, the identification of the metabolites that correspond to the

m/z values associated with the highest VIP values will be important to understanding the impacts produced by experimental factors.

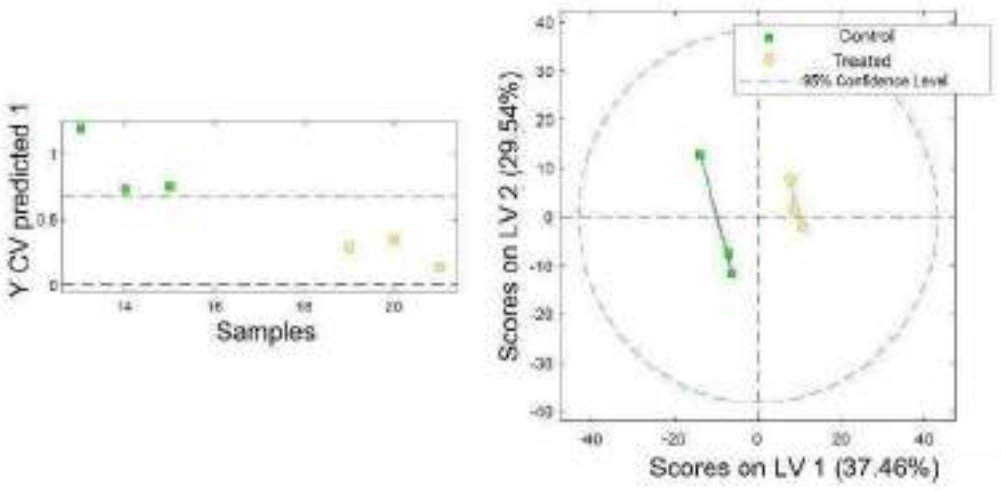


Figure 2.23. Example of the scores plot and predictions on a PLS-DA model.

The validation of the results is crucial when building the PLS-DA model and choosing the number of LVs needed. The ideal case is to have a large number of samples to have two well-populated datasets: one for the calibration and other for validation. In the case of a fewer number of samples, cross-validation strategies should be employed. Among cross validation strategies, this PhD Thesis includes two cross validation strategies: leave-one-out (datasets with less than 20 samples) and random subsets (for more than 20 samples). During leave-one-out, data are separated in a way that one sample per each iteration acts as a test dataset, while the rest becomes the training set. In contrast, when using random subsets, the training set is divided randomly, and test dataset size varies with the total number of samples.

There are two main parameters to assess the PLS-DA model quality related to the correct assignment of samples to their belonging groups, which are provided in the so-called 'confusion' matrix of the model [308]: sensibility and selectivity. Sensibility is related to the probability of correctly classifying one sample to the class it belongs to, whereas selectivity is the probability of correctly assigning that a sample does not belong to the class. There are four possible scenarios: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), as summarized in **Table 2.6**. Treated or exposed classes are labelled positive (P) and

control class, negative (N). Hence, VP and TP refer to the correct predictions, whereas TN and FN refer to wrong predictions.

Table 2.6. Matrix confusion for PLS-DA diagnostic.

		Calculated class	
		Treated/Exposed	Control
Experimental class	Treated/Exposed	VP	FN
	Control	FP	VN

Both sensitivity and selectivity can be estimated from the confusion matrix parameters, as shown in **Equations 5 and 6**, respectively, and in both cases, the optimal scenario is that values are equal to 1.

Equation 5 $$\boldsymbol{Sensitivity} = \frac{\boldsymbol{VP}}{\boldsymbol{VP+FN}}$$

Equation 6 $$\boldsymbol{Selectivity} = \frac{\boldsymbol{VN}}{\boldsymbol{VN+FP}}$$

A third parameter, called Matthews Correlation Coefficient (MCC), estimates the quality of the binary classification, and whose value should be in the range -1 and 1 [309]. A coefficient of 1 represents a perfect prediction model, 0 is associated with a random prediction and -1 is assigned to a totally wrong prediction according to the observation. **Equation 7** shows how MCC is calculated.

Equations 7 $$\boldsymbol{MCC} = \frac{\boldsymbol{TP \times TN - FP \times FN}}{\sqrt{(\boldsymbol{TP+FP})(\boldsymbol{TP+FN})(\boldsymbol{TN+FP})(\boldsymbol{TN+FP})(\boldsymbol{TN+FN})}}$$

Clustering methods

Apart from PCA, other clustering methods are designed to organize and characterize samples into groups by sets of variables. These methods aim to display the possible similarities and differences between groups of samples and variables and arrange them accordingly. Clustering methods can be grouped into hierarchical or non-hierarchical [305]. In this PhD Thesis, only hierarchical clustering methods have been employed.

Hierarchical clustering analysis (HCA) methods are characterized by tree structures in which samples are nested according to their similarities at different levels. The main advantage is that prior information on the data is not required (i.e., they are unsupervised methods). The sample relationship arrangement can be performed bottom-up or top-down, depending on whether the trend is to join or

separate the samples. The usual output of these methods is a **dendrogram**, which is a tree-structured graph that groups samples according to their similarities. The distance between the objects is measured, and the groups are established from more to less similar, ending in a global and unique group. **Heatmaps** are also helpful tools to provide complementary information about the variables. For instance, peak areas values or fold changes (ratios exposed/control) with colors assigned by magnitude. **Figure 2.24** shows an example of a combination of dendrogram and heatmap (clustergram), with a color scale based on the fold changes values of the features.

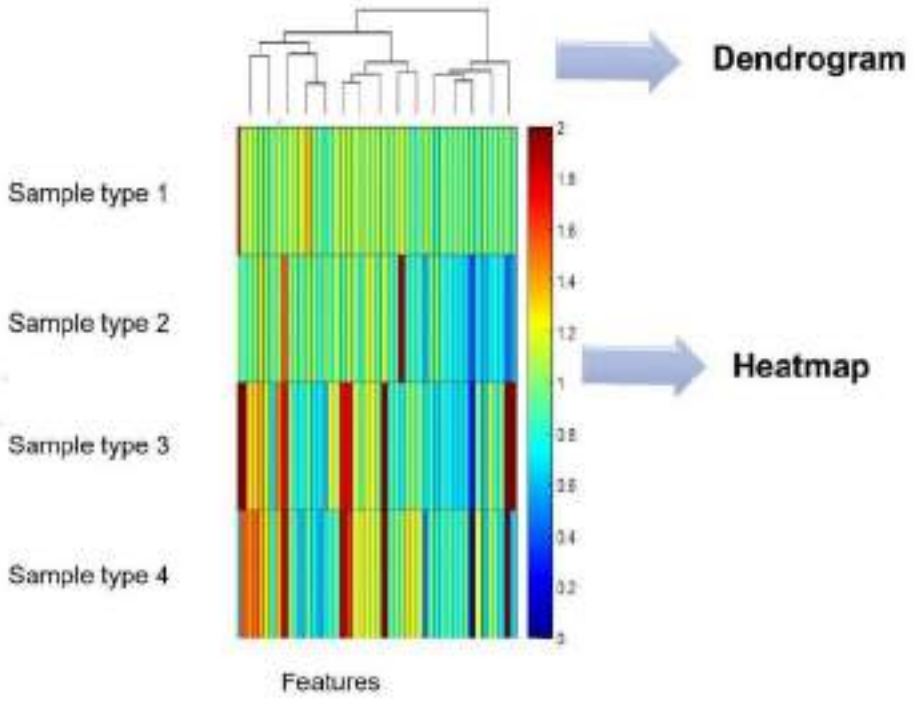


Figure 2.24. Example of dendrogram plus heatmap, where the variables are clustered according to their similarities while the samples remained ordered by experiment. The color bar indicates the intensity of the fold changes (in red when the concentration of the variables higher in the treated samples than in the controls, and in blue, the opposite scenario).

Statistical tests

There are two main categories of statistical tests: univariate and multivariate. The first group is applied when only one variable is considered at a time, whereas the second group allows for the determination of contribution from multiple variables simultaneously.

The two most common univariate methods are the **T-test** and analysis of variance or **ANOVA**. If data follow a normal distribution, then these two parametric

tests can be applied to evaluate differences between two or more groups of samples. Both determine if the mean of a variable (e.g., the concentration of a certain metabolite) is significantly different between sample groups or populations.

T-test [310] compares the mean values of a variable between two groups of samples or two populations. On the other side, ANOVA [311] can be applied to more than two sample groups. The importance of the factors and their interactions are evaluated from a unique or univariate response (i.e., for each variable individually) rather than providing the significance of the factor for the whole set of variables simultaneously (see **Table 2.7**).

Univariate tests are recommended when a reduced number of variables are studied, for instance, in targeted analysis. The interpretation of these tests is usually straightforward. However, when the number of hypotheses increases, the probability of making false positives (i.e., the mistaken rejection of a null hypoPhD Thesis, also called type I error) also augments. In multiple hypoPhD Thesis testing, different corrections can be applied to avoid both overestimating (false positives) and underestimating (false negatives) significant variables. For instance, Bonferroni correction [312] reduces type I errors when several comparisons are performed in a single dataset. The new *p*-value is calculated by dividing the original *p*-value by the number of simultaneous comparisons performed. The main disadvantage is that this correction is too strict, and true positives can also be considered false positives. Another correction is the false discovery rate or FDR [313], which in a softer manner evaluates multiple hypotheses, and minimizes the number of false discoveries by reducing the number of false positives.

Consequently, when hundreds or thousands of variables are evaluated simultaneously (a common situation in untargeted analysis), multivariate methods are preferred. Otherwise, univariate methods should be corrected to avoid the multiple hypothesis testing problem.

Multivariate analysis of variance or MANOVA is a form of ANOVA applied in multivariate analysis [314]. This method allows the comparison of means from different samples when there are two or more dependent variables. However, the main drawback of MANOVA is that it cannot be applied in datasets where the number of variables is higher than the number of samples, which becomes impractical for metabolomic studies. Among the different approaches that have arisen in the recent years to overcome this limitation, the most successful ones are based on PCA analysis or similar techniques (e.g., Simultaneous Component Analysis, SCA), as is the case of the three multivariate statistical tests evaluated in this PhD Thesis:

ANOVA-simultaneous component analysis (ASCA) [315,316], Regularized Multivariate ANOVA (rMANOVA) [317], and Group-wise ASCA (GASCA) [318]. The comparison and evaluation of metabolomic results are included in **scientific publication IV** (**Chapter 3**). These three multivariate methods also allow detecting the relevant metabolites (or the *m/z* values of importance) associated with the factors from the design of the experiment (DOE), as summarized in **Table 2.7**. A brief explanation of each of these methods is given below.

ANOVA-simultaneous component analysis or ASCA [315] is a powerful method when dealing with complex experimental designs (e.g., multiple experimental factors and variables). Therefore, it has been frequently used in metabolomic studies. ASCA combines the advantages of ANOVA with SCA. In ASCA, individual factors and their multiple interactions are analyzed to establish whether they are significant or not. If every factor level has the same number of replicates, the experimental design is known as balanced. The opposite case is called unbalanced and requires an alternative manner of calculating the sum of squares because factors can be correlated (i.e., are not orthogonal) [315].

In the case that 3 factors are employed, the ANOVA model can be described with **Equation 8**:

Equation 8 $$\mathbf{X} = \bar{\mathbf{X}} + \mathbf{X_A} + \mathbf{X_B} + \mathbf{X_C} + \mathbf{X_{AB}} + \mathbf{X_{AC}} + \mathbf{X_{BC}} + \mathbf{X_{ABC}} + \mathbf{E}$$

$\mathbf{X}$ refers to the original experimental matrix, and $\bar{\mathbf{X}}$ is the mean matrix of the original. $\mathbf{X_A}, \mathbf{X_B}$ and $\mathbf{X_C}$ are the matrices having the observed data variations due to the individual effects of the different factors, whereas the other matrices ($\mathbf{X_{AB}}, \mathbf{X_{BC}}$ and $\mathbf{X_{ABC}}$) are the matrices having the variations caused by the interactions of the factors involved. $\mathbf{E}$ is the residual matrix that contains the data variation not explained by the model.

Equation 9 describes how SCA is applied in every factor and interaction, as follows:

Equation 9 $$\mathbf{X} = \bar{\mathbf{X}} + \mathbf{T_A}\boldsymbol{P}_A^T + \mathbf{T_B}\boldsymbol{P}_B^T + \mathbf{T_C}\boldsymbol{P}_C^T + \mathbf{T_{AB}}\boldsymbol{P}_{AB}^T + \mathbf{T_{AC}}\boldsymbol{P}_{AC}^T + \mathbf{T_{BC}}\boldsymbol{P}_{BC}^T + +\mathbf{T_{ABC}}\boldsymbol{P}_{ABC}^T + \mathbf{E}$$

Each $\mathbf{T}$ and $\mathbf{P}$ factor matrices gives the scores and loadings of the different factors and interactions. Thus, the significance of the effects of every single factor and their interactions can be assessed, and the statistical relevance can be estimated with a *p*-value calculated with a permutation test (normally between 1000 and 10000 permutations). The null hypothesis of the permutation test is that the factor does not produce any effect, and the alternative hypothesis is that the factor does have a significant experimental effect [305]. However, the main limitation of ASCA is that it

considers no correlation between the variables, which is not true in metabolomic studies, where the different variables are metabolites that can be associated with the same metabolic pathways.

Regularized MANOVA or rMANOVA [317] is described as the weighted average between ASCA and MANOVA. rMANOVA establishes an optimal regulation factor (ranged between 0 and 1). If this factor is equal to 0, then the model is the same as a MANOVA model, but if it is equal to 1, the result is an ASCA model. rMANOVA models are placed in an intermediate situation (between o and 1), which is the most commonly encountered scenario. In practice, if compared to ASCA, rMANOVA assumes that variables can correlate (which provides a more realistic scenario). Compared to MANOVA, rMANOVA models are applicable when there are less samples than variables (the common case in metabolomics). The rMANOVA factors and interactions significance is also calculated with a permutation test, similarly to ASCA.

Group-wise ASCA or GASCA [318] is a sparse implementation of ASCA, which means that can be applied in the presence of a large number of variables that do not present a response for some of samples (e.g., metabolites found in controls but not in treated samples, or *viceversa*) and, in the case, when variables are correlated. In addition, GASCA only considers the significant variables to be included in the final model, instead of considering all variables, like ASCA. GASCA replaces the PCA step of ASCA by group-wise PCA (GPCA), which simplifies the interpretation. Finally, significance of the factors and interactions is also evaluated by means of permutation tests.

Table 2.7. Comparison of univariate, multivariate and classification methods regarding their ability to determine the statistical significance of factors from DOE considering all variables at the same time, and the significant variables associated with these factors from DOE.

Statistical methods		Statistical significance of factors from DOE for all variables simultaneously	Identification of significant variables associated with factors from DOE
Univariate methods	ANOVAs	-----	X
	T test	-----	X
Multivariate methods	ASCA	X	X
	rMANOVA	X	X
	GASCA	X	X
Classification methods	PLS-DA - VIPs	-----	X

References

[1] E.H. Perspectives, What ' s Happening Downstream of DNA, Environmental Health. 112 (2004) 411–415.

[2] O. Fiehn, Metabolomics - The link between genotypes and phenotypes, Plant Molecular Biology. 48 (2002) 155–171. https://doi.org/10.1023/A:1013713905833.

[3] T. Jendoubi, Approaches to integrating metabolomics and multi-omics data: A primer, Metabolites. 11 (2021). https://doi.org/10.3390/metabo11030184.

[4] M.R. Belhaj, N.G. Lawler, N.J. Hoffman, Metabolomics and lipidomics: Expanding the molecular landscape of exercise biology, Metabolites. 11 (2021). https://doi.org/10.3390/metabo11030151.

[5] B.C. Muthubharathi, T. Gowripriya, K. Balamurugan, Metabolomics: small molecules that matter more, Molecular Omics. 17 (2021) 210–229. https://doi.org/10.1039/d0mo00176g.

[6] M. Jacob, A.L. Lopata, M. Dasouki, A.M. Abdel Rahman, Metabolomics toward personalized medicine, Mass Spectrometry Reviews. 38 (2019) 221–238. https://doi.org/10.1002/mas.21548.

[7] M. Szeremeta, K. Pietrowska, A. Niemcunowicz-Janica, A. Kretowski, M. Ciborowski, Applications of metabolomics in forensic toxicology and forensic medicine, International Journal of Molecular Sciences. 22 (2021) 1–16. https://doi.org/10.3390/ijms22063010.

[8] E. Olesti, V. González-Ruiz, M.F. Wilks, J. Boccard, S. Rudaz, Approaches in metabolomics for regulatory toxicology applications, Analyst. 146 (2021) 1820–1834. https://doi.org/10.1039/d0an02212h.

[9] M. Hernández-Mesa, B. le Bizec, G. Dervilly, Metabolomics in chemical risk analysis – A review, Analytica Chimica Acta. 1154 (2021). https://doi.org/10.1016/j.aca.2021.338298.

[10] B.P. Lankadurai, E.G. Nagato, M.J. Simpson, Environmental metabolomics: An emerging approach to study organism responses to environmental stressors, Environmental Reviews. 21 (2013) 180–205. https://doi.org/10.1139/er-2013-0011.

[11] M.R. Viant, Applications of metabolomics to the environmental sciences, Metabolomics. 5 (2009) 1–2. https://doi.org/10.1007/s11306-009-0157-3.

[12] M.C. Kido Soule, K. Longnecker, W.M. Johnson, E.B. Kujawinski, Environmental metabolomics: Analytical strategies, Marine Chemistry. 177 (2015) 374–387. https://doi.org/10.1016/j.marchem.2015.06.029.

[13] G.D. Tredwell, B. Edwards-Jones, D.J. Leak, J.G. Bundy, The development of metabolomic sampling procedures for Pichia pastoris, and baseline metabolome data, PLoS ONE. 6 (2011). https://doi.org/10.1371/journal.pone.0016286.

[14] J.F. Xiao, B. Zhou, H.W. Ressom, Metabolite identification and quantitation in LC-MS/MS-based metabolomics, TrAC - Trends in Analytical Chemistry. 32 (2012) 1–14. https://doi.org/10.1016/j.trac.2011.08.009.

[15] H.E. Johnson, D. Broadhurst, D.B. Kell, M.K. Theodorou, R.J. Merry, G.W. Griffith, High-Throughput Metabolic Fingerprinting of Legume Silage Fermentations via Fourier Transform Infrared Spectroscopy and Chemometrics, Applied and Environmental Microbiology. 70 (2004) 1583–1592. https://doi.org/10.1128/AEM.70.3.1583-1592.2004.

[16] A.H. Emwas, R. Roy, R.T. McKay, L. Tenori, E. Saccenti, G.A. Nagana Gowda, D. Raftery, F. Alahmari, L. Jaremko, M. Jaremko, D.S. Wishart, Nmr spectroscopy for metabolomics research, Metabolites. 9 (2019). https://doi.org/10.3390/metabo9070123.

[17] K. Sasaki, H. Sagawa, M. Suzuki, H. Yamamoto, M. Tomita, T. Soga, Y. Ohashi, Metabolomics Platform with Capillary Electrophoresis Coupled with High-Resolution Mass Spectrometry for Plasma Analysis, Analytical Chemistry. 91 (2019) 1295–1301. https://doi.org/10.1021/acs.analchem.8b02994.

[18] G. Paglia, A.J. Smith, G. Astarita, Ion mobility mass spectrometry in the omics era: Challenges and opportunities for metabolomics and lipidomics, Mass Spectrometry Reviews. (2021). https://doi.org/10.1002/mas.21686.

[19] M.E. Dueñas, E.A. Larson, Y.J. Lee, Toward mass spectrometry imaging in the metabolomics scale: Increasing metabolic coverage through multiple on-tissue chemical modifications, Frontiers in Plant Science. 10 (2019) 1–11. https://doi.org/10.3389/fpls.2019.00860.

[20] M.J. Taylor, J.K. Lukowski, C.R. Anderton, Spatially Resolved Mass Spectrometry at the Single Cell: Recent Innovations in Proteomics and Metabolomics, J Am Soc Mass Spectrom. 32 (2021) 872–894. https://doi.org/10.1021/jasms.0c00439.

[21] D. Miura, Y. Fujimura, H. Wariishi, In situ metabolomic mass spectrometry imaging: Recent advances and difficulties, Journal of Proteomics. 75 (2012) 5052–5060. https://doi.org/10.1016/j.jprot.2012.02.011.

[22] D.Y. Lee, B.P. Bowen, T.R. Northen, Mass spectrometry-based metabolomics, analysis of metabolite-protein interactions, and imaging, Biotechniques. 49 (2010) 557–565. https://doi.org/10.2144/000113451.
[23] M.S. Monteiro, M. Carvalho, M.L. Bastos, P. Guedes de Pinho, Metabolomics Analysis for Biomarker Discovery: Advances and Challenges, Current Medicinal Chemistry. 20 (2013) 257–271. https://doi.org/10.2174/092986713804806621.
[24] R. Wehrens, J.A. Hageman, F. van Eeuwijk, R. Kooke, P.J. Flood, E. Wijnker, J.J.B. Keurentjes, A. Lommen, H.D.L.M. van Eekelen, R.D. Hall, R. Mumm, R.C.H. de Vos, Improved batch correction in untargeted MS-based metabolomics, Metabolomics. 12 (2016). https://doi.org/10.1007/s11306-016-1015-8.
[25] R. Tautenhahn, G.J. Patti, D. Rinehart, G. Siuzdak, XCMS online: A web-based platform to process untargeted metabolomic data, Analytical Chemistry. 84 (2012) 5035–5039. https://doi.org/10.1021/ac300698c.
[26] Y. Guitton, M. Tremblay-Franco, G. le Corguillé, J.F. Martin, M. Pétéra, P. Roger-Mele, A. Delabrière, S. Goulitquer, M. Monsoor, C. Duperier, C. Canlet, R. Servien, P. Tardivel, C. Caron, F. Giacomoni, E.A. Thévenot, 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, International Journal of Biochemistry and Cell Biology. 93 (2017) 89–101. https://doi.org/10.1016/j.biocel.2017.07.002.
[27] T. Pluskal, S. Castillo, A. Villar-Briones, M. Orešič, MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data, BMC Bioinformatics. 11 (2010). https://doi.org/10.1186/1471-2105-11-395.
[28] and M.A. Hiroshi Tsugawa, Tomas Cajka, Tobias Kind, Yan Ma, Brendan Higgins, Kazutaka Ikeda, Mitsuhiro Kanazawa, Jean VanderGheynst, Oliver Fiehn, MS-DIAL: Data Independent MS/MS Deconvolution for Comprehensive, Nat Methods. 12 (2015) 523–526. https://doi.org/10.1038/nmeth.3393.MS-DIAL.
[29] H. Tsugawa, K. Ikeda, M. Takahashi, A. Satoh, Y. Mori, H. Uchino, N. Okahashi, Y. Yamada, I. Tada, P. Bonini, Y. Higashi, Y. Okazaki, Z. Zhou, Z.J. Zhu, J. Koelmel, T. Cajka, O. Fiehn, K. Saito, M. Arita, M. Arita, A lipidome atlas in MS-DIAL 4, Nature Biotechnology. 38 (2020) 1159–1163. https://doi.org/10.1038/s41587-020-0531-2.
[30] J. Xia, I. v. Sinelnikov, B. Han, D.S. Wishart, MetaboAnalyst 3.0-making metabolomics more meaningful, Nucleic Acids Research. 43 (2015) W251–W257. https://doi.org/10.1093/nar/gkv380.
[31] Z. Pang, J. Chong, G. Zhou, D.A. de Lima Morais, L. Chang, M. Barrette, C. Gauthier, P.É. Jacques, S. Li, J. Xia, MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights, Nucleic Acids Research. 49 (2021) W388–W396. https://doi.org/10.1093/NAR/GKAB382.
[32] H. Tsugawa, Advances in computational metabolomics and databases deepen the understanding of metabolisms, Current Opinion in Biotechnology. 54 (2018) 10–17. https://doi.org/10.1016/j.copbio.2018.01.008.
[33] Directive 2010/63/EU of the European Parliament, (n.d.). https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32010L0063 (accessed April 13, 2022).
[34] U. Strähle, S. Scholz, R. Geisler, P. Greiner, H. Hollert, S. Rastegar, A. Schumacher, I. Selderslaghs, C. Weiss, H. Witters, T. Braunbeck, Zebrafish embryos as an alternative to animal experiments-A commentary on the definition of the onset of protected life stages in animal welfare regulations, Reproductive Toxicology. 33 (2012) 128–132. https://doi.org/10.1016/j.reprotox.2011.06.121.
[35] S. Fields, M. Johnston, Whither model organism research?, Science (1979). 307 (2005) 1885–1886. https://doi.org/10.1126/SCIENCE.1108872/ASSET/E624AAFF-B89C-47DD-8EA5-5CDEC54CB737/ASSETS/SCIENCE.1108872.FP.PNG.
[36] A.S. Edison, R.D. Hall, C. Junot, P.D. Karp, I.J. Kurland, R. Mistrik, L.K. Reed, K. Saito, R.M. Salek, C. Steinbeck, L.W. Sumner, M.R. Viant, The time is right to focus on model organism metabolomes, Metabolites. 6 (2016). https://doi.org/10.3390/metabo6010008.
[37] D. Botstein, G.R. Fink, Yeast: An experimental organism for 21st century biology, Genetics. 189 (2011) 695–704. https://doi.org/10.1534/genetics.111.130765.
[38] M. Ramirez-Gaona, A. Marcu, A. Pon, A.C. Guo, T. Sajed, N.A. Wishart, N. Karu, Y.D. Feunang, D. Arndt, D.S. Wishart, YMDB 2.0: A significantly expanded version of the yeast metabolome database, Nucleic Acids Research. 45 (2017) D440–D445. https://doi.org/10.1093/nar/gkw1058.
[39] L.S. Roca, A.F.G. Gargano, P.J. Schoenmakers, Development of comprehensive two-dimensional low-flow liquid-chromatography setup coupled to high-resolution mass spectrometry for shotgun proteomics, Analytica Chimica Acta. 1156 (2021) 338349. https://doi.org/10.1016/j.aca.2021.338349.
[40] H. Schoeny, E. Rampler, Y. el Abiead, F. Hildebrand, O. Zach, G. Hermann, G. Koellensperger, A combined flow injection/reversed-phase chromatography-high-resolution mass spectrometry

workflow for accurate absolute lipid quantification with 13C internal standards, Analyst. 146 (2021) 2591–2599. https://doi.org/10.1039/d0an02443k.

[41] R. Spence, G. Gerlach, C. Lawrence, C. Smith, The behaviour and ecology of the zebrafish, Danio rerio, Biological Reviews. 83 (2008) 13–34. https://doi.org/10.1111/j.1469-185X.2007.00030.x.

[42] K. Howe, M.D. Clark, C.F. Torroja, J. Torrance, C. Berthelot, M. Muffato, J.E. Collins, S. Humphray, K. McLaren, L. Matthews, S. McLaren, I. Sealy, M. Caccamo, C. Churcher, C. Scott, J.C. Barrett, R. Koch, G.J. Rauch, S. White, W. Chow, B. Kilian, L.T. Quintais, J.A. Guerra-Assunção, Y. Zhou, Y. Gu, J. Yen, J.H. Vogel, T. Eyre, S. Redmond, R. Banerjee, J. Chi, B. Fu, E. Langley, S.F. Maguire, G.K. Laird, D. Lloyd, E. Kenyon, S. Donaldson, H. Sehra, J. Almeida-King, J. Loveland, S. Trevanion, M. Jones, M. Quail, D. Willey, A. Hunt, J. Burton, S. Sims, K. McLay, B. Plumb, J. Davis, C. Clee, K. Oliver, R. Clark, C. Riddle, D. Eliott, G. Threadgold, G. Harden, D. Ware, B. Mortimer, G. Kerry, P. Heath, B. Phillimore, A. Tracey, N. Corby, M. Dunn, C. Johnson, J. Wood, S. Clark, S. Pelan, G. Griffiths, M. Smith, R. Glithero, P. Howden, N. Barker, C. Stevens, J. Harley, K. Holt, G. Panagiotidis, J. Lovell, H. Beasley, C. Henderson, D. Gordon, K. Auger, D. Wright, J. Collins, C. Raisen, L. Dyer, K. Leung, L. Robertson, K. Ambridge, D. Leongamornlert, S. McGuire, R. Gilderthorp, C. Griffiths, D. Manthravadi, S. Nichol, G. Barker, S. Whitehead, M. Kay, J. Brown, C. Murnane, E. Gray, M. Humphries, N. Sycamore, D. Barker, D. Saunders, J. Wallis, A. Babbage, S. Hammond, M. Mashreghi-Mohammadi, L. Barr, S. Martin, P. Wray, A. Ellington, N. Matthews, M. Ellwood, R. Woodmansey, G. Clark, J. Cooper, A. Tromans, D. Grafham, C. Skuce, R. Pandian, R. Andrews, E. Harrison, A. Kimberley, J. Garnett, N. Fosker, R. Hall, P. Garner, D. Kelly, C. Bird, S. Palmer, I. Gehring, A. Berger, C.M. Dooley, Z. Ersan-Ürün, C. Eser, H. Geiger, M. Geisler, L. Karotki, A. Kirn, J. Konantz, M. Konantz, M. Oberländer, S. Rudolph-Geiger, M. Teucke, K. Osoegawa, B. Zhu, A. Rapp, S. Widaa, C. Langford, F. Yang, N.P. Carter, J. Harrow, Z. Ning, J. Herrero, S.M.J. Searle, A. Enright, R. Geisler, R.H.A. Plasterk, C. Lee, M. Westerfield, P.J. de Jong, L.I. Zon, J.H. Postlethwait, C. Nüsslein-Volhard, T.J.P. Hubbard, H.R. Crollius, J. Rogers, D.L. Stemple, The zebrafish reference genome sequence and its relationship to the human genome, Nature. 496 (2013) 498–503. https://doi.org/10.1038/nature12111.

[43] D. Raldúa, B. Thienpont, P.J. Babin, Zebrafish eleutheroembryos as an alternative system for screening chemicals disrupting the mammalian thyroid gland morphogenesis and function, Reproductive Toxicology. 33 (2012) 188–197. https://doi.org/10.1016/j.reprotox.2011.09.001.

[44] G.R. Garcia, P.D. Noyes, R.L. Tanguay, Advancements in zebrafish applications for 21st century toxicology, Pharmacology and Therapeutics. 161 (2016) 11–21. https://doi.org/10.1016/j.pharmthera.2016.03.009.

[45] C. Parng, W.L. Seng, C. Semino, P. McGrath, Zebrafish: a preclinical model for drug screening., Assay Drug Dev Technol. 1 (2002) 41–48. https://doi.org/10.1089/154065802761001293.

[46] H. Sukardi, H.T. Chng, E.C.Y. Chan, Z. Gong, S.H. Lam, Zebrafish for drug toxicity screening: Bridging the in vitro cell-based models and in vivo mammalian models, Expert Opinion on Drug Metabolism and Toxicology. 7 (2011) 579–589. https://doi.org/10.1517/17425255.2011.562197.

[47] M. Hölttä-Vuori, V.T.V. Salo, L. Nyberg, C. Brackmann, A. Enejder, P. Panula, E. Ikonen, Zebrafish: Gaining popularity in lipid research, Biochemical Journal. 429 (2010) 235–242. https://doi.org/10.1042/BJ20100293.

[48] H. Löhr, M. Hammerschmidt, Zebrafish in endocrine systems: Recent advances and implications for human disease, Annual Review of Physiology. 73 (2011) 183–211. https://doi.org/10.1146/annurev-physiol-012110-142320.

[49] S. Jarque, J. Ibarra, M. Rubio-Brotons, J. García-Fernández, J. Terriente, Multiplex analysis platform for endocrine disruption prediction using zebrafish, International Journal of Molecular Sciences. 20 (2019). https://doi.org/10.3390/ijms20071739.

[50] S. Bertoli, A. Leone, A. Battezzati, Human bisphenol a exposure and the “diabesity phenotype,” Dose-Response. 13 (2015). https://doi.org/10.1177/1559325815599173.

[51] P. Mirmira, C. Evans-Molina, Bisphenol A, obesity, and type 2 diabetes mellitus: Genuine concern or unnecessary preoccupation?, Translational Research. 164 (2014) 13–21. https://doi.org/10.1016/j.trsl.2014.03.003.

[52] S. Siddique, G. Zhang, C. Kubwabo, Exposure to bisphenol a and risk of developing type 2 diabetes: A mini review, Emerging Contaminants. 6 (2020) 274–282. https://doi.org/10.1016/j.emcon.2020.07.005.

[53] C.B. Kimmel, W.W. Ballard, S.R. Kimmel, B. Ullmann, T.F. Schilling, Stages of embryonic development of the zebrafish, Developmental Dynamics. 203 (1995) 253–310. https://doi.org/10.1002/aja.1002030302.

[54] ZFIN - The Zebrafish Information Networ, (n.d.). https://zfin.org/ (accessed April 13, 2022).

[55] P. Borrill, Blurring the boundaries between cereal crops and model plants, New Phytologist. 228 (2020) 1721–1727. https://doi.org/10.1111/nph.16229.

[56] T. Sasaki, B. Burr, International Rice Genome Sequencing Project: The effort to completely sequence the rice genome, Current Opinion in Plant Biology. 3 (2000) 138–142. https://doi.org/10.1016/S1369-5266(99)00047-3.
[57] R.M. Welch, R.D. Graham, Breeding for micronutrients in staple food crops from a human nutrition perspective, Journal of Experimental Botany. 55 (2004) 353–364. https://doi.org/10.1093/jxb/erh064.
[58] W.J. Hong, Y.J. Kim, A.K.N. Chandran, K.H. Jung, Infrastructures of systems biology that facilitate functional genomic study in rice, Rice. 12 (2019). https://doi.org/10.1186/s12284-019-0276-z.
[59] A. Oikawa, F. Matsuda, M. Kusano, Y. Okazaki, K. Saito, Rice metabolomics, Rice. 1 (2008) 63–71. https://doi.org/10.1007/s12284-008-9009-4.
[60] U. Uawisetwathana, N. Karoonuthaisiri, Metabolomics for rice quality and traceability: feasibility and future aspects, Current Opinion in Food Science. 28 (2019) 58–66. https://doi.org/10.1016/j.cofs.2019.08.008.
[61] N. Rana, M.S. Rahim, G. Kaur, R. Bansal, S. Kumawat, J. Roy, R. Deshmukh, H. Sonah, T.R. Sharma, Applications and challenges for efficient exploration of omics interventions for the enhancement of nutritional quality in rice (Oryza sativa L.), Critical Reviews in Food Science and Nutrition. 60 (2020) 3304–3320. https://doi.org/10.1080/10408398.2019.1685454.
[62] K. Thi, X. Vo, M. Rahman, M. Rahman, T. Trinh, S.T. Kim, J.-S. Jeon, Proteomics and Metabolomics Studies on the Biotic Stress Responses of Rice: an Update, (n.d.). https://doi.org/10.1186/s12284-021-00461-4.
[63] F.R. Castro-moretti, I.N. Gentzel, D. Mackey, A.P. Alonso, Metabolomics as an emerging tool for the study of plant–pathogen interactions, Metabolites. 10 (2020) 1–23. https://doi.org/10.3390/metabo10020052.
[64] S. Schaarschmidt, L.M.F. Lawas, J. Kopka, S.V.K. Jagadish, E. Zuther, Physiological and molecular attributes contribute to high night temperature tolerance in cereals, Plant Cell and Environment. 44 (2021) 2034–2048. https://doi.org/10.1111/pce.14055.
[65] N. Fàbregas, A.R. Fernie, The metabolic response to drought, Journal of Experimental Botany. 70 (2019) 1077–1085. https://doi.org/10.1093/jxb/ery437.
[66] S.A. Ganie, K.A. Molla, R.J. Henry, K. v. Bhat, T.K. Mondal, Advances in understanding salt tolerance in rice, Theoretical and Applied Genetics. 132 (2019) 851–870. https://doi.org/10.1007/s00122-019-03301-8.
[67] X. Wu, Y. Liu, S. Yin, K. Xiao, Q. Xiong, S. Bian, S. Liang, H. Hou, J. Hu, J. Yang, Metabolomics revealing the response of rice (Oryza sativa L.) exposed to polystyrene microplastics, Environmental Pollution. 266 (2020) 115159. https://doi.org/10.1016/j.envpol.2020.115159.
[68] V. Mahdavi, M.M. Farimani, F. Fathi, A. Ghassempour, A targeted metabolomics approach toward understanding metabolic variations in rice under pesticide stress, Analytical Biochemistry. 478 (2015) 65–72. https://doi.org/10.1016/j.ab.2015.02.021.
[69] N. Arif, N.C. Sharma, V. Yadav, N. Ramawat, N.K. Dubey, D.K. Tripathi, D.K. Chauhan, S. Sahi, Understanding Heavy Metal Stress in a Rice Crop: Toxicity, Tolerance Mechanisms, and Amelioration Strategies, Journal of Plant Biology. 62 (2019) 239–253. https://doi.org/10.1007/s12374-019-0112-4.
[70] K.K. Rai, N. Pandey, R.P. Meena, S.P. Rai, Biotechnological strategies for enhancing heavy metal tolerance in neglected and underutilized legume crops: A comprehensive review, Ecotoxicology and Environmental Safety. 208 (2021) 111750. https://doi.org/10.1016/j.ecoenv.2020.111750.
[71] S. Kumar, R.S. Dubey, R.D. Tripathi, D. Chakrabarty, P.K. Trivedi, Omics and biotechnology of arsenic stress and detoxification in plants: Current updates and prospective, Environment International. 74 (2015) 221–230. https://doi.org/10.1016/j.envint.2014.10.019.
[72] C. Hu, J. Shi, S. Quan, B. Cui, S. Kleessen, Z. Nikoloski, T. Tohge, D. Alexander, L. Guo, H. Lin, J. Wang, X. Cui, J. Rao, Q. Luo, X. Zhao, A.R. Fernie, D. Zhang, Metabolic variation between japonica and indica rice cultivars as revealed by non-targeted metabolomics, Scientific Reports. 4 (2014). https://doi.org/10.1038/srep05067.
[73] P. Mirabelli, L. Coppola, M. Salvatore, Cancer cell lines are useful model systems for medical research, Cancers (Basel). 11 (2019). https://doi.org/10.3390/cancers11081098.
[74] S. Moscato, F. Ronca, D. Campani, S. Danti, Poly(vinyl alcohol)/gelatin Hydrogels Cultured with HepG2 Cells as a 3D Model of Hepatocellular Carcinoma: A Morphological Study, Journal of Functional Biomaterials. 6 (2015) 16–32. https://doi.org/10.3390/jfb6010016.
[75] OECD/OCDE, OECD guideline for testing of chemicals - Daphnia sp., acute immobilisation test and reproduction test., Oecd. (2004). http://www.oecd.org/chemicalsafety/testing/46507550.pdf.
[76] D. Krewski, D. Acosta Jr, M. Andersen, H. Anderson, J.C. Bailar III, K. Boekelheide, R. Brent, G. Charnley, V.G. Cheung, S. Green Jr, K.T. Kelsey, N.I. Kerkvliet, A.A. Li, L. McCray, O. Meyer, TOXICITY TESTING IN THE 21ST CENTURY: A VISION AND A STRATEGY Staff of Committee on Toxicity Testing and Assessment of Environmental Agents 1 R. Samuel McLaughlin Centre for

Population Health Risk Assessment, Institute of Population HHS Public Access, Toxicology and Environmental Health. 13 (2010) 51–138. https://doi.org/10.1080/10937404.2010.483176.TOXICITY.

[77] N. Geng, H. Zhang, B. Zhang, P. Wu, F. Wang, Z. Yu, J. Chen, Effects of short-chain chlorinated paraffins exposure on the viability and metabolism of human hepatoma HepG2 cells, Environmental Science and Technology. 49 (2015) 3076–3083. https://doi.org/10.1021/es505802x.

[78] N.B. Javitt, Hep G2 cells as a resource for metabolic studies: lipoprotein, cholesterol, and bile acids, FASEB J. 4 (1990) 161–168. https://doi.org/10.1096/FASEBJ.4.2.2153592.

[79] F. Li, L. Cao, S. Parikh, R. Zuo, Three-Dimensional Spheroids With Primary Human Liver Cells and Differential Roles of Kupffer Cells in Drug-Induced Liver Injury, Journal of Pharmaceutical Sciences. 109 (2020) 1912–1923. https://doi.org/10.1016/j.xphs.2020.02.021.

[80] H. Ramírez-Malule, D.H. Quiñones-Murillo, D. Manotas-Duque, Emerging contaminants as global environmental hazards. A bibliometric analysis, Emerging Contaminants. 6 (2020) 179–193. https://doi.org/10.1016/j.emcon.2020.05.001.

[81] T. Rasheed, M. Bilal, F. Nabeel, M. Adeel, H.M.N. Iqbal, Environmentally-related contaminants of high concern: Potential sources and analytical modalities for detection, quantification, and treatment, Environment International. 122 (2019) 52–66. https://doi.org/10.1016/j.envint.2018.11.038.

[82] J.K. Saha, R. Selladurai, S. Kundu, A.K. Patra, Water, Agriculture, Soil and Environmental, 1989. https://doi.org/10.1016/S0167-9244(08)70239-7.

[83] M. Perrett, B. Sivarajah, C.L. Cheney, J.B. Korosi, L. Kimpe, J.M. Blais, J.P. Smol, Impacts on aquatic biota from salinization and metalloid contamination by gold mine tailings in sub-Arctic lakes, Environmental Pollution. 278 (2021) 116815. https://doi.org/10.1016/j.envpol.2021.116815.

[84] C.L. Cheney, K.M. Eccles, L.E. Kimpe, J.R. Thienpont, J.B. Korosi, J.M. Blais, Determining the effects of past gold mining using a sediment palaeotoxicity model, Science of the Total Environment. 718 (2020) 137308. https://doi.org/10.1016/j.scitotenv.2020.137308.

[85] J.J. Aristizabal-Henao, A. Ahmadireskety, E.K. Griffin, B. Ferreira Da Silva, J.A. Bowden, Lipidomics and environmental toxicology: Recent trends, Current Opinion in Environmental Science and Health. 15 (2020) 26–31. https://doi.org/10.1016/j.coesh.2020.04.004.

[86] WHO - Arsenic, (n.d.). https://www.who.int/news-room/fact-sheets/detail/arsenic (accessed April 13, 2022).

[87] S. Shankar, U. Shanker, Shikha, Arsenic contamination of groundwater: A review of sources, prevalence, health risks, and strategies for mitigation, Scientific World Journal. 2014 (2014). https://doi.org/10.1155/2014/304524.

[88] J. Bundschuh, J. Schneider, M.A. Alam, N.K. Niazi, I. Herath, F. Parvez, B. Tomaszewska, L.R.G. Guilherme, J.P. Maity, D.L. López, A.F. Cirelli, A. Pérez-Carrera, N. Morales-Simfors, M.T. Alarcón-Herrera, P. Baisch, D. Mohan, A. Mukherjee, Seven potential sources of arsenic pollution in Latin America and their environmental and health impacts, Science of the Total Environment. 780 (2021). https://doi.org/10.1016/j.scitotenv.2021.146274.

[89] K.C. Saha, Diagnosis of arsenicosis, Journal of Environmental Science and Health - Part A Toxic/Hazardous Substances and Environmental Engineering. 38 (2003) 255–272. https://doi.org/10.1081/ESE-120016893.

[90] G. Sun, Arsenic contamination and arsenicosis in China, Toxicology and Applied Pharmacology. 198 (2004) 268–271. https://doi.org/10.1016/j.taap.2003.10.017.

[91] E. Shaji, M. Santosh, K. v. Sarath, P. Prakash, V. Deepchand, B. v. Divya, Arsenic contamination of groundwater: A global synopsis with focus on the Indian Peninsula, Geoscience Frontiers. 12 (2021) 101079. https://doi.org/10.1016/j.gsf.2020.08.015.

[92] FDA, Arsenic in Rice and Rice Products Risk Assessment Report, Center for Food Safety and Applied Nutrition of the Food and Drug Administration. 1 (2016) 1–284.

[93] H.M. Anawar, J. Akai, K.M.G. Mostofa, S. Safiullah, S.M. Tareq, Arsenic poisoning in groundwater: Health risk and geochemical sources in Bangladesh, Environment International. 27 (2002) 597–604. https://doi.org/10.1016/S0160-4120(01)00116-7.

[94] L. Ramsay, M.M. Petersen, B. Hansen, J. Schullehner, P. van der Wens, D. Voutchkova, S.M. Kristiansen, Drinking Water Criteria for Arsenic in High-Income, Low-Dose Countries: The Effect of Legislation on Public Health, Environmental Science and Technology. 55 (2021) 3483–3493. https://doi.org/10.1021/acs.est.0c03974.

[95] F.J. Zhao, S.P. McGrath, A.A. Meharg, Arsenic as a food chain contaminant: Mechanisms of plant uptake and metabolism and mitigation strategies, Annual Review of Plant Biology. 61 (2010) 535–559. https://doi.org/10.1146/annurev-arplant-042809-112152.

[96] J. feng Gu, H. Zhou, H. ling Tang, W. tao Yang, M. Zeng, Z. ming Liu, P. qin Peng, B. han Liao, Cadmium and arsenic accumulation during the rice growth period under in situ remediation,

Ecotoxicology and Environmental Safety. 171 (2019) 451–459. https://doi.org/10.1016/j.ecoenv.2019.01.003.
[97] L.D.B. Suriyagoda, K. Dittert, H. Lambers, Mechanism of arsenic uptake, translocation and plant resistance to accumulate arsenic in rice grains, Agriculture, Ecosystems and Environment. 253 (2018) 23–37. https://doi.org/10.1016/j.agee.2017.10.017.
[98] P. Kumarathilaka, S. Seneweera, Y.S. Ok, A. Meharg, J. Bundschuh, Arsenic in cooked rice foods: Assessing health risks and mitigation options, Environment International. 127 (2019) 584–591. https://doi.org/10.1016/j.envint.2019.04.004.
[99] A. Roel, F. Campos, M. Verger, R. Huertas, G. Carracelas, Regional variability of arsenic content in Uruguayan polished rice, Chemosphere. 288 (2022) 132426. https://doi.org/10.1016/j.chemosphere.2021.132426.
[100] C. Casals-Casas, B. Desvergne, Endocrine disruptors: From endocrine to metabolic disruption, Annual Review of Physiology. 73 (2011) 135–162. https://doi.org/10.1146/annurev-physiol-012110-142200.
[101] A. Pereira-Fernandes, C. Vanparys, L. Vergauwen, D. Knapen, P.G. ermaines Jorens, R. Blust, Toxicogenomics in the 3T3-L1 cell line, a new approach for screening of obesogenic compounds, Toxicol Sci. 140 (2014) 352–363. https://doi.org/10.1093/toxsci/kfu092.
[102] J.J. Heindel, B. Blumberg, M. Cave, R. Machtinger, A. Mantovani, M.A. Mendez, A. Nadal, P. Palanza, G. Panzica, R. Sargis, L.N. Vandenberg, F. vom Saal, Metabolism disrupting chemicals and metabolic disorders, Reproductive Toxicology. 68 (2017) 3–33. https://doi.org/10.1016/j.reprotox.2016.10.001.
[103] K. Yoon, S.J. Kwack, H.S. Kim, B.M. Lee, Estrogenic endocrine-disrupting chemicals: Molecular mechanisms of actions on putative human diseases, Journal of Toxicology and Environmental Health - Part B: Critical Reviews. 17 (2014) 127–174. https://doi.org/10.1080/10937404.2014.882194.
[104] L. Mao, S. Fang, M. Zhao, W. Liu, H. Jin, Effects of Bisphenol A and Bisphenol S Exposure at Low Doses on the Metabolome of Adolescent Male Sprague-Dawley Rats, Chemical Research in Toxicology. 34 (2021) 1578–1587. https://doi.org/10.1021/acs.chemrestox.1c00018.
[105] S. Yue, J. Yu, Y. Kong, H. Chen, M. Mao, C. Ji, S. Shao, J. Zhu, J. Gu, M. Zhao, Metabolomic modulations of HepG2 cells exposed to bisphenol analogues, Environment International. 129 (2019) 59–67. https://doi.org/10.1016/j.envint.2019.05.008.
[106] J. Corrales, L.A. Kristofco, W. Baylor Steele, B.S. Yates, C.S. Breed, E. Spencer Williams, B.W. Brooks, Global assessment of bisphenol a in the environment: Review and analysis of its occurrence and bioaccumulation, Dose-Response. 13 (2015) 1–29. https://doi.org/10.1177/1559325815598308.
[107] ECHA - BPA, (n.d.). https://echa.europa.eu/es/substance-information/-/substanceinfo/100.001.133 (accessed April 13, 2022).
[108] J. Michałowicz, Bisphenol A - Sources, toxicity and biotransformation, Environmental Toxicology and Pharmacology. 37 (2014) 738–758. https://doi.org/10.1016/j.etap.2014.02.003.
[109] C.F.V. Scopel, C. Sousa, M.R.F. Machado, W.G. dos Santos, Bpa toxicity during development of zebrafish embryo, Brazilian Journal of Biology. 81 (2021) 437–447. https://doi.org/10.1590/1519-6984.230562.
[110] Y. Han, Y. Fei, M. Wang, Y. Xue, H. Chen, Y. Liu, Study on the Joint Toxicity of BPZ, BPS, BPC and BPF to Zebrafish, Molecules. 26 (2021). https://doi.org/10.3390/MOLECULES26144180.
[111] H. Tapiero, G. Nguyen Ba, K.D. Tew, Estrogens and environmental estrogens, Biomedicine and Pharmacotherapy. 56 (2002) 36–44. https://doi.org/10.1016/S0753-3322(01)00155-X.
[112] L.S. Shore, M. Shemesh, Estrogen as an Environmental Pollutant, Bulletin of Environmental Contamination and Toxicology. 97 (2016) 447–448. https://doi.org/10.1007/s00128-016-1873-9.
[113] M. Adeel, X. Song, Y. Wang, D. Francis, Y. Yang, Environmental impact of estrogens on human, animal and plant life: A critical review, Environment International. 99 (2017) 107–119. https://doi.org/10.1016/j.envint.2016.12.010.
[114] C.J. Martyniuk, R. Martínez, L. Navarro-Martín, J.H. Kamstra, A. Schwendt, S. Reynaud, L. Chalifour, Emerging concepts and opportunities for endocrine disruptor screening of the non-EATS modalities, Environmental Research. 204 (2022) 111904. https://doi.org/10.1016/j.envres.2021.111904.
[115] E. Gracia-Lor, J. v. Sancho, R. Serrano, F. Hernández, Occurrence and removal of pharmaceuticals in wastewater treatment plants at the Spanish Mediterranean area of Valencia, Chemosphere. 87 (2012) 453–462. https://doi.org/10.1016/j.chemosphere.2011.12.025.
[116] B.F. da Silva, A. Jelic, R. López-Serna, A.A. Mozeto, M. Petrovic, D. Barceló, Occurrence and distribution of pharmaceuticals in surface water, suspended solids and sediments of the Ebro river basin, Spain, Chemosphere. 85 (2011) 1331–1339. https://doi.org/10.1016/j.chemosphere.2011.07.051.

[117] L. Wiest, T. Chonova, A. Bergé, R. Baudot, F. Bessueille-Barbier, L. Ayouni-Derouiche, E. Vulliet, Two-year survey of specific hospital wastewater treatment and its impact on pharmaceutical discharges, Environmental Science and Pollution Research. 25 (2018) 9207–9218. https://doi.org/10.1007/s11356-017-9662-5.
[118] H. al Qarni, P. Collier, J. O'Keeffe, J. Akunna, Investigating the removal of some pharmaceutical compounds in hospital wastewater treatment plants operating in Saudi Arabia, Environmental Science and Pollution Research. 23 (2016) 13003–13014. https://doi.org/10.1007/s11356-016-6389-7.
[119] A. Shraim, A. Diab, A. Alsuhaimi, E. Niazy, M. Metwally, M. Amad, S. Sioud, A. Dawoud, Analysis of some pharmaceuticals in municipal wastewater of Almadinah Almunawarah, Arabian Journal of Chemistry. 10 (2017) S719–S729. https://doi.org/10.1016/j.arabjc.2012.11.014.
[120] Z. Xie, G. Lu, J. Liu, Z. Yan, B. Ma, Z. Zhang, W. Chen, Occurrence, bioaccumulation, and trophic magnification of pharmaceutically active compounds in Taihu Lake, China, Chemosphere. 138 (2015) 140–147. https://doi.org/10.1016/J.CHEMOSPHERE.2015.05.086.
[121] R.C. Pivetta, C. Rodrigues-Silva, A.R. Ribeiro, S. Rath, Tracking the occurrence of psychotropic pharmaceuticals in Brazilian wastewater treatment plants and surface water, with assessment of environmental risks, Science of the Total Environment. 727 (2020). https://doi.org/10.1016/J.SCITOTENV.2020.138661.
[122] R. Anjali, S. Shanthakumar, Insights on the current status of occurrence and removal of antibiotics in wastewater by advanced oxidation processes, Journal of Environmental Management. 246 (2019) 51–62. https://doi.org/10.1016/j.jenvman.2019.05.090.
[123] J.P. Fernandes, C.M.R. Almeida, M.A. Salgado, M.F. Carvalho, A.P. Mucha, Pharmaceutical compounds in aquatic environments— occurrence, fate and bioremediation prospective, Toxics. 9 (2021) 1–26. https://doi.org/10.3390/toxics9100257.
[124] S. Lacorte, C. Gómez-Canela, C. Calas-Blanchard, Pharmaceutical residues in senior residences wastewaters: High loads, emerging risks, Molecules. 26 (2021) 1–18. https://doi.org/10.3390/molecules26165047.
[125] C. Gómez-Canela, T. Sala-Comorera, V. Pueyo, C. Barata, S. Lacorte, Analysis of 44 pharmaceuticals consumed by elderly using liquid chromatography coupled to tandem mass spectrometry, Journal of Pharmaceutical and Biomedical Analysis. 168 (2019) 55–63. https://doi.org/10.1016/j.jpba.2019.02.016.
[126] M.E. Valdés, B. Huerta, D.A. Wunderlin, M.A. Bistoni, D. Barceló, S. Rodriguez-Mozaz, Bioaccumulation and bioconcentration of carbamazepine and other pharmaceuticals in fish under field and controlled laboratory experiments. Evidences of carbamazepine metabolization by fish, Science of The Total Environment. 557–558 (2016) 58–67. https://doi.org/10.1016/J.SCITOTENV.2016.03.045.
[127] S. Yan, M. Wang, J. Zha, L. Zhu, W. Li, Q. Luo, J. Sun, Z. Wang, Environmentally Relevant Concentrations of Carbamazepine Caused Endocrine-Disrupting Effects on Nontarget Organisms, Chinese Rare Minnows (Gobiocypris rarus), Environmental Science and Technology. 52 (2018) 886–894. https://doi.org/10.1021/ACS.EST.7B06476/SUPPL_FILE/ES7B06476_SI_001.PDF.
[128] J. Xin, S. Yan, X. Hong, H. Zhang, J. Zha, Environmentally relevant concentrations of carbamazepine induced lipid metabolism disorder of Chinese rare minnow (Gobiocypris rarus) in a gender-specific pattern, Chemosphere. 265 (2021) 129080. https://doi.org/10.1016/J.CHEMOSPHERE.2020.129080.
[129] E. Garreta-Lara, A. Checa, D. Fuchs, R. Tauler, S. Lacorte, C.E. Wheelock, C. Barata, Effect of psychiatric drugs on Daphnia magna oxylipin profiles, (2018). https://doi.org/10.1016/j.scitotenv.2018.06.333.
[130] I. Fuertes, B. Piña, C. Barata, Changes in lipid profiles in Daphnia magna individuals exposed to low environmental levels of neuroactive pharmaceuticals, Science of the Total Environment. 733 (2020) 139029. https://doi.org/10.1016/j.scitotenv.2020.139029.
[131] C. Gómez-Canela, V. Pueyo, C. Barata, S. Lacorte, R.M. Marcé, Development of predicted environmental concentrations to prioritize the occurrence of pharmaceuticals in rivers from Catalonia, Science of the Total Environment. 666 (2019) 57–67. https://doi.org/10.1016/j.scitotenv.2019.02.078.
[132] L. Meng, X. Li, X. Wang, K. Ma, G. Liu, J. Zhang, Amoxicillin effects on functional microbial community and spread of antibiotic resistance genes in amoxicillin manufacture wastewater treatment system, Journal of Environmental Sciences. 61 (2017) 110–117. https://doi.org/10.1016/J.JES.2017.09.020.
[133] R. Andreozzi, V. Caprio, C. Ciniglia, M. de Champdoré, R. lo Giudice, R. Marotta, E. Zuccato, Antibiotics in the environment: Occurrence in Italian STPs, fate, and preliminary assessment on algal toxicity of amoxicillin, Environmental Science and Technology. 38 (2004) 6832–6838. https://doi.org/10.1021/ES049509A/SUPPL_FILE/ES049509ASI20040324_101936.PDF.

[134] DrugBank - Trazodone, (n.d.). https://go.drugbank.com/drugs/DB00656.
[135] C.S. Voican, E. Corruble, S. Naveau, G. Perlemuter, Antidepressant-induced liver injury: A review for clinicians, American Journal of Psychiatry. 171 (2014) 404–415. https://doi.org/10.1176/appi.ajp.2013.13050709.
[136] Drugbank - Carbamazepine, (n.d.). https://go.drugbank.com/drugs/DB00564 (accessed April 14, 2022).
[137] FDA - DILIrank, (n.d.). https://www.fda.gov/science-research/liver-toxicity-knowledge-base-ltkb/drug-induced-liver-injury-rank-dilirank-dataset (accessed April 14, 2022).
[138] D.S. Wishart, A.C. Guo, E. Oler, F. Wang, A. Anjum, H. Peters, R. Dizon, Z. Sayeeda, S. Tian, B.L. Lee, M. Berjanskii, R. Mah, M. Yamamoto, J. Jovel, C. Torres-Calzada, M. Hiebert-Giesbrecht, V.W. Lui, D. Varshavi, D. Varshavi, D. Allen, D. Arndt, N. Khetarpal, A. Sivakumaran, K. Harford, S. Sanford, K. Yee, X. Cao, Z. Budinski, J. Liigand, L. Zhang, J. Zheng, R. Mandal, N. Karu, M. Dambrova, H.B. Schiöth, R. Greiner, V. Gautam, HMDB 5.0: the Human Metabolome Database for 2022, Nucleic Acids Res. 50 (2022) D622–D631. https://doi.org/10.1093/NAR/GKAB1062.
[139] D.S. Wishart, T. Jewison, A.C. Guo, M. Wilson, C. Knox, Y. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong, S. Bouatra, I. Sinelnikov, D. Arndt, J. Xia, P. Liu, F. Yallou, T. Bjorndahl, R. Perez-Pineiro, R. Eisner, F. Allen, V. Neveu, R. Greiner, A. Scalbert, HMDB 3.0-The Human Metabolome Database in 2013, Nucleic Acids Research. 41 (2013) 801–807. https://doi.org/10.1093/nar/gks1065.
[140] Smith, G. O'Maille, E.J. Want, C. Qin, S.A. Trauger, T.R. Brandon, D.E. Custodio, R. Abagyan, G. Siuzdak, METLIN: a metabolite mass spectral database, Ther Drug Monit. 27 (2005).
[141] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M.Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, T. Nishioka, MassBank: A public repository for sharing mass spectral data for life sciences, Journal of Mass Spectrometry. 45 (2010) 703–714. https://doi.org/10.1002/jms.1777.
[142] T. Schulze, R. Meier, N. Alygizakis, E. Schymanski, E. Bach, D.H. LI, lauperbe, raalizadeh, S. Tanaka, M. Witting, MassBank/MassBank-data: Release version 2021.12, (2021). https://doi.org/10.5281/ZENODO.5775684.
[143] E. Fahy, S. Subramaniam, R.C. Murphy, M. Nishijima, C.R.H. Raetz, T. Shimizu, F. Spener, G. van Meer, M.J.O. Wakelam, E.A. Dennis, Update of the LIPID MAPS comprehensive classification system for lipids, Journal of Lipid Research. 50 (2009) 9–14. https://doi.org/10.1194/jlr.R800095-JLR200.
[144] E. Fahy, M. Sud, D. Cotter, S. Subramaniam, LIPID MAPS online tools for lipid research, Nucleic Acids Research. 35 (2007). https://doi.org/10.1093/nar/gkm324.
[145] NIST - Standard reference database, (n.d.). https://www.nist.gov/srd/nist-standard-reference-database-1a) (accessed April 14, 2022).
[146] P.R. Haddad, M. Taraji, R. Szücs, Prediction of Analyte Retention Time in Liquid Chromatography, Analytical Chemistry. 93 (2021) 228–256. https://doi.org/10.1021/acs.analchem.0c04190.
[147] P. Bonini, T. Kind, H. Tsugawa, D.K. Barupal, O. Fiehn, Retip: Retention Time Prediction for Compound Annotation in Untargeted Metabolomics, Analytical Chemistry. 92 (2020) 7515–7522. https://doi.org/10.1021/acs.analchem.9b05765.
[148] B.C. Naylor, J. Leon Catrow, J. Alan Maschek, J.E. Cox, QSRR automator: A tool for automating retention time prediction in lipidomics and metabolomics, Metabolites. 10 (2020). https://doi.org/10.3390/metabo10060237.
[149] J. FOLCH, M. LEES, G.H. SLOANE STANLEY, A simple method for the isolation and purification of total lipides from animal tissues., J Biol Chem. 226 (1957) 497–509. https://doi.org/10.1016/s0021-9258(18)64849-5.
[150] W.J. Bligh, E.G. and Dyer, Canadian Journal of Biochemistry and Physiology, Canadian Journal of Biochemistry and Physiology. 37 (1959).
[151] T. Züllig, M. Trötzmüller, H.C. Köfeler, Lipidomics from sample preparation to data analysis: a primer, Analytical and Bioanalytical Chemistry. 412 (2020) 2191–2209. https://doi.org/10.1007/s00216-019-02241-y.
[152] R. Martínez, L. Navarro-Martín, M. van Antro, I. Fuertes, M. Casado, C. Barata, B. Piña, Changes in lipid profiles induced by bisphenol A (BPA) in zebrafish eleutheroembryos during the yolk sac absorption stage, Chemosphere. 246 (2020). https://doi.org/10.1016/j.chemosphere.2019.125704.
[153] N. Dalmau, J. Jaumot, R. Tauler, C. Bedia, Epithelial-to-mesenchymal transition involves triacylglycerol accumulation in DU145 prostate cancer cells, Mol. BioSyst. 11 (2015) 3397–3406. https://doi.org/10.1039/C5MB00413F.

[154] M. Navarro-Reig, J. Jaumot, R. Tauler, An untargeted lipidomic strategy combining comprehensive two-dimensional liquid chromatography and chemometric analysis, Journal of Chromatography A. 1568 (2018) 80–90. https://doi.org/10.1016/j.chroma.2018.07.017.
[155] M.A. López-Bascón, M. Calderón-Santiago, J. Sánchez-Ceinos, A. Fernández-Vega, R. Guzmán-Ruiz, J. López-Miranda, M.M. Malagon, F. Priego-Capote, Influence of sample preparation on lipidomics analysis of polar lipids in adipose tissue, Talanta. 177 (2018) 86–93. https://doi.org/10.1016/j.talanta.2017.09.017.
[156] N. Akawi, A. Checa, A.S. Antonopoulos, I. Akoumianakis, E. Daskalaki, C.P. Kotanidis, H. Kondo, K. Lee, D. Yesilyurt, I. Badi, M. Polkinghorne, N. Akbar, J. Lundgren, S. Chuaiphichai, R. Choudhury, S. Neubauer, K.M. Channon, S.S. Torekov, C.E. Wheelock, C. Antoniades, Fat-Secreted Ceramides Regulate Vascular Redox State and Influence Outcomes in Patients With Cardiovascular Disease, J Am Coll Cardiol. 77 (2021) 2494–2513. https://doi.org/10.1016/j.jacc.2021.03.314.
[157] E. Ortiz-Villanueva, J. Jaumot, R. Martínez, L. Navarro-Martín, B. Piña, R. Tauler, Assessment of endocrine disruptors effects on zebrafish (Danio rerio) embryos by untargeted LC-HRMS metabolomic analysis, Science of the Total Environment. 635 (2018) 156–166. https://doi.org/10.1016/j.scitotenv.2018.03.369.
[158] L. Perez de Souza, S. Alseekh, F. Scossa, A.R. Fernie, Ultra-high-performance liquid chromatography high-resolution mass spectrometry variants for metabolomics research, Nature Methods. 18 (2021) 733–746. https://doi.org/10.1038/s41592-021-01116-4.
[159] E.M. Harrieder, F. Kretschmer, S. Böcker, M. Witting, Current state-of-the-art of separation methods used in LC-MS based metabolomics and lipidomics, Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences. 1188 (2022). https://doi.org/10.1016/j.jchromb.2021.123069.
[160] E. Cífková, R. Hájek, M. Lísa, M. Holčapek, Hydrophilic interaction liquid chromatography-mass spectrometry of (lyso)phosphatidic acids, (lyso)phosphatidylserines and other lipid classes, Journal of Chromatography A. 1439 (2016) 65–73. https://doi.org/10.1016/j.chroma.2016.01.064.
[161] P. Jandera, Stationary and mobile phases in hydrophilic interaction chromatography: A review, Analytica Chimica Acta. 692 (2011) 1–25. https://doi.org/10.1016/j.aca.2011.02.047.
[162] Y. Guo, S. Gaiki, Retention and selectivity of stationary phases for hydrophilic interaction chromatography, Journal of Chromatography A. 1218 (2011) 5920–5938. https://doi.org/10.1016/j.chroma.2011.06.052.
[163] P. Appelblad, T. Jonsson, E. Pontén, C. Viklund, W. Jiang, <MerckSeQuant.ZIC-HILIC.Technical_Guide.A_Practical_Guide_to_HILIC.pdf>, Merck SeQuant. (2005) 3–27.
[164] A. Lioupi, M. Marinaki, State-of-the-art in LC – MS Approaches for Probing the Polar Metabolome, (n.d.) 1–26.
[165] S. Cubbon, C. Antonio, J. Wilson, J. Thomas-Oates, Metabolomic applications of HILIC–LC–MS, Mass Spectrometry Reviews. 29 (2010) 671–684. https://doi.org/10.1002/MAS.20252.
[166] D.Q. Tang, L. Zou, X.X. Yin, C.N. Ong, HILIC-MS for metabolomics: An attractive and complementary approach to RPLC-MS, Mass Spectrom Rev. 35 (2016) 574–600. https://doi.org/10.1002/mas.21445.
[167] L. Cui, H. Lu, Y.H. Lee, Challenges and emergent solutions for LC-MS/MS based untargeted metabolomics in diseases, Mass Spectrometry Reviews. 37 (2018) 772–792. https://doi.org/10.1002/mas.21562.
[168] F. Yuan, S. Kim, X. Yin, X. Zhang, I. Kato, Integrating two-dimensional gas and liquid chromatography-mass spectrometry for untargeted colorectal cancer metabolomics: A proof-of-principle study, Metabolites. 10 (2020) 1–17. https://doi.org/10.3390/metabo10090343.
[169] L. Montero, M. Herrero, Two-dimensional liquid chromatography approaches in Foodomics – A review, Analytica Chimica Acta. 1083 (2019) 1–18. https://doi.org/10.1016/j.aca.2019.07.036.
[170] W. Lv, X. Shi, S. Wang, G. Xu, Multidimensional liquid chromatography-mass spectrometry for metabolomic and lipidomic analyses, TrAC - Trends in Analytical Chemistry. 120 (2019) 115302. https://doi.org/10.1016/j.trac.2018.11.001.
[171] L.W. Manuel Schlimpert, D.P. Christoph Bauer, J.T. Sonja Krieger, Comprehensive Two-Dimensional Liquid Chromatography in Metabolome Analysis, Journal of Chromatography & Separation Techniques. 06 (2015). https://doi.org/10.4172/2157-7064.1000288.
[172] P.F. Brandão, A.C. Duarte, R.M.B.O. Duarte, Comprehensive multidimensional liquid chromatography for advancing environmental and natural products research, TrAC - Trends in Analytical Chemistry. 116 (2019) 186–197. https://doi.org/10.1016/j.trac.2019.05.016.
[173] F. Stilo, C. Bicchi, A.M. Jimenez-Carvelo, L. Cuadros-Rodriguez, S.E. Reichenbach, C. Cordero, Chromatographic fingerprinting by comprehensive two-dimensional chromatography: Fundamentals and tools, TrAC - Trends in Analytical Chemistry. 134 (2021) 116133. https://doi.org/10.1016/j.trac.2020.116133.

[174] F.A. Franchina, D. Zanella, L.M. Dubois, J.F. Focant, The role of sample preparation in multidimensional gas chromatographic separations for non-targeted analysis with the focus on recent biomedical, food, and plant applications, Journal of Separation Science. 44 (2021) 188–210. https://doi.org/10.1002/jssc.202000855.
[175] X. Wang, X. Song, L. Zhu, X. Geng, F. Zheng, Q. Zhao, X. Sun, D. Zhao, S. Feng, M. Zhao, B. Sun, Unraveling the acetals as ageing markers of Chinese Highland Qingke Baijiu using comprehensive two-dimensional gas chromatography-Time-of-flight mass spectrometry combined with metabolomics approach, Food Quality and Safety. 5 (2021) 1–8. https://doi.org/10.1093/fqsafe/fyab014.
[176] P. Zhang, S. Carlin, C. Lotti, F. Mattivi, U. Vrhovsek, On sample preparation methods for fermented beverage VOCs profiling by GCxGC-TOFMS, Metabolomics. 16 (2020) 1–10. https://doi.org/10.1007/s11306-020-01718-7.
[177] T.J. Trinklein, S. Schöneich, P.E. Sudol, C.G. Warren, D. v. Gough, R.E. Synovec, Total-transfer comprehensive three-dimensional gas chromatography with time-of-flight mass spectrometry, Journal of Chromatography A. 1634 (2020) 461654. https://doi.org/10.1016/j.chroma.2020.461654.
[178] V.R. Curovic, T. Suvitaival, I. Mattila, L. Ahonen, K. Trošt, S. Theilade, T.W. Hansen, C. Legido-Quigley, P. Rossing, Circulating metabolites and lipids are associated to diabetic retinopathy in individuals with type 1 diabetes, Diabetes. 69 (2020) 2217–2226. https://doi.org/10.2337/db20-0104.
[179] S. Kistner, M. Döring, R. Krüger, M.J. Rist, C.H. Weinert, D. Bunzel, B. Merz, K. Radloff, R. Neumann, S. Härtel, A. Bub, Sex-specific relationship between the cardiorespiratory fitness and plasma metabolite patterns in healthy humans—results of the karmen study, Metabolites. 11 (2021). https://doi.org/10.3390/metabo11070463.
[180] T. van der Laan, H. Elfrink, F. Azadi-Chegeni, A.C. Dubbelman, A.C. Harms, D.M. Jacobs, U. Braumann, A.H. Velders, J. van Duynhoven, T. Hankemeier, Fractionation platform for target identification using off-line directed two-dimensional chromatography, mass spectrometry and nuclear magnetic resonance, Analytica Chimica Acta. 1142 (2021) 28–37. https://doi.org/10.1016/j.aca.2020.10.054.
[181] B.W.J. Pirok, D.R. Stoll, P.J. Schoenmakers, Recent Developments in Two-Dimensional Liquid Chromatography: Fundamental Improvements for Practical Applications, Analytical Chemistry. 91 (2019) 240–263. https://doi.org/10.1021/acs.analchem.8b04841.
[182] Y. Chen, L. Montero, O.J. Schmitz, Advance in on-line two-dimensional liquid chromatography modulation technology, TrAC - Trends in Analytical Chemistry. 120 (2019) 115647. https://doi.org/10.1016/j.trac.2019.115647.
[183] B.W.J. Pirok, A.F.G. Gargano, P.J. Schoenmakers, Optimizing separations in online comprehensive two-dimensional liquid chromatography, Journal of Separation Science. 41 (2018) 68–98. https://doi.org/10.1002/jssc.201700863.
[184] Y. Chen, J. Li, O.J. Schmitz, Development of an At-Column Dilution Modulator for Flexible and Precise Control of Dilution Factors to Overcome Mobile Phase Incompatibility in Comprehensive Two-Dimensional Liquid Chromatography, Analytical Chemistry. 91 (2019) 10251–10257. https://doi.org/10.1021/acs.analchem.9b02391.
[185] Y. Chen, L. Montero, J. Luo, J. Li, O.J. Schmitz, Application of the new at-column dilution (ACD) modulator for the two-dimensional RP×HILIC analysis of Buddleja davidii, Analytical and Bioanalytical Chemistry. 412 (2020) 1483–1495. https://doi.org/10.1007/s00216-020-02392-3.
[186] L. Willmann, T. Erbes, S. Krieger, J. Trafkowski, M. Rodamer, B. Kammerer, Metabolome analysis via comprehensive two-dimensional liquid chromatography: identification of modified nucleosides from RNA metabolism, Analytical and Bioanalytical Chemistry 2015 407:13. 407 (2015) 3555–3566. https://doi.org/10.1007/S00216-015-8516-6.
[187] K. Arena, F. Cacciola, L. Dugo, P. Dugo, L. Mondello, Determination of the metabolite content of Brassica juncea cultivars using comprehensive two-dimensional liquid chromatography coupled with a photodiode array and mass spectrometry detection, Molecules. 25 (2020) 1–12. https://doi.org/10.3390/molecules25051235.
[188] L. Montero, S.W. Meckelmann, H. Kim, J.F. Ayala-Cabrera, O.J. Schmitz, Differentiation of industrial hemp strains by their cannabinoid and phenolic compounds using LC × LC-HRMS, Analytical and Bioanalytical Chemistry. (2022). https://doi.org/10.1007/S00216-022-03925-8.
[189] M. Xu, J. Legradi, P. Leonards, Evaluation of LC-MS and LC×LC-MS in analysis of zebrafish embryo samples for comprehensive lipid profiling, Analytical and Bioanalytical Chemistry. 412 (2020) 4313–4325. https://doi.org/10.1007/s00216-020-02661-1.
[190] M. Holčapek, M. Ovčačíková, M. Lísa, E. Cífková, T. Hájek, Continuous comprehensive two-dimensional liquid chromatography-electrospray ionization mass spectrometry of complex lipidomic samples, Anal Bioanal Chem. 407 (2015) 5033–5043. https://doi.org/10.1007/s00216-015-8528-2.

[191] M. Olfert, S. Bäurer, M. Wolter, S. Buckenmaier, E. Brito-de la Fuente, M. Lämmerhofer, Comprehensive profiling of conjugated fatty acid isomers and their lipid oxidation products by two-dimensional chiral RP×RP liquid chromatography hyphenated to UV- and SWATH-MS-detection, Analytica Chimica Acta. 1202 (2022). https://doi.org/10.1016/J.ACA.2022.339667.

[192] W. Ma, S. Wang, T. Zhang, E.Y. Zhang, L. Zhou, C. Hu, J.J. Yu, G. Xu, Activation of choline kinase drives aberrant choline metabolism in esophageal squamous cell carcinomas, Journal of Pharmaceutical and Biomedical Analysis. 155 (2018) 148–156. https://doi.org/10.1016/j.jpba.2018.03.062.

[193] W. Lv, L. Wang, Q. Xuan, X. Zhao, X. Liu, X. Shi, G. Xu, Pseudotargeted Method Based on Parallel Column Two-Dimensional Liquid Chromatography-Mass Spectrometry for Broad Coverage of Metabolome and Lipidome, Analytical Chemistry. 92 (2020) 6043–6050. https://doi.org/10.1021/acs.analchem.0c00372.

[194] M.A.I. Prodhan, B. Shi, M. Song, L. He, F. Yuan, X. Yin, P. Bohman, C.J. McClain, X. Zhang, Integrating comprehensive two-dimensional gas chromatography mass spectrometry and parallel two-dimensional liquid chromatography mass spectrometry for untargeted metabolomics, Analyst. 144 (2019) 4331–4341. https://doi.org/10.1039/c9an00560a.

[195] T.S. Bos, W.C. Knol, S.R.A. Molenaar, L.E. Niezen, P.J. Schoenmakers, G.W. Somsen, B.W.J. Pirok, Recent applications of chemometrics in one- and two-dimensional chromatography, Journal of Separation Science. 43 (2020) 1678–1727. https://doi.org/10.1002/jssc.202000011.

[196] D.M. Makey, V. Shchurik, H. Wang, H.R. Lhotka, D.R. Stoll, A. Vazhentsev, I. Mangion, E.L. Regalado, I.A.H. Ahmad, Mapping the Separation Landscape in Two-Dimensional Liquid Chromatography: Blueprints for Efficient Analysis and Purification of Pharmaceuticals Enabled by Computer-Assisted Modeling, Analytical Chemistry. 93 (2021) 964–972. https://doi.org/10.1021/acs.analchem.0c03680.

[197] X. Jiang, A. van der Horst, P.J. Schoenmakers, Breakthrough of polymers in interactive liquid chromatography, Journal of Chromatography A. 982 (2002) 55–68. https://doi.org/10.1016/S0021-9673(02)01483-8.

[198] S. CHAPEL, F. Rouvière, V. Peppermans, G. Desmet, S. Heinisch, A comprehensive study on the phenomenon of total breakthrough in liquid chromatography, Journal of Chromatography A. 1653 (2021) 462399. https://doi.org/10.1016/j.chroma.2021.462399.

[199] H.C. van de Ven, J. Purmova, G. Groeneveld, T.S. Bos, A.F.G. Gargano, S. van der Wal, Y. Mengerink, P.J. Schoenmakers, Living with breakthrough: Two-dimensional liquid-chromatography separations of a water-soluble synthetically grafted bio-polymer, Separations. 7 (2020) 1–16. https://doi.org/10.3390/separations7030041.

[200] S. Chapel, F. Rouvière, S. Heinisch, Comparison of existing strategies for keeping symmetrical peaks in on-line Hydrophilic Interaction Liquid Chromatography x Reversed-Phase Liquid Chromatography despite solvent strength mismatch, Journal of Chromatography A. 1642 (2021). https://doi.org/10.1016/j.chroma.2021.462001.

[201] M. Navarro-Reig, J. Jaumot, A. Baglai, G. Vivó-Truyols, P.J. Schoenmakers, R. Tauler, Untargeted Comprehensive Two-Dimensional Liquid Chromatography Coupled with High-Resolution Mass Spectrometry Analysis of Rice Metabolome Using Multivariate Curve Resolution, Analytical Chemistry. 89 (2017) 7675–7683. https://doi.org/10.1021/acs.analchem.7b01648.

[202] D.R. Stoll, K. Shoykhet, P. Petersson, S. Buckenmaier, Active Solvent Modulation: A Valve-Based Approach to Improve Separation Compatibility in Two-Dimensional Liquid Chromatography, Analytical Chemistry. 89 (2017) 9260–9267. https://doi.org/10.1021/acs.analchem.7b02046.

[203] D.R. Stoll, H.R. Lhotka, D.C. Harmes, B. Madigan, J.J. Hsiao, G.O. Staples, High resolution two-dimensional liquid chromatography coupled with mass spectrometry for robust and sensitive characterization of therapeutic antibodies at the peptide level, Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences. 1134–1135 (2019) 121832. https://doi.org/10.1016/j.jchromb.2019.121832.

[204] M. Pursch, A. Wegener, S. Buckenmaier, Evaluation of active solvent modulation to enhance two-dimensional liquid chromatography for target analysis in polymeric matrices, Journal of Chromatography A. 1562 (2018) 78–86. https://doi.org/10.1016/j.chroma.2018.05.059.

[205] P. Yang, W. Gao, T. Zhang, M. Pursch, J. Luong, W. Sattler, A. Singh, S. Backer, Two-dimensional liquid chromatography with active solvent modulation for studying monomer incorporation in copolymer dispersants, Journal of Separation Science. 42 (2019) 2805–2815. https://doi.org/10.1002/jssc.201900283.

[206] S. Bäurer, W. Guo, S. Polnick, M. Lämmerhofer, Simultaneous Separation of Water- and Fat-Soluble Vitamins by Selective Comprehensive HILIC × RPLC (High-Resolution Sampling) and Active Solvent Modulation, Chromatographia. 82 (2019) 167–180. https://doi.org/10.1007/s10337-018-3615-0.

[207] F. Li, X. Su, S. Bäurer, M. Lämmerhofer, Multiple heart-cutting mixed-mode chromatography-reversed-phase 2D-liquid chromatography method for separation and mass spectrometric characterization of synthetic oligonucleotides, Journal of Chromatography A. 1625 (2020). https://doi.org/10.1016/j.chroma.2020.461338.
[208] S.L. Weatherbee, T. Brau, D.R. Stoll, S.C. Rutan, M.M. Collinson, Simulation of elution profiles in liquid chromatography – IV: Experimental characterization and modeling of solute injection profiles from a modulation valve used in two-dimensional liquid chromatography, Journal of Chromatography A. 1626 (2020) 1–10. https://doi.org/10.1016/j.chroma.2020.461373.
[209] P.W. Carr, D.R. Stoll, Two-dimensional liquid chromatography: Principles, practical implementation and applications, Agilent Technical Note. (2015) 1–163.
[210] P. Dugo, F. Cacciola, T. Kumm, G. Dugo, L. Mondello, Comprehensive multidimensional liquid chromatography: Theory and applications, Journal of Chromatography A. (2008). https://doi.org/10.1016/j.chroma.2007.06.074.
[211] S. Alseekh, A. Aharoni, Y. Brotman, K. Contrepois, J. D'Auria, J. Ewald, J. C. Ewald, P.D. Fraser, P. Giavalisco, R.D. Hall, M. Heinemann, H. Link, J. Luo, S. Neumann, J. Nielsen, L. Perez de Souza, K. Saito, U. Sauer, F.C. Schroeder, S. Schuster, G. Siuzdak, A. Skirycz, L.W. Sumner, M.P. Snyder, H. Tang, T. Tohge, Y. Wang, W. Wen, S. Wu, G. Xu, N. Zamboni, A.R. Fernie, Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices, Nature Methods. 18 (2021) 747–756. https://doi.org/10.1038/s41592-021-01197-1.
[212] J.L. Ren, A.H. Zhang, L. Kong, X.J. Wang, Advances in mass spectrometry-based metabolomics for investigation of metabolites, RSC Advances. 8 (2018) 22335–22350. https://doi.org/10.1039/c8ra01574k.
[213] S.V. Hoffmann Edmond de, Mass Spectrometry: principles and applications, 3rd Edition - Edmond de Hoffmann, Vincent Stroobant, (2007) 1–4.
[214] F. Tugizimana, P.A. Steenkamp, L.A. Piater, I.A. Dubery, Mass spectrometry in untargeted liquid chromatography/mass spectrometry metabolomics: Electrospray ionisation parameters and global coverage of the metabolome, Rapid Communications in Mass Spectrometry. 32 (2018) 121–132. https://doi.org/10.1002/rcm.8010.
[215] Z. Lei, D. v. Huhman, L.W. Sumner, Mass spectrometry strategies in metabolomics, Journal of Biological Chemistry. 286 (2011) 25435–25442. https://doi.org/10.1074/jbc.R111.238691.
[216] D.R. Allen, B.C. McWhinney, Quadrupole Time-of-Flight Mass Spectrometry: A Paradigm Shift in Toxicology Screening Applications, Clinical Biochemist Reviews. 40 (2019) 135–146. https://doi.org/10.33176/AACB-19-00023.
[217] S. Han, W. van Treuren, C.R. Fischer, B.D. Merrill, B.C. DeFelice, J.M. Sanchez, S.K. Higginbottom, L. Guthrie, L.A. Fall, D. Dodd, M.A. Fischbach, J.L. Sonnenburg, A metabolomics pipeline for the mechanistic interrogation of the gut microbiome, Springer US, 2021. https://doi.org/10.1038/s41586-021-03707-9.
[218] M. Oh, S. Park, H. Kim, G.J. Choi, S.H. Kim, Application of uplc-qtof-ms based untargeted metabolomics in identification of metabolites induced in pathogen-infected rice, Plants. 10 (2021) 1–13. https://doi.org/10.3390/plants10020213.
[219] C. di Poto, X. Tian, X. Peng, H.M. Heyman, M. Szesny, S. Hess, L.H. Cazares, Metabolomic Profiling of Human Urine Samples Using LC-TIMS-QTOF Mass Spectrometry, J Am Soc Mass Spectrom. 32 (2021) 2072–2080. https://doi.org/10.1021/jasms.0c00467.
[220] G. Castro, M. Ramil, R. Cela, I. Rodríguez, Identification and determination of emerging pollutants in sewage sludge driven by UPLC-QTOF-MS data mining, Science of the Total Environment. 778 (2021). https://doi.org/10.1016/j.scitotenv.2021.146256.
[221] The Switch Is On: From Triple Quadrupoles and Q-TOF to Orbitrap High Resolution Mass Spectrometry | Thermo Fisher Scientific - ES, (n.d.). https://www.thermofisher.com/es/es/home/global/forms/industrial/switch-triple-quadrupoles-qtof-orbitrap-hrms.html (accessed April 14, 2022).
[222] T.N. Clark, J. Houriet, W.S. Vidar, J.J. Kellogg, D.A. Todd, N.B. Cech, R.G. Linington, Interlaboratory Comparison of Untargeted Mass Spectrometry Data Uncovers Underlying Causes for Variability, Journal of Natural Products. 84 (2021) 824–835. https://doi.org/10.1021/acs.jnatprod.0c01376.
[223] D. Szabó, G. Schlosser, K. Vékey, L. Drahos, Á. Révész, Collision energies on QTof and Orbitrap instruments: How to make proteomics measurements comparable?, Journal of Mass Spectrometry. 56 (2021) e4693. https://doi.org/10.1002/JMS.4693.
[224] J. Guo, T. Huan, Comparison of Full-Scan, Data-Dependent, and Data-Independent Acquisition Modes in Liquid Chromatography-Mass Spectrometry Based Untargeted Metabolomics, Analytical Chemistry. 92 (2020) 8072–8080. https://doi.org/10.1021/acs.analchem.9b05135.
[225] I. Ten-Doménech, T. Martínez-Sena, M. Moreno-Torres, J.D. Sanjuan-Herráez, J. v. Castell, A. Parra-Llorca, M. Vento, G. Quintás, J. Kuligowski, Comparing targeted vs. untargeted MS2 data-

dependent acquisition for peak annotation in LC–MS metabolomics, Metabolites. 10 (2020). https://doi.org/10.3390/metabo10040126.
[226] E. Defossez, J. Bourquin, S. von Reuss, S. Rasmann, G. Glauser, Eight key rules for successful data-dependent acquisition in mass spectrometry-based metabolomics, Mass Spectrometry Reviews. (2021). https://doi.org/10.1002/MAS.21715.
[227] I. Tada, R. Chaleckis, H. Tsugawa, I. Meister, P. Zhang, N. Lazarinis, B. Dahlén, C.E. Wheelock, M. Arita, Correlation-Based Deconvolution (CorrDec) to Generate High-Quality MS2 Spectra from Data-Independent Acquisition in Multisample Studies, Analytical Chemistry. 92 (2020) 11310–11317. https://doi.org/10.1021/acs.analchem.0c01980.
[228] J. Godzien, A. Gil de la Fuente, A. Otero, C. Barbas, Metabolite Annotation and Identification, 1st ed., Elsevier B.V., 2018. https://doi.org/10.1016/bs.coac.2018.07.004.
[229] C. Hawkins, D. Ginzburg, K. Zhao, W. Dwyer, B. Xue, A. Xu, S. Rice, B. Cole, S. Paley, P. Karp, S.Y. Rhee, Plant Metabolic Network 15: A resource of genome-wide metabolism databases for 126 plants and algae, Journal of Integrative Plant Biology. 63 (2021) 1888–1905. https://doi.org/10.1111/JIPB.13163.
[230] M. Kanehisa, S. Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes, Nucleic Acids Research. 28 (2000) 27–30. https://doi.org/10.1093/NAR/28.1.27.
[231] B.B. Misra, New software tools, databases, and resources in metabolomics: updates from 2020, Metabolomics. 17 (2021) 1–24. https://doi.org/10.1007/s11306-021-01796-1.
[232] J. Trygg, E. Holmes, T. Lundstedt, Chemometrics in metabonomics, Journal of Proteome Research. 6 (2007) 469–479. https://doi.org/10.1021/pr060594q.
[233] S.R.A. Molenaar, T.A. Dahlseid, G.M. Leme, D.R. Stoll, P.J. Schoenmakers, B.W.J. Pirok, Peak-tracking algorithm for use in comprehensive two-dimensional liquid chromatography – Application to monoclonal-antibody peptides, Journal of Chromatography A. 1639 (2021) 461922. https://doi.org/10.1016/j.chroma.2021.461922.
[234] A. Paul, P. de Boves Harrington, Chemometric applications in metabolomic studies using chromatography-mass spectrometry, TrAC Trends in Analytical Chemistry. 135 (2021) 116165. https://doi.org/10.1016/j.trac.2020.116165.
[235] M. Pérez-Cova, C. Bedia, D.R. Stoll, R. Tauler, J. Jaumot, MSroi: A pre-processing tool for mass spectrometry-based studies, Chemometrics and Intelligent Laboratory Systems. 215 (2021). https://doi.org/10.1016/j.chemolab.2021.104333.
[236] E. Gorrochategui, J. Jaumot, R. Tauler, ROIMCR: A powerful analysis strategy for LC-MS metabolomic datasets, BMC Bioinformatics. 20 (2019) 1–17. https://doi.org/10.1186/s12859-019-2848-8.
[237] A. de Juan, R. Tauler, Multivariate Curve Resolution: 50 years addressing the mixture analysis problem – A review, Analytica Chimica Acta. 1145 (2021) 59–78. https://doi.org/10.1016/j.aca.2020.10.051.
[238] M. Navarro-Reig, J. Jaumot, A. García-Reiriz, R. Tauler, Evaluation of changes induced in rice metabolome by Cd and Cu exposure using LC-MS with XCMS and MCR-ALS data analysis strategies, Analytical and Bioanalytical Chemistry. (2015). https://doi.org/10.1007/s00216-015-9042-2.
[239] F. Puig-Castellví, C. Bedia, I. Alfonso, B. Piña, R. Tauler, Deciphering the Underlying Metabolomic and Lipidomic Patterns Linked to Thermal Acclimation in Saccharomyces cerevisiae, J Proteome Res. (2018). https://doi.org/10.1021/acs.jproteome.7b00921.
[240] N. Dalmau, C. Bedia, R. Tauler, Validation of the Regions of Interest Multivariate Curve Resolution (ROIMCR) procedure for untargeted LC-MS lipidomic analysis, Analytica Chimica Acta. 1025 (2018) 80–91. https://doi.org/10.1016/j.aca.2018.04.003.
[241] C. Perez-Lopez, A. Ginebreda, M. Carrascal, D. Barcelò, J. Abian, R. Tauler, Non-target protein analysis of samples from wastewater treatment plants using the regions of interest-multivariate curve resolution (ROIMCR) chemometrics method, Journal of Environmental Chemical Engineering. 9 (2021). https://doi.org/10.1016/j.jece.2021.105752.
[242] E. Garreta-Lara, C. Gómez-Canela, B. Campos, C. Barata, R. Tauler, S. Lacorte, Combined targeted/untargeted analytical and chemometric approaches in the characterization of Daphnia magna metabolomic changes under bisphenol A exposure, Microchemical Journal. 165 (2021). https://doi.org/10.1016/j.microc.2021.106150.
[243] E. Figueira, D. Matos, P. Cardoso, C. Sá, C. Fernandes, R. Tauler, C. Bedia, An underground strategy to increase mercury tolerance in the salt marsh halophyte Juncus maritimus Lam.: Lipid remodelling and Hg restriction, Environmental and Experimental Botany. 191 (2021). https://doi.org/10.1016/j.envexpbot.2021.104619.
[244] A. Cerrato, C. Bedia, A.L. Capriotti, C. Cavaliere, V. Gentile, M. Maggi, C.M. Montone, S. Piovesana, A. Sciarra, R. Tauler, A. Laganà, Untargeted metabolomics of prostate cancer

zwitterionic and positively charged compounds in urine, Analytica Chimica Acta. 1158 (2021). https://doi.org/10.1016/j.aca.2021.338381.
[245] C. Bedia, M. Badia, L. Muixí, T. Levade, R. Tauler, A. Sierra, GM2-GM3 gangliosides ratio is dependent on GRP94 through down-regulation of GM2-AP cofactor in brain metastasis cells, Scientific Reports. 9 (2019) 1–12. https://doi.org/10.1038/s41598-019-50761-5.
[246] F. Puig-Castellví, C. Bedia, I. Alfonso, B. Piña, R. Tauler, Deciphering the Underlying Metabolomic and Lipidomic Patterns Linked to Thermal Acclimation in Saccharomyces cerevisiae, Journal of Proteome Research. 17 (2018) 2034–2044. https://doi.org/10.1021/acs.jproteome.7b00921.
[247] C. Bedia, À. Sierra, R. Tauler, Multimodal multisample spectroscopic imaging analysis of tumor tissues using multivariate curve resolution, Chemometrics and Intelligent Laboratory Systems. 215 (2021) 104366. https://doi.org/10.1016/J.CHEMOLAB.2021.104366.
[248] R. Lotfi Khatoonabadi, M. Vosough, L.L. Hohrenk, T.C. Schmidt, Employing complementary multivariate methods for a designed nontarget LC-HRMS screening of a wastewater-influenced river, Microchemical Journal. 160 (2021) 105641. https://doi.org/10.1016/j.microc.2020.105641.
[249] L.L. Hohrenk, M. Vosough, T.C. Schmidt, Implementation of Chemometric Tools to Improve Data Mining and Prioritization in LC-HRMS for Nontarget Screening of Organic Micropollutants in Complex Water Matrixes, Analytical Chemistry. 91 (2019) 9213–9220. https://doi.org/10.1021/acs.analchem.9b01984.
[250] F. Rezaei, M. Sheikholeslami, M. Vosough, M. Maeder, Handling of highly coeluted chromatographic peaks by multivariate curve resolution for a complex bioanalytical problem: Quantitation of selected corticosteroids and mycophenolic acid in human plasma, Talanta. 187 (2018) 1–12. https://doi.org/10.1016/j.talanta.2018.04.089.
[251] L.L. Hohrenk, F. Itzel, N. Baetz, J. Tuerk, M. Vosough, T.C. Schmidt, Comparison of Software Tools for Liquid Chromatography-High-Resolution Mass Spectrometry Data Processing in Nontarget Screening of Environmental Samples, Analytical Chemistry. 92 (2020) 1898–1907. https://doi.org/10.1021/acs.analchem.9b04095.
[252] N.J. Nielsen, G. Tomasi, R.J.N. Frandsen, M.B. Kristensen, J. Nielsen, H. Giese, J.H. Christensen, A pre-processing strategy for liquid chromatography time-of-flight mass spectrometry metabolic fingerprinting data, Metabolomics. 6 (2010) 341–352. https://doi.org/10.1007/s11306-010-0211-1.
[253] R. Stolt, R.J.O. Torgrip, J. Lindberg, L. Csenki, J. Kolmert, I. Schuppe-Koistinen, S.P. Jacobsson, Second-order peak detection for multicomponent high-resolution LC/MS data, Analytical Chemistry. 78 (2006) 975–983. https://doi.org/10.1021/ac050980b.
[254] R. Tautenhahn, C. Bottcher, S. Neumann, Highly sensitive feature detection for high resolution LC/MS, BMC Bioinformatics. 9 (2008) 1–16. https://doi.org/10.1186/1471-2105-9-504.
[255] C.A. Smith, E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification, Analytical Chemistry. 78 (2006) 779–787. https://doi.org/10.1021/ac051437y.
[256] K. Segers, W. Zhang, N. Aourz, J. Bongaerts, S. Declerck, D. Mangelings, T. Hankemeier, D. de Bundel, Y. vander Heyden, I. Smolders, R. Ramautar, A. van Eeckhaut, CE-MS metabolic profiling of volume-restricted plasma samples from an acute mouse model for epileptic seizures to discover potentially involved metabolomic features, Talanta. 217 (2020) 121107. https://doi.org/10.1016/j.talanta.2020.121107.
[257] G.H. Golub, C. Reinsch, Singular value decomposition and least squares solutions, Numerische Mathematik. 14 (1970) 403–420. https://doi.org/10.1007/BF02163027.
[258] W. Windig, N.B. Gallagher, J.M. Shaver, B.M. Wise, A new approach for interactive self-modeling mixture analysis, Chemometrics and Intelligent Laboratory Systems. 77 (2005) 85–96. https://doi.org/10.1016/j.chemolab.2004.06.009.
[259] A. de Juan, J. Jaumot, R. Tauler, Multivariate Curve Resolution (MCR). Solving the mixture analysis problem, Analytical Methods. 6 (2014) 4964–4976. https://doi.org/10.1039/c4ay00571f.
[260] A.C. Olivieri, R. Tauler, The effect of data matrix augmentation and constraints in extended multivariate curve resolution–alternating least squares, Journal of Chemometrics. 31 (2017) 1–10. https://doi.org/10.1002/cem.2875.
[261] J. Jaumot, R. Gargallo, A. de Juan, R. Tauler, A graphical user-friendly interface for MCR-ALS: A new tool for multivariate curve resolution in MATLAB, Chemometrics and Intelligent Laboratory Systems. 76 (2005) 101–110. https://doi.org/10.1016/j.chemolab.2004.12.007.
[262] M. Navarro-Reig, J. Jaumot, T.A. van Beek, G. Vivó-Truyols, R. Tauler, Chemometric analysis of comprehensive LC×LC-MS data: Resolution of triacylglycerol structural isomers in corn oil, Talanta. 160 (2016) 624–635. https://doi.org/10.1016/j.talanta.2016.08.005.
[263] H.P. Bailey, S.C. Rutan, P.W. Carr, Factors that affect quantification of diode array data in comprehensive two-dimensional liquid chromatography using chemometric data analysis, Journal of Chromatography A. 1218 (2011) 8411–8422. https://doi.org/10.1016/j.chroma.2011.09.057.

[264] R. Tauler, A. de Juan, Chapter 5 - Multivariate Curve Resolution for Quantitative Analysis, in: A.M. de la Peña, H.C. Goicoechea, G.M. Escandar, A.C.B.T.-D.H. in S. and T. Olivieri (Eds.), Fundamentals and Analytical Applications of Multiway Calibration, Elsevier, 2015: pp. 247–292. https://doi.org/https://doi.org/10.1016/B978-0-444-63527-3.00005-9.
[265] H.P. Bailey, S.C. Rutan, Chemometric resolution and quantification of four-way data arising from comprehensive 2D-LC-DAD analysis of human urine, Chemometrics and Intelligent Laboratory Systems. 106 (2011) 131–141. https://doi.org/10.1016/j.chemolab.2010.07.008.
[266] H.P. Bailey, S.C. Rutan, Chemometric resolution and quantification of four-way data arising from comprehensive 2D-LC-DAD analysis of human urine, Chemometrics and Intelligent Laboratory Systems. 106 (2011) 131–141. https://doi.org/10.1016/j.chemolab.2010.07.008.
[267] S.E.G. Porter, D.R. Stoll, S.C. Rutan, P.W. Carr, J.D. Cohen, Analysis of four-way two-dimensional liquid chromatography-diode array data: Application to metabolomics, Analytical Chemistry. 78 (2006) 5559–5569. https://doi.org/10.1021/ac0606195.
[268] E. Peré-Trepat, S. Lacorte, R. Tauler, Alternative calibration approaches for LC-MS quantitative determination of coeluted compounds in complex environmental mixtures using multivariate curve resolution, Analytica Chimica Acta. 595 (2007) 228–237. https://doi.org/10.1016/j.aca.2007.04.011.
[269] M. Bayat, M. Marín-García, J.B. Ghasemi, R. Tauler, Application of the area correlation constraint in the MCR-ALS quantitative analysis of complex mixture samples, Analytica Chimica Acta. 1113 (2020) 52–65. https://doi.org/10.1016/j.aca.2020.03.057.
[270] A.C. de O. Neves, R. Tauler, K.M.G. de Lima, Area correlation constraint for the MCR−ALS quantification of cholesterol using EEM fluorescence data: A new approach, Analytica Chimica Acta. 937 (2016) 21–28. https://doi.org/10.1016/j.aca.2016.08.011.
[271] J. Jaumot, B. Igne, C.A. Anderson, J.K. Drennen, A. de Juan, Blending process modeling and control by multivariate curve resolution, Talanta. 117 (2013) 492–504. https://doi.org/10.1016/j.talanta.2013.09.037.
[272] R. Goodacre, D. Broadhurst, A.K. Smilde, B.S. Kristal, J.D. Baker, R. Beger, C. Bessant, S. Connor, G. Capuani, A. Craig, T. Ebbels, D.B. Kell, C. Manetti, J. Newton, G. Paternostro, R. Somorjai, M. Sjöström, J. Trygg, F. Wulfert, Proposed minimum reporting standards for data analysis in metabolomics, Metabolomics. 3 (2007) 231–241. https://doi.org/10.1007/s11306-007-0081-3.
[273] X. Zhang, J. Dong, D. Raftery, Five Easy Metrics of Data Quality for LC-MS-Based Global Metabolomics, Analytical Chemistry. 92 (2020) 12925–12933. https://doi.org/10.1021/acs.analchem.0c01493.
[274] D. Broadhurst, R. Goodacre, S.N. Reinke, J. Kuligowski, I.D. Wilson, M.R. Lewis, W.B. Dunn, Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies, Metabolomics. 14 (2018) 1–17. https://doi.org/10.1007/s11306-018-1367-3.
[275] J. Kuligowski, D. Pérez-Guaita, I. Lliso, J. Escobar, Z. León, L. Gombau, R. Solberg, O.D. Saugstad, M. Vento, G. Quintás, Detection of batch effects in liquid chromatography-mass spectrometry metabolomic data using guided principal component analysis, Talanta. 130 (2014) 442–448. https://doi.org/10.1016/j.talanta.2014.07.031.
[276] Á. Sánchez-Illana, J.D. Piñeiro-Ramos, J.D. Sanjuan-Herráez, M. Vento, G. Quintás, J. Kuligowski, Evaluation of batch effect elimination using quality control replicates in LC-MS metabolite profiling, Analytica Chimica Acta. 1019 (2018) 38–48. https://doi.org/10.1016/j.aca.2018.02.053.
[277] J. Kuligowski, Á. Sánchez-Illana, D. Sanjuán-Herráez, M. Vento, G. Quintás, Intra-batch effect correction in liquid chromatography-mass spectrometry using quality control samples and support vector regression (QC-SVRC), Analyst. 140 (2015) 7810–7817. https://doi.org/10.1039/c5an01638j.
[278] F. Dieterle, A. Ross, G. Schlotterbeck, H. Senn, Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in1H NMR metabonomics, Analytical Chemistry. 78 (2006) 4281–4290. https://doi.org/10.1021/ac051632c.
[279] B. Li, J. Tang, Q. Yang, X. Cui, S. Li, S. Chen, Q. Cao, W. Xue, N. Chen, F. Zhu, Performance evaluation and online realization of data-driven normalization methods used in LC/MS based untargeted metabolomics analysis, Scientific Reports. 6 (2016) 1–13. https://doi.org/10.1038/srep38881.
[280] R.E. Mohler, K.M. Dombek, J.C. Hoggard, E.T. Young, R.E. Synovec, Comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry analysis of metabolites in fermenting and respiring yeast cells, Analytical Chemistry. 78 (2006) 2700–2709. https://doi.org/10.1021/ac052106o.

[281] R.E. Mohler, K.M. Dombek, J.C. Hoggard, K.M. Pierce, E.T. Young, R.E. Synovec, Comprehensive analysis of yeast metabolite GC×GC-TOFMS data: Combining discovery-mode and deconvolution chemometric software, Analyst. 132 (2007) 756–767. https://doi.org/10.1039/b700061h.
[282] R.E. Mohler, B.P. Tu, K.M. Dombek, J.C. Hoggard, E.T. Young, R.E. Synovec, Identification and evaluation of cycling yeast metabolites in two-dimensional comprehensive gas chromatography-time-of-flight-mass spectrometry data, Journal of Chromatography A. 1186 (2008) 401–411. https://doi.org/10.1016/j.chroma.2007.10.063.
[283] S.E. Prebihalo, D.K. Pinkerton, R.E. Synovec, Impact of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry experimental design on data trilinearity and parallel factor analysis deconvolution, Journal of Chromatography A. 1605 (2019) 460368. https://doi.org/10.1016/j.chroma.2019.460368.
[284] P.T. Chun, R.J. McPherson, L.C. Marney, S.Z. Zangeneh, B.A. Parsons, A. Shojaie, R.E. Synovec, S.E. Juul, Serial plasma metabolites following hypoxic-ischemic encephalopathy in a nonhuman primate model, Developmental Neuroscience. 37 (2015) 161–171. https://doi.org/10.1159/000370147.
[285] L.R. Snyder, J.C. Hoggard, T.J. Montine, R.E. Synovec, Development and application of a comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry method for the analysis of l-β-methylamino-alanine in human tissue, Journal of Chromatography A. 1217 (2010) 4639–4647. https://doi.org/10.1016/j.chroma.2010.04.065.
[286] B. Khakimov, J.M. Amigo, S. Bak, S.B. Engelsen, Plant metabolomics: Resolution and quantification of elusive peaks in liquid chromatography-mass spectrometry profiles of complex plant extracts using multi-way decomposition methods, Journal of Chromatography A. 1266 (2012) 84–94. https://doi.org/10.1016/j.chroma.2012.10.023.
[287] F.C. Guizellini, G.G. Marcheafave, M. Rakocevic, R.E. Bruns, I.S. Scarminio, P.K. Soares, PARAFAC HPLC-DAD metabolomic fingerprint investigation of reference and crossed coffees, Food Research International. 113 (2018) 9–17. https://doi.org/10.1016/j.foodres.2018.06.070.
[288] P.K. Soares, G.G. Marcheafave, A. de A. Gomes, I.S. Scarminio, R.E. Bruns, Mixture Design PARAFAC HPLC-DAD Metabolomic Fingerprints of Fractionated Organic and Basic Extracts from Erythrina speciosa Andrews Leaves, Chromatographia. 81 (2018) 1189–1200. https://doi.org/10.1007/s10337-018-3554-9.
[289] N. Escudero, F.C. Marhuenda-Egea, R. Ibanco-Cañete, E.A. Zavala-Gonzalez, L. v. Lopez-Llorca, A metabolomic approach to study the rhizodeposition in the tritrophic interaction: Tomato, Pochonia chlamydosporia and Meloidogyne javanica, Metabolomics. 10 (2014) 788–804. https://doi.org/10.1007/s11306-014-0632-3.
[290] A.J. Lawaetz, R. Bro, M. Kamstrup-Nielsen, I.J. Christensen, L.N. Jørgensen, H.J. Nielsen, Fluorescence spectroscopy as a potential metabonomic tool for early detection of colorectal cancer, Metabolomics. 8 (2012) 111–121. https://doi.org/10.1007/s11306-011-0310-7.
[291] R. Bro, PARAFAC. Tutorial and applications, Chemometrics and Intelligent Laboratory Systems. 38 (1997) 149–171. https://doi.org/10.1016/S0169-7439(97)00032-4.
[292] D.W. Cook, S.C. Rutan, Chemometrics for the analysis of chromatographic data in metabolomics investigations, Journal of Chemometrics. 28 (2014) 681–687. https://doi.org/10.1002/cem.2624.
[293] R. Bro, C.A. Andersson, H.A.L. Kiers, PARAFAC2 - Part II. Modeling chromatographic data with retention time shifts, Journal of Chemometrics. 13 (1999) 295–309. https://doi.org/10.1002/(SICI)1099-128X(199905/08)13:3/4<295::AID-CEM547>3.0.CO;2-Y.
[294] H. Parastar, J.R. Radović, J.M. Bayona, R. Tauler, Solving chromatographic challenges in comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry using multivariate curve resolution-alternating least squares ABC Highlights: Authored by Rising Stars and Top Experts, Analytical and Bioanalytical Chemistry. 405 (2013) 6235–6249. https://doi.org/10.1007/s00216-013-7067-y.
[295] H.P. Bailey, S.C. Rutan, Chemometric resolution and quantification of four-way data arising from comprehensive 2D-LC-DAD analysis of human urine, Chemometrics and Intelligent Laboratory Systems. 106 (2011) 131–141. https://doi.org/10.1016/j.chemolab.2010.07.008.
[296] M. Pérez-Cova, R. Tauler, J. Jaumot, Chemometrics in comprehensive two-dimensional liquid chromatography: A study of the data structure and its multilinear behavior, Chemometrics and Intelligent Laboratory Systems. 201 (2020). https://doi.org/10.1016/j.chemolab.2020.104009.
[297] B.Q. Li, J. Chen, J.J. Li, X. Wang, H.L. Zhai, The application of a Tchebichef moment method to the quantitative analysis of multiple compounds based on three-dimensional HPLC fingerprint spectra, Analyst. 140 (2015) 630–636. https://doi.org/10.1039/c4an01736f.
[298] R. Tauler, Multivariate curve resolution of multiway data using the multilinearity constraint, Journal of Chemometrics. 35 (2021) 1–24. https://doi.org/10.1002/cem.3279.

[299] M. Marín-García, R. Tauler, Chemometrics characterization of The Llobregat river dissolved organic matter, Chemometrics and Intelligent Laboratory Systems. 201 (2020). https://doi.org/10.1016/j.chemolab.2020.104018.
[300] H. Parastar, J.R. Radović, M. Jalali-Heravi, S. Diez, J.M. Bayona, R. Tauler, Resolution and quantification of complex mixtures of polycyclic aromatic hydrocarbons in heavy fuel oil sample by means of GC × GC-TOFMS combined to multivariate curve resolution, Analytical Chemistry. 83 (2011) 9289–9297. https://doi.org/10.1021/ac201799r.
[301] M. Navarro-Reig, J. Jaumot, T.A. van Beek, G. Vivó-Truyols, R. Tauler, Chemometric analysis of comprehensive LC×LC-MS data: Resolution of triacylglycerol structural isomers in corn oil, Talanta. 160 (2016) 624–635. https://doi.org/10.1016/j.talanta.2016.08.005.
[302] G.S. Ochoa, P.E. Sudol, T.J. Trinklein, R.E. Synovec, Class comparison enabled mass spectrum purification for comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry, Talanta. 236 (2022) 122844. https://doi.org/10.1016/j.talanta.2021.122844.
[303] N. Feizi, F.S. Hashemi-Nasab, F. Golpelichi, N. Sabouruh, H. Parastar, Recent trends in application of chemometric methods for GC-MS and GC×GC-MS-based metabolomic studies, TrAC - Trends in Analytical Chemistry. 138 (2021) 116239. https://doi.org/10.1016/j.trac.2021.116239.
[304] L. Yi, N. Dong, Y. Yun, B. Deng, D. Ren, S. Liu, Y. Liang, Chemometric methods in data processing of mass spectrometry-based metabolomics: A review, Analytica Chimica Acta. 914 (2016) 17–34. https://doi.org/10.1016/j.aca.2016.02.001.
[305] A. Checa, C. Bedia, J. Jaumot, Lipidomic data analysis: Tutorial, practical guidelines and applications, Analytica Chimica Acta. 885 (2015) 1–16. https://doi.org/10.1016/j.aca.2015.02.068.
[306] M. Barker, W. Rayens, Partial least squares for discrimination, Journal of Chemometrics. 17 (2003) 166–173. https://doi.org/10.1002/cem.785.
[307] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: A basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems. 58 (2001) 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1.
[308] E. Szymańska, E. Saccenti, A.K. Smilde, J.A. Westerhuis, Double-check: Validation of diagnostic statistics for PLS-DA models in metabolomics studies, Metabolomics. 8 (2012) 3–16. https://doi.org/10.1007/s11306-011-0330-3.
[309] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, BBA - Protein Structure. 405 (1975) 442–451. https://doi.org/10.1016/0005-2795(75)90109-9.
[310] T.K. Kim, T test as a parametric statistic, Korean Journal of Anesthesiology. 68 (2015) 540. https://doi.org/10.4097/KJAE.2015.68.6.540.
[311] H.-Y. Kim, Analysis of variance (ANOVA) comparing means of more than two groups, Restorative Dentistry & Endodontics. 39 (2014) 74. https://doi.org/10.5395/rde.2014.39.1.74.
[312] O. Jean Dunn, MULTIPLE COMPARISONS AMONG MEANS, (n.d.).
[313] Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing Author (s): Yoav Benjamini and Yosef Hochberg Source : Journal of the Royal Statistical Society . Series B (Methodological), Vol . 57 , No . 1 (1995), Publi, Journal of the Royal Statistical Society. 57 (1995) 289–300.
[314] J.G. Scheiner, S.M., MANOVA: multiple response variables and multispecies interactions. Design and Analysis of Ecological Experiments, in: C. Press (Ed.), Design and Analysis of Ecological Experiments, 1993: pp. 94–112.
[315] A.K. Smilde, J.J. Jansen, H.C.J. Hoefsloot, R.J.A.N. Lamers, J. van der Greef, M.E. Timmerman, ANOVA-simultaneous component analysis (ASCA): A new tool for analyzing designed metabolomics data, Bioinformatics. 21 (2005) 3043–3048. https://doi.org/10.1093/bioinformatics/bti476.
[316] C. Bertinetto, J. Engel, J. Jansen, ANOVA simultaneous component analysis: A tutorial review, Analytica Chimica Acta: X. 6 (2020) 100061. https://doi.org/10.1016/j.acax.2020.100061.
[317] J. Engel, L. Blanchet, B. Bloemen, L.P. van den Heuvel, U.H.F. Engelke, R.A. Wevers, L.M.C. Buydens, Regularized MANOVA (rMANOVA) in untargeted metabolomics, Analytica Chimica Acta. 899 (2015) 1–12. https://doi.org/10.1016/j.aca.2015.06.042.
[318] E. Saccenti, A.K. Smilde, J. Camacho, Group-wise ANOVA simultaneous component analysis for designed omics experiments, Metabolomics. 14 (2018) 1–18. https://doi.org/10.1007/s11306-018-1369-1.

Chapter three

Evaluation and comparison of chemometric strategies for chromatography-mass spectrometry-based data

3.1 Introduction

As already mentioned, data analysis is still a major bottleneck in metabolomics. Hence, there is a need for tools that allow in-depth characterization of the increasingly bigger and more complex datasets. This requirement is even more evident when LC × LC-MS is applied, in comparison, for instance, with LC-MS data. All in all, a considerable effort has been made in recent years to implement different chemometric strategies to deal with chromatography-mass-spectrometry datasets from different perspectives, such as method development and optimization, data pre-processing and resolution, or compound discovery and annotation, among others [1].

The metabolomic data analysis workflow usually presents three main steps [2–4], as shown in **Figure 3.1**:

1) **Pre-processing** methods that aim to reduce the dimensions of the datasets and enhance their quality.

2) **Resolution** methods that obtain the spectra and elution profiles of the chemical constituents of a mixture.

3) **Post-processing** methods, including several univariate and multivariate statistical, exploratory, feature detection or classification methods, that allow pattern recognition and their statistical inference.

Although the second step is not as frequently emphasized as the other two steps in metabolomic studies, it is highly recommended in the case of overlapping signals in the analysis of complex mixtures. Besides, it provides qualitative and quantitative information about the analytes. This Chapter focuses on the contributions of this PhD Thesis to the metabolomic data analysis workflow, although the strategies proposed can be applied to any other field of study.

The pre-processing step aims to compress huge datasets by applying filters on spectral and retention time modes, in addition to improve data quality by eliminating noise and baseline contributions and by aligning, modelling and smoothing chromatographic peaks. The regions of interest (ROI) strategy is proposed in this PhD Thesis as a pre-processing step. The main advantage of this procedure is the ability of compressing the large MS datasets in the spectral mode without losing spectral accuracy, on contraposition to other compression strategies such as binning [3], as discussed later in this Chapter.

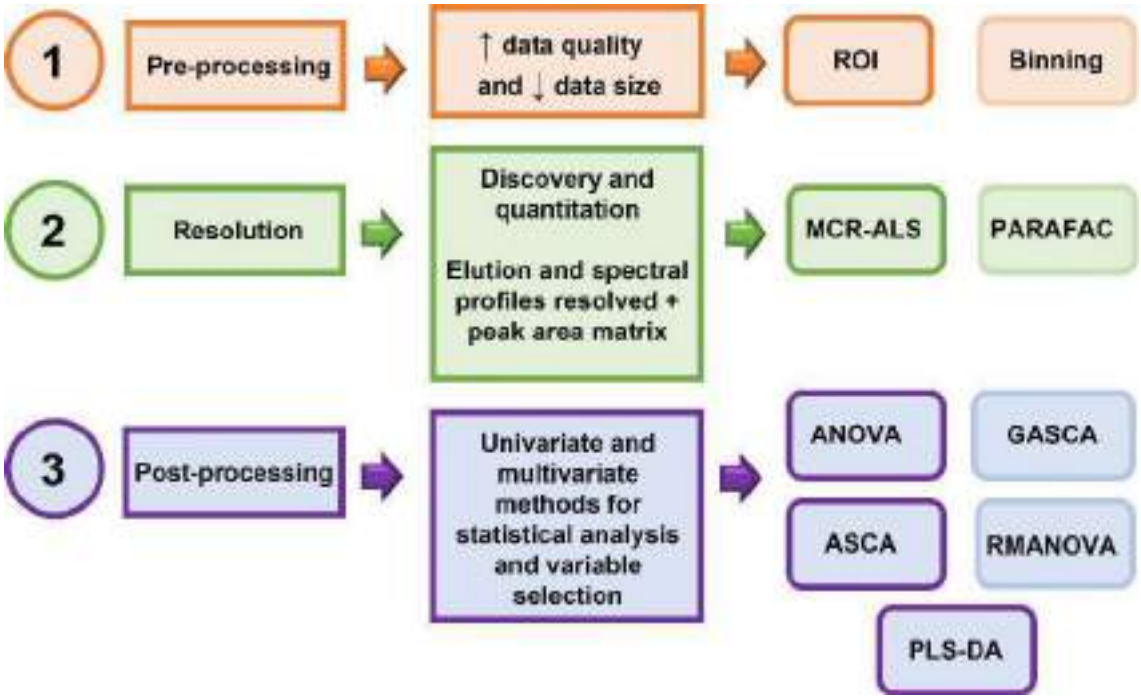


Figure 3.1. Data analysis workflow proposed for metabolomic studies.

Recently, or research group has launched the MSroi MATLAB GUI application that implements this ROI approach to mass spectrometry-based datasets [5]. The MSroi GUI has been employed in this PhD Thesis for filtering and compressing one and two-dimensional liquid chromatography coupled to mass spectrometry datasets (LC-MS and LC×LC-MS). More specifically, this software has been successfully applied in **scientific publications III, IV, V, VI, VII.** Its main functionalities are summarized in **Figure 3.2**. Firstly, this GUI allows the visualization of total ion chromatograms (TIC), as well as the generation of 2D contour plots in the case of LC ×LC-MS data. The elution profiles for each ROI can be examined in 1D or 2D formats. The preselected ROI should be further considered if it has an approximately Gaussian peak shape or discarded if not. Secondly, the data compression step can be performed both in the spectral or chromatographic(s) modes. In the case of this PhD Thesis, its main use has been the compression in the mass-to-charge (*m/z*) direction. The dimensionality of the datasets is reduced significantly by eliminating the *m/z* associated with instrumental noise, while keeping those signals above an intensity threshold. Discrete averaged *m/z* values for each ROI are obtained. Thirdly, each peak of each ROI can be integrated for all samples. Thus, quantitative

information is also provided by this software, as will be discussed in **the following Chapter** (in **scientific publication VI**).

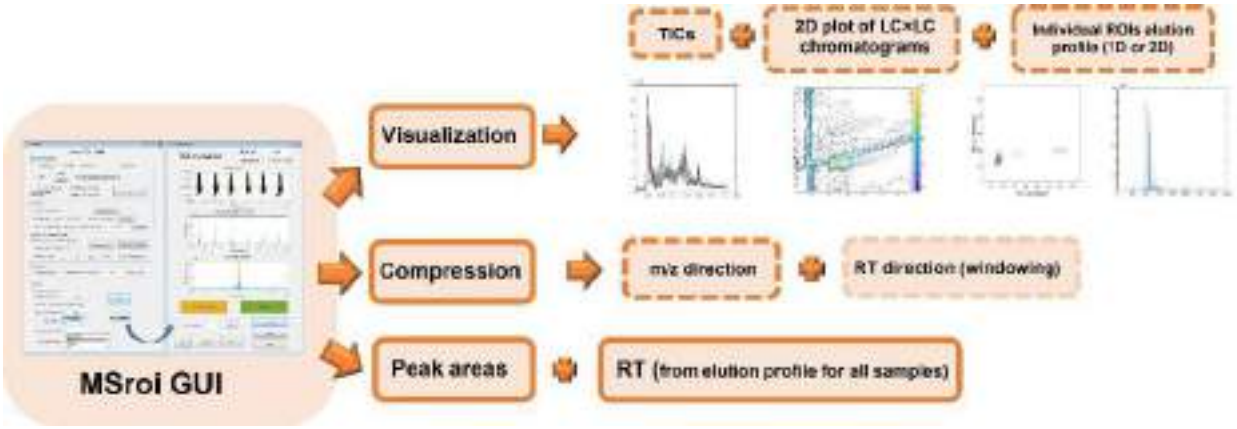


Figure 3.2. Functionalities of the MSroi GUI.

The second resolution step aims to discover and quantify the different constituents present in the analyzed mixtures using a multivariate approach. This means that this resolution step involves multivariate signals, i.e., full scan mass spectrometry, where the whole MS spectra are acquired within a specific mass range. When univariate overlapped signals (i.e., signal intensities at individual *m/z* values or wavelengths in the case of ultraviolet-visible spectroscopy, UV) are resolved, then the deconvolution term is preferred. In the chemometrics field, two resolution methods are commonly proposed for the analysis of hyphenated chromatographic multivariate data, as shown in **Figure 3.3**: the multivariate curve resolution alternating least squares (MCR-ALS) method [6] and the parallel factor analysis (PARAFAC) method [7]. There is also a variant of this last method known as PARAFAC2 whose purpose is the processing of chromatographic-spectral data in the presence of retention time shifts of chromatographic peaks among samples [8]. The choice of the resolution method relies on the fulfillment of the trilinear model of the data analyzed [9]. For instance, PARAFAC will only produce satisfactory results from a chemical point of view if trilinear model requirements are totally fulfilled. This constraint implies that each resolved component is described by a single dyad of profiles in each mode. Instead, PARAFAC2 allows for the relaxation of the trilinear model requirements. Therefore, peaks of the elution profiles can be time-shifted in the different analyzed samples (chromatographic runs) [7,10,11]. In contrast, MCR-

ALS is proposed for the analysis of data fulfilling a bilinear model, which can be easily adapted to datasets fulfilling a trilinear model, for all or only some of the components. MCR-ALS also allows the possibility of time-shifting and changes in shape of the resolved elution profiles. The MCR model can impose the fulfilment of the trilinear model in a much more flexible way than PARAFAC and even PARAFAC2 [12,13]. In **scientific publication III,** the application of the trilinear model to LC × LC-MS and LC × LC-UV data is assessed, and different resolution methods are compared, including PARAFAC, MCR-ALS with trilinearity constraint (trilinear approach) and bilinear MCR-ALS.

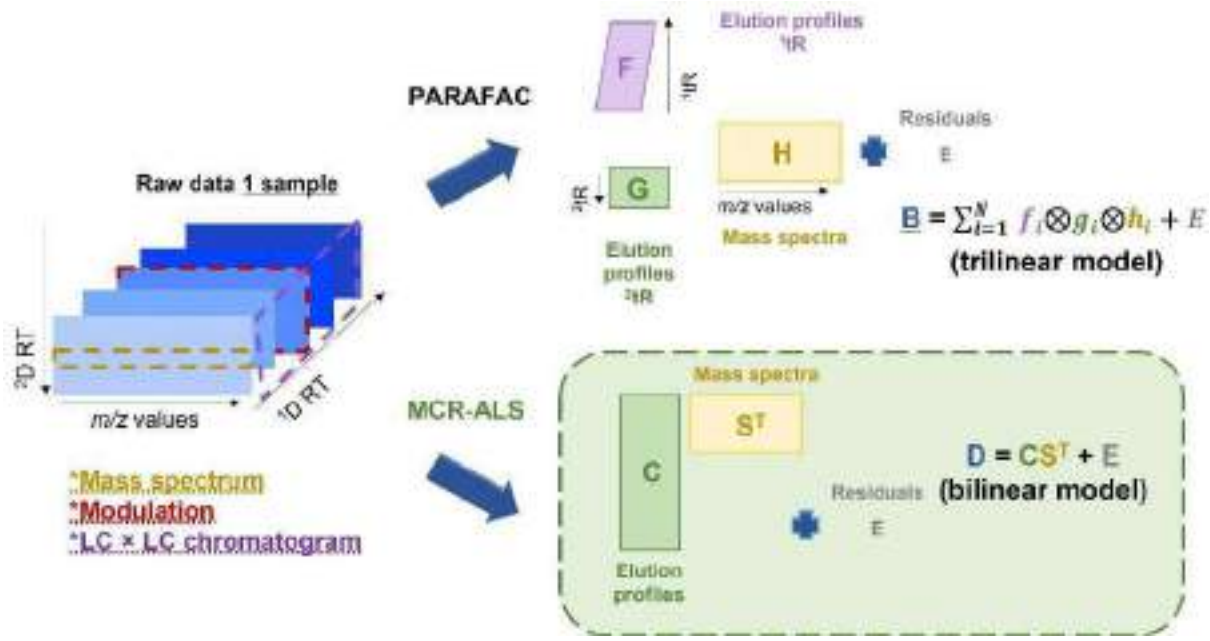


Figure 3.3. Visualization of the decomposition of a LC × LC-MS sample according to PARAFAC or MCR-ALS methods.

A wide variety of post-processing methods is currently available for metabolomic data, including unsupervised methods (e.g., principal component analysis, clustering methods or self-organizing maps), supervised methods (e.g., PLS regression-based methods, support vector machines), or pathway analysis methods (e.g., over-representation analysis, functional class scoring, Gaussian graphical models) as reviewed in [14–16]. In this PhD Thesis, these post-processing methods were applied to the peak area data matrices obtained from the MSroi GUI, in the case of **scientific publications IV, V and VI** or to peak area data matrices after MCR-ALS resolution,

in the case of **scientific publications VI and VII.** Alternatively, in **scientific publication VIII**, the post-processing step was performed on the peak area data matrices provided by vendor software or other metabolomic software such as MS-DIAL (see **Chapter 5** for more details).

The novelty of this PhD Thesis regarding post-processing analysis relies on the comparison of different ANOVA-based multivariate statistical methods in the framework of metabolomic studies. Statistical analysis is a critical step in metabolomics to assess the statistical differences between control samples and treated (or exposed) ones, according to the different factors included in the design of the experiment (DOE). In some cases, the simultaneous evaluation of multiple classes of treated samples is also pursued (e.g., several concentration levels). However, using the traditional approach for multivariate analysis of variance (MANOVA) presents serious limitations. The main problem is that this method cannot be applied if the number of variables exceeds the number of samples, which is the most common scenario in metabolomic studies [17]. For this reason, different approaches have arisen for overcoming this difficulty, such as ANOVA-simultaneous component analysis (ASCA) [18], regularized multivariate ANOVA (rMANOVA) [19] and group-wise ANOVA simultaneous component analysis (GASCA) [20]. Besides, in metabolomic studies, it is also crucial to identify the significant variables related to the factors from DOE, which can be considered potential markers of the specific treatment or experimental condition. Therefore, the ability of these methods to select these variables was also investigated in **scientific publication IV**.

Hence, an additional goal of this Chapter is to evaluate statistical methods for selecting the more significant variables (markers) in metabolomic studies.

3.2 Scientific publications

This section includes **scientific publications III and IV**, with a brief introduction and discussion of each of them:

SCIENTIFIC PUBLICATION III

This publication focuses on the chemometric evaluation of high-dimensional data, i.e., LC×LC-MS, LC×LC-UV and LC×LC-UV-MS (fused data). The multilinear behavior of the multidimensional chromatography data is assessed by considering:

1) The comparison of the singular value decomposition (SVD) of the different augmentation strategies.

2) The assessment of the core consistency diagnostic of the PARAFAC decomposition.

3) The evaluation of the data fitting using MCR bilinear and trilinear approaches. The resolving ability of the MCR-ALS method in the case of highly overlapping signals is also tested for this type of data.

Lastly, due to the rather strong deviations of the trilinear model by LC×LC datasets, bilinear MCR-ALS is proposed as the most suitable resolution method for this type of data. The quality of the results obtained when (UV and MS) are simultaneously analyzed (data fusion) is also discussed.

SCIENTIFIC PUBLICATION IV

This publication aims to assess the performance of three statistical methods based on ANOVA for metabolomics studies: 1) ASCA, 2) rMANOVA, and 3) GASCA). The evaluation is performed according to their ability to determine whether a factor from the experimental design is statistically significant, and which are the most relevant variables associated with these factors. These variables are potential markers of the experimental factors (i.e., control vs treated), which can be a key aspect of metabolomic studies. These potential markers were compared with those obtained using the PLS-DA method, which is considered the reference method in metabolomic studies for selecting significant variables.

III. SCIENTIFIC PUBLICATION III

Title: Chemometrics in comprehensive two-dimensional liquid chromatography: A study of the data structure and its multilinear behavior

Authors: Miriam Pérez-Cova, Romà Tauler, Joaquim Jaumot

Citation reference: Chemometrics and Intelligent Laboratory Systems 201 (2020) 104009

DOI: 10.1016/j.chemolab.2020.104009

Chemometrics and Intelligent Laboratory Systems 201 (2020) 104009

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Chemometrics in comprehensive two-dimensional liquid chromatography: A study of the data structure and its multilinear behavior

Miriam Pérez-Cova [a,b], Romà Tauler [a], Joaquim Jaumot [a,*]

[a] Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, 08034, Barcelona, Spain
[b] Department of Chemical Engineering and Analytical Chemistry, University of Barcelona, Diagonal 647, 08028, Barcelona, Spain

ARTICLE INFO

Keywords:
Comprehensive
Two-dimensional liquid chromatography
Multilinear
Multiset
Multivariate curve resolution

ABSTRACT

Comprehensive multidimensional chromatographic techniques, such as GC×GC coupled to FID or MS and LC×LC coupled to UV or MS, have gained popularity in recent years. From the analytical perspective, these techniques allow obtaining higher peak capacities and resolution power, as well as adding selectivity from the second orthogonal dimension. From the chemometric point of view, these multidimensional techniques generate highly complex datasets, which present several challenges for their analysis. On the one hand, the selection of the appropriate chemometric data analysis tool requires the understanding of the underlying data structure and its multilinear behavior. On the other hand, peak resolution in complex samples is still a challenge, because of their possible overlapping in one or two chromatographic dimensions despite the increased resolution power.

In this work, a comprehensive two-dimensional liquid chromatography method hyphenated simultaneously to PDA and MS detectors was employed for the analysis of a mixture of 31 pharmaceutical compounds. Chemometric evaluation of the obtained two-dimensional chromatograms focuses on two different goals. First, the assessment of the multilinear behavior of the high-dimensional data for each of the two detection modes (LC×LC-UV and LC×LC-MS) and, also, for the multiset data obtained by fusion of the data coming from both detectors. In addition, the chemometric resolution of peaks of overlapping compounds was evaluated using the multivariate curve resolution alternating least squares (MCR-ALS) method. Finally, the advantages of data fusion from UV and MS detectors were discussed, such as the increased ability for compound identification.

1. Introduction

The choice of multidimensional separation platforms for the analysis of highly complex samples have significantly increased in the last two decades [1]. Compared to mono-dimensional approaches, they can achieve higher peak capacities and, consequently, greater separation power, which leads to a more in-depth characterization of challenging mixtures [1,2]. The addition of another dimension to the separation can be performed through the combination of many different analytical techniques (e.g., ion-mobility spectrometry, IMS [3]; supercritical fluid chromatography, SFC [4]; two-dimensional gas chromatography, 2D-GC [5]; or two-dimensional liquid chromatography, 2D-LC [6]). Among them, comprehensive two-dimensional liquid chromatography, LC×LC, has arisen as an attractive approach for nonvolatile analytes in complex mixtures from different fields of application (e.g., biopharmaceuticals [7], natural medicines [8], polymers [9], proteins [10] and peptides [11], plant metabolomics [12], environmental [13] or food industries [14]). The combination of different retention mechanisms offered by LC adds extra selectivity to the separation when two uncorrelated chromatographic modes (i.e. orthogonal mechanisms) are joint. Besides higher peak capacities and peak production rates can also be achieved, which considerably facilitates peak annotation and identification [2] and is especially useful in untargeted analysis. The most frequent detectors employed in LC×LC analysis are high-resolution mass spectrometry, HRMS [7,10,15], and tandem mass spectrometry, MS/MS [10,13–15]; and ultraviolet–visible spectroscopy, UV–Vis [8], or even the combination of both [7,10,12,14]. Other possibilities include ion-mobility spectrometry, IMS [3]; evaporative light scattering detector, ELSD [9]; and fluorescence detector, FLD [11]. Some of the drawbacks still associated with comprehensive LC×LC analysis versus 1D-LC and other types of 2D-LC approaches (i.e. heart-cut or selective comprehensive) are: conceptual and instrumental complexity, long analysis time and increased difficulties in method development and data analysis [2].

The analysis of these multidimensional datasets remains a challenge

* Corresponding author.
E-mail address: joaquim.jaumot@idaea.csic.es (J. Jaumot).

https://doi.org/10.1016/j.chemolab.2020.104009
Received 16 December 2019; Received in revised form 16 March 2020; Accepted 23 March 2020
Available online 26 March 2020
0169-7439/© 2020 Elsevier B.V. All rights reserved.

due to their size and complexity [16]. To overcome the first issue, compression techniques can be used in either the chromatographic or spectral modes (especially in the case of working with data acquired using a mass spectrometry detector) [17]. In both cases, the binning method can be used for the data reduction, but there are alternatives such as wavelets in the chromatographic dimension or regions of interest in the spectral dimension which could have advantages such as the high-rate compression without losing information [18,19].

More critical is the pitfall related to the inherent complexity and structure of the multidimensional data [20]. For this reason, one of the primary aims when dealing with this type of data is to fully understand which is the inner structure of these datasets to be able to select the optimal model for the data exploration and resolution. Within the chemometrics field, there is extensively work with multidimensional chromatographic data obtained using GC×GC (comprehensive two-dimensional gas-chromatography). In this case, the trilinear behavior of the data has been demonstrated in several works and, consequently, the use of methods forcing the trilinear model (i.e. PARAFAC) gives successful results [21,22]. However, the use of other approaches that do not force the fulfilment of the trilinear model, such as the MCR-ALS method, using a bilinear decompositon model, has been proposed and also provided noteworthy results.

More recently, multidimensional liquid chromatography techniques have increased their use and popularity [23,24]. Alongside, some chemometric tools have been proposed to deal with this kind of data [25–27]. So far, most of the work has focused on the analysis of data acquired through UV–vis detectors, whereas fewer examples can still be found using mass spectrometry. In the UV–vis case, the work by Rutan group should be mentioned, as they proposed different strategies to evaluate the structure of this data and to perform the quantification of the analytes of interest [25,26,28,29]. As a general conclusion, the trilinear behavior of the data was doubtful, in particular, when a gradient was applied in the second chromatographic dimension. In the case of the MS acquired data, multilinear behavior is still a field of study. In previous work, the multilinear nature of a small region of the chromatogram was evaluated, and the fulfilment of the trilinear model was discarded [27]. However, the analyzed region was highly complex with overlapped signals, both in the chromatographic and spectral modes. For this reason, this work aims to deepen in this study by extending the analysis to the entire chromatogram and, also, considering the possible advantages that the fusion of the data acquired by UV and MS detectors can provide.

In this work, this extensive analysis of comprehensive two-dimensional liquid chromatography data was attempted pursuing two primary goals. First, the evaluation of the multilinear behavior of the comprehensive LC×LC-UV and LC×LC-MS data. Besides, the assessment of the evaluation of the fused dataset combining the information coming from both spectral modes. Second, the MCR-ALS resolution of each one of the generated datasets (LC×LC-UV, LC×LC-MS and fused) to retrieve the maximum qualitative and quantitative information.

2. Materials and methods

2.1. Chemicals and reagents

Thirty-one pharmaceutical compounds of 98–99% purity supplied by Sigma-Aldrich (St. Louis, USA) including acetaminophen, aspirin, polyethylene glycol, metformin, L-ascorbic acid, amoxicillin, levofloxacin, thiobarbituric acid, venlafaxine, ranitidine hydrochloride, caffeine, 2,4-dichlorobenzyl, chlorpheniramine, escitalopram, lidocaine, valsartan, estrone, erythromycin, ciprofloxacin, diclofenac, ibuprofen, carbamazepine, trazodone, budesonide, N-acetyl-L-cysteine, dextromethorphan, cloperastine, losartan, tramadol, sulphametoxazole, and hydrochlorothiazole were used as a complex drug mixture. Ammonium acetate (≥99.0%), acetic acid (≥95.0%) and formic acid (≥95.0%) were also purchased from Sigma-Aldrich. HPLC grade water and acetonitrile were supplied by Merck KGaA (Darmstadt, Germany).

Stock standards solutions of the pharmaceutical compounds were prepared at a concentration of 1000 mg L^{-1} in water or acetonitrile according to its solubility. A mixture working solution for chromatographic analysis including all compounds was prepared at 50 mg L^{-1} in acetonitrile. Different replicates of this sample were injected into the comprehensive liquid chromatographic system.

2.2. LC × LC–DAD–MS analysis

LC × LC analyses were carried out on an Acquity UHPLC system (Waters, Milford, MA, US) equipped with a quaternary pump and an autosampler. For the chromatographic second dimension separation, an additional LC pump (Waters 1525 binary HPLC pump) was coupled to this instrument. An Acquity UPLC Column Manager (Waters, Milford, MA, US), equipped with two 6-port two-position valves (see Fig. S1) was used as an interface between the two columns.

For the first-dimension separation an RP Zorbax Eclipse XDB-C18 (150 mm × 2.1 mm i.d.; 5 μm) provided by Agilent (Santa Clara, CA, US) was employed, under the following conditions: flow rate 39 μL min^{-1}, injection volume 20 μL, and (A) acetonitrile, (B) water, both containing 0.1% formic acid. The gradient program of the separation used started with 10% A and increased to 100% A during 30 min. Then, initial conditions were re-established and held for a 30-min column re-equilibration (see schematic representation in Fig. S2a). Total chromatographic analysis time was 80 min. In the second chromatographic dimension, a Kinetex HILIC (30 mm × 3 mm i.d.; 2.6 μm) column provided by Phenomenex (Torrance, CA, US) was used, at a flow rate of 0.5 mL min^{-1}, and 40 °C. Mobile phases consisted of (A) 5 mM ammonium acetate, adjusted with acetic acid at pH 5.5, and (B) acetonitrile. Full in fraction gradient was employed, starting at 95%B 0 min; 0–1.5 min, from 95% to 75% B; 1.5–1.6 min back to 95% B, and held until 1.8 min (Fig. S2b). Modulation time was 1.80 min.

These chromatographic conditions were established taking into account the work previously done in our research group. The RP dimension conditions were based on the work by Gomez-Canela [30] for the analysis of a complex mixture of drugs. In addition, HILIC conditions were optimized starting with the second dimension settings determined by Navarro-Reig for an LC×LC lipidomic application [31].

The first detector employed was a UV–visible diode array Waters Acquity PDA Detector (Waters Corporation, MA, USA), under the following conditions: sampling rate at 20 points sec^{-1}, pulse width at 0.1 s, 1.2 nm of resolution, and wavelength scan range from 190 to 600 nm.

The second detector, a mass spectrometer triple quadrupole detector (TQD, Waters, Milford, MA, US) equipped with an electrospray (ESI) as ionisation source, working in both negative and positives modes, was connected serially to the PDA detector. Nitrogen (purity > 99.98%) was used as desolvation gas at the flow rate of 800 L h^{-1}. Desolvation temperature and cone voltage were set at 450 °C, and 50 V, respectively. Samples were analyzed in full scan mode with a mass acquisition range from 100 to 800 Da.

For control, data acquisition and initial data preprocessing MassLynx 4.1, from the Mass Spectrometer (Waters Corporation, MA, USA), was employed.

2.3. Data import and formatting

LC×LC-UV-MS raw files (.raw) were converted to Common Data Format files (.CDF) for each acquired function (UV, MS polarities), with the DataBridge file converter provided by the MassLynx® software from Waters. Next, these CDF files for the different data functions were imported into the MATLAB computing environment (Release 2018b, The Mathworks Inc, Natick, MA, USA) by using the *mzcdfread.m* and *mzcdf2peak.m* functions available in the MATLAB Bioinformatics Toolbox (version 4.11).

ROIMCR approach was employed for importing these complex datasets [19,32]. First, the ROI strategy was used to compress the spectral

data (both UV and MS). This method is especially useful for the MS data as it allowed the selection of the most interesting mass traces [33]: those m/z values whose intensity signals were higher than a fixed signal-to-noise ratio threshold (*SNRThr*) and appeared a minimum number of times in the chromatographic mode. In the case of MS datasets, the ROI parameters were fixed to a *SNRThr* set at 0.1% of the maximum MS signal intensity, the mass accuracy of the spectrometer set at 0.5 Da/e for the TQD analyzer used in this work, and the minimum number of occurrences to be considered as a chromatographic peak was set at 25. In the case of the UV dataset, the spectral compression was not mandatory, but the ROI method also allowed a straightforward import of the data. More details about ROI strategy can be found at the work of Gorrochategui [32].

Using the ROI approach, a regular data structure (i.e. a data cube) was obtained for each sample and spectral technique (see Fig. 1). This data cube contained the chromatographic and spectral information as the three axes correspond to the two chromatographic modes, and the third mode comprised the spectral information for each of the used techniques [16]. However, for the subsequent chemometric analysis, it was more convenient to unfold this data cube into a two-way data matrix. The most common approach is to unfold the data cube keeping in common the spectral dimension. This way, several cube slices containing the spectral information for each chromatographic modulation are taken. In this strategy, the augmented matrix is built up by setting every two-dimensional chromatogram one on the top of each other in the column-wise direction. For instance, in the MS case, the data cube could be unfolded into a data matrix containing all measured retention times considering both chromatographic dimensions in the rows. In contrast, the selected number of m/z values corresponded to the number of columns of this unfolded data matrix. In the UV case, an analogous matrix was generated containing in the rows all retention times and in the columns all measured wavelengths. However, there were other options to unfold the data cube into a two-way data matrix maintaining as independent either the first or the second chromatographic dimension [27, 34].

2.4. Chemometric analysis strategy

This work pursues the assessment of different chemometric based strategies for the analysis of these complex multidimensional chromatography datasets. Therefore, the first step was the evaluation of the structure and multilinear behavior of the generated experimental data. The use of two hyphenated detectors allowed the comparison of this multilinear behavior for each one of the spectral modes and, also, for both spectral modes simultaneously, considering a classical data fusion approach of matrix augmentation by concatenation of LC×LC-UV and LC×LC-MS matrices [35]. Then, the performance of the MCR-ALS resolution method to the analysis of this mixture of drugs sample was assessed.

2.4.1. Evaluation of the multilinear behavior of LC×LC-UV and LC×LC-MS datasets

The assessment of the multilinear nature of the multidimensional chromatography data was performed on the previously described unfolded data matrices. First, the data structure of each spectral technique (either UV or MS) was evaluated by comparing the effect of cube unfolding in the singular value decomposition. There are different options to perform the unfolding of the data cube into a two-day data matrix [34]. These possibilities depend on what dimension is kept in common spectral (wavelengths or m/z values) or each one of the chromatographic dimensions. In the case of a trilinear dataset, the SVD decomposition should be the same independently of the data unfolding strategy. If deviations from this behavior are observed, it could be supposed that the dataset under study is not trilinear and, therefore, the application of pure trilinear methods could not be recommended. Another option to evaluate the multilinear nature of a dataset is to calculate the core consistency diagnostic of a trilinear method, such as PARAFAC [36]. In this case, this core consistency value should be close to 100% if the dataset under study is trilinear, and deviations from this ideal behavior will cause a decrease in this value. Finally, a third approach to assess the multilinear nature of a dataset is to evaluate the comparison of the fitting between models using a bilinear and a trilinear approach [37]. These different strategies can be easily implemented in, for instance, the MCR-ALS method, which allows the resolution of the dataset without applying (i.e. bilinear decomposition) and applying these multilinear constraints (i.e. trilinear decomposition) [38]. If fitting using these two approaches remains approximately the same, a multilinear behavior can be assumed because the imposed trilinear model fulfills the trilinear constraints [39]. In contrast, if a clear decrease in the fitting is observed when the multilinear constraints are applied, a bilinear nature should be supposed. This worsening on the data fitting can be easily explained as the multilinear constraint forces a behavior of the data that is not fulfilled.

2.4.2. Resolution of LC×LC-UV and LC×LC-MS datasets

After assessing the multilinear behavior of LC×LC-DAD and LC×LC-MS datasets, the chemometric resolution of the data was attempted. There are different options to perform this resolution but, considering the results obtained for the evaluation of the multilinear behavior, the MCR-ALS method was chosen to perform this study. This selection was motivated by the suitability of the method for the analysis of each dataset independently and the fusion of both datasets (i.e. multiset analysis). The strategy used for these two approaches is described below.

2.4.2.1. *Single dataset analysis.* The analysis of a single LC×LC dataset (either using acquisition by UV or MS) followed the typical procedure for the MCR-ALS resolution, extensively described elsewhere and briefly introduced here [40,41] (see Fig. 2a). From the different options of data cube unfolding, the MCR-ALS was performed on the two-way data matrix keeping the spectral information in common by taking the different two-dimensional chromatograms one on the top of each other (as described above in Fig. 1). This data matrix is decomposed according to a

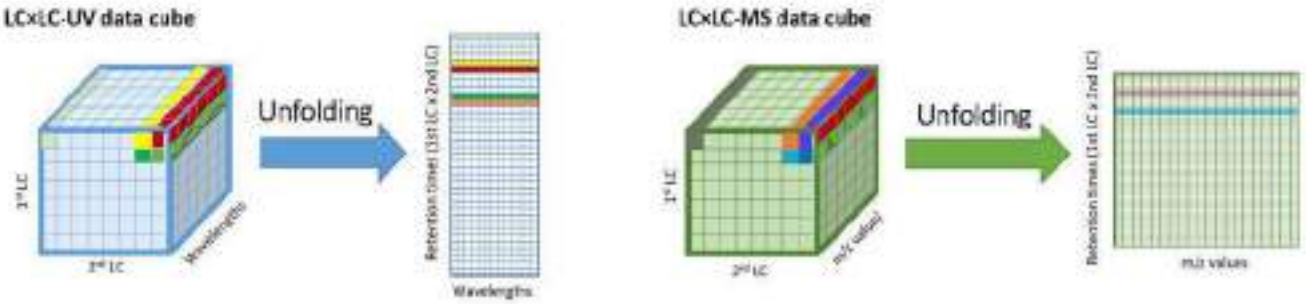


Fig. 1. Data cube obtained for both LC×LC-UV and LC×LC-MS datasets. It is also shown the usual unfolding approach to a two-way data matrix keeping constant the spectral dimension.

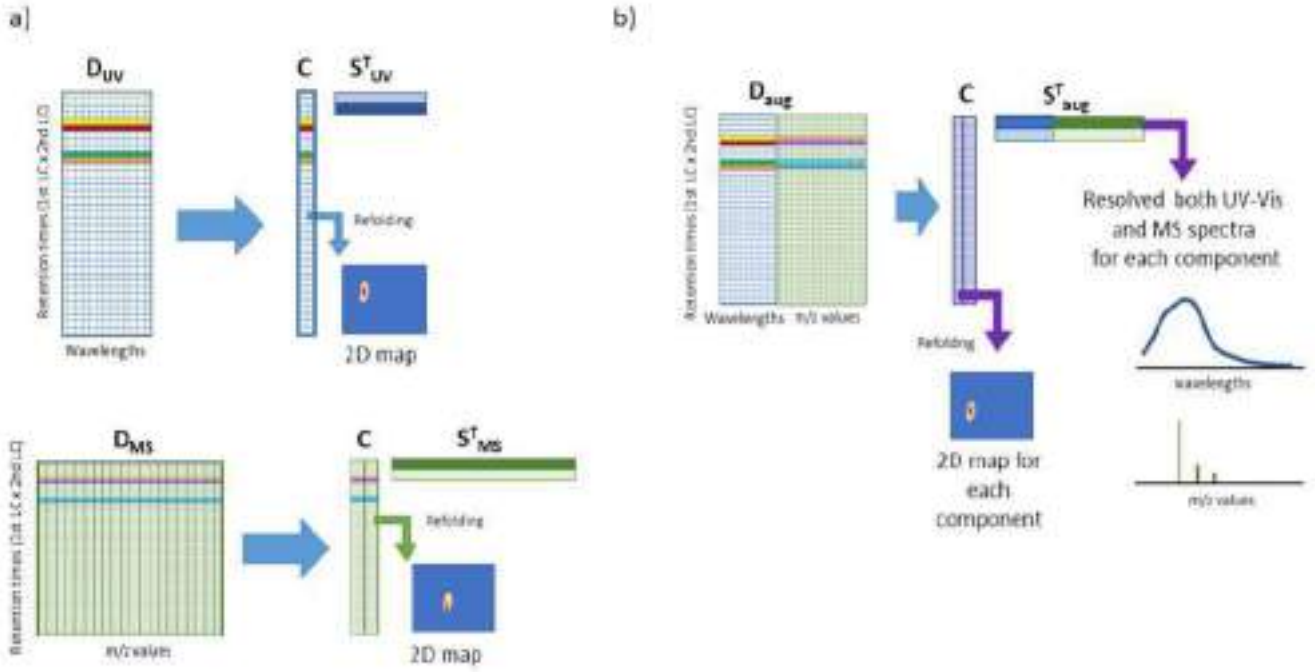


Fig. 2. a) MCR-ALS resolution procedure for the analysis of a single LC×LC-UV (up) and LC×LC-MS (down) dataset. b) MCR-ALS resolution procedure for the analysis of a multiset built up by concatenation of synchronized LC×LC-UV and LC×LC-MS datasets.

bilinear additive model defined by the generalization of Lambert-Beer's law. This bilinear model can be written as:

$$\mathbf{D} = \mathbf{CS}^T + \mathbf{E}$$

where **D** (size of I rows × J columns) is the previously described LC×LC unfolded data matrix, I corresponds to the number of retention times for all the modulations of the second chromatographic dimension, and J corresponds to the number of spectral measurements (either wavelengths in the UV case or m/z values in the MS case). **C** (I×N) is the matrix containing the resolved 2D chromatographic profiles for the N considered number of components, and $\mathbf{S}^T$ (N×J) their corresponding spectra. Finally, **E** (I×J) contains the residuals not explained by the bilinear model using this number of components.

The first step of the MCR-ALS pipeline consists of the determination of the number of components, usually performed employing a singular value decomposition approach. This procedure allows determining an estimated number of components for the analysis [38]. However, in the analysis of this multidimensional chromatographic data, this determination is not straightforward (especially in the case of the MS acquired data) as it is difficult to differentiate between those components related to the chemical signals with minor contributions and those related to random contributions such as experimental noise or baseline drifts [32]. For this reason, several models using the expected number of components (and other with one or two components or less) were built up before selecting the final number of components. The evaluation of all these models allowed refining the optimal number of components taking into account the amount of variance explained, the visual assessment of the resolved chromatographic and spectral profiles, and trying to use the minimum number of components to avoid overfitting.

Next step consists of the determination of initial estimates for the ALS optimization. In this case, an approach based on the SIMPLISMA method was employed to generate initial estimates for the spectra for the appropriate number of components [42]. Finally, ALS optimization was carried out applying some constraints to give chemical meaning to the pure mathematical solution [40]. In this case, non-negativity constraints for both chromatographic and spectral profiles and spectral normalization were employed. As stated above, in some cases, the trilinear constraint was also applied to assess the multilinear behavior of the considered dataset.

Explained variance (Equation (3)) was used as a figure of merit to evaluate the goodness of the different MCR-ALS resolution analyses [38].

$$R^2(\%) = 100\frac{\sum_{ij} d_{ij}^2 - \sum_{ij} e_{ij}^2}{\sum_{ij} d_{ij}^2}$$

where d_{ij}^2 represents the signal of the **D** matrix at a particular i-th time and j-th spectral value, and e_{ij}^2 corresponds to the residual of $\mathbf{E}_{aug}$ at the same time and spectral data.

2.4.2.2. Multiset analysis. MCR-ALS allows the analysis of the dataset containing the information from both the UV and MS acquired data. This data fusion strategy has been largely employed in the past by concatenating both datasets in a row-wise manner (see Fig. 2b) [35,40]. However, as the acquisition rate of the MS and DAD detectors was different (i.e. the same chromatographic run acquired 96001 UV spectra but only acquired 2552 MS spectra), a preliminary compression of the UV information was mandatory. The application of a binning approach allowed the synchronization of both datasets that was verified by checking some selected chromatographic peaks. Therefore, no further alignment treatments were required. In addition, an intensity scaling within the augmented matrix was performed to give both datasets (LC×LC-UV and LC×LC-MS) the same weight as differences in the magnitude of the signal intensities can be observed.

Then, the fused dataset was analyzed using the MCR-ALS method following the same strategy as in the previous case. The dataset can be decomposed following an extended bilinear model keeping the information for each one of the two spectroscopic techniques:

$$\mathbf{D}_{aug} = [\mathbf{D}_{UV}\mathbf{D}_{MS}] = \mathbf{C}[\mathbf{S}_{UV}^T\mathbf{S}_{MS}^T] + \mathbf{E}_{aug}$$

In this case, $\mathbf{D}_{aug}$ is the row-wise augmented containing the chromatographic information obtained from both detectors (UV and MS). The MCR-ALS method allows the resolution of the elution profiles in matrix **C** fusing the information from both detectors. In contrast, the MCR-ALS method recovers the spectra profiles for each spectroscopic mode

independently (S^T_{UV} and S^T_{MS}). This allows obtaining qualitative information for compound identification combining the results from UV and MS techniques [43]. This type of simultaneous analysis has demonstrated to be useful for the resolution of complex datasets such as those generated in comprehensive liquid chromatography.

3. Results

Fig. 3 depicts the data generated in the LC×LC-UV-MS analysis of the drugs mixture sample. The chromatographic details are shown in the upper panels. Fig. 3a shows the representation of the D_{MS} matrix, where the elution of compounds during all the analysis can be observed. Similarly, Fig. 3b represents the contour plot of the total ion current (TIC) 2D chromatogram in which several spots can be distinguished (some of them overlapped) corresponding to the several drugs present in the mixture. Spectral details are depicted in the lower panels. Fig. 3c shows the molecular absorption spectra acquired by the diode-array detector. In contrast, Fig. 3d represents the MS spectra in which signals at the m/z values of the analyzed drugs can be observed. Other examples of 2D chromatograms can be found in the supplementary material.

The first step in the chemometric analysis is the evaluation of the number of components present in the dataset. Fig. S3 shows the graphical results of the SVD analysis in which the comparison of the number of components detected for UV, MS and the fused UV-MS datasets can be performed. The evaluation of the values of the normalized eigenvalues shows that the MS detector allowed the reliable determination of a higher number of components; i.e. more than twenty components can be easily differentiated considering the MS data whereas only five emerged above noise considering the UV detector. When UV and MS data were fused, the SVD analysis showed an intermediate trend. The first eigenvalues showed an average behavior between both detector modes. In contrast, when considering a larger number of components, the results were closer to the eigenvalues obtained for the analysis of the MS dataset showing one of the advantages of the data fusion strategy. Therefore, the joint analysis of UV and MS datasets helped to quantify a larger number of components which may be adequately resolved in the subsequent steps of the analysis.

3.1. Trilinear nature assessment

Next, the trilinear behavior of the LC×LC-UV and LC×LC-MS datasets should be assessed. Different strategies could be employed to verify if the data fulfils the trilinear model and, here, three approaches were considered: i) the evaluation of the SVD behavior for different unfolding schemes of the data cube; ii) the evaluation of the PARAFAC core consistency of each dataset; and iii) the effect on the figures of merit of the application of the multilinear (trilinear) constraints during the MCR-ALS resolution. The comparison and integration of the results obtained by these different approaches provided a guide for the selection of the model to use for the resolution of this type of data. Fig. 4 depicts the results obtained for the different methods.

The evaluation of the results obtained for the SVD decomposition considering the different unfolding strategies indicated a non-trilinear behavior (Fig. 4a). However, MS dataset only showed minor deviations for the diverse unfolding approaches which could indicate that LC×LC-MS data was not far from a trilinear nature. In contrast, UV dataset showed significant variations in the profiles obtained for the different unfolding approaches indicating that the LC×LC-DAD data was far from showing a trilinear behavior.

Similar results were obtained when evaluating the PARAFAC core consistency (Fig. 4b). In both UV and MS datasets, only very small models (three components in the case of LC×LC-MS and two components in the case of LC×LC-UV) showed an acceptable fulfillment of the trilinear model. However, these small models did not represent most of the variation present in the dataset, as a larger number of components could be expected (i.e. mixture of 31 drugs).

Finally, the last approach to assess the trilinear nature of a dataset is the evaluation of the effect of the application of the multilinear constraint in a soft-modelling method such as MCR-ALS (Fig. 4c). In this case, a worsening of the MCR-ALS resolution results (lower explained variance)

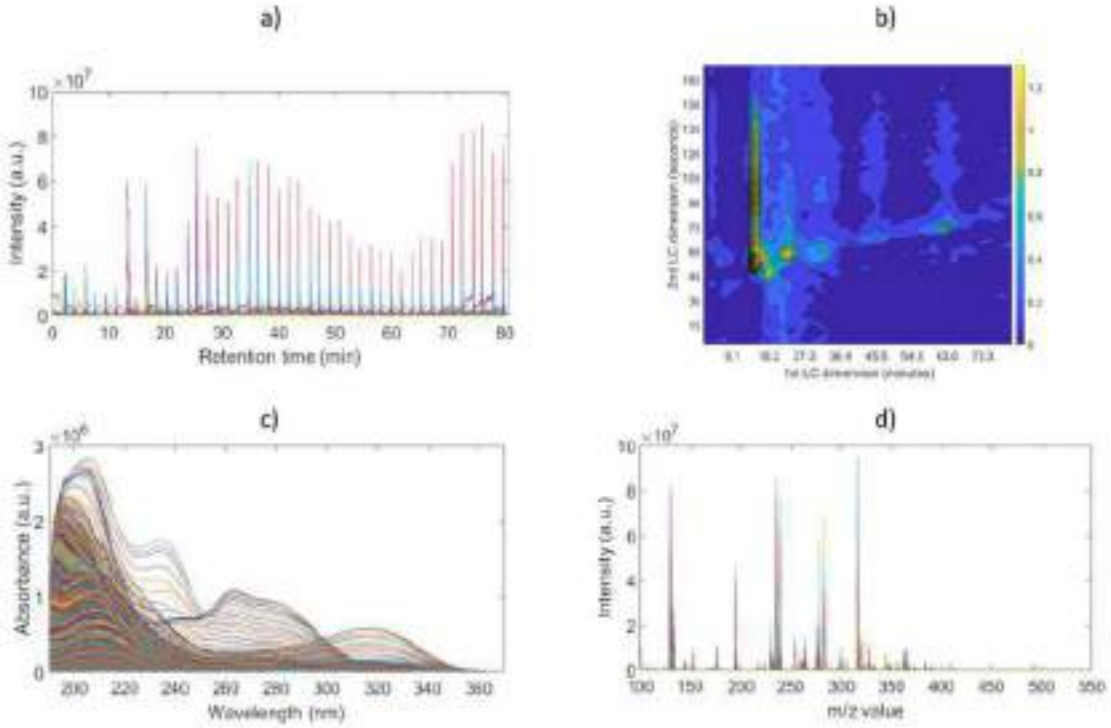


Fig. 3. Representation of the one- and two-dimensional chromatograms obtained in the analysis of the drugs sample together with their corresponding spectral information. a) Representation of the D_{MS} matrix showing all the ion chromatograms recovered after the ROI processing), b) 2D contour plot of the total ion chromatogram, c) UV–Vis molecular absorption spectra, d) MS spectra.

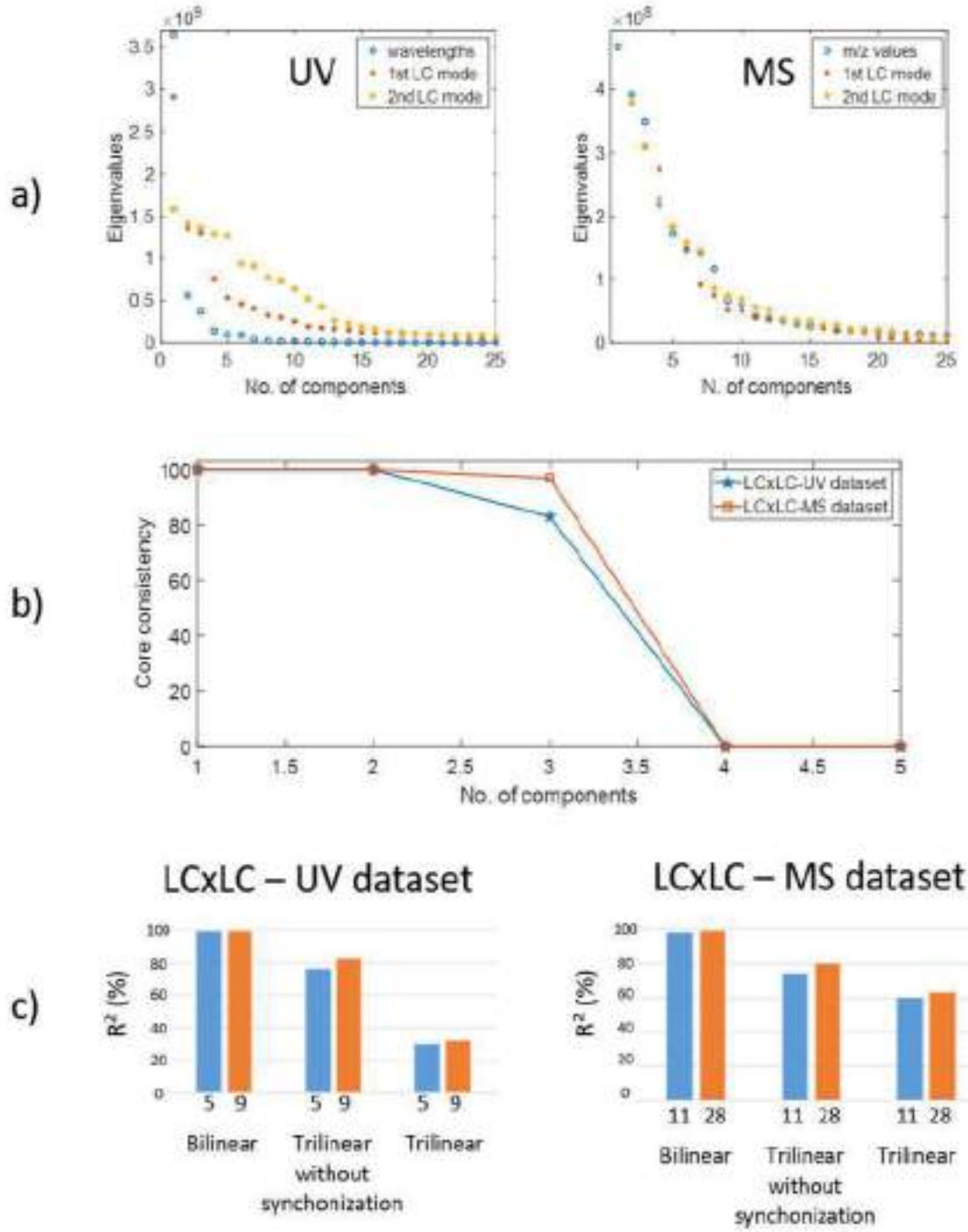


Fig. 4. Results of the evaluation of the trilinear behavior of the LC×LC datasets: a) SVD analysis, b) Core consistency analysis, c) Assessment of the effect of multilinear constraints on the MCR-ALS resolution for LC×LC-UV (right) and LC×LC-MS datasets. Barplots indicate the amount of explained variance when imposing multilinear constraints in the MCR-ALS optimization. In each case, the results obtained for a two number of resolved components are shown: LC×LC-UV (5 and 9 MCR-ALS resolved components) and LC×LC-MS (11 and 28 MCR-ALS resolved components).

can be observed for both datasets. For the LC×LC-UV data, a clear diminution of the explained variance can be seen for models built using a wide range of components from the pure bilinear behavior [almost 100% of the variance explained for the two depicted cases using five and nine components) to the application of a soft-trilinear constraint (without synchronization around an 80% of the explained variance, where using a larger number of components provided slightly better results) or a hard-trilinear constraint (less than a 40% of explained variance with similar behavior to that observed in the previous case). This trend indicated that the data does not fulfil the trilinear model and, for this reason, most of the variance remains unexplained when the model was imposed. A similar tendency can be observed when considering the LC×LC-MS dataset. In this case, the diminution of the explained variance was not so pronounced (from almost 100% of the bilinear to model to a 60% after applying the hard-trilinear constraint), but again the application of a trilinear model to resolve this dataset should be disregarded.

In conclusion, subsequent studies were performed using a bilinear model and, in particular, the bilinear approach of the MCR-ALS method.

3.2. LC×LC-UV and LC×LC-MS single dataset analysis

The bilinear MCR-ALS method was applied to solve the LC×LC-UV and LC×LC-MS datasets independently. Each one of the data cubes was unfolded into a two-way data matrix (keeping in common the spectral dimension as stated above) and, then, the classical pipeline for the MCR-ALS approach was applied. As the determination of the number of components was not straightforward, several MCR-ALS resolutions were performed using a broad range of components.

First, a small number of signals was detected in the LC×LC-UV chromatogram and the SVD analysis. Therefore, the number of components checked ranged from 5 to 14. After the evaluation of the different resolved MCR-ALS models, the minor number of components providing the maximum amount of information was nine (additional components seemed to overfit the data including mostly noise contributions, whereas some chemical information seemed to be lost when fewer components were employed). During these MCR-ALS resolutions, non-negativity constraints for both chromatographic and spectral profiles were employed as well as spectral normalization. Using these constraints and nine components, the explained variance was 99.9%. Visual inspection allowed the characterization of most of the resolved components but, in some cases, it was clear the overlapping between the signals of more than one compound. For instance, Fig. 5a-b shows the recovered two-dimensional chromatogram (after unfolding), and the corresponding UV spectrum of one of these components, respectively. In this case, a single spot can be observed in the two-dimensional chromatogram, but the shape of the resolved UV spectrum seems to be the result of more than a single compound. Therefore, the resolution of LC×LC-UV dataset did not permit to recover most of the analytical information due to the lack of selectivity of the technique (i.e. similar spectra of several of the analyzed drugs).

The analysis of the LC×LC-MS dataset overcame this limitation. In this case, the number of components tested was much larger (i.e. range from 10 to 34 components). Finally, the selected number of components was 28, which gave an explained variance of 99.6% using the same constraints as in the LC×LC-UV analysis. In this case, the MCR-ALS resolution allowed the characterization of a larger number of compounds as can be seen in Fig. 5c. MS spectra of the resolved components revealed that most of the drugs could be detected (Fig. 5d) and, therefore, identified. In addition, most of the resolved components showed a single spot on the two-dimensional chromatogram as can be seen in the reproduced chromatogram overlaying some of the resolved MCR-ALS components (noisy contributions were disregarded from this plot). However, there were still some components not resolved with more than a single contribution both in the chromatogram and the spectra profiles. For this reason, the fusion from the information coming from the MS and UV detectors could help to improve the results and detect almost all the drugs present in the mixture.

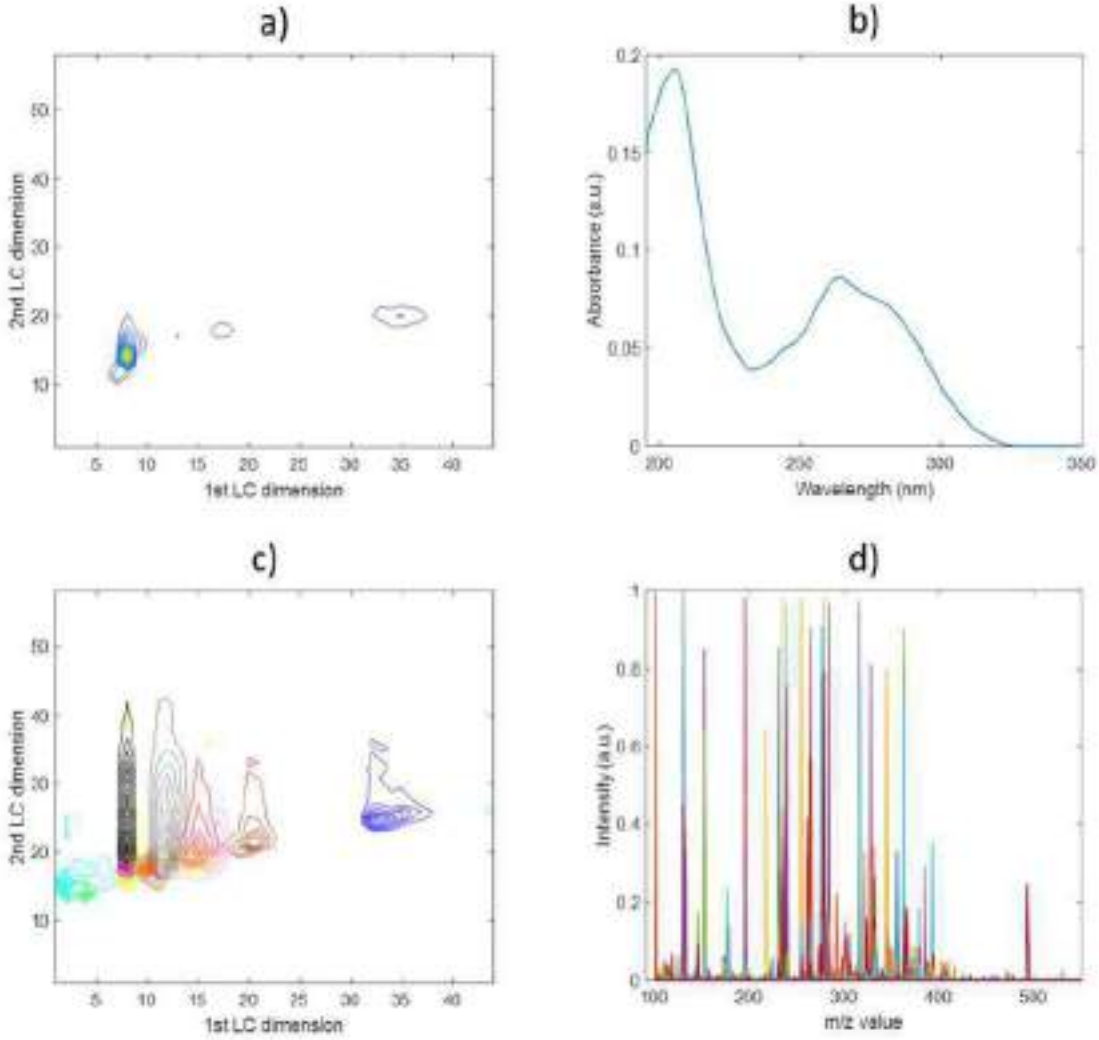


Fig. 5. Results of the MCR-ALS analysis for the individual datasets. Example component resolved in the LC×LC-UV dataset: a) 2D chromatogram, b) UV–Vis spectrum. Example component resolved in the LC×LC-MS dataset: c) Reconstruction of a 2D chromatogram with seven overlaid MCR-ALS resolved components, d) MS resolved spectra of seven MCR-ALS resolved components.

3.3. LC×LC-UV and LC×LC-MS fused dataset analysis

The fusion of LC×LC-UV and LC×LC-MS for their simultaneous analysis using the MCR-ALS method required building-up a row-wise augmented matrix. However, as the acquisition rate of the MS and DAD detectors was different (fastest in the case of the UV detector giving signals at 96001 retention times, whereas the MS detector only measured 2552 retention times) the number of rows of the unfolded two-way data matrices was not equal, and a preliminary step of data compression was mandatory. In this case, the selected approach for data synchronization was the binning of the LC×LC-UV data to reduce the total number of retention times to 2552, which allowed the concatenation of both data matrices. Once the augmented data matrix was built up, the MCR-ALS resolution was performed, as in the previous cases, following the classical pipeline. Firstly, several models using a different number of components were generated. Here, the final number of components used was 33 (a range from 28 to 36 components was tested) and the same constraints as in the previous single analysis were used: non-negativity in both chromatographic and spectral modes, and spectral normalization. Using these parameters, the explained variance was more than 99%, and almost all the drugs could be determined. Fig. 6 contains an example of the results obtained for a resolved MCR-ALS component which provided three sources of information. The refolded chromatographic profiles show the two-dimensional chromatogram of the considered component, where in many cases a single spot can be observed corresponding to a single compound (Fig. 6a). In addition, the spectral information of this component (in many cases, a single compound) was also retrieved. This component is characterized by both its UV (Fig. 6b) and MS (Fig. 6c) spectra. This improved definition is one of the main advantages of the data fusion strategies as the combination of the information coming from multiple (two, in this case) sources allows us to obtain a better model and, consequently, interpretation of the data. Here, the data fusion strategy allowed to resolve the UV spectrum of many components which were impossible to solve in the independent analysis. Moreover, the information coming from the UV data also helped to resolve the MS and chromatographic profiles that, in some cases, presented a bad characterization when analyzing the LC×LC-MS data alone.

The analysis of the spectral results also provided interesting knowledge. The resolved MCR-ALS components provided the UV and MS spectral signatures that could allow the identification of the detected compounds. In Fig. 6b and c the resolved spectra for an MCR-ALS component is shown. The MS spectrum indicates that the compound has a molecular peak at an m/z value of 315 (see in the inset the isotopic peaks) together with other minor contribution (probably fragments of the molecule). This spectral information is complemented by the UV spectrum, with a broad band around 320 nm. Using this information, a tentative identification of the resolved compounds can be performed.

The combination of the chromatographic and spectral knowledge provided by MCR-ALS permits the identification of the resolved compounds. In this case, the evaluation of the spectral results (both MS and UV data) allowed the identification of most of the drugs present in the mixture. For instance, the MS and UV spectra previously described on Fig. 6 can be tentatively assigned to ranitidine that has an exact mass of

Table 1
Summary of the identification of compounds from the MCR-ALS resolved profiles of the fused datasets.

Detected compounds in MCR-ALS resolved profiles	Not detected compounds in MCR-ALS resolved profiles
Acetaminophen	Aspirin
Amoxicillin	2,4-dichlorobenzyl
Caffeine	Budesonide
Carbamazepine	Chlorpheniramine
Ciprofloxacin	Esomeprazole
Clopenastine	Metformin
Dextromethorphan	N-Acetyl-L-Cysteine
Diclofenac	
Erythromycin	
Estrone	
Hydrochlorothiazide	
Ibuprofen	
L-ascorbic acid	
Levofloxacin	
Lidocaine	
Losartan	
Polyethylene glycol	
Ranitidine	
Sulphamethoxazole	
Thiobarbituric acid	
Tramadol	
Trazodone	
Valsartan	
Venlafaxine	

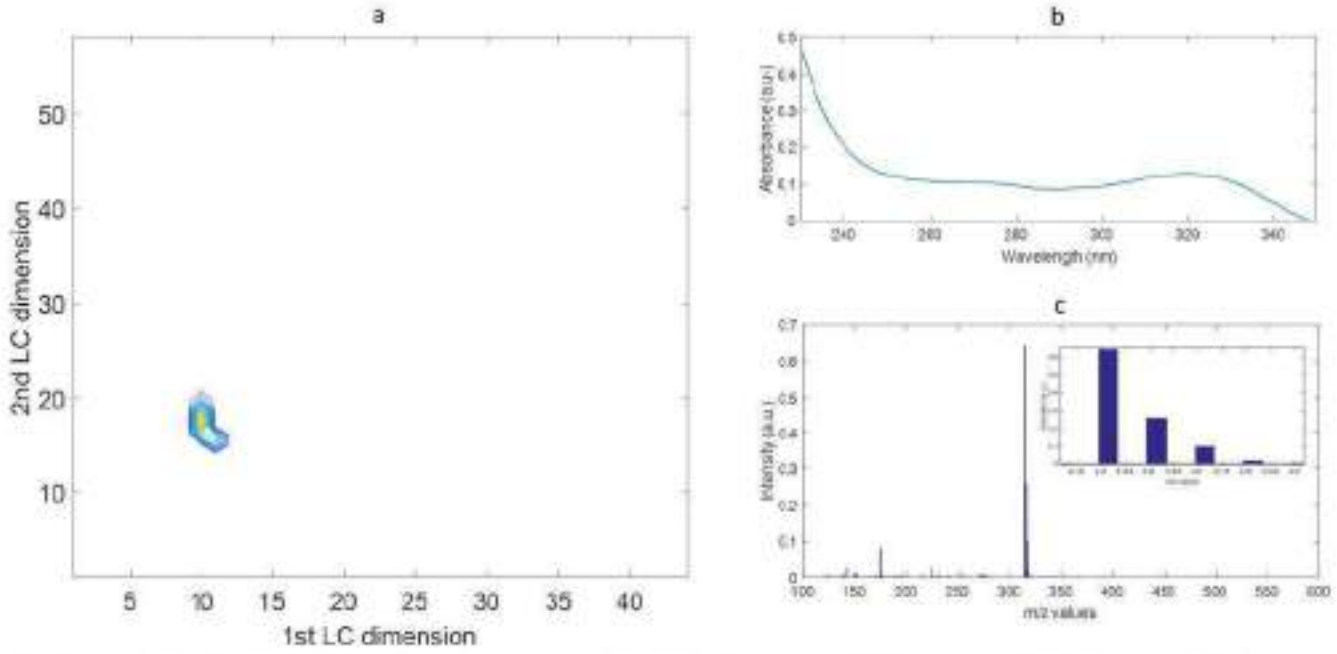


Fig. 6. Results of the MCR-ALS analysis for the fused datasets. Example of an MCR-ALS component: a) 2D chromatogram, b) UV-vis spectrum, c) MS spectrum.

314.1 Da (the detected peak was the adduct $[M+H]^+$) and, as can be found in the literature, its UV spectrum shows a broad maximum around 320 nm (depending on the solvent environment). Table 1 shows the entire list of identified compounds from the initial drug mixture.

Here, the identification of compounds from the resolved MCR-ALS components was quite straightforward as the number of compounds is reduced and *a priori* known. However, the powerfulness of the proposed strategy makes it useful in more complex cases, such as those encountered in untargeted metabolomics studies.

4. Conclusions

In this work, the evaluation of the multilinear performance of comprehensive LC×LC-DAD and LC×LC-MS data revealed deviations from the ideal trilinear behavior. These divergences cause that pure trilinear methods should be avoided for the analysis of these datasets. Therefore, multiset analysis based on matrix-augmentation strategies fulfilling a bilinear method could be recommended. For instance, the MCR-ALS method allows the resolution of these complex datasets using a large number of components, describing most of the variance present in the data. In addition, the flexibility of this method for building and processing augmented data matrices allows the simultaneous analysis of the data coming from two (or more) spectroscopic techniques. As a consequence, this data fusion approach allows obtaining higher quality results as different sources of information are resolved together (chromatographic, UV spectra and MS spectra) so that both quantitative and qualitative results (i.e. compound identification) are more reliable.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Miriam Pérez-Cova: Investigation, Visualization, Writing - original draft. **Romà Tauler:** Conceptualization, Writing - review & editing. **Joaquim Jaumot:** Conceptualization, Investigation, Writing - review & editing, Funding acquisition.

Acknowledgements

The research leading to these results has received funding from the Spanish Ministry of Science and Innovation (MCI, Grant CTQ2017-82598-P). The authors also want to grant support from the Catalan Agency for Management of University and Research Grants (AGAUR, Grant 2017SGR753) and Spanish Ministry of Science and Innovation (MCI, Project CEX2018-000794-S). MPC acknowledges a predoctoral FPU 16/02640 scholarship from Spanish Ministry of Education and Vocational Training (MEFP).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2020.104009.

References

[1] A. Bagdi, A.F.G. Gargano, J. Jacobs, Y. Meijestis, M. Honing, R. van der Wal, P.J. Schoenmakers, Comprehensive lipidomic analysis of human plasma using multidimensional liquid- and gas-phase separations: two-dimensional liquid chromatography-mass spectrometry vs. liquid chromatography-trapped-ion-mobility-mass spectrometry, J. Chromatogr. A 1530 (2017) 90–103.

[2] B.W.J. Pirok, D.R. Stoll, P.J. Schoenmakers, Recent developments in two-dimensional liquid chromatography: fundamental improvements for practical applications, Anal. Chem. 91 (2019) 240–263.

[3] P. Vivo, M. Muller, J. Vivo, M.A. Brander, A.G.J. Tredoux, H. Pasch, A. De Villiers, Comprehensive three-dimensional LC × LC – ion mobility spectrometry separation combined with high-resolution MS for the analysis of complex samples, Anal. Chem. 90 (2018) 11643–11650.

[4] M. Serrut, A. Corgier, G. Crétier, A. le Masle, S. Dubant, S. Heinisch, Potential and limitations of on-line comprehensive reversed phase liquid chromatography × supercritical fluid chromatography for the separation of neutral compounds: an approach to separate an aqueous extract of bio-oil, J. Chromatogr. A 1402 (2015) 124–133.

[5] Z. Yu, K. Huang, A. Reim, P.D. Charles, A. Northage, D. Jackson, T. Perry, R.M. Kessler, Optimizing 2D gas chromatography mass spectrometry for robust tissue, serum and urine metabolite profiling, Talanta 165 (2017) 685–691.

[6] D.R. Stoll, Recent advances in 2D LC for bioanalysis, Bioanalysis 7 (2015) 2125–2142.

[7] K. Sandra, G. Vanhoenacker, I. Vandenheede, M. Steenbeke, M. Joseph, P. Sandra, Multiple heart-cutting and comprehensive two-dimensional liquid chromatography hyphenated to mass spectrometry for the characterization of the antibody-drug conjugate ado-trastuzumab emtansine, J. Chromatogr. B: Anal. Technol. Biomed. Life Sci. 1032 (2016) 119–130.

[8] F.Q. Wang, X. Tang, J.F. Li, Y.L. Wu, Y.Y. Sun, M.J. Fang, J. Wu, X.M. Wang, Y.K. Qin, Development of an on-line mixed-mode gel liquid chromatography × reversed phase liquid chromatography method for separation of water extract from Flos Carthami, J. Chromatogr. A 1519 (2017) 145–151.

[9] A. Ndiripo, H. Pasch, Comprehensive analysis of oxidized waxes by solvent and thermal gradient interaction chromatography and two-dimensional liquid chromatography, Anal. Chem. 90 (2018) 7626–7634.

[10] B.J. Vonk, S. Wouters, A. Barzan, G. Vivo-Truyols, S. Eeltink, L.J. de Koning, P.J. Schoenmakers, Post-polymerization photografting on methacrylate-based monoliths for separation of intact proteins and protein digests with comprehensive two-dimensional liquid chromatography hyphenated with high-resolution mass spectrometry, Anal. Bioanal. Chem. 407 (2015) 3817–3829.

[11] C. Chang, C.F. Liao, Novel dual two-dimensional liquid chromatography online coupled to ultraviolet detector, fluorescence detector, ion-trap mass spectrometer for short peptide amino acid sequence determination with bottom-up strategy, J. Chin. Chem. Soc. 65 (2018) 714–725.

[12] L. Marrero, E. Ibáñez, M. Russo, E. di Sanzo, I. Restelli, A.L. Piccinelli, R. Celano, A. Cifuentes, M. Herrero, Metabolite profiling of licorice (Glycyrrhiza glabra) from different locations using comprehensive two-dimensional liquid chromatography coupled to diode array and tandem mass spectrometry detection, Anal. Chim. Acta 913 (2016) 145–159.

[13] J. Leonhardt, T. Teutenberg, G. Bandemann, O. Gassner, T.C. Schmidt, A new method for the determination of peak distribution across a two-dimensional separation space for the identification of optimal column combinations, Anal. Bioanal. Chem. 408 (2016) 8079–8088.

[14] P. Donato, F. Rigano, F. Cacciola, M. Schure, E. Farnetti, M. Russo, P. Dugo, L. Mondello, Comprehensive two-dimensional liquid chromatography- tandem mass spectrometry for the simultaneous determination of wine polyphenols and target contaminants, J. Chromatogr. A 1450 (2016) 54–62.

[15] M. Holčapek, M. Ovčačíková, M. Lísa, S. Cífková, T. Hájek, Continuous comprehensive two-dimensional liquid chromatography-electrospray ionization mass spectrometry of complex lipidomic samples, Anal. Bioanal. Chem. 407 (2015) 5033–5043.

[16] M. Navarro-Reig, L. Berka, R. Tauler, J. Jaumot, Chemometric strategies for peak detection and profiling from multidimensional chromatography, Processes (2018) 18.

[17] J.C. May, J.A. McLean, Advanced multidimensional separations in mass spectrometry: navigating the big data deluge, Annu. Rev. Anal. Chem. 9 (2016) 387–409.

[18] M. Daszykowski, B. Walczak, Use and abuse of chemometrics in chromatography, Trac Trends Anal. Chem. 25 (2006) 1081–1096.

[19] E. Gorrochategui, J. Jaumot, S. Lacorte, R. Tauler, Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: overview and workflow, Trac Trends Anal. Chem. 82 (2016) 425–442.

[20] D.E. Pinkerton, R.M. Pierce, R.E. Synovec, Chemometric Resolution of Complex Higher Order Chromatographic Data with Spectral Detection, Data Handling in Science and Technology, Elsevier, Amsterdam, Netherlands, 2016, pp. 323–352.

[21] R.M. Pierce, B. Kehimkar, J.C. Marney, J.C. Haggard, R.E. Synovec, Review of chemometric analysis techniques for comprehensive two dimensional separations data, J. Chromatogr. A 1255 (2012) 3–11.

[22] S.E. Prebihalo, K.E. Berrier, C.E. Freye, F.D. Bahaghighat, N.R. Moore, D.E. Pinkerton, R.E. Synovec, Multidimensional gas chromatography: advances in instrumentation, chemometrics, and applications, Anal. Chem. 90 (2018) 505–532.

[23] Y. Sadowski, E. Garreta-Lara, J.B. Gilmore, S. Lacorte, V. Matamoros, R. Tauler, Chemometric analysis of comprehensive two dimensional gas chromatography-mass spectrometry metabolomics data, J. Chromatogr. A 1488 (2017) 113–125.

[24] H. Puxtaur, R. Tauler, Multivariate curve resolution of hyphenated and multidimensional chromatographic measurements: a new insight to address current chromatographic challenges, Anal. Chim. Acta 86 (2014) 286–297.

[25] D.W. Cook, M.L. Hamilton, H.C. Harmes, D.R. Stoll, S.C. Rutan, Comparison of multivariate curve resolution strategies in quantitative LC×LC: application to the quantification of furanocoumarins in aqueous vegetables, Anal. Chim. Acta 961 (2017) 49–58.

[26] D.W. Cook, S.C. Rutan, D.R. Stoll, P.W. Carr, Two dimensional resolved liquid chromatography – a chemometric approach to improve accuracy and precision of

quantitation in liquid chromatography using 2D separation, dual detectors, and multivariate curve resolution, Anal. Chim. Acta 659 (2010) 87–95.
[27] M. Navarro-Reig, J. Jaumot, T.A. van Beek, G. Vivó-Truyols, R. Tauler, Chemometric analysis of comprehensive LC×LC-MS data: resolution of triacylglycerol structural isomers in corn oil, Talanta 160 (2016) 624–635.
[28] H.P. Bailey, S.C. Rutan, Chemometric resolution and quantification of four-way data arising from comprehensive 2D-LC-DAD analysis of human urine, Chemometr. Intell. Lab. Syst. 106 (2011) 131–141.
[29] H.P. Bailey, S.C. Rutan, P.W. Carr, Factors that affect quantification of diode array data in comprehensive two-dimensional liquid chromatography using chemometric data analysis, J. Chromatogr. A 1218 (2011) 8411–8422.
[30] C. Gómez-Canela, X. Rovira García, F. Martínez-Jerónimo, R.M. Marcé, C. Barata, Analysis of neurotransmitters in Daphnia magna affected by neuroactive pharmaceuticals using liquid chromatography-high resolution mass spectrometry, Environ. Pollut. 254 (2019).
[31] M. Navarro-Reig, J. Jaumot, R. Tauler, An untargeted lipidomic strategy combining comprehensive two-dimensional liquid chromatography and chemometric analysis, J. Chromatogr. A 1568 (2018) 80–90.
[32] E. Gorrochategui, J. Jaumot, R. Tauler, ROIMCR: a powerful analysis strategy for LC-MS metabolomic datasets, BMC Bioinf. 20 (2019).
[33] R. Tautenhahn, C. Böttcher, S. Neumann, Highly sensitive feature detection for high resolution LC/MS, BMC Bioinf. 9 (2008).
[34] R. Tauler, I. Marqués, E. Casassas, Multivariate curve resolution applied to three-way trilinear data: study of a spectrofluorimetric acid-base titration of salicylic acid at three excitation wavelengths, J. Chemometr. 12 (1998) 55–75.
[35] A. de Juan, R. Tauler, Data Fusion by Multivariate Curve Resolution, Data Handling in Science and Technology, 2019, pp. 205–233.
[36] R. Bro, H.A.L. Kiers, A new efficient method for determining the number of components in PARAFAC models, J. Chemometr. 17 (2003) 274–286.
[37] A. Malik, R. Tauler, Performance and validation of MCR-ALS with quadrilinear constraint in the analysis of noisy datasets, Chemometr. Intell. Lab. Syst. 135 (2014) 223–234.
[38] J. Jaumot, A. de Juan, R. Tauler, MCR-ALS GUI 2.0: new features and applications, Chemometr. Intell. Lab. Syst. 140 (2015) 1–12.
[39] R. Tauler, A. Smilde, B. Kowalski, Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution, J. Chemometr. 9 (1995) 31–58.
[40] A. De Juan, J. Jaumot, R. Tauler, Multivariate Curve Resolution (MCR). Solving the mixture analysis problem, Anal. Methods 6 (2014) 4964–4976.
[41] R. Tauler, Multivariate curve resolution applied to second order data, Chemometr. Intell. Lab. Syst. 30 (1995) 133–146.
[42] W. Windig, J. Guilment, Interactive self-modeling mixture analysis, Anal. Chem. 63 (1991) 1425–1432.
[43] M. Marín-García, G. Ioele, H. Franquet-Griell, S. Lacorte, G. Ragno, R. Tauler, Investigation of the photodegradation profile of tamoxifen using spectroscopic and chromatographic analysis and multivariate curve resolution, Chemometr. Intell. Lab. Syst. 174 (2018) 128–141.

Supplementary Material

Chemometrics in comprehensive two-dimensional liquid chromatography: a study of the data structure and its multilinear behaviour

Miriam Pérez-Cova[1,2], Romà Tauler[1] Joaquim Jaumot[1*]

[1]*Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, E08034 Barcelona, Spain.*

[2]*Department of Chemical Engineering and Analytical Chemistry, University of Barcelona, Diagonal 647, E08028 Barcelona, Spain.*

Corresponding author: Dr. Joaquim Jaumot

Postal address: Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, E08034 Barcelona, Spain.

E-mail: joaquim.jaumot@idaea.csic.es

Figure S1. Scheme of the LC × LC–DAD–MS system, showing how valves change in subsequent modulations, from Position 1 to 2, and then 1 again, consecutively.

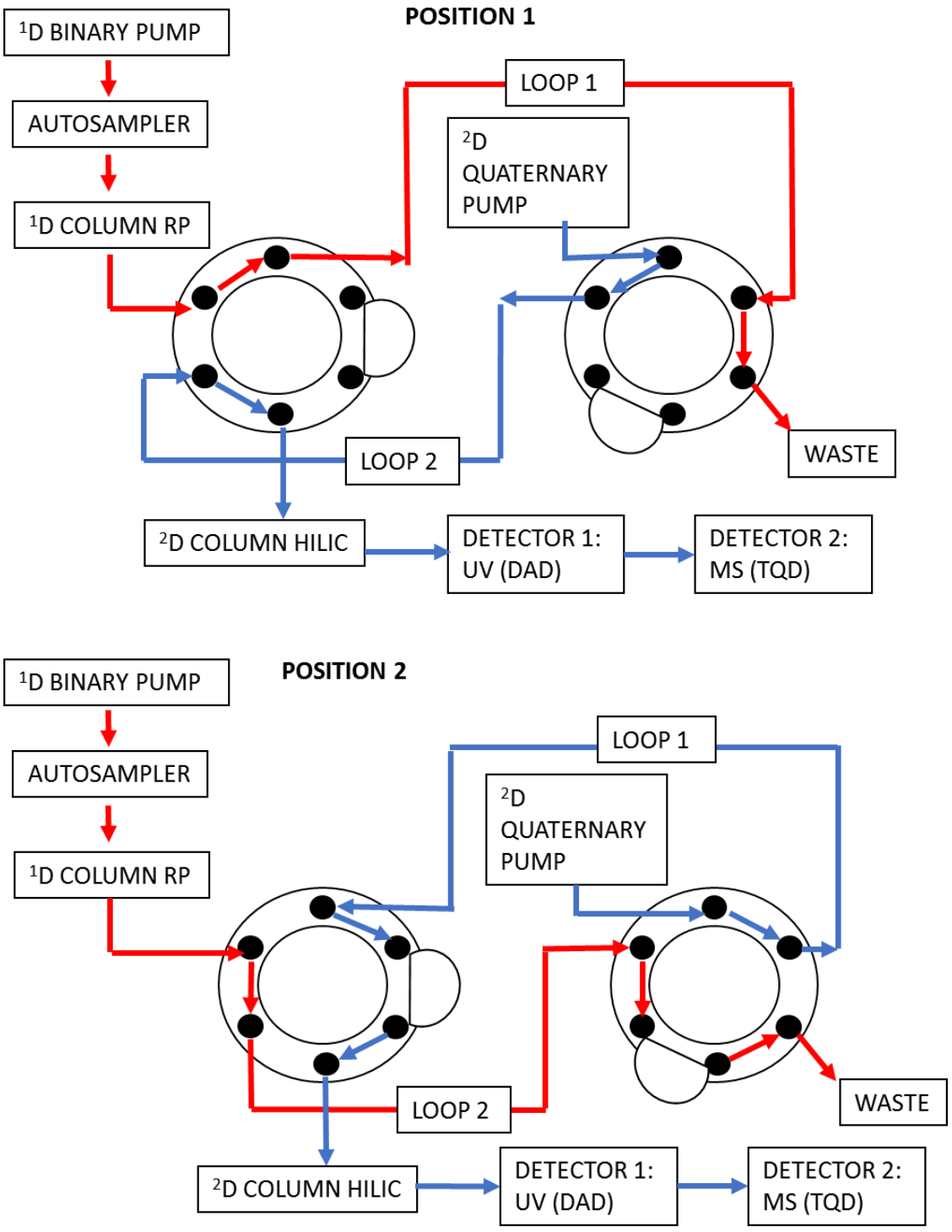


Figure S2. Gradients of the first (a) and second (b) dimensions (^{1}D and ^{2}D), increasing the percentage of the organic phase with time, AcN + 0.1% Formic acid, and AcN, respectively.

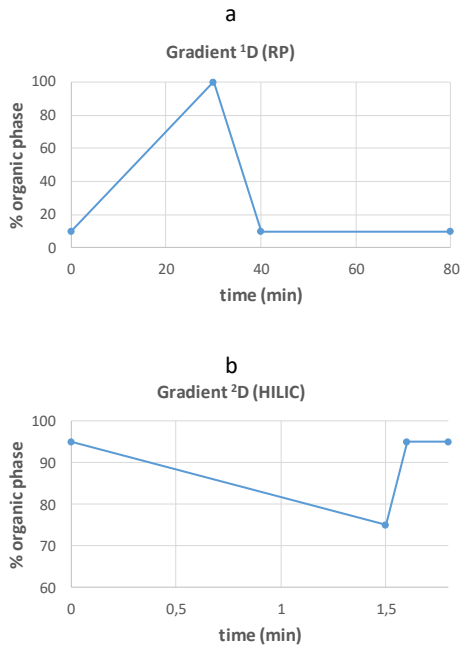


Figure S3. SVD analysis of the different LC × LC datasets. (circle) MS data, (star) UV-vis data, (asterisk) fused MS and UV-Vis data.

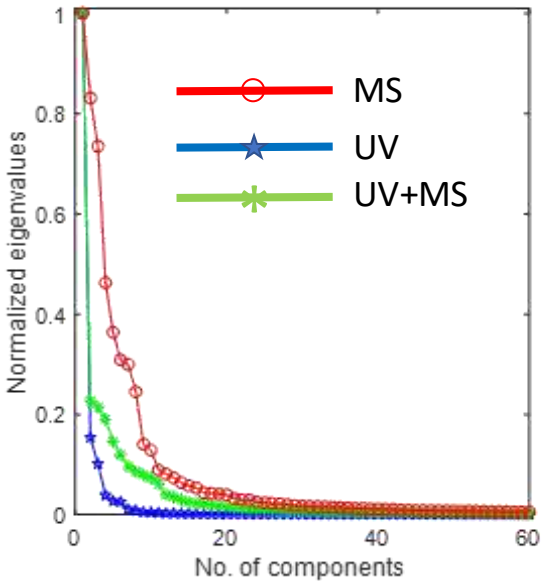
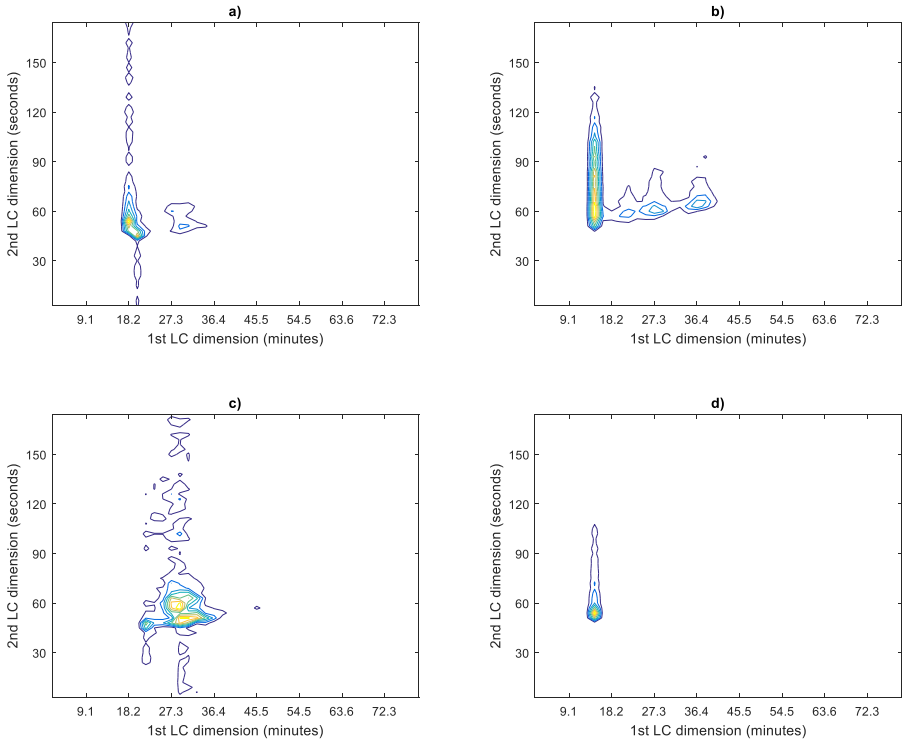


Figure S4. 2D-LC chromatograms obtained for selected *m/z* values after the ROI compression. These chromatograms can be tentatively assigned to: a) Thiobarbituric acid; b) L-ascorbic acid; c) Estrone; and d) Amoxicillin.



IV. SCIENTIFIC PUBLICATION IV

Title: Comparison of multivariate ANOVA-based approaches for the determination of relevant variables in experimentally designed metabolomic studies

Miriam Pérez-Cova, Stefan Platikanov, Dwight R. Stoll, Romà Tauler, Joaquim Jaumot

Citation reference: Molecules 27 (2022), 3304

DOI: 10.3390/molecules27103304

molecules

MDPI

Article

Comparison of Multivariate ANOVA-Based Approaches for the Determination of Relevant Variables in Experimentally Designed Metabolomic Studies

Miriam Pérez-Cova [1,2], Stefan Platikanov [1], Dwight R. Stoll [3], Romà Tauler [1] and Joaquim Jaumot [1,*]

[1] Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, E08034 Barcelona, Spain; mpcqam@idaea.csic.es (M.P.-C.); stefan.platikanov@idaea.csic.es (S.P.); rtqam@idaea.csic.es (R.T.)
[2] Department of Chemical Engineering and Analytical Chemistry, University of Barcelona, Diagonal 647, E08028 Barcelona, Spain
[3] Department of Chemistry, Gustavus Adolphus College, 800 West College Avenue, Saint Peter, MN 56082, USA; dstoll@gustavus.edu
* Correspondence: joaquim.jaumot@idaea.csic.es

Abstract: The use of chemometric methods based on the analysis of variances (ANOVA) allows evaluation of the statistical significance of the experimental factors used in a study. However, classical multivariate ANOVA (MANOVA) has a number of requirements that make it impractical for dealing with metabolomics data. For this reason, in recent years, different options have appeared that overcome these limitations. In this work, we evaluate the performance of three of these multivariate ANOVA-based methods (ANOVA simultaneous component analysis—ASCA, regularized MANOVA–rMANOVA, and Group-wise ANOVA-simultaneous component analysis—GASCA) in the framework of metabolomics studies. Our main goals are to compare these various ANOVA-based approaches and evaluate their performance on experimentally designed metabolomic studies to find the significant factors and identify the most relevant variables (potential markers) from the obtained results. Two experimental data sets were generated employing liquid chromatography coupled to mass spectrometry (LC-MS) with different complexity in the design to evaluate the performance of the statistical approaches. Results show that the three considered ANOVA-based methods have a similar performance in detecting statistically significant factors. However, relevant variables pointed by GASCA seem to be more reliable as there is a strong similarity with those variables detected by the widely used partial least squares discriminant analysis (PLS-DA) method.

Keywords: feature detection; ANOVA; ASCA; rMANOVA; GASCA; metabolomics; biomarkers

Citation: Pérez-Cova, M.; Platikanov, S.; Stoll, D.R.; Tauler, R.; Jaumot, J. Comparison of Multivariate ANOVA-Based Approaches for the Determination of Relevant Variables in Experimentally Designed Metabolomic Studies. *Molecules* **2022**, *27*, 3304. https://doi.org/10.3390/molecules27103304

Academic Editor: Irini Doytchinova

Received: 26 March 2022
Accepted: 19 May 2022
Published: 20 May 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, chemometric tools have been used to analyze omic data and, in particular, metabolomic data obtained mainly through the hyphenation of chromatographic and mass spectrometric techniques [1]. These studies have different goals for which these chemometric tools are helpful, as introduced below [2].

First, chemometric approaches such as classification (or discrimination) methods allow the differentiation of groups of samples (i.e., case and control samples) and, more interestingly, the detection of variables that discriminate between these groups [3]. These variables are usually called markers, and can be associated in metabolomic studies with specific molecules that, for instance, are altered due to a certain exposure or treatment. The selection of these relevant variables that allow the characterization of the different groups of samples (also known as the feature selection step) is critical in the analysis of metabolomic datasets. In fact, the biological interpretation of the metabolic changes observed between sample groups is often based exclusively on the selected variables. A

widely used example of these tools is the partial least squares discriminant analysis (PLS-DA) [4], a supervised method (i.e., it uses information about the identity of samples when building up the calibration model) focused on the differences between the sample types [5]. In addition, the use of variable selection methods helps to distinguish between the variables that are the most related to each type of sample and those that have a more significant influence on achieving a correct differentiation [6,7]. The two most used variable selection methods are the Selectivity Ratio (SR) [8] and the Variable Influence on Projection (VIP) scores [9]. However, as highlighted in the literature, the incorrect use of these methods can lead to misleading results because PLS-DA tends to overfit data [10,11]. For this reason, interest in alternative methods for sample discrimination and variable selection is increasing (e.g., principal component analysis in combination with linear discriminant analysis (PCA-LDA) or principal components-discriminant function analysis (PC-DFA)). Nevertheless, most of these alternative methods are non-linear approaches (i.e., machine learning approaches such as neural networks or random forests), allowing both the samples' discrimination and the determination of the variables that most strongly influence the model [11,12]. On the one hand, these methods can often handle datasets of thousands of variables as well as missing values without pre-processing required. They are also robust to overfitting and outliers. On the other hand, visualization is rather complex and difficult to interpret. The selection of the appropriate parameters is crucial, and a worse classification is encountered when compared with PLS-DA. Additionally, PLS-DA provides a better dimensionality reduction. This problem has already been addressed, and previous works have reported comparisons between different discriminant methods [11].

In contrast, methods focused on evaluating the statistical significance of the studied experimental factors have arisen in the last years. Various approaches have appeared with the common characteristic of relying on ANOVA to decompose the data variance as a function of the experimental design and the considered factors [13]. First, multivariate ANOVA (MANOVA) was proposed [14]. However, its main limitation is primarily related to the required sample size: MANOVA has a strong requirement of having more samples than variables. In metabolomic studies, the most common scenario is to have more variables than samples, which limits the success of this approach [15]. For this reason, alternative approaches were proposed allowing the multivariate data analysis without the need for meeting these strict MANOVA requirements (i.e., sample size and independency, variables multivariate normality, equal group covariance matrices) due to a previous step of data compression. These approaches can be divided into those that perform the data compression step using principal component analysis (PCA, or similar techniques such as Simultaneous Component Analysis, SCA) and those using regression-based methods such as Partial Least Squares (such as ANOVA-PLS [16] or ANOVA Target Projection [17]). Finally, PCA-based approaches seem to be more successful, which has led to the proposal of a variety of methods with this common feature.

The main difference between these PCA-based approaches is how the compression step is implemented to the factor matrices obtained after ANOVA decomposition. The ANOVA-PCA method was initially proposed [18]. In ANOVA-PCA, the residuals were added to the effects' matrix before their evaluation. Later, ANOVA Simultaneous Component Analysis (ASCA) was presented, with remarkable success [19,20]. The main difference between ASCA and the previous methods is that ASCA does not consider the residuals for modelling the ANOVA-decomposed matrices of the effects. In addition, ASCA assumes both equal variance and no correlation between the considered variables, which could affect the obtained models and hinder their interpretation. Alternative methods such as rMANOVA (regularized MANOVA) have been proposed to overcome these limitations [21]. rMANOVA is a kind of intermediate method with features between MANOVA and ASCA, since it allows the variable correlation without forcing all variance equality. Similarly, the GASCA (group-wise ANOVA-simultaneous component analysis) method has been presented [22]. GASCA attempts to overcome limitations of ASCA by

using an approximation based on group-wise sparsity in the presence of correlated variables to facilitate interpretation. These last two methods have been proposed for the analysis of omics data that are characterized by their high dimensionality in the direction of the variables (and a reduced number of samples) and their sparsity due to the presence of a large number of variables that do not present a response for certain samples (i.e., large number of zero elements) [23].

In addition to providing information on the statistical significance of the experimental factors studied, these ANOVA-based methods can also determine the variables most related to the considered experimental factor (i.e., molecules that can be considered markers for the different sample groups). Knowledge related to potential markers can be obtained similarly to that described above when PLS-DA is used. However, only some implementations of these ANOVA-based methods enable this variable selection in a straightforward way, and, in some cases, a reliable determination of potential markers is difficult to achieve [21]. Hence, an in-depth comparison of the performance of the main aforementioned ANOVA-based methods is needed, on the one hand, to evaluate the significant factors of the experimental design, and on the other hand, to assess the identification of the most relevant variables that discriminate sample groups. The ideal scenario would be to find the method that best accomplishes both goals in a single analysis.

In this work, we have evaluated the ability of these ANOVA-based methods to detect the variables responsible for the differences between groups of samples. In this way, the ASCA, rMANOVA and GASCA results are compared, taking as reference the most relevant variables determined by standard methods such as univariate statistical tests and multivariate PLS-DA analysis using VIP Scores as the variable selection method. This study was carried out using experimental datasets of different complexity obtained by liquid chromatography coupled to mass spectrometry (LC-MS). Two experiments were performed: a case with only one factor in the design (yeast samples with two extraction protocols), and a more complex case with multiple factors (zebrafish embryos samples exposed to two endocrine disruptor chemicals (EDCs), each at two concentration levels). In these examples, the effects of the design factors are analyzed both using the chromatograms (total ion current chromatograms) and from the areas of the different analytes (mass values) observed in the data.

2. Results

The performance of the different multivariate ANOVA-based methods has been compared considering the two following aspects: the statistical significance of the experimental factors (i.e., lipid extraction on yeast growth dataset and exposure level on zebrafish dataset) and the relevant variables selected for characterization of samples. This list of selected relevant variables was then compared with the results obtained by other widely used approaches (particularly PLS-DA variable selection methods).

2.1. Statistical Assessment of Experimental Factors Effects

First, the three ANOVA-based methods were compared using the TIC chromatograms of the yeast samples both in positive and negative ionization modes. In this case, the variables are retention times at which relevant compounds are eluting (e.g., these molecules are presented exclusively in only one sample group, or the peak height is different according to the various sample groups). Since a replicate of the sphingolipid samples was lost (Extraction B), a balanced data set could not be generated (eight samples of Extraction A and seven samples of Extraction B). Thus, when required, one of the samples of Extraction A was removed to allow the study to consider a balanced data set.

Table 1 shows the results obtained for the three multivariate ANOVA-based methods. In most cases, the experimental design factor studied (i.e., lipid extraction on yeast growth dataset and exposure level on zebrafish dataset) could be regarded as statistically significant. In ASCA and rMANOVA, the obtained p-values were very close to the lower threshold marked by the number of performed permutations set to 10,000, indicating a

large significant effect. In contrast, GASCA results showed some differences between the results obtained for positive or negative MS ionization. The calculated *p*-values for TICs in the positive ionization mode (0.001) were lower than those calculated for the negative ionization mode (0.039). This outcome was reasonable because the lipids extracted in both extractions provided several different signals in the positive mode (due to the variety of families of extracted lipids giving more variability in the measured signal). Still, fewer differences were observed for the negative ionization mode (i.e., fewer lipids provided an observable signal in the MS spectra). If the features matrix is considered, similar results were obtained between the two ionization modes. The ASCA results did not show relevant differences between positive and negative ionization modes, and the *p*-values obtained were similar to those for rMANOVA and GASCA. In this case, all methods detected a clear statistical significance effect for the factor representing the type of extraction (probably due to the major variability in the features matrix over the TICs matrix when considering the values for the chromatograms and the ROI determined areas).

Table 1. Summary of the statistical assessment study for the considered datasets showing obtained *p*-values for the different ANOVA-based approaches.

	Dataset	Experimental Factor	ASCA	rMANOVA	GASCA
TIC Matrix	Yeast–MS negative ionization mode	Type of lipid extraction	0.0001 (0.0007 *)	0.0001	0.039 *
	Yeast–MS positive ionization mode	Type of lipid extraction	0.0001 (0.0001 *)	0.0001	0.001 *
Features Matrix	Yeast–MS negative ionization mode	Type of lipid extraction	0.0001 (0.0001 *)	0.0001	0.002 *
	Yeast–MS positive ionization mode	Type of lipid extraction	0.0001 (0.0001 *)	0.0001	0.002 *
	Zebrafish embryos–BPA exposure	Exposure concentration			
		Control vs. Low	0.0001	0.0001	0.09
		Control vs. High	0.0001	0.0001	0.10
		Control vs. Low vs. High	0.0001	0.0001	0.01
	Zebrafish embryos–E2 exposure	Exposure concentration			
		Control vs. Low	0.4472	0.0001	0.47
		Control vs. High	0.0001	0.0001	0.22
		Control vs. Low vs. High	0.0093	0.0001	0.35

[1] Balanced data (a sample was eliminated from the set).

The study of the zebrafish samples allowed for a more in-depth study. Here, the feature matrix contains the areas of the variables filtered after ROI procedure. These variables are expressed as *m/z* values and can be associated with metabolites by their accurate mass and their fragmentation pattern (the matches between the experimental and theoretical MS/MS spectra). Regarding the zebrafish dataset, there were two possible comparisons at two dose/concentration levels for each chemical defining the studied experimental factors: control vs. low and control vs. high. In addition, there was also a three-level study considering control vs. low vs. high. In this study, it was observed that the effects caused by BPA or E2 are different. In the case of BPA, almost all comparisons provided significant *p*-values (except the two-level studies evaluated by GASCA). In contrast, for E2, only rMANOVA found a statistically significant effect of the chemical exposure in two-level studies (both low and high doses). We noted that ASCA did not detect the effect of E2 at low concentration as statistically significant, which from a biological point of view makes sense (i.e., E2 is a natural estrogenic hormone, whereas BPA is an exogenous endocrine disruptor [24]). Thus, it could be expected that E2 would have smaller effects compared to BPA. In contrast, rMANOVA found a statistically significant effect even in the case of E2, which seemed to point out it was the more sensitive ANOVA-based multivariate with

regards to detecting differences between the considered groups. However, it was not clear if these differences were caused by real potential markers, or could be related to experimental error (e.g., background contributions, badly detected metabolites). This hypothesis could be reinforced when considering the list of potential markers detected by each method (see below) and the significant differences between potential candidates detected by rMANOVA and the PLS-DA VIP scores approach. If the ternary systems were considered, all three methods identified statistically significant effects in the case of BPA, whereas there were divergent results with E2. ASCA and rMANOVA provided statistically significant *p*-values, but GASCA did not determine a ternary effect. These results agreed with what was observed in the individual two-level studies (control vs. low and control vs. high) in which there was no significant effect for GASCA in any case. In contrast, ASCA and rMANOVA gave a significant effect when the control vs. high dose was considered.

From these results, it seemed clear that they did not always provide analogous results, despite some similarities being observed between the three ANOVA-based methods. If we consider ASCA as the reference ANOVA-based method due to its most common use, a direct relationship with rMANOVA or GASCA results cannot be established. It seemed that, in general, rMANOVA tended to determine more statistically significant effects with results similar to ASCA. This behavior could be expected as there is a clear relationship between ASCA and rMANOVA established by the regularization factor (δ). In contrast, GASCA (especially in the case of the feature matrix analysis, probably due to the sparser data structure) did not detect these minor effects and, consequently, the design factor was not identified as statistically significant. For example, in the case of zebrafish samples exposed to low-level E2, no relevant effects were expected from a biological viewpoint. Furthermore, these results were confirmed by the PCA analysis of zebrafish exposed to E2 (see Figure S1 in Supplementary Materials). The PC1 vs. PC2 scores diagram shows that high-dose exposed samples grouped together, far from the control and low-dose samples. In contrast, the control and low-dose samples were much closer and, therefore, were not identified as statistically different.

2.2. Impact on Variable Selection

In addition to the previous statistical significance study, applying methods based on the combination of ANOVA and factor analysis allowed the exploration of the distribution of samples and variables in the new dimensional space defined by the principal components. Considering the scores diagrams, in all cases, the ANOVA-based methods enabled differentiation of the samples based on the factor studied, including in cases where the statistical study did not identify statistically significant factors (for example, the exposure to low concentration E2).

Figure 1 shows as an example the results obtained in two cases. In the first row, the results obtained in the study of the TICs of the yeast experiment in positive mode are shown (the factor was identified as significant in all cases). For the three methods (ASCA, rMANOVA, and GASCA), the first component differentiated by sample type (Figure 1A). The largest within-group difference was observed for GASCA, and to a minor extent for ASCA. In contrast, rMANOVA showed a significant difference among the different types of samples, but almost no differences between the different samples of a particular type. This may reflect the impact of the ANOVA decomposition in the different approaches. This decomposition seemed to force a major similarity within group samples in rMANOVA, whereas ASCA and GASCA could leave more variability.

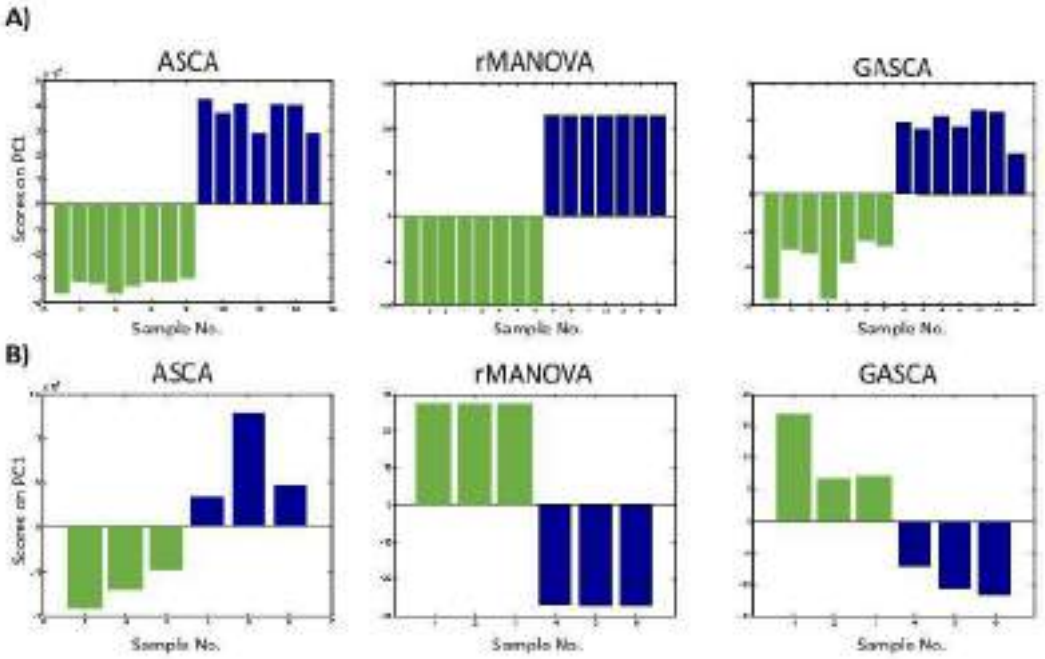


Figure 1. Exploration of the scores generated by the ANOVA-based methods: values of the first component. (**A**) TICs yeast positive. Sample colouring depending on the factor studied: green bars–extraction A: phospholipids, and blue bars—extraction B: sphingolipids; (**B**) Zebrafish embryos exposed to low–dose BPA. Sample colouring depending on the factor studied: green bars—control samples and blue bars—low-dose BPA treatment.

In the second row of plots (Figure 1.B), results for the study of the features matrix of the treatment of fish with low-dose BPA compared with controls are shown. In this case, the effect detected by GASCA was not significant (p-value > 0.05) and, although the scores plot discriminated between the control and exposed samples, this difference was minor when compared to rMANOVA, in which the behaviour of the sample types was much more distinct. ASCA showed an intermediate performance giving statistical significance to the factor, but with a representation of the score values similar to GASCA.

Figure 2 shows how the variables behave in the cases discussed above (i.e., TICs yeast positive, A and B panels, and zebrafish embryos exposed to low-dose BPA, C, and D panels). In addition to the profiles provided by the ANOVA-based methods and to compare with a widely used method in the field of metabolomics, profiles of the variable selection approaches (e.g., Selectivity Ratio and the VIP Scores) obtained by PLS-DA are also shown. In these PLS-DA models, classes were defined according to the used experimental design. For instance, in the yeast growth studies, samples from extraction A were set as a class (0) whereas samples from extraction B were set as another class (1). Figure 2A shows the profiles obtained by PLS-DA, and Figure 2B shows those obtained by methods based on different multivariate ANOVA methods previously used for the analysis of the TICs yeast study in positive ionization mode. In both cases, variable channels (i.e., retention times of the TIC chromatograms) between 200 and 250 were highlighted as the chromatographic regions enabling differentiation between sample types. Moreover, when considering the first loadings profiles corresponding to the ANOVA-based methods factor decomposed matrices, different patterns could be distinguished. The ASCA loading profile resembled the TIC chromatogram of the sample and the PLS-DA selectivity ratio profile (see similarity between these profiles in Figure 2A,B). Instead, the GASCA loading profile was more similar to the profile obtained for the PLS-DA VIPs scores. Finally, the rMANOVA profile was the most different to the other methods, but positive and negative features were observed in the profiles.

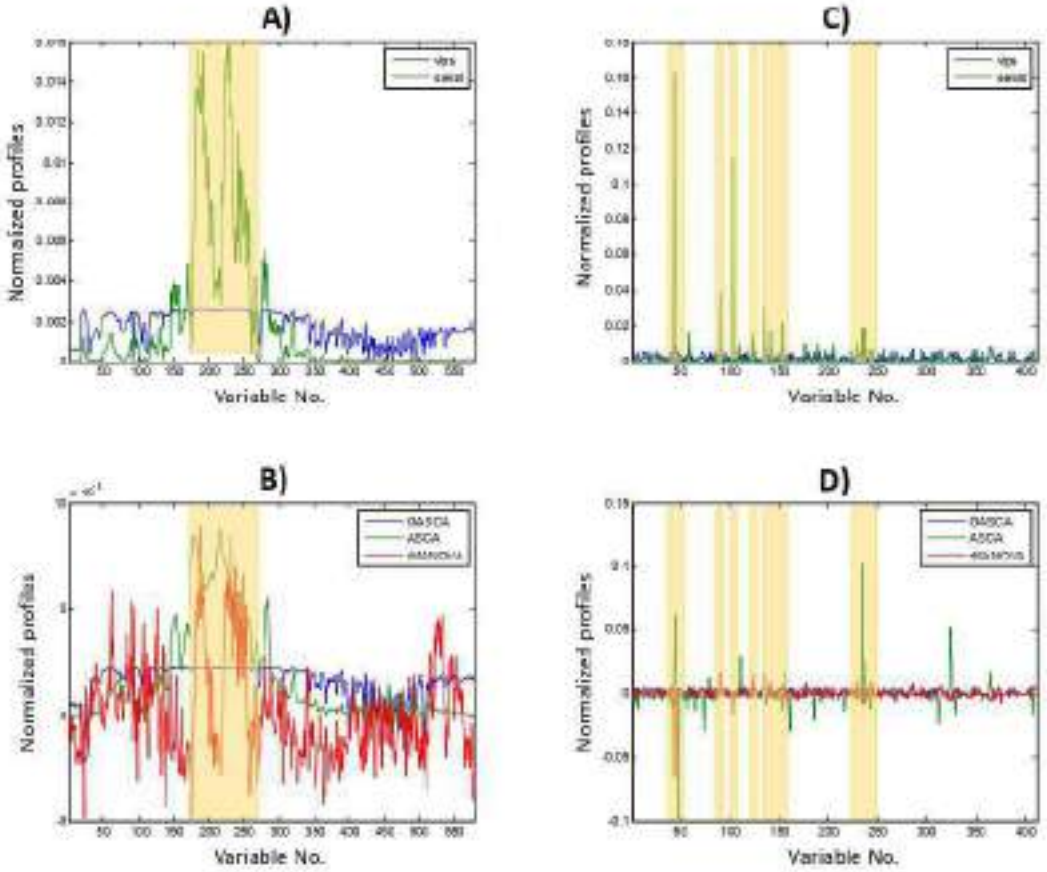


Figure 2. Comparison of the loadings obtained by PLS-DA variable selection methods and ANOVA-based methods. (**A**) TICs yeast positive PLS-DA profiles: VIP scores and selectivity ratio; (**B**) TICs yeast positive ANOVA–based approaches: ASCA, rMANOVA, and CASCA loadings; (**C**) Zebrafish embryos exposed to low–dose BPA PLS-DA profiles: VIP scores and selectivity ratio; (**D**) Zebrafish embryos exposed to low–dose BPA ANOVA–based approaches: ASCA, rMANOVA, and GASCA profiles. In each plot, profiles were normalized to an equal area for representation in the same scale. Shadowed boxes represent regions with a high number of relevant variables.

In the case of the features matrices from the ROI analysis for the study of zebrafish embryos with low-dose exposure, the loadings profiles are shown in Figure 2C for the PLS-DA based methods and Figure 2D for the ANOVA-based methods. Similarly, there is an observable link between the variables relevant for both PLS-DA and ANOVA-based methods (i.e., regions that showed larger positive contributions for PLS-DA methods and positive or negative contributions for ASCA methods, as highlighted by the shadowed boxes in each figure). Focusing on the ANOVA-based profiles, the ASCA profile was the most different from the other approaches. A quantitative evaluation of the similarity of the profiles can also be performed by calculating the correlation coefficient between the different sets of profiles (Table S1). For example, in the case of the study of yeast TICs in the positive ionization mode, the ASCA profile was more similar to that obtained by the selectivity ratio approach (0.83), while GASCA was more similar to the profiles obtained using the VIPs scores (0.93). In contrast, rMANOVA was more different to the other profiles (lower coefficient values). Additionally, the same trend was also observed for the rest of the studies when considering the entire Table S1. In general, a good similarity was found between the loading profiles obtained for the different approaches with relatively large correlation coefficients. However, the similarity of the loading profiles resolved for CASCA and the PLS-DA VIP scores could be highlighted because, in all cases, they showed the highest correlation values (all had a value above 0.89).

Next, the matching variables selected as relevant from the different approaches were compared, taking as a benchmark the variables determined by the field's reference (i.e., PLS-DA VIP Scores). Figure 3 shows the logical relations between these selected relevant variables for the different approaches using Venn diagrams. Here, four cases were considered: TICs and features matrices of the yeast study in the positive mode and at two levels of BPA exposure to zebrafish embryos. In all cases, the total number of relevant variables has been limited to 50 to focus on the variables with greater importance, and in an attempt to avoid coincidences by chance. In the case of PLS-DA VIP scores and Selectivity Ratio methods, those 50 variables with the higher values were selected. For the ANOVA-based methods, these variables showing the 50 largest loadings values in absolute value for the first component were selected. As shown above, this first component was enough to distinguish between the various sample types for the studied cases when considering the related factor matrix from the ANOVA decomposition. The results obtained in the analysis of the data from the rest of the examples are shown in the Supplementary Materials, giving concordant results (Figure S2).

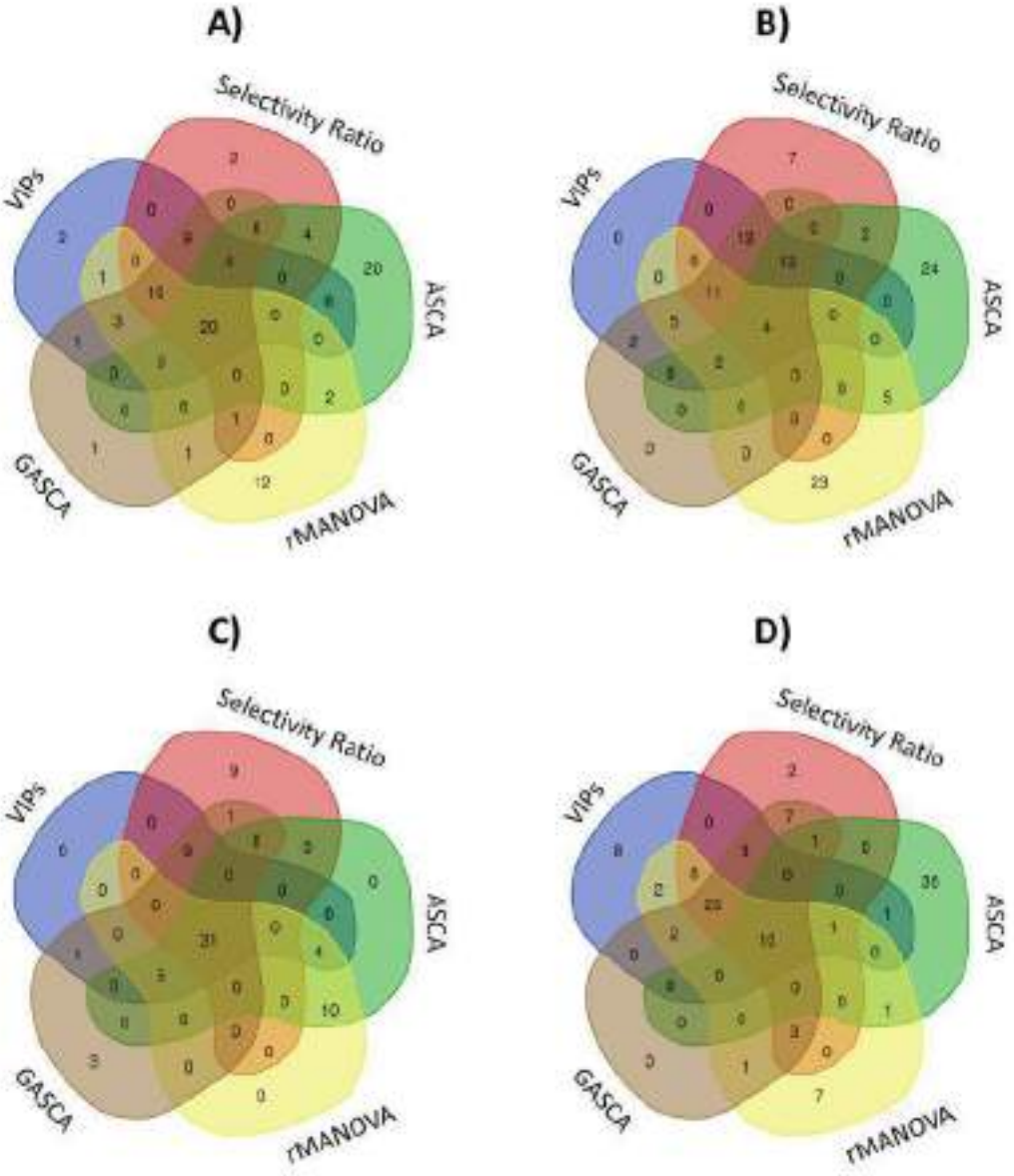


Figure 3. Venn diagrams summarizing the relationships among the 50 different selected variables detected for each data set. (**A**) TICs matrix for yeast positive MS ionization mode; (**B**) Features matrix for yeast positive MS ionization mode; (**C**) Zebrafish embryos exposed to low-dose BPA; and (**D**) Zebrafish embryos exposed to high-dose BPA.

Figure 3A shows results from the study using the TICs obtained for yeast in positive ionization mode. The Venn diagram showed that 20 variables are common to all considered approaches. Only ASCA (20) and rMANOVA (12) presented a relevant number of unique variables detected only by one method. These results confirmed the previous evidence in which the variable selection profiles or loadings associated with each method were evaluated. GASCA was the ANOVA method that provided the most similar results compared with PLS-DA. Figure 3B shows the evaluation of the corresponding selected variables obtained after preprocessing the LC-MS yeast samples in the positive mode. In this case, the number of variables common to all approaches is much lower (4). Again, only ASCA and rMANOVA have many unique variables (causing this low number of coincident variables). When considering the study of zebrafish embryos treated with BPA at two dose levels (Figure 3C,D), the obtained results led to similar conclusions. However, ASCA showed a different behaviour compared to all the other methods. For instance, in the control vs. high BPA exposure study, ASCA had many unique variables that avoided the coincidence from other methods. A list of identified metabolites present in zebrafish embryos is included in Table S5 (only the compounds that were characterized at MS/MS level are included). The significance obtained with each of the methods tested (VIPs, selectivity ratio, GASCA, ASCA, and rMANOVA) is included for each compound. Again, ASCA provided higher statistical values to different compounds than the rest of the methods, in agreement with the analysis from Figure 3.

Finally, the univariate ANOVA and multivariate ASCA-based profiles were individually compared with those retrieved by the PLS-DA VIP-Scores approach (Supplementary materials Table S2). In the case of yeast TICs data, the coincidence of the detected variables with PLS-DA was larger with GASCA (47 of 50), followed by rMANOVA (34), and finally ASCA (24 of 50). When comparing the selected variables with the PLS-DA, this trend occurred in most studied cases (Figure 4). In summary, rMANOVA and GASCA could be better options if the main goal of the study is variable selection after the ANOVA decomposition stage. This fact confirmed the theoretical basis from which the ASCA method had the initial purpose of statistical assessment of factors in experimental design and data exploration by SCA. However, the newly developed methods such as rMANOVA or GASCA showed advantages when the aim of the study was to perform feature detection to characterize the experimental design factors.

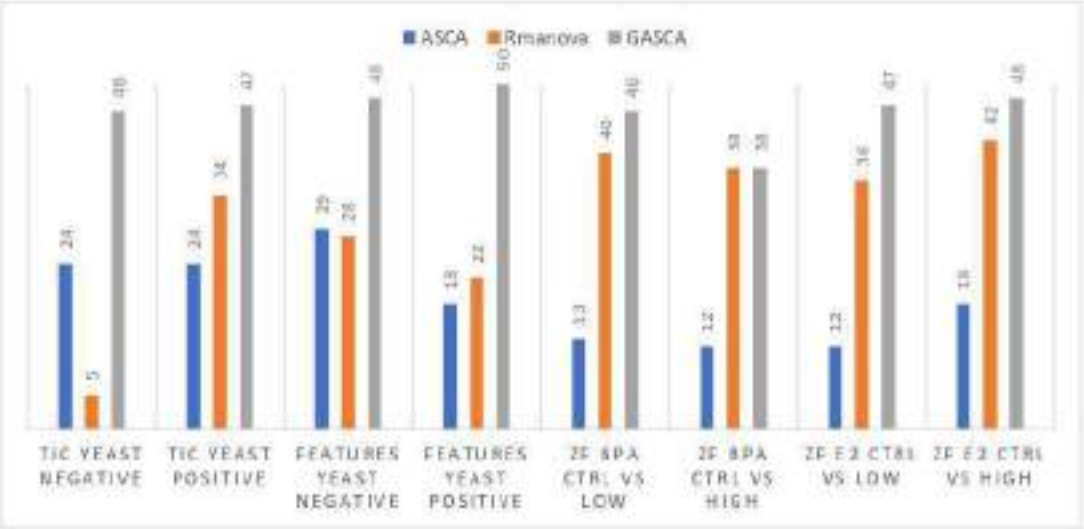


Figure 4. Comparison of the number of coincident variables detected by PLS-DA and the considered ANOVA-based methods. Maximum possible number of coincidences is 50.

3. Materials and Methods

3.1. Chemicals and Reagents

Bisphenol A (BPA, ≥99.0%), 17-β-estradiol (E2, ≥98.0%) methylene blue (certified by the Biological Stain Commission, ≥82.0%), calcium sulphate ($CaSO_4 \cdot 2H_2O$, ≥99.0%), dimethyl sulfoxide (DMSO, for molecular biology, ≥99.9%), potassium hydroxide (KOH, ≥85.0%), ammonium acetate (NH_4Ac, ≥99.0%), formic acid (HForm, ≥95.0%), acetic acid (HAc, ≥95.0%), phosphate buffered saline (PBS), yeast extract, bacteriological peptone, and D-glucose were purchased from Sigma-Aldrich (Merck, Darmstadt, Germany). Ammonium formate (NH_4Form, ≥99%) was obtained from Fluka Analytical (Honeywell, Muskegon, MI, USA). Chloroform ($CHCl_3$, ≥99.0%) was provided by Carlo-Erba reagents (Dasti Group, Milan, Italy). Instant Ocean sea salt was purchased from Aquarium Systems (Sarrebourg, France), whereas dried flakes were obtained from TetraMin (Tetra, Melle, Germany). HPLC grade water and acetonitrile (AcN) were supplied by Merck KGaA (Merck, Darmstadt, Germany), and methanol (MeOH) HPLC grade from Fisher Chemical (Thermo Fisher Scientific, Fair Lawn, NJ, USA). L-methionine sulfone was purchased from Sigma-Aldrich (St. Louis, MI, USA). Lipid standards used were purchased from Avanti Polar Lipids (Alabaster, AL, US). The glycerophospholipids and triacylglycerides (PL) standards mix included: 1,2,3-17:0 triglyceride (TG), 1,3-17:0 (d5) diglyceride (DG), 17:0 cholesteryl ester (CE), 16:0 D31-18:1 phosphatidylethanolamine (PE), 16:0 D31-18:1 phosphatidylserine (PS), 16:0 D31-18:1 phosphatidylglycerol (PG), 16:0D31-18:1 phosphatidylcholine (PC), 17:1 lyso PC (LPC), 17:1 lyso PE (LPE), 17:1 lyso PG (LPG), and 17:1 lyso PS (LPS). The sphigolipids (SL) standards mix included: N-dodecanoylsphingosine, N-dodecanoylglucosyl-sphingosine, and N-dodecanoylsphingosylphosphorylcholine.

3.2. Yeast Experiments

First, the performance of the different statistical methods was explored by studying the growth of yeast culture and considering the method used for lipid extraction (either for glycerophospholipids or sphingolipids) as an experimental factor. Therefore, results obtained from the two lipid types of extractions are compared.

3.2.1. Culture Growth

A preculture of *Saccharomyces cerevisiae* (BY4741 strain) was kept at 30 °C and agitated at 150 rpm in yeast extract peptone dextrose (YPD) medium (composed of 20 $g \cdot L^{-1}$ bacteriological peptone, 10 $g \cdot L^{-1}$ yeast extract, 20 $g \cdot L^{-1}$ glucose at 40%) for 48 h [25]. Inoculation with the preculture was performed with a fresh YPD medium to an absorbance of 0.1 at 600 nm (A_{600}). When an A_{600} of 0.6 was reached, 20 mL of each culture were taken, including five biological replicates for each condition. Fractions were centrifuged (3 min at 3000 rpm at 4 °C) and washed twice with PBS. The supernatant was discarded, and pellets transferred to Eppendorf tubes were kept at −80 °C until extraction.

3.2.2. Lipid Extractions

Lipids from yeast samples were extracted by two different procedures, based on the previous work from Puig-Castellví [25] and Dalmau [26], with minor modifications. The first approach was a general lipid extraction, mainly targeting glycerophospholipids and triacylglycerides (Extraction A). The second extraction included a saponification step focused on the analysis of sphingolipids (Extraction B).

Extraction A started with the addition of 400 µL of Milli-Q water to the frozen samples. Then, samples were vortexed and transferred to glass vials. Next, 1 mL of MeOH, 2 mL of $CHCl_3$, and 40 µL of PL standard mix at a concentration of 20 µM were added. Vortex and ultrasonication steps were applied to the vials in cycles of 3 min and 15 min.

respectively. Glass beads were added, and the previous step was repeated twice. Samples were left overnight at 48 °C in a thermostatic bath, then evaporated to dryness under nitrogen gas and stored at −80 °C until use. Before analysis, extracts were re-suspended in 800 µL of MeOH, centrifuged 3 min at 10,000 rpm at 4 °C, and aliquots of 200 µL were transferred to chromatographic vials, where a 10 µL aliquote of SL standards mix was added to each vial.

Extraction B differed in the initial proportion of the employed MeOH/$CHCl_3$ mixture, which was 2/1 in this case. The SL standard mix was added at the beginning in the same proportion as PL for set A. After overnight incubation, 75 µL of KOH of 1 M in MeOH were added to the samples, that were then sonicated for 15 min and kept for 2 h at 37 °C (i.e., a saponification step). Next, KOH was neutralized by adding 75 µL of 1 M HAc, and samples were evaporated until dryness under nitrogen gas and stored at −80 °C until further use. Before analysis, samples were resuspended as for set A, adding 10 µL of PL standards mix instead of the SL standards mix.

Quality controls samples (QCs) were composed of 25 µL of each biological replicate from each set of samples (extracted samples A and B).

3.2.3. LC-MS Analysis

A total of five samples (biological replicates) of each extraction were randomly analyzed; one sample per set of extraction was also analyzed in triplicate (instrumental replicates). In total, 16 chromatograms were obtained (eight for each extraction batch). QCs and blanks were also interspersed in the chromatographic sequence. LC-MS analysis was carried out with a Waters Acquity UPLC system coupled to a Waters LCT Premier orthogonal accelerated time-of-flight mass spectrometer (Waters), operated in both positive (ESI+) and negative (ESI-) electrospray ionization modes. Full-scan spectra from 50 to 1500 Da were acquired at a scan cycle time of 0.3 s. The chromatographic method employed was described previously [25]. Briefly, an RP C8 Acquity UPLC bridged ethylene hybrid (Waters) column of 100 mm × 2.1 mm i.d. (1.7 µm) was employed. Mobile phases were: A) MeOH 1 mM NH_4Form and 0.2% HForm; B) H_2O 2 mM NH_4Form and 0.2% HForm. The solvent elution gradient started at 80%A, increased until 90%A at 3 min, held at 90%A until minute 6 min, increased to 99%A at 15 min, and held at 99% until 18 min. Then, the column was re-equilibrated during 2 min. Flow rate, injection volume and column temperature were set at 0.3 mL min^{-1}, 10 µL and 30 °C, respectively.

3.3. *Zebrafish Embryos Experiments*

The second dataset aimed to evaluate of the effects caused by two endocrine-disrupting chemicals (BPA and E2) in zebrafish embryos. In this study, we assessed the impact of the different concentration exposure levels in the development of zebrafish embryos by considering the changes in the metabolome.

3.3.1. Zebrafish Maintenance

Adult wild-type zebrafish (*Danio rerio*) were fed twice a day with dried flakes and maintained at a temperature of 28 ± 1 °C, with photoperiods (light-night) of 12 h. Fish water, prepared in Milli-Q water, contained 90 µg·mL^{-1} of Instant Ocean sea salt and calcium sulphate (100 µg·mL^{-1}), as previously reported [27]. Zebrafish embryos were obtained by natural mating placing five females and three males in 4-L breeding tanks. Eggs were separated from adults through a bottom mesh. At 2 h post-fertilization (hpf), eggs were collected and rinsed. At 24 hpf, fertilized eggs were washed three times with 0.0002% methylene blue and randomly distributed in 6-well multiplates as follows: 15 individuals per 5.0 mL of fish water, 8 replicates of each condition, in different plates, to account for possible "tank" effects.

All experiments were approved by the Institutional Animal Care and Use Committees at the Research and Development Centre of the Spanish National Research Council

(CID-CSIC) and were also conducted under the institutional guidelines under a license from the local government (DAMM 7669, 7964).

3.3.2. Exposure Protocols

BPA and E2 working solutions were prepared daily in fish water at a final concentration of 0.2% DMSO by diluting from stock solutions at higher concentrations in DMSO, previously prepared and kept at 4 °C until use. Exposure concentrations were chosen by a preliminary range-finding test and based on previous studies [27,28].

Until 48 hpf, embryos were kept in fish water to avoid early embryonic processes. Then, exposure started, and solutions were changed daily to ensure continuous exposure to the contaminant until embryo collection at 120 hpf. Control samples in 0.2% DMSO (without treatment) were also included in the multiplates. The following concentrations were used as low and high nominal exposure concentrations and using DMSO as a vehicle; BPA concentration levels were set to 4.4 and 17.5 µM, respectively, whereas, for E2, 1 and 4 µM concentrations were used, respectively. Pools of 30 zebrafish embryos were gathered (15 + 15 from different wells from the same plate) for each biological replicate. A total number of three biological replicates per treatment were used for LC-MS.

3.3.3. Metabolite Extraction

The frozen Eppendorf tubes containing the embryos were kept in dry ice. Then, 0.900 mL of methanol and 90 µL of L-methionine sulfone at 50 mg L^{-1} were added to each sample. Samples were vortexed, sonicated for 15 min, and centrifuged at 14,500 rpm for 10 min at 4 °C. Next, the supernatant was isolated, and 500 µL of water and 300 µL of $CHCl_3$ were added. Samples were vortexed again, placed on ice at 4 °C, and centrifuged under the same conditions. Aqueous fractions (upper layer) were collected and evaporated to dryness under nitrogen gas. Samples were re-suspended in 100 µL of AcN:H_2O (1:1), centrifuged, and transferred to a chromatographic vial, where they were evaporated until dryness and kept at −80 °C. Finally, extracts were re-suspended before injection with 100 µL of AcN:H_2O (1:1).

Quality control (QC) samples were generated by pooling 10 µL of an extract from each condition studied (two concentration levels of both compounds, BPA and E2, plus control samples).

3.3.4. LC-MS Analysis

Three biological replicates were analyzed for each sample condition (control, low, and high exposure concentrations) and each treatment (BPA or E2). In total, 18 samples were randomly analyzed, with QCs and blanks added in the sequence.

Chromatographic separation was carried on a 1290 Infinity II HPLC system (Agilent Technologies, Santa Clara, CA, USA), using a HILIC column (TSK Gel Amide-80 column, 250×, 2.1 mm; 5 µm) from Tosoh Bioscience (Tokyo, Japan) at room temperature. The chromatographic method was adapted from a previous work [27]. Briefly, mobile phases composition were: (A) 5 mM of NH_4Ac adjusted to pH 5.5 with HAc, and (B) AcN. The solvent elution gradient started at 25% of A, increased to 30% of A at 8 min, then to 60% A at 10 min, and held until 12 min. Then, the column was re-equilibrated for 8 min. The flow rate was set at 0.15 mL min^{-1}, the injection volume was 5 µL, and the autosampler temperature was 4 °C.

A 6545XT AdvanceBio LC/Q-TOF (Agilent Technologies, Santa Clara, CA, USA) with a Dual AJS ESI source was employed in negative ionization mode. High-resolution mass spectrometry conditions were set as follows: gas temperature, 250 °C; drying gas, 13 L min^{-1}; nebulizer, 35 psi; sheath gas temperature and flow, 350 °C and 12 L min^{-1}, respectively. Mass range was set from 50 to 1700 Da, with an acquisition frequency of 333.33 ms/spectrum. An auto MS/MS protocol was set for obtaining iterative MS/MS fragmentations of the QCs and collision energy was set to 20 eV.

3.4. Data Analysis

Figure 5 summarizes the main steps of the strategy followed to analyze the LC-MS data sets and is described in detail in the subsections below.

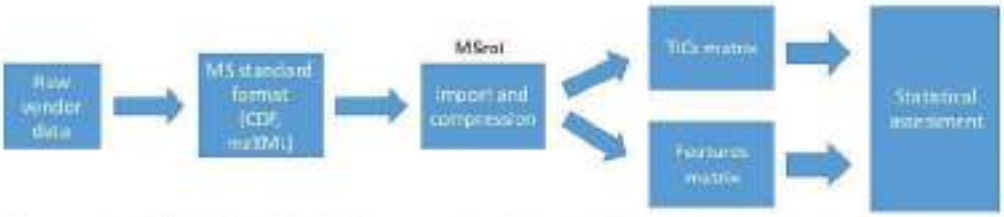


Figure 5. Workflow of the data analysis strategy from the MS raw data acquisition to the statistical assessment.

3.4.1. Data Import and Compression

First, LC-MS raw data acquired using the vendor software was transformed into MS open data formats (the first step in the workflow from Figure 5). Waters LC-MS chromatograms (.raw) from yeast samples were transformed into the CDF format using the Databridge function (MassLynx 4.1 software, Waters, Milford, MA, USA). However, Agilent LC-HRMS chromatograms (.d) from zebrafish embryos exposure study were transformed into the mzXML format in centroid mode using the MSConvert tool from the ProteoWizard suite (64-bit, 3.0.20361 version) [29].

The next step consisted of importing these files into the selected computing platform (MATLAB, Release 2020b, The Mathworks Inc, Natick, MA, USA). Here, total ion current (TIC) chromatograms were directly obtained. In addition, a features matrix containing only those signals with intensity over a pre-defined threshold was also generated. In this work (the second step of Figure 5), the MSROI approach was applied to perform this data importing procedure and, simultaneously, spectral compression [30,31]. Regions of interest (ROI) parameters used in each case are shown in Table S3.

After this procedure, the matrices to be analyzed were built up (third step of Figure 5). On the one hand, the TIC chromatogram for every sample allowed to build up a matrix including all TIC information (size of this matrix was the number of samples by the number of points in the time axis, i.e., retention times). On the other hand, the MSROI procedure generated a features matrix containing the peak areas of the detected features (defined by a *m/z* value) for each sample [32].

Then, these TIC chromatograms and feature matrices were independently normalized to correct the instrumental intensity drifts among injections. This normalization procedure was performed by dividing all the variables' areas (by sample) by the mean area of surrogates and internal standards for each sample (SL and PL lipid standards mixture for the yeast and L-methionine for the zebrafish embryos studies) and the amount of sample considered (A_{600} values for yeast and number of embryos for zebrafish studies).

3.4.2. Statistical Assessment

The statistical evaluation of the TIC and features matrices followed a common workflow (the last step of the workflow depicted in Figure 5).

First, principal component analysis (PCA) was applied to perform a preliminary data exploration. PCA scores enabled a visual comparison of the samples according to the experimental design employed in each study (i.e., the family of lipids considered in the case of yeast samples and exposure level in the case of zebrafish embryo samples) and detection of potential outliers. In addition, the evaluation of PCA loadings can also provide preliminary insights regarding the variables more related to a particular sample type. However, in general, the determination of these variables is somewhat arbitrary and analyst-dependent. PCA was applied to mean-centered (TICs) and autoscaled (features) matrices.

Next, different approaches were tested to identify the most relevant features linked to the experimental design. In the omics field (and, in particular, metabolomics), the determination of these most relevant features (i.e., potential biomarkers) has been widely carried out using univariate techniques based on statistical hypothesis testing. Depending on the experimental design (i.e., number of groups) and the properties of the data, parametric (i.e., Student's *t* and univariate ANOVA) or non-parametric tests (i.e., Wilcoxon test or Kruskal-Wallis) are used. However, when many features are considered, multiple hypotheses testing can lead to an uncontrolled number of false-positives [33]. To overcome this problem, different approaches have been proposed to minimize the number of false-positives in the selection of these potential markers. Here, *t*-tests were performed for binary (two types of samples) comparison whereas ANOVA tests were employed for studies involving ternary comparisons. The list of variables selected by these statistical hypothesis approaches was corrected in this work by the Benjamini-Hochberg procedure [34]. Only these variables with a corrected *p*-value lower than 0.05 were considered statistically significant.

An alternative to multiple hypothesis tests (i.e., one test for each feature) is to adopt a multivariate approach. The standard approach in MS-based metabolomics is the application of partial least squares discriminant analysis (PLS-DA). The most relevant variables were identified from the generated model using approaches such as the selectivity ratio (SR) [8] or VIP scores [9]. SR method is based on calculating the ratio between explained and unexplained variances for each variable in the target projection vector. This approach combines the regression vector and the variance/covariance of the data matrix to identify which variables are more relevant in the classification model. In contrast, VIP scores are calculated as the weighted sum of the squares of the PLS weights relating each latent variable with the amount of explained variance for the correct class classification. Therefore, variables with a large VIP score were associated with a better description of the class belonging. Usually, variables with a VIP score greater than one are selected as relevant, considering that the average of the squared VIP scores equals one. However, in the literature, several papers describe the benefits of this approach, as well as its potential limitations [5,11,35]. In this work, PLS-DA models were built on the mean-centered total ion current chromatograms (TICs) and autoscaled for features (i.e., defined by a particular *m/z* value from the MSROI approach) matrices. The reliability of the obtained features was assessed by means of the calculation of 1000 replicate PLS-DA models, randomly removing between 1% and 10% of the total number of variables as described in Deng et al. [36]. Selected variables after VIP scores or selectivity ratio determination were almost the same for the different considered conditions (see Supplementary Materials for more details).

Since many omic studies are based on statistically designed experiments, several chemometric methods have been proposed in recent years to extract the statistically relevant information related to the factors used in the experimental design. Here, three different approaches were evaluated.

First, ASCA analysis was applied to statistically assess the significance of the design factors used in both studies and to determine the most relevant variables associated with these factors. ASCA combines the variance decomposition power of ANOVA according to the experimental design, with the ability to explore the effects caused for all variables through Simultaneous Component Analysis (SCA) [37]. This analysis strategy enables independent evaluation of the statistical significance of each experimental factor (and possible factor interactions). It is recommended that ASCA is applied to well-balanced sample designs [15,19]. Only in this case, the sum of squares (SSQ) of elements of the ANOVA decomposed matrices represents appropriately the amount of variance of the original matrix explained by each factor and by their interaction. When the experimental design is unbalanced, corrections for the calculation of these sums of squares are required to define the type II SSQ and type III SSQ. Next, the statistical significance of each factor (and of their interaction) is estimated by means of a permutation test, evaluating the null hypothesis H_0 (no experimental effect) against the alternative hypothesis H_1 (experimental effect).

This test is performed by calculating the SSQ of the data in the considered matrix and of the SSQ values obtained when rows of the matrix are permuted [15]. A *p*-value was then calculated by considering the number of permuted SSQ values larger than the original SSQ and the total number of permutations performed. In addition, the evaluation of SCA scores and loadings provide information regarding sample and variable distribution and the importance for each considered factor. The ASCA loadings obtained for each factor show the more relevant variables for its modelling. Here, TICs and features' area matrices were mean-centered before ASCA analysis, and the number of iterations for the permutation test was set to 10,000.

The assumption of non-correlation between variables means that ASCA might not be a reliable option for feature detection in metabolomics studies, since the behavior of some of the studied variables (metabolite concentrations) might be correlated.

Next, rMANOVA was used to evaluate TIC and features data matrices for both studies. This method proposed by Engel in 2015 [21] overcomes the limitations of sample size (MANOVA) and the correlation between variables (ASCA). The critical step of rMANOVA is determining the optimal regularization factor (δ, in a range between 0 and 1) that is calculated according to the Ledoit-Wolf theorem [38]. Depending on the value of this regularization factor, the rMANOVA model will be equal to a MANOVA model (δ = 0) or to an ASCA model (δ = 1). However, the most common situation is that this factor adopts intermediate values in which the advantages of rMANOVA models are more relevant. Finally, the statistical assessment of the experimental factors is performed using a permutation test, as described above for ASCA. However, compared with ASCA, in some circumstances, rMANOVA can allow more straightforward determination of the most relevant features. Engel's implementation of the rMANOVA algorithm has been used in this work. TICs and features matrices were mean-centered before the analysis, and the number of permutations for the permutation test was set to 10,000.

Finally, the last method used in this work is group-wise ANOVA simultaneous component analysis (GASCA) proposed by Saccenti [22]. GASCA attempts to overcome some ASCA limitations by applying the group-wise PCA (GPCA) [39] in the second step after ANOVA decomposition. The GPCA algorithm relies on the sparsity of loadings to increase the simplicity and interpretation of the generated model by considering relationships between variables (metabolites). Due to the impact of the GPCA model on the obtained loadings for each factor, the potential usefulness of this approach for feature detection should be tested. As in the previous cases, balanced experimental designs are preferred to simplify the analysis, and the statistical assessment is performed through a permutation test (10,000 permutations used). Data were mean-centered before the analysis.

3.4.3. Software Used

Univariate statistical tests were performed by using t-test and anova1 functions available at the MATLAB Statistics and Machine Learning Toolbox (MATLAB 2020b, The Mathworks Inc, Natick, MA, USA). Obtained *p*-values were adjusted by the Benjamini-Hochberg algorithm available at the FalseDiscovery library published at the github.com/carbocation/falsediscovery (accessed on 9 May 2022). ANOVA PLS-DA and ASCA were performed using PLS Toolbox 8.9.1 (Eigenvector Research Inc, Wenatchee, WA, USA), working under MATLAB 2020b. The MATLAB source code of the regularized MANOVA is available at the following github repository: github.com/JasperE/regularized-MANOVA (accessed on 9 May 2022). The GASCA algorithm is also freely available in the MATLAB MEDA toolbox and can be downloaded from the address: github.com/josecamachop/MEDA-Toolbox (accessed on 9 May 2022). Venn diagrams were generated using the tool from the Bioinformatics & Evolutionary Genomics group at VIB/UGent (bioinformatics.psb.ugent.be/webtools/Venn/, accessed on 9 May 2022).

3.4.4. Metabolite Identification

Metabolites in zebrafish QC samples were identified based on the MS/MS spectral matches using public metabolite libraries from the MS-DIAL website [40]. The parameters employed for MS-DIAL software are included in Table S4. The identified compounds, their significance, and other relevant information (e.g., HMDB code, chemical formula, retention time) are included in Table S5.

4. Conclusions

In this work, we have evaluated the ability of three multivariate ANOVA-based methods to determine the statistical significance of the experimental design factors (e.g., lipid extraction protocol, pollutants dose of exposure) and their ability to select relevant variables linked to these factors.

On the one hand, the evaluation of the statistical assessment indicated that ASCA determined the statistical significance where it was expected to exist based on the previous biological knowledge of the experiment and its experimental design. In contrast, GASCA provided some inconsistent results as, in some cases, factors were not pointed as statistically significant when they were expected to be. One possibility to improve these statistical significance results and the interpretation of multivariate ANOVA-based methods could be the use of resolved elution profiles (or areas derived from them) of the different sample constituents resolved by chemometric methods, such as MCR-ALS.

On the other hand, GASCA was the ANOVA-based method that provided a list of relevant variables most similar to the variable list provided, considering the VIP scores obtained by the PLS-DA method. In addition, this variable selection step was the major weakness of ASCA since the obtained variables list was the most dissimilar when compared to variables pointed by all the other methods.

In both cases (i.e., considering the statistical significance and variable selection), rMANOVA showed acceptable results. Therefore, rMANOVA could be an option if both statistical assessment and feature detection studies are performed. In contrast, ASCA and GASCA could be employed for only statistical assessment or variable selection, respectively. Table 2 summarizes the main advantages and limitations of each multivariate ANOVA-based method, as well as gives some recommendations regarding the use of each method. In addition, a more comprehensive study to elucidate the impact of the different methods will require an experimental design with a larger number of samples to reinforce the obtained conclusions. Finally, it should be noted that the results obtained for each method are dataset-dependent and, despite that the main trends should be conserved, different results regarding the statistical significance or variable selections could be obtained depending on the data structure.

Table 2. Summary of the main advantages, limitations, and opportunities of the considered ANOVA-based methods.

	ASCA	rMANOVA	GASCA
Advantages	Widespread use in metabolomics (reference multivariate statistical method) Best match between experimental and expected significance	Best of both worlds (model depending on data I MANOVA and ASCA)	A good option for sparse data (i.e., metabolomic datasets) Best match with VIPs from PLS-DA for identifying significant variables
Limitations	Most dissimilar matches identifying significant variables compared to VIPs from PLS-DA	Dissimilar matches with VIPs from PLS-DA in selection of relevant variables	Very strict for determination of significant factors (only factors with very low *p*-values in other methods will appear as significant)

	It assumes metabolites are not correlated and that they all have the same variance.		
Opportunities	Good choice when combined with PLS-DA (VIPs) for the determination of the significant variables	Good choice when aiming one method for statistical analysis and selecting relevant variables (but further validation on the variables is desirable)	Good option for assessing the significance of variables and factors when big effects are encountered (very significant factors in the DOE)

Supplementary Materials: The following supporting information can be downloaded at: www.mdpi.com/article/10.3390/molecules27103304/s1; Table S1. Comparison of profiles obtained for variable selection. Values are the correlation coefficient of the absolute values of the vector profiles. Table S2. Logical relationships between the features detected by PLS-DA and FDR-corrected statistical tests, Selectivity ratio, ASCA, rMANOVA and GASCA. Shadowed columns represent common features between the two compared methods and Bold characters highlights those comparison with a number of coincidences higher than 80%. Table S3. ROI parameters selected for each dataset. Table S4. Parameters employed for MS-DIAL analysis. Table S5. Tentative identification of metabolites of endocrine disruption on zebrafish embryos and their significance with the different statistical methods. Figure S1. Zebrafish embryos exposed to a low-dose of estradiol. PCA analysis: PC1 vs. PC2 scores plot. Figure S2. Venn diagrams summarizing the relationships on the variables detected for each data set. A) TICs matrix for yeast negative; B) Features matrix for yeast negative; C) Zebrafish embryos exposed to low-dose estradiol; and D) Zebrafish embryos exposed to high-dose estradiol.

Author Contributions: Conceptualization, J.J. and M.P.-C.; methodology, M.P.-C. and S.P.; formal analysis, S.P.; data curation, M.P.-C. and S.P.; writing—original draft preparation, M.P.-C. and J.J.; writing—review and editing, M.P.-C., S.P., D.R.S., R.T. and J.J.; supervision, R.T. and JJ.; funding acquisition, D.R.S. and J.J. All authors have read and agreed to the published version of the manuscript.

Funding: The research leading to these results have received funding from grants CTQ2017-82598-P and CEX2018-000794-S funded by MCIN/AEI/10.13039/501100011033. The authors also want to grant support from the Catalan Agency for Management of University and Research Grants (AGAUR, Grant 2017SGR753). MPC acknowledges a predoctoral FPU 16/02640 scholarship from the Spanish Ministry of Education and Vocational Training (MEFP).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in Zenodo (https://zenodo.org/record/6384813, accessed 8 May 2022).

Acknowledgments: The authors would like to thank Francesc Puig-Castellvi, Marta Casado and Laia Navarro-Martin for the help with the yeast and zebrafish experiments. The 1290 LC system and 6545 XT QTOF instrumentation were provided to DS as gifts by Agilent Technologies through their Thought Leader program.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Jaumot, J.; Bedia, C. Introduction to Data Analysis in Omics Sciences. In *Comprehensive Foodomics*, 1st ed.; Cifuentes, A., Ed.; Elsevier: Amsterdam, Netherlands, 2020; pp. 226–240.
2. Chong, J.; Wishart, D.S.; Xia, J. Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis. *Curr. Protoc. Bioinform.* **2019**, *68*, e86. https://doi.org/10.1002/cpbi.86.
3. Oliveri, P. Class-modelling in food analytical chemistry: Development, sampling, optimization and validation issues—A tutorial. *Anal. Chim. Acta* **2017**, *982*, 9–19. https://doi.org/10.1016/j.aca.2017.05.013.
4. Barker, M.; Rayens, W. Partial least squares for discrimination. *J. Chemom.* **2003**, *17*, 166–173. https://doi.org/10.1002/cem.785.

5. Brereton, R.G.; Lloyd, G.R. Partial least squares discriminant analysis: Taking the magic away. *J. Chemom.* **2014**, *28*, 213–225. https://doi.org/10.1002/cem.2609.
6. Andersen, C.M.; Bro, R. Variable selection in regression-a tutorial. *J. Chemom.* **2010**, *24*, 728–737. https://doi.org/10.1002/cem.1360.
7. Mehmood, T.; Liland, K.H.; Snipen, L.; Sæbø, S. A review of variable selection methods in Partial Least Squares Regression. *Chemom. Intell. Lab. Syst.* **2012**, *118*, 62–69. https://doi.org/10.1016/j.chemolab.2012.07.010.
8. Rajalahti, T.; Arneberg, R.; Berven, F.S.; Myhr, K.M.; Ulvik, R.J.; Kvalheim, O.M. Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. *Chemom. Intell. Lab. Syst.* **2009**, *95*, 35–48. https://doi.org/10.1016/j.chemolab.2008.08.004.
9. Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1.
10. Farrés, M.; Platikanov, S.; Tsakovski, S.; Tauler, R. Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation. *J. Chemom.* **2015**, *29*, 528–536. https://doi.org/10.1002/cem.2736.
11. Gromski, P.S.; Muhamadali, H.; Ellis, D.I.; Xu, Y.; Correa, E.; Turner, M.L.; Goodacre, R. A tutorial review: Metabolomics and partial least squares-discriminant analysis—A marriage of convenience or a shotgun wedding. *Anal. Chim. Acta* **2015**, *879*, 10–23. https://doi.org/10.1016/j.aca.2015.02.012.
12. Ferreira, A.J.; Figueiredo, M.A.T. Efficient feature selection filters for high-dimensional data. *Pattern Recognit. Lett.* **2012**, *33*, 1794–1804. https://doi.org/10.1016/j.patrec.2012.05.019.
13. Jansen, J.; Engel, J. ASCA: The Implementation of Design of Experiments Into Multivariate Modelling in Chemometrics. In *Comprehensive Analytical Chemistry*, 2nd ed.; Jaumot, J., Bedia, C., Tauler, R. Eds.; Elsevier: Amsterdam, Netherlands, 2018; Volume 82, pp. 301–335. https://doi.org/10.1016/bs.coac.2018.08.007.
14. Stahle, L.; Wold, S. Multivariate analysis of variance (MANOVA). *Chemom. Intell. Lab. Syst.* **1990**, *9*, 127–141. https://doi.org/10.1016/0169-7439(90)80094-M.
15. Bertinetto, C.; Engel, J.; Jansen, J. ANOVA simultaneous component analysis: A tutorial review. *Anal. Chim. Acta X* **2020**, *6*, 100061. https://doi.org/10.1016/j.acax.2020.100061.
16. Angélina, E.C.; Qannari, E.M.; Moyon, T.; Alexandre-Gouabau, M.C. AoV-PLS: A new method for the analysis of multivariate data depending on several factors. *Electron. J. Appl. Stat. Anal.* **2015**, *8*, 214–235. https://doi.org/10.1285/i20705948v8n2p214.
17. Marini, F.; de Beer, D.; Joubert, E.; Walczak, B. Analysis of variance of designed chromatographic data sets: The analysis of variance-target projection approach. *J. Chromatogr. A* **2015**, *1405*, 94–102. https://doi.org/10.1016/j.chroma.2015.05.060.
18. Harrington, P.D.B.; Vieira, N.E.; Espinoza, J.; Nien, J.K.; Romero, R.; Yergey, A.L. Analysis of variance-principal component analysis: A soft tool for proteomic discovery. *Anal. Chim. Acta* **2005**, *544*, 118–127. https://doi.org/10.1016/j.aca.2005.02.042.
19. Jansen, J.J.; Hoefsloot, H.C.J.; Van Der Greef, J.; Timmerman, M.E.; Westerhuis, J.A.; Smilde, A.K. ASCA: Analysis of multivariate data obtained from an experimental design. *J. Chemom.* **2005**, *19*, 469–481. https://doi.org/10.1002/cem.952.
20. Smilde, A.K.; Jansen, J.J.; Hoefsloot, H.C.J.; Lamers, R.J.A.N.; van der Greef, J.; Timmerman, M.E. ANOVA-simultaneous component analysis (ASCA): A new tool for analyzing designed metabolomics data. *Bioinformatics* **2005**, *21*, 3043–3048. https://doi.org/10.1093/bioinformatics/bti476.
21. Engel, J.; Blanchet, L.; Bloemen, B.; Van den Heuvel, L.P.; Engelke, U.H.F.; Wevers, R.A.; Buydens, L.M.C. Regularized MANOVA (rMANOVA) in untargeted metabolomics. *Anal. Chim. Acta* **2015**, *899*, 1–12. https://doi.org/10.1016/j.aca.2015.06.042.
22. Saccenti, E.; Smilde, A.K.; Camacho, J. Group-wise ANOVA simultaneous component analysis for designed omics experiments. *Metabolomics* **2018**, *14*, 73. https://doi.org/10.1007/s11306-018-1369-1.
23. Tinnevelt, G.H.; Engelke, U.F.H.; Wevers, R.A.; Veenhuis, S.; Willemsen, M.A.; Coene, K.L.M.; Kulkarni, P.; Jansen, J.J. Variable selection in untargeted metabolomics and the danger of sparsity. *Metabolites* **2020**, *10*, 470. https://doi.org/10.3390/metabo10110470.
24. Martínez, R.; Herrero-Nogareda, L.; Van Antro, M.; Campos, M.P.; Casado, M.; Barata, C.; Piña, B.; Navarro-Martín, L. Morphometric signatures of exposure to endocrine disrupting chemicals in zebrafish eleutheroembryos. *Aquat. Toxicol.* **2019**, *214*, 105232. https://doi.org/10.1016/j.aquatox.2019.105232.
25. Puig-Castellví, F.; Bedia, C.; Alfonso, I.; Piña, B.; Tauler, R. Deciphering the Underlying Metabolomic and Lipidomic Patterns Linked to Thermal Acclimation in Saccharomyces cerevisiae. *J. Proteome Res.* **2018**, *17*, 2034–2044. https://doi.org/10.1021/acs.jproteome.7b00921.
26. Dalmau, N.; Jaumot, J.; Tauler, R.; Bedia, C. Epithelial-to-mesenchymal transition involves triacylglycerol accumulation in DU145 prostate cancer cells. *Mol. BioSystems* **2015**, *11*, 3397–3406. https://doi.org/10.1039/c5mb00413f.
27. Ortiz-Villanueva, E.; Jaumot, J.; Martínez, R.; Navarro-Martín, L.; Piña, B.; Tauler, R. Assessment of endocrine disruptors effects on zebrafish (Danio rerio) embryos by untargeted LC-HRMS metabolomic analysis. *Sci. Total Environ.* **2018**, *635*, 156–166. https://doi.org/10.1016/j.scitotenv.2018.03.369.
28. Ortiz-Villanueva, E.; Navarro-Martín, L.; Jaumot, J.; Benavente, F.; Sanz-Nebot, V.; Piña, B.; Tauler, R. Metabolic disruption of zebrafish (Danio rerio) embryos by bisphenol A. An integrated metabolomic and transcriptomic approach. *Environ. Pollut.* **2017**, *231*, 22–36. https://doi.org/10.1016/j.envpol.2017.07.095.
29. Chambers, M.C.; MacLean, B.; Burke, R.; Amodei, D.; Ruderman, D.L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **2012**, *30*, 918–920. https://doi.org/10.1038/nbt.2377.

30. Pérez-Cova, M.; Jaumot, J.; Tauler, R. Untangling comprehensive two-dimensional liquid chromatography data sets using regions of interest and multivariate curve resolution approaches. *TrAC Trends Anal. Chem.* **2021**, *137*, 116207. https://doi.org/10.1016/j.trac.2021.116207.
31. Gorrochategui, E.; Jaumot, J.; Tauler, R. ROIMCR: A powerful analysis strategy for LC-MS metabolomic datasets. *BMC Bioinform.* **2019**, *20*, 256. https://doi.org/10.1186/s12859-019-2848-8.
32. Navarro-Reig, M.; Bedia, C.; Tauler, R.; Jaumot, J. Chemometric Strategies for Peak Detection and Profiling from Multidimensional Chromatography. *Proteomics* **2018**, *18*, 1700327. https://doi.org/10.1002/pmic.201700327.
33. Storey, J.D. A direct approach to false discovery rates. *J. R. Stat. Society. Ser. B Stat. Methodol.* **2002**, *64*, 479–498. https://doi.org/10.1111/1467-9868.00346.
34. Benjamini, Y.; Hochberg, Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Stat.* **2000**, *25*, 60–83. https://doi.org/10.3102/10769986025001060.
35. Cocchi, M.; Biancolillo, A.; Marini, F. Chemometric Methods for Classification and Feature Selection. In *Comprehensive Analytical Chemistry*, 2nd ed.; Jaumot, J., Bedia, C., Tauler, R., Eds.; Elsevier: Amsterdam, Netherlands, 2018; Volume 82, pp. 265–299. https://doi.org/10.1016/bs.coac.2018.08.006.
36. Deng, B.C.; Yun, Y.H.; Liang, Y.Z. Model population analysis in chemometrics *Chemom. Intell. Lab. Syst.* **2015**, *149*, 166–176. https://doi.org/10.1016/j.chemolab.2017.11.016.
37. Zwanenburg, G.; Hoefsloot, H.C.J.; Westerhuis, J.A.; Jansen, J.J.; Smilde, A.K. ANOVA–principal component analysis and ANOVA–simultaneous component analysis: A comparison. *J. Chemom.* **2011**, *25*, 561–567. https://doi.org/10.1002/cem.1400.
38. Ledoit, O.; Wolf, M. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J. Empir. Financ.* **2003**, *10*, 603–621. https://doi.org/10.1016/S0927-5398(03)00007-0.
39. Camacho, J.; Rodriguez-Gómez, R.A.; Saccenti, E. Group-Wise Principal Component Analysis for Exploratory Data Analysis. *J. Comput. Graph. Stat.* **2017**, *26*, 501–512. https://doi.org/10.1080/10618600.2016.1265527.
40. Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; VanderGheynst, J.; Fiehn, O.; Arita, M. MS-DIAL: Data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat. Methods* **2015**, *12*, 523–526. https://doi.org/10.1038/nmeth.3393.

Supplementary Material

Comparison of multivariate ANOVA-based approaches for the determination of relevant variables in experimentally designed metabolomic studies

Miriam Pérez-Cova[1,2], Stefan Platikanov[1], Dwight R. Stoll[3], Romà Tauler[1] and Joaquim Jaumot[1,*]

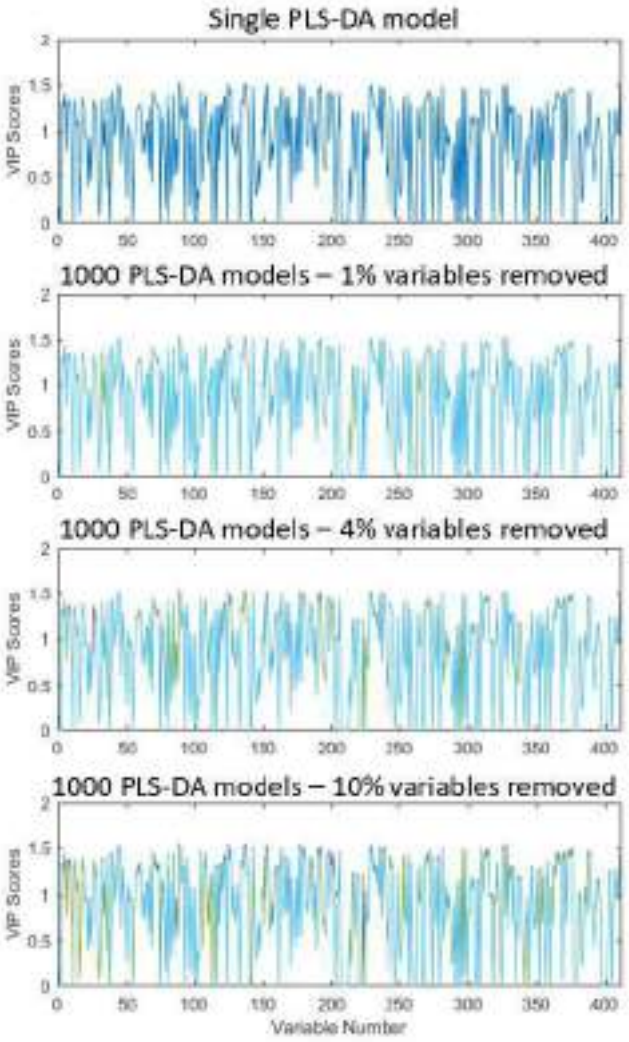
[1] Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, E08034 Barcelona, Spain
[2] Department of Chemical Engineering and Analytical Chemistry, University of Barcelona, Diagonal 647, Barcelona, E08028, Barcelona, Spain
[3] Department of Chemistry, Gustavus Adolphus College, 800 West College Avenue, Saint Peter, MN 56082, United States

Contents

A. Evaluation of the performance of PLS-DA models for feature selection.

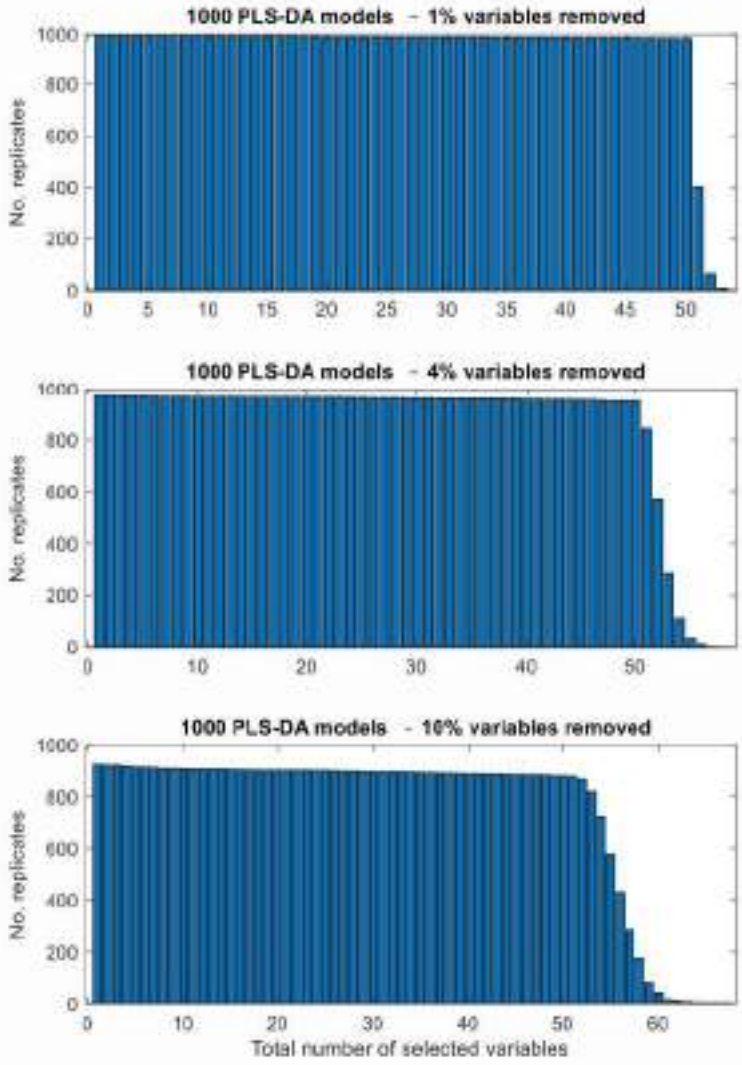
We have calculated a large number of PLS-DA models removing a different number of variables according to these conditions:
- Datasets considered: 1) Yeast Tic Matrix positive ionization mode; and 2) Zebrafish embryos BPA exposure control vs high.
- Number of replications: 1000 (total number of PLS-DA models in each case).
- Number of variables eliminated in each model. We have tested three different levels: 1) 1% of the total variables; 2) 4% of the total variables; 3) 10% of the total variables.

Results show that the used feature selection process is reliable since the selected variables are practically identical. Below, a graphical representation of the obtained VIP scores is shown, focusing on the zebrafish dataset.



It can be seen that the obtained VIP scores profiles are almost identical despite that in the 1000 models of the bottom plot (10% of the samples removed in each permutation), subtle differences can be observed.

Analogous results were obtained when the selectivity ratio selected variables were considered. For this reason, we first evaluated the number of times that each variable is selected as one of the 50 more relevant VIP scores (considering that, especially in the last case, it will be removed several times from the analysis).



Finally, we represented a Venn diagram to evaluate the similarity between the variables determined in each case. It can be seen that 86% of the variables are detected in all cases, and if we consider the models calculated, removing less than 5% of the variables is 100%. Only, in the case of removing 10% of the variables, there are some non-coincident variables with the original model, although most of them are ranked between the 51 and 68 higher VIP scores.

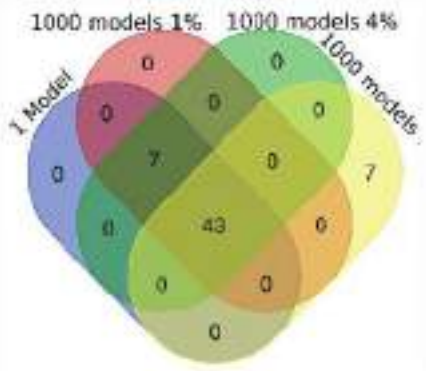


Table S1. Comparison of profiles obtained for variable selection. Values are the correlation coefficient of the absolute values of the vector profiles.

		VIPS	Selrat	ASCA	rMANOVA	GASCA
TIC Yeast negative	VIPS					
	Selrat	0.87				
	ASCA	0.80	0.86			
	rMANOVA	0.06	-0.05	-0.02		
	GASCA	0.89	0.84	0.76	0.04	
TIC yeast positive	VIPS					
	Selrat	0.55				
	ASCA	0.65	0.83			
	rMANOVA	0.32	0.56	0.41		
	GASCA	0.93	0.43	0.53	0.29	
Features yeast negative	VIPS					
	Selrat	0.90				
	ASCA	0.23	0.14			
	rMANOVA	0.54	0.31	0.18		
	GASCA	0.95	0.83	0.24	0.49	
Features yeast positive	VIPS					
	Selrat	0.57				
	ASCA	0.26	0.34			
	rMANOVA	0.30	0.35	0.21		
	GASCA	0.96	0.48	0.25	0.25	
Feature zebrafish BPA Ctrl vs Low	VIPS					
	Selrat	0.42				
	ASCA	0.28	0.26			
	rMANOVA	0.60	0.83	0.22		
	GASCA	0.95	0.33	0.26	0.48	
Feature zebrafish BPA Ctrl vs High	VIPS					
	Selrat	0.36				
	ASCA	0.28	0.08			
	rMANOVA	0.59	0.86	0.15		
	GASCA	0.94	0.29	0.27	0.47	
Feature zebrafish E2 Ctrl vs Low	VIPS					
	Selrat	0.76				
	ASCA	0.29	0.27			
	rMANOVA	0.83	0.61	0.23		
	GASCA	0.94	0.65	0.27	0.72	
Feature zebrafish E2 Ctrl vs High	VIPS					
	Selrat	0.70				
	ASCA	0.27	0.23			
	rMANOVA	0.88	0.82	0.24		
	GASCA	0.95	0.57	0.24	0.81	

Table S2. Logical relationships between the features detected by PLS-DA and FDR-corrected statistical tests, Selectivity ratio, ASCA, rMANOVA and GASCA. Shadowed columns represent common features between the two compared methods and Bold characters highlights those comparison with a number of coincidences higher than 80%.

	FDR vs VIPS			Selectivity Ratio vs VIPs		ASCA vs VIPs		rMANOVA vs VIPS		GASCA vs VIPs	
	FDR	Common	VIPs	SelRat VIPs	Common	ASCA VIPs	Common	rMANOVA VIPs	Common	GASCA VIPs	Common
TICs yeast negative	155	44	6	7	**43**	26	24	45	5	4	**46**
TICs yeast positive	411	50	0	7	**43**	20	24	16	34	3	**47**
Features yeast negative	3	19	11	5	**65**	21	20	22	28	2	**68**
Features yeast positive	342	50	0	9	**41**	32	18	28	22	0	**50**
Zebrafish BPA Ctrl vs Low	0	14	36	10	**40**	37	13	10	**40**	4	**46**
Zebrafish BPA Ctrl vs High	0	22	28	13	37	38	12	12	**48**	12	38
Zebrafish E2 Ctrl vs Low	0	2	48	13	36	38	12	14	36	3	**47**
Zebrafish E2 Ctrl vs High	0	0	50	4	**46**	34	16	8	**42**	2	**48**

Table S3. ROI parameters selected for each dataset.

Dataset	Signal-to-noise ratio threshold (%)	Min-max signal factor	Mass error tolerance	Minimum number of occurences	*m/z* values calculation
Yeast (ESI +)	0.4	4	30	70	Median
Yeast (ESI -)	1.4	2	30	10	Median
Zebrafish embryos (ESI -)	0.3	1	15	50	Median

Figure S1. Zebrafish embryos exposed to a low-dose of estradiol. PCA analysis: PC1 vs PC2 scores plot.

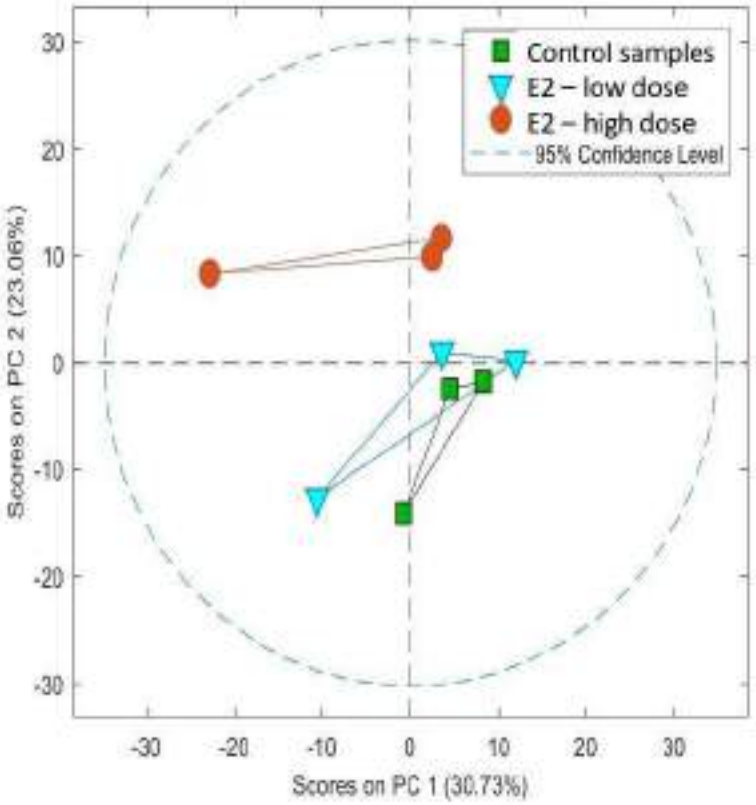
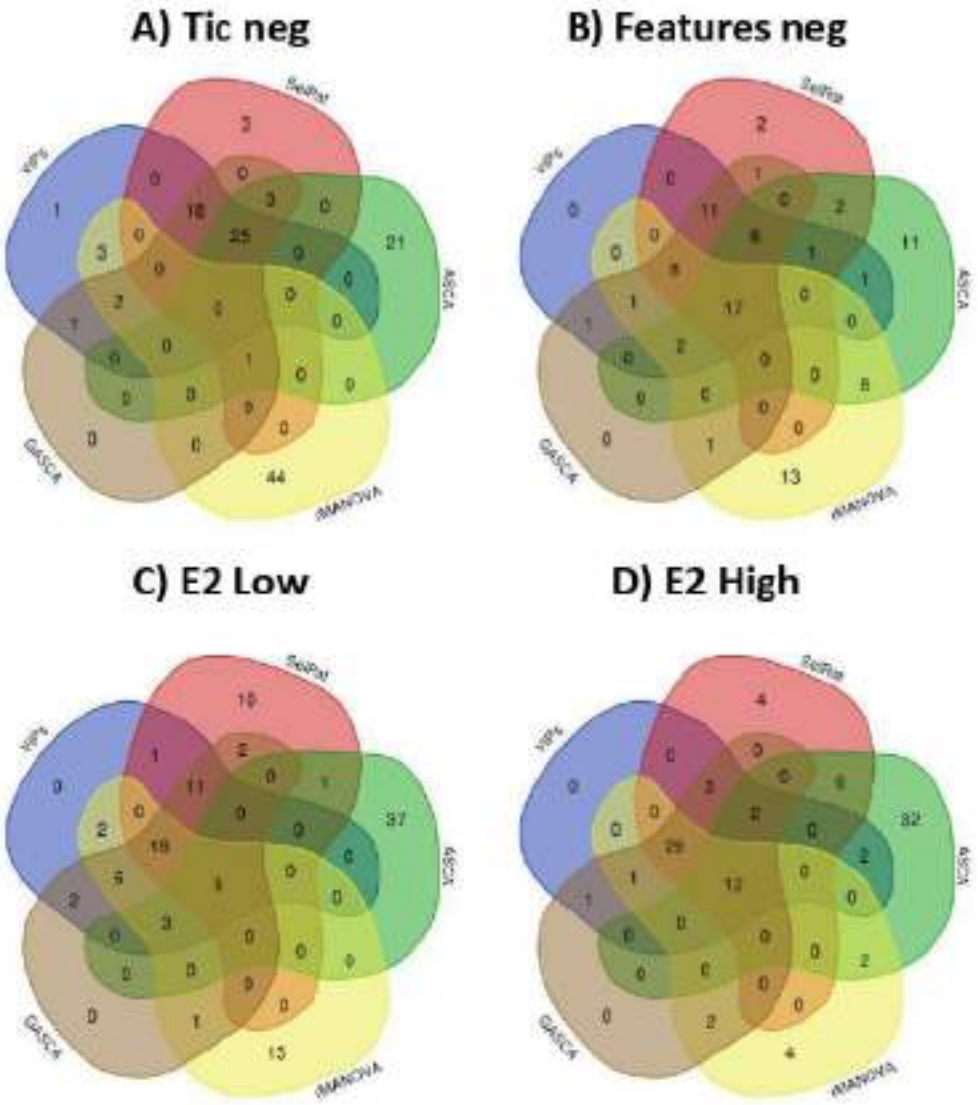


Figure S2. Venn diagrams summarizing the relationships on the variables detected for each data set. A) TICs matrix for yeast negative; B) Features matrix for yeast negative; C) Zebrafish embryos exposed to low-dose estradiol; and D) Zebrafish embryos exposed to high-dose estradiol.



Supplementary Material B

Comparison of multivariate ANOVA-based approaches for the determination of relevant variables in experimentally designed metabolomic studies

Miriam Pérez-Cova[1,2], Stefan Platikanov[1], Dwight R. Stoll[3], Romà Tauler[1] and Joaquim Jaumot[1]*

[1] Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, 08034 Barcelona, Spain

[2] Department of Chemical Engineering and Analytical Chemistry, University of Barcelona, Diagonal 647, Barcelona, E08028, Barcelona, Spain

[3] Department of Chemistry, Gustavus Adolphus College, 800 West College Avenue, Saint Peter, MN 56082, United States

MS-DIAL parameters

Start up a project	**HILIC-HRMS method**
Ionization type	Soft ionization
Separation type	Chromatography (LC)
Method type	Data dependent MS/MS
Data type (MS1)	Profile
Data type (MS/MS)	Profile
Ion mode	Negative ion mode
Target omics	Metabolomics
Data collection	
MS1 tolerance	0.01
MS2 tolerance	0.01
Retention time begin	0
Retention time end	20
Mass range begin	50
Mass range end	1700
Maximum charged number	2
Consider Cl and Br elements	Unchecked
Number of threads	20
Execute retention time corrections	Unchecked
Peak detection	
Minimum peak height	500
Mass slice width	0.1
Smoothing method	Linear weighted moving average
Smoothing level	3
Minimum peak width	5
Exclusion mass list (tolerance: 0.01Da)	922.0098
MS2Dec	
Sigma window value	0.5
MS2Dec amplitude cut off	100
Exclude after precursor	Checked
Keep isotope until	0.5
Keep the isotopic ion w/o MS2Dec	Unchecked
Identification	
Retention time tolerance	0.5
Accurate mass tolerance (MS1)	0.01
Accurate mass tolerance (MS2)	0.015
Identification score cut off	70
Use retention time for scoring	Unchecked
Use retention time for filtering	Unchecked
Postidentification	Not used
Adduct	
Molecular species	[M-H]-, [M+CH3OH-H]-, [M-H-H2O]-
Alignment	
Retention time tolerance	0.05

MS1 tolerance	0.015
Retention time factor	0
MS1 factor	1
Peak count filter	0
N% detected in at least one group	0
Remove feature based on blank information	Unchecked
Sample average / blank average	5
Keep "reference matched" metabolite features	Checked
Keep "suggested (w/o MS2)" metabolite features	Unchecked
Keep removable features and assign the tag	Checked
Gap filling by compulsion	Checked
Isotope tracking	
	Not used

Metabolite Identification and significance

Average Rt(min)	Average Mz	Metabolite name	Adduct	HMDB	KEGG	Chemical formula	delta (ppm)
13.99	104.036	Serine	[M-H]-	HMDB0062263	C00716	C3H7NO3	15.21
13.69	118.052	Threonine	[M-H]-	HMDB0000167	C00188	C4H9NO3	13.57
3.57	121.029	Benzoic acid	[M-H]-	HMDB0001870	C00539	C7H6O2	0.61
10.93	124.009	Taurine	[M-H]-	HMDB0000251	C00245	C2H7NO3S	14.85
11.04	128.036	Pyroglutamic acid	[M-H]-	HMDB0000267	C01879	C5H7NO3	12.12
13.51	130.060	Creatine	[M-H]-	HMDB0000064	C00300	C4H9N3O2	9.08
9.14	130.088	L-Leucine	[M-H]-	HMDB0000687	C00123	C6H13NO2	11.20
6.91	135.031	Hypoxanthine	[M-H]-	HMDB0000157	C00262	C5H4N4O	3.81
7.62	140.010	O-Phosphoethanolamine	[M-H]-	HMDB0000224	C00346	C2H8NO4P	8.35
10.96	146.047	L-Glutamic acid	[M-H]-	HMDB0000148	C00025	C5H9NO4	10.59
10.94	152.083	N-omega-Acetylhistamine	[M-H]-	HMDB0013253	C05135	C7H11N3O	6.00
16.10	154.061	Histidine	[M-H]-	HMDB0000177	C00135	C6H9N3O2	5.79
8.89	166.973	Phospho(enol)pyruvic acid	[M-H]-	HMDB0000263	C00074	C3H5O6P	11.97
3.10	171.138	Decanoic acid	[M-H]-	HMDB0000511	C01571	C10H20O2	2.31
12.56	174.038	N-Acetylaspartic acid	[M-H]-	HMDB0000812	C01042	C6H9NO5	11.13
8.20	178.036	Isoxanthopterin	[M-H]-	HMDB0000704	C03975	C6H5N5O2	2.13
10.39	180.064	Tyrosine	[M-H]-	HMDB0000158	C00082	C9H11NO3	8.82
13.93	191.021	Citric acid	[M-H]-	HMDB0000094	C00158	C6H8O7	9.70
10.96	196.070	N-Acetylhistidine	[M-H]-	HMDB0032055	C02997	C8H11N3O3	9.62
4.88	218.100	Pantothenate (vitamine B5)	[M-H]-	HMDB0000210	C00864	C9H17NO5	13.29
15.25	221.058	Cystathionine	[M-H]-	HMDB0000099	C02291	C7H14N2O4S	5.80

10.60	229.010	D-Arabinose-5-phosphate	[M-H]-	HMDB0011734	C01112	C5H11O8P	7.11
12.77	259.019	D-Fructose-6-phosphate	[M-H]-	HMDB0000124	C00085	C6H13O9P	10.90
3.91	261.134	9-(2,3-dihydroxypropoxy)-9-oxononanoic acid	[M-H]-	NA	NA	C12H22O6	0.59
8.01	267.072	Inosine	[M-H]-	HMDB0000195	C00294	C10H12N4O5	4.81
2.93	269.248	Heptadecanoic acid	[M-H]-	HMDB0002259	NA	C17H34O2	0.95
2.92	281.249	Oleic acid	[M-H]-	HMDB0000207	C00712	C18H34O2	5.07
9.18	282.086	Guanosine	[M-H]-	HMDB0000133	C00387	C10H13N5O5	6.65
8.20	303.080	N-Acetylaspartylglutamic acid	[M-H]-	HMDB0001067	C12270	C11H16N2O8	10.03
2.90	303.231	Arachidonic acid	[M-H]-	HMDB0001043	C00219	C20H32O2	3.68
2.55	311.169	Triptophenolide	[M-H]-	NA	NA	C20H24O3	15.04
9.18	323.029	Uridine-5-monophosphate	[M-H]-	HMDB0000288	C00105	C9H13N2O9P	3.62
2.55	325.185	Hydroquinidine	[M-H]-	NA	C10696	C20H26N2O2	21.45
8.74	346.055	Adenosine Monophosphate	[M-H]-	HMDB0000045	C00020	C10H14N5O7P	0.54
9.63	347.041	Inosine 5'-phosphate	[M-H]-	HMDB0000175	C00130	C10H13N4O8P	4.47
2.62	349.113	Estrone-3-sulfate	[M-H]-	HMDB0001425	C02538	C18H22O5S	5.47
11.30	362.051	Guanosine 5'-monophosphate	[M-H]-	HMDB0001397	C00144	C10H14N5O8P	1.26
13.95	383.113	S-Adenosyl-homocysteine	[M-H]-	HMDB0000939	C00021	C14H20N6O5S	1.58
11.11	426.023	Adenosine 5'-diphosphate	[M-H]-	HMDB0001341	C00008	C10H15N5O10P2	2.32
3.30	445.187	Estrone-3-(beta-D-glucuronide)	[M-H]-	HMDB0004483	C11133	C24H30O8	2.13
3.91	452.275	LPE(16:0)	[M-H]-	HMDB0011473	C05973	C21H44NO7P	5.40
7.90	455.098	Riboflavin-5'-monophosphate	[M-H]-	NA	NA	C17H21N4O9P	1.84
11.63	579.025	UDP-D-Glucuronic acid	[M-H]-	HMDB0000935	C00167	C15H22N2O18P2	3.00
10.04	606.076	Uridine 5'-diphospho-N-acetylgalactosamine	[M-H]-	HMDB0000304	C00203	C17H27N3O17P2	3.90
13.50	611.146	Oxidized glutathione	[M-H]-	HMDB0003337	C00127	C20H32N6O12S2	3.05
12.85	662.100	beta-Nicotinamide adenine dinucleotide	[M-H]-	HMDB0000902	C00003	C21H27N7O14P2	1.90

	Control vs High - BPA exposure				
Metabolite name	**gasca**	**asca**	**rmanova**	**vips**	**selrat**
Serine	0.055	0.014	0.016	1.116	1.115
Threonine	0.042	0.014	0.012	0.782	0.421
Benzoic acid	0.066	0.003	0.093	1.577	2.583
Taurine	0.062	0.004	0.022	1.389	2.079
Pyroglutamic acid	0.015	0.006	0.015	0.308	0.058
Creatine	0.067	**0.050**	0.037	1.609	4.084
L-Leucine	0.040	**0.058**	0.013	0.791	0.384
Hypoxanthine	0.058	**0.077**	0.007	1.255	1.527
O-Phosphoethanolamine	0.069	0.004	0.049	1.714	5.543
L-Glutamic acid	0.059	**0.090**	0.012	1.261	1.574
N-omega-Acetylhistamine	0.062	0.007	0.018	1.398	2.091
Histidine	0.051	**0.214**	0.058	0.930	0.682
Phospho(enol)pyruvic acid	0.047	0.013	0.012	0.921	0.618
Decanoic acid	0.037	0.001	0.034	0.732	0.367
N-Acetylaspartic acid	0.055	**0.191**	0.003	1.130	1.129
Isoxanthopterin	0.065	0.014	0.021	1.498	2.811
Tyrosine	0.000	0.000	0.044	0.482	0.000
Citric acid	0.008	0.003	0.021	0.216	0.007
N-Acetylhistidine	0.000	0.000	0.000	0.000	0.000
Pantothenate (vitamine B5)	0.052	0.007	0.032	0.989	0.786
Cystathionine	0.051	0.004	0.023	1.004	0.916
D-Arabinose-5-phosphate	0.054	0.004	0.042	1.074	1.002
D-Fructose-6-phosphate	0.054	0.004	0.030	1.052	1.123
9-(2,3-dihydroxypropoxy)-9-oxononanoic acid	0.066	0.002	0.020	1.570	3.902
Inosine	0.061	**0.077**	0.067	1.343	2.034
Heptadecanoic acid	0.003	0.000	0.013	0.077	0.000
Oleic acid	0.033	0.023	0.035	0.668	0.259
Guanosine	0.019	0.002	0.039	0.142	0.041
N-Acetylaspartylglutamic acid	0.057	0.002	0.001	1.189	1.371
Arachidonic acid	0.041	0.010	0.028	0.791	0.451
Triptophenolide	**0.072**	0.022	0.082	**1.896**	**11.294**
Uridine-5-monophosphate	0.054	0.010	0.003	1.066	0.986
Hydroquinidine	**0.071**	0.022	0.056	1.807	**7.589**
Adenosine Monophosphate	0.050	0.026	0.012	0.941	0.727
Inosine 5'-phosphate	0.050	0.002	0.035	1.025	0.958
Estrone-3-sulfate	0.000	0.000	0.000	0.000	0.000
Guanosine 5'-monophosphate	0.059	0.012	0.009	1.277	1.751
S-Adenosyl-homocysteine	0.040	0.001	0.042	0.769	0.431
Adenosine 5'-diphosphate	0.061	0.004	0.006	1.311	2.139
Estrone-3-(beta-D-glucuronide)	0.000	0.000	0.000	0.000	0.000
LPE(16:0)	0.062	0.007	0.027	1.380	2.282
Riboflavin-5'-monophosphate	0.046	0.001	0.037	1.164	0.653
UDP-D-Glucuronic acid	0.000	0.000	0.000	0.000	0.000

Uridine 5'-diphospho-N-acetylgalactosamine	0.000	0.000	0.000	0.000	0.000
Oxidized glutathione	0.026	0.001	0.005	0.447	0.111
beta-Nicotinamide adenine dinucleotide	0.061	0.002	0.020	1.321	1.979

	Control vs High - E2 exposure				
Metabolite name	**gasca**	**asca**	**rmanova**	**vips**	**selrat**
Serine	0.016	0.003	0.003	0.192	0.024
Threonine	0.013	0.003	0.021	0.214	0.018
Benzoic acid	0.028	0.001	0.012	0.303	0.058
Taurine	0.001	0.000	0.011	0.124	0.000
Pyroglutamic acid	**0.083**	**0.057**	0.086	**2.714**	**3.048**
Creatine	0.037	**0.015**	0.030	0.596	0.148
L-Leucine	0.046	**0.086**	0.053	1.007	0.260
Hypoxanthine	0.013	**0.012**	0.003	0.172	0.026
O-Phosphoethanolamine	0.035	0.001	0.027	0.537	0.130
L-Glutamic acid	0.014	**0.012**	0.016	0.213	0.027
N-omega-Acetylhistamine	0.046	0.003	0.041	0.843	0.246
Histidine	0.016	**0.067**	0.010	0.212	0.026
Phospho(enol)pyruvic acid	0.059	0.009	0.062	1.409	0.550
Decanoic acid	0.028	0.001	0.077	0.309	0.060
N-Acetylaspartic acid	0.046	**0.085**	0.054	0.863	0.240
Isoxanthopterin	0.031	0.004	0.022	0.449	0.098
Tyrosine	0.028	0.008	0.020	0.457	0.093
Citric acid	0.049	**0.025**	0.030	1.035	0.407
N-Acetylhistidine	0.000	0.000	0.000	0.000	0.000
Pantothenate (vitamine B5)	0.007	0.001	0.014	0.124	0.010
Cystathionine	0.017	0.001	0.000	0.203	0.026
D-Arabinose-5-phosphate	0.017	0.001	0.056	0.290	0.019
D-Fructose-6-phosphate	0.061	0.004	0.059	1.479	0.568
9-(2,3-dihydroxypropoxy)-9-oxononanoic acid	0.032	0.001	0.020	0.454	0.101
Inosine	0.058	**0.043**	0.099	1.503	0.499
Heptadecanoic acid	0.063	0.002	0.059	1.571	0.793
Oleic acid	0.051	**0.032**	0.028	1.031	0.435
Guanosine	**0.075**	0.006	0.094	2.295	1.163
N-Acetylaspartylglutamic acid	0.012	0.001	0.023	0.153	0.021
Arachidonic acid	0.045	**0.017**	0.032	0.839	0.308
Triptophenolide	0.015	0.001	0.007	0.207	0.025
Uridine-5-monophosphate	**0.075**	**0.015**	**0.133**	2.199	1.096
Hydroquinidine	0.015	0.001	0.004	0.204	0.025
Adenosine Monophosphate	**0.072**	**0.041**	**0.136**	2.042	0.921
Inosine 5'-phosphate	0.054	0.003	0.007	1.197	0.490
Estrone-3-sulfate	**0.096**	**0.133**	**0.304**	**3.703**	**12.059**
Guanosine 5'-monophosphate	0.041	0.006	0.044	0.706	0.185
S-Adenosyl-homocysteine	0.001	0.000	0.012	0.141	0.000
Adenosine 5'-diphosphate	0.055	0.002	0.053	1.212	0.418

Estrone-3-(beta-D-glucuronide)	**0.090**	**0.027**	**0.165**	**3.242**	**4.013**
LPE(16:0)	**0.078**	0.004	**0.115**	**2.384**	1.304
Riboflavin-5'-monophosphate	0.000	0.000	0.000	0.000	0.000
UDP-D-Glucuronic acid	0.000	0.000	0.000	0.000	0.000
Uridine 5'-diphospho-N-acetylgalactosamine	0.000	0.000	0.000	0.000	0.000
Oxidized glutathione	0.037	0.002	0.040	0.543	0.166
beta-Nicotinamide adenine dinucleotide	0.062	0.001	0.090	1.511	0.571

Metabolite name	**MS/MS spectrum**
Serine	74.02399:561
Threonine	74.0238:359
Benzoic acid	77.03995:189 91.0181:190 92.02766:442 120.02243:191 121.02903:1640
Taurine	124.00857:389
Pyroglutamic acid	128.03227:187
Creatine	88.03874:2153
L-Leucine	130.08809:210
Hypoxanthine	65.01337:739 66.01033:112 75.00942:237 92.02406:3123 133.01642:233 135.0312:672
O-Phosphoethanolamine	140.00606:104
L-Glutamic acid	102.05522:1244 128.03622:227
N-omega-Acetylhistamine	58.03081:248 80.0386:370 81.04467:348 110.07127:2522
Histidine	72.00764:115 80.03858:416 81.04465:173 93.04604:2083 102.13209:251 108.055:127 110.07124:303 137.03307:116 154.06078:152
Phospho(enol)pyruvic acid	78.95869:5935
Decanoic acid	171.13811:263
N-Acetylaspartic acid	58.03087:857 70.03085:105 71.01293:147 78.95894:265 88.03902:903 115.00421:117 130.04869:109
Isoxanthopterin	65.0136:959 65.99932:158 90.01084:147 92.02438:112 108.0192:172 135.03169:227 136.0145:1489 161.00896:385 178.03676:460
Tyrosine	72.00779:120 106.0398:116 119.05033:1010 163.03937:358
Citric acid	57.0336:493 67.01793:507 85.02973:821 87.00967:2033 111.00991:1083
N-Acetylhistidine	59.01285:146 80.03868:257 81.04475:614 93.04615:919 108.05514:144 110.07138:10294 134.07243:240 137.03728:305 154.06096:1416
Pantothenate (vitamine B5)	71.05093:312 80.96388:215 88.03913:1080 99.044:116 146.08057:568
Cystathionine	120.01102:234 134.02832:410
D-Arabinose-5-phosphate	78.95895:1865 96.96817:856 138.98021:296 210.99992:143
D-Fructose-6-phosphate	78.95869:992 96.96785:2647
9-(2,3-dihydroxypropoxy)-9-oxononanoic acid	61.01776:109 125.09753:162 187.09627:282
Inosine	92.02406:205 135.0312:6010 149.043:134

Heptadecanoic acid	57.03363:131 62.96278:130 102.9451:105 225.21893:123 269.24753:1423
Oleic acid	136.09113:104 281.1395:112 281.24973:3310
Guanosine	108.01913:275 133.0168:520 150.04198:5342 282.08002:258
N-Acetylaspartylglutamic acid	58.02821:273 96.0097:860 102.05527:197 111.02082:185 128.03629:1366 135.0316:265 146.04688:278 267.07159:332
Arachidonic acid	59.01283:117 80.96375:122 259.24191:202 303.23126:561
Triptophenolide	80.96382:104 130.76787:102 183.01089:337 274.88031:402 292.89182:100 311.16962:9053
Uridine-5-monophosphate	78.95889:2732 96.96809:2364 111.02078:456 138.9801:172 150.98003:271 192.98659:159 210.99976:413 280.02075:156 323.02924:394
Hydroquinidine	183.01089:245 243.10075:102 325.18484:8669
Adenosine Monophosphate	78.95895:4548 96.96818:1732 134.04842:605 150.98015:276 210.99994:308
Inosine 5'-phosphate	78.95889:4734 96.96809:2148 135.03156:752 150.98003:358 210.99976:306 347.04065:632
Estrone-3-sulfate	80.96374:267 269.15643:866 349.11282:7065
Guanosine 5'-monophosphate	78.95894:4997 96.96816:278 150.04202:260 319.04324:111 362.05048:840
S-Adenosyl-homocysteine	74.99174:231 86.99358:135 105.00195:121 134.04446:3510 188.03932:255 248.05653:216 383.11285:576
Adenosine 5'-diphosphate	78.95888:1511 134.04428:1015 158.92577:2074 272.95587:357 328.04205:1452 408.00635:796
Estrone-3-(beta-D-glucuronide)	59.01266:366 75.00942:180 85.02946:457 89.0219:155 99.00912:234 103.00441:188 113.02363:2662 175.02048:729 269.15002:603 445.18591:1197
LPE(16:0)	140.00998:196 192.03139:124 196.03644:571 214.04646:124 255.22987:5888 452.27475:1590
Riboflavin-5'-monophosphate	78.95898:183 96.96822:1067 135.03171:125 198.99791:350 213.01517:185 241.0712:107 255.08632:181 455.09024:245
UDP-D-Glucuronic acid	78.95869:212 158.92537:193 254.98592:499 305.01041:117 323.02841:1186 384.98285:292 402.98886:3172 579.02325:2606
Uridine 5'-diphospho-N-acetylgalactosamine	158.92154:174 272.95612:108 282.03366:318 300.04654:228 362.00443:168 384.98413:732 402.99023:295 606.07623:5290
Oxidized glutathione	82.02901:103 128.03258:324 143.04376:127 177.03192:174 197.85272:119 203.05011:175 254.07626:402 272.08826:1255 304.05777:672 306.07623:6179 338.04431:682 482.10379:149 593.13129:134 611.14722:4465
beta-Nicotinamide adenine dinucleotide	272.9559:191 328.04208:182 408.01337:243 426.02264:314 540.05621:6973

Range of signficance for each method and study:		
BPA C vs H analysis	Max	Min
Vips	2.290	1.888
Selectivity ratio (selrat)	224.512	6.061
asca	0.589	0.026
rmanova	0.908	0.100
gasca	0.075	0.070
E2 C vs H analysis	Max	Min
Vips	3.7563	2.3191
Selectivity ratio (selrat)	23.1375	1.4347
asca	0.7836	0.0174
rmanova	0.3672	0.1182
gasca	0.0974	0.0759

3.3 Discussion

This section discusses the results obtained in **scientific publications III and IV**. The different chemometric approaches for each of the steps of the data analysis workflow are assessed, and the most suitable ones are proposed for further use in future studies. At the end of this section, a brief discussion and prospects of the proposed workflow are included.

3.3.1 Regions of interest for spectral compression

Untargeted metabolomic datasets usually require a preliminary step of data pre-processing, including filtering and compression prior to the analysis, due to their big size. This is especially crucial when high-resolution mass spectrometry (HRMS) is employed, as data reduction is mandatory before their analysis with desktop computers.

Binning was the traditional data compression approach [21]. In this strategy, the *m/z* axis from the raw mass spectra is split according to specific bin sizes, i.e., a multiple of the mass accuracy of the mass spectrometer used is selected as bin size. The data is compressed and transformed into a data matrix. However, this binning approach has clear disadvantages. Mass spectral accuracy is usually significantly reduced when establishing the bins. If bin sizes are too small, i.e., 0.001 *m/z* units, computation storage needs will be too big, and processing time will be very slow. On the other side if bin sizes are bigger, like 0.1 *m/z* units, computation times will be faster, but *m/z* resolution is lost. In addition, if the bin size is not selected properly, then several peaks can be merged into the same bin, increasing the noise level as well as the chances of neglecting small peaks. The opposite scenario should also be avoided. If the same peak is divided into different bins, the usual peak shape is lost, and the peak cannot be correctly determined. Thus, if data were acquired in HRMS mode, but bins are too large, the spectral accuracy is lost during the bin compression, and then the pre-processing step is counterproductive. Besides, it is often encountered that only binning does not allow a sufficient reduction in data dimensionality (especially if small bin sizes are used), and compression in the chromatographic time mode is also required (e.g., time-windowing [22] or wavelets [23]). On the one hand, time-windowing splits the chromatogram according to the retention time direction and the different fractions are analyzed separately, which increases the total analysis time. On the other hand, wavelets are another possibility

[23], which reduce the chromatographic mode into a new time scale without any loss of significant information, such as the intensities and shapes of the chromatographic peaks and their temporal location. Although useful for compressing and denoising, the wavelets approach adds an extra step to the binning process, which, again lengthens the whole data analysis workflow.

As a general conclusion, binning is not recommended for the analysis of mass spectrometry data. In contrast, in the case of UV data, as the scan speed is high, the application of binning can be easily performed without any dramatic decreases in spectral resolution. As proved in **scientific publication III**, this can be an effective way to synchronize MS and UV datasets when performing data fusion.

Other more suitable compression strategies have been considered for MS data, such as the regions of interest or ROI strategy. Compared to binning, the main benefit of the ROI strategy is that there is no loss of mass spectral accuracy [3]. In addition, no longer compression in the chromatographic mode is usually required afterwards, as the resultant ROI data matrices are enough compressed even in LC × LC-HRMS, where very large datasets are produced. A single LC × LC-HRMS file can be reduced from GB to MB with the ROI procedure. Besides. the ROI approach is also faster than binning if the intensity threshold is placed above the noise level.

In recent years, the regions of interest approach has replaced the binning strategy for spectral compression in mass spectrometry-based metabolomics. It was initially proposed and included in the *centWave* algorithm of the XCMS software, probably the most popular algorithm nowadays for metabolomic pre-processing [24]. Thus, there is a need for software development to perform spectral compression for any MS data, including LC × LC-MS data. For the analysis of metabolomic data, pre-processing can be performed directly on an LC × LC generic vendor software such as LC × LC Edition Software from GC Image™, or AnalyzerPro® XD from SpectralWorks. However, specific metabolomics software able to manage LC × LC data is, however, not very common. An exception would be MS-DIAL [25], although peak integration is performed only in a one-dimensional chromatographic basis, not recognizing multiple peaks (from the subsequent modulations) as the same compound. Moreover, an additional step of peak alignment is required in all this type of software. As stated in **scientific publication III**, LC × LC-MS does not usually fulfill trilinear model requirements. The practical implications of this model deviation are that peak shifts in both retention dimensions are commonly encountered. In the case of targeted

analysis, where all analytes are known *a priori*, it is relatively easy to check if the alignment has worked properly for all the compounds of interest. On the contrary, in the case of untargeted analysis, this checking is not straightforward. The advantage of the ROI strategy is that it does not need this previous chromatographic peak alignment of the data, and, therefore, this step can be avoided. Hence, the MSroi GUI proposes an alternative way for the analysis of MS data, including LC × LC datasets [5]. This application of the ROI procedure to untargeted LC × LC-MS metabolomic data has been successfully applied in previous studies from the research group where this PhD Thesis has been carried out [22,26]. In **scientific publication III**, the MSroi GUI is applied to LC × LC-MS, LC × LC-UV (only for importing into MATLAB purposes) and LC × LC-UV-MS (fused data). Analogously to binning, ROI also transforms the dataset (a data cube in the case of LC × LC-MS or LC × LC-UV data) into a column-wise augmented data matrix. The ROI strategy is adequate also for the data unfolding steps employed in the trilinearity assessment, which will be further described in the following section of this Discussion.

3.3.2 Multivariate curve resolution alternating least squares as a resolution method

The application of resolution methods required a preliminary investigation of whether the data follows an ideal trilinear behavior or not. Rutan and coworkers raised this question in their work on LC × LC-UV datasets, and deviations from trilinearity behavior were encountered [27–29]. In these cases, retention time shifts between the different modulations (i.e., ^{2}D chromatograms) within each sample and between samples were observed. The latter were observed when analyzing multiple chromatograms simultaneously. Hence, the use of PARAFAC is discouraged unless a peak alignment is performed before the analysis and trilinearity is tested beforehand. A semi-automated method was proposed by Allen et al. for peak alignment without a reference injection [27]. However, in the presence of strong coelutions, changes in the peak shape are commonly found, which are not solved with a mere alignment step. Consequently, approaches such as PARAFAC2 could also fail due to these changes in peak shape [30]. MCR-ALS was then proposed as a bilinear approach. The main advantage of this method compared to PARAFAC and PARAFAC2 is that it can deal with both intra-run and inter-run retention time shifts and peak shape distortions without any prior alignment or peak modelling, and only

spectra alignment and reproducibility are required (which is the general situation). Although MCR-ALS allows the application of a trilinearity constraint (trilinear model approach), a shift correction and modelling [12] could be implemented in the trilinear model approach. However, in the case of LC × LC, all the modulations corresponding to the same peak should be modelled, and this correction may fail.

Despite the fact that the MCR-ALS bilinear model was preferred for LC × LC-UV datasets, the question of whether LC × LC-MS presented deviations from trilinearity was still an open question in the literature. A previous study evaluated the trilinearity behavior in the case of a specific region of a LC × LC-MS chromatogram. Results obtained with MCR (bilinear model), MCR (trilinear model), PARAFAC and PARAFAC 2 were compared [9]. The quality of the model was evaluated, among other parameters, using the lack of fit (LOF) parameter, which asseses the fitting of the model, and the core consistency of the derived trilinear models (i.e., PARAFAC). Regarding the LOF, the lowest LOF values, and consequently the best models, were obtained for MCR-ALS as a bilinear model and PARAFAC2 (3% for both). The other models, i.e., MCR-ALS trilinear and PARAFAC, presented worse LOF values from 14% up to 21%. PARAFAC2 assumed a trilinear behavior allowing for peak shifting, but even using this approach the method did not properly model the LC × LC-MS chromatographic regions data. Significant deviations from trilinearity were found, as the core consistency diagnostic of PARAFAC was not 100% as expected for trilinear data, but significantly lower below (80%). These deviations were caused not only by retention time shifts but also by changes in the peak shapes. Although the core consistency of PARAFAC2 was 99%, this method presents serious disadvantages. One important difference between bilinear MCR-ALS approach and PARAFAC2 is that, until very recently, the second did not allow the applications of non-negativity constraints in the elution profile of the ^{2}D separation, which led to profiles that were not easy to interpret compared to those resolved by the other methods. This fact had already been pointed out by different authors [31] in the analysis of liquid chromatographic data when coupled to UV and fluorescence detectors. In their work, Bortolato *et al.* describe PARAFAC2 limitations, such as that it cannot deal with peak distortions nor peak time shifts for coeluting compounds. Also, in the work of Navarro-Reig et al. [9], PARAFAC2 was not recommended for LC × LC, especially in the presence of interferences and coelutions. However, a recently published study showed a new implementation of PARAFAC2 in which non-negativity constraint can be imposed in the three modes [32], which was tested for fluorescence spectroscopic

data. The potential applications of this new version of PARAFAC2 for LC × LC data as well as of MCR-ALS trilinear allowing peak shifting are being explored at present.

To confirm these results [9,27–29], the fulfillment of the trilinear model was assessed for LC × LC-MS and LC × LC-UV datasets (both detectors individually and fused) in **scientific publication III**. **Figure 3.4** summarizes the three strategies employed:

A) The comparison of the singular value decomposition, SVD, to the unfolded three-way data in the three modes [33].

B) The evaluation of PARAFAC core consistency [34].

C) The assessment of data fitting and explained variance of bilinear and trilinear MCR-ALS approaches [35].

Firstly, the results of the singular value decomposition (SVD) on the unfolded three-way data were evaluated. An LC × LC sample, regardless of which of the detectors is employed (MS or UV), can be unfolded in three different manners, according to the three modes of the three-way data cube, i.e., along with the ^{1}D and ^{2}D chromatographic dimensions and the spectral mode (composed by *m/z* values or wavelengths). For instance, in the ROI strategy, the data is first arranged in a column-wise data matrix with the common *m/z* values in the columns and the retention times at which MS spectra are acquired in the rows. All subsequent data modulations are concatenated vertically. In the ideal case that trilinearity was achieved, the SVD analysis of the three augmented data matrices would give the same number of significant singular values (chemical rank or mathematical rank in the absence of noise) [36]. The analysis of the SVD decomposition showed that in the case of LC × LC-UV data, strong deviations from trilinear behavior were encountered because large differences were found for the three unfolding strategies tested. In contrast, only small deviations were appreciated in the case of LC × LC-MS. From the three unfolding strategies, the column-wise data matrix (i.e., the unfolding strategy employed in the ROI evaluation) was the option that led to a lower number of significant compounds for explaining the same percentage of variance. This behavior was found for both UV and MS datasets, although it was more accentuated for the first case (see blue graphs in **Figure 3.4.A**). These results agree well with previous results by Navarro-Reig, using this SVD approach to evaluate this type of data [9] and justify the unfolding strategy employed in the ROI procedure.

Secondly, the PARAFAC core consistency was evaluated. The trilinear behavior was only determined for small models (<3 components, see **Figure 3.4.B**), but the analyzed mixture was composed of 31 compounds. Therefore, a larger number of components was expected. A remarkable decrease in the core consistency when increasing the number of components was also reported in the previous work by Navarro-Reig et al. [9], which is explained by the fact that the trilinear model cannot describe well the evaluated data [34].

Thirdly, bilinear and trilinear MCR-ALS approaches were compared. When the trilinearity constraint was applied, the explained variances decreased from 100% in a pure bilinear model to 60% and 40% for MS and UV, respectively (see **Figure 3.4.C**). This is a very large difference which cannot be only justified by overfitting in the case of using the bilinear model. In the case that a trilinear constraint is relaxed, allowing elution profiles to be shifted among modulations and samples (runs), the explained variance still diminished from 100% (bilinear) to 80% (trilinear), in agreement with the results reported by Navarro-Reig [9]. Since, in all cases, data fitting was worsening significantly from bilinear to trilinear modelling approaches for both datasets (i.e., LC×LC-MS and LC×LC-UV), it was finally decided that bilinear modelling performed better due to the lack of fulfillment of the trilinear model by the LC×LC data.

Hence, the results agreed with those from Rutan and coworkers and our research group previous work. The lack of trilinearity is more tangible in the case of UV detection, but it is also present in LC×LC-MS data. Thus, bilinear MCR-ALS is recommended as a resolution method for LC×LC data.

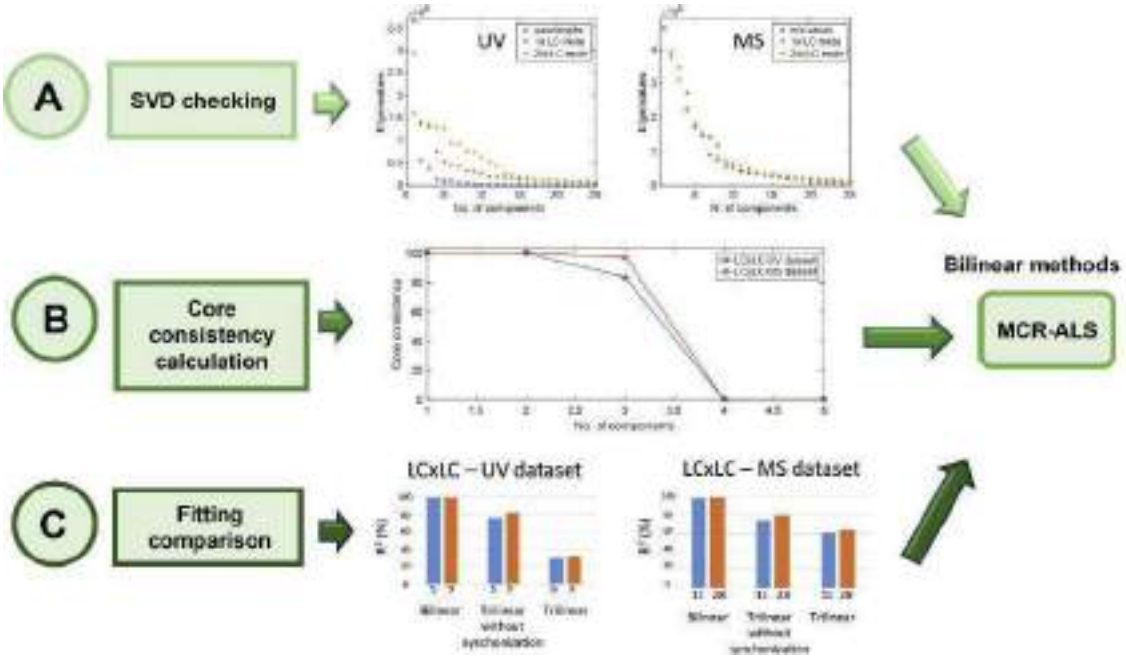


Figure 3.4. Strategies for the evaluation of trilinearity of LC × LC-MS and LC × LC-UV and the outputs provided: A) comparison of the single value decomposition when different unfolding strategies are used; B) evaluation of PARAFAC core consistency; C) assessment of the data fitting and explained variances of bilinear or trilinear MCR-ALS. *Figure extracted from* ***scientific publication III***.

MCR-ALS was therefore successfully applied to the resolution of LC × LC-MS, LC × LC-UV and fused LC × LC-UV-MS data, using a bilinear model for factor decomposition with a sufficient number of components to explain a significative amount of data variance. In all cases, non-negativity constraints were applied in both chromatographic and spectral modes, as well as spectral normalization. Unimodality constraint has also been used sometimes in the analysis of LC-UV data [28,37–39], but not needed in the case of LC-MS as unimodal elution profiles were usually already obtained directly, with little rotation ambiguity associated [6,40].

LC × LC-UV results resulted in fewer components due to the lack of selectivity compared to MS. In contrast, the fused UV and MS information required an increase in the number of components and, consequently, of identified compounds. The reason was that UV spectra of additional components could be finally resolved using the data fusion approach, thanks to the additional information provided by MS. The UV and MS fused data analysis also improved the quality of the MS profiles which were poorly resolved in the individual analysis of MS data. Hence, UV and MS data fusion provided more reliable qualitative and quantitative results due to improved

resolution of the component spectra. More information for the annotation of the resolved compounds was also available with this approach. Although this is a simple case example with a reduced number of known *a priori* compounds, the MCR-ALS as bilinear approach (i.e., no multilinearity constraint applied) was validated for the analysis of LC×LC-MS, LC×LC-UV and fused of LC×LC-UV-MS datasets in the study of more complex cases in untargeted metabolomic analysis of biological samples. Examples of applications of this bilinear method to LC×LC-MS data will be further described in **the following Chapter**, in **scientific publications V and VI**.

3.3.3 Statistical assessment and variables selection for metabolomic datasets

PLS-DA is the most employed method in metabolomic studies for discriminant analysis of different classes of samples (e.g., control *versus* treated) and to identify the variables that present significant changes between these classes (i.e., potential markers of the treatments). Other alternative methods include principal component-discriminant function analysis (PC-DFA), support vector machines (SVM) and random forests (RF). A comparison among the four (including PLS-DA) has already been performed by Gromski et al. [41]. In **scientific publication IV,** the ability of three multivariate ANOVA-based methods to select these potential markers is evaluated. The goal of this study was to look for a statistical method that provided the significance of experimental factors and what were the significant variables associated with these factors, simultaneously. **Figure 3.6** summarizes the main findings in **scientific publication IV** with an example of one of the evaluated datasets (metabolomic analysis of zebrafish embryos exposed to estradiol (E2)). On the one hand, the significance of "dose" factor is assessed (**Figure 3.6.A**). The PCA scores plot of these three sample classes (Control, High and Low) shows a clear differentiation between controls and the high dose, although controls and the lower dose cluster are very close. Consequently, Control vs High comparison was expected to be significant regarding "dose" factor (p-value lower than 0.05), which agreed with ASCA and rMANOVA but not with GASCA results. Differentiation between Control and Low samples is less evident, but as both sample classes cluster together, "dose factor" is less likely to be significant at this low level. ASCA results were the most reliable according to the observed PCA samples separation. On the other hand, the list of significant features that can be obtained with the different methods was also

evaluated (**Figure 3.6.B**). In this case, GASCA furnished the highest number of significant variables, in agreement with the first 50 VIP values from PLS-DA, with only two of these variables being not coincident. Eight and thirty-four variables were not matching for rMANOVA and ASCA, respectively. Thus, GASCA identified practically the same features as the PLS-DA method. These coincident results between GASCA and PLS-DA could be explained due to the ability of GASCA models to cope with the sparsity of the datasets.

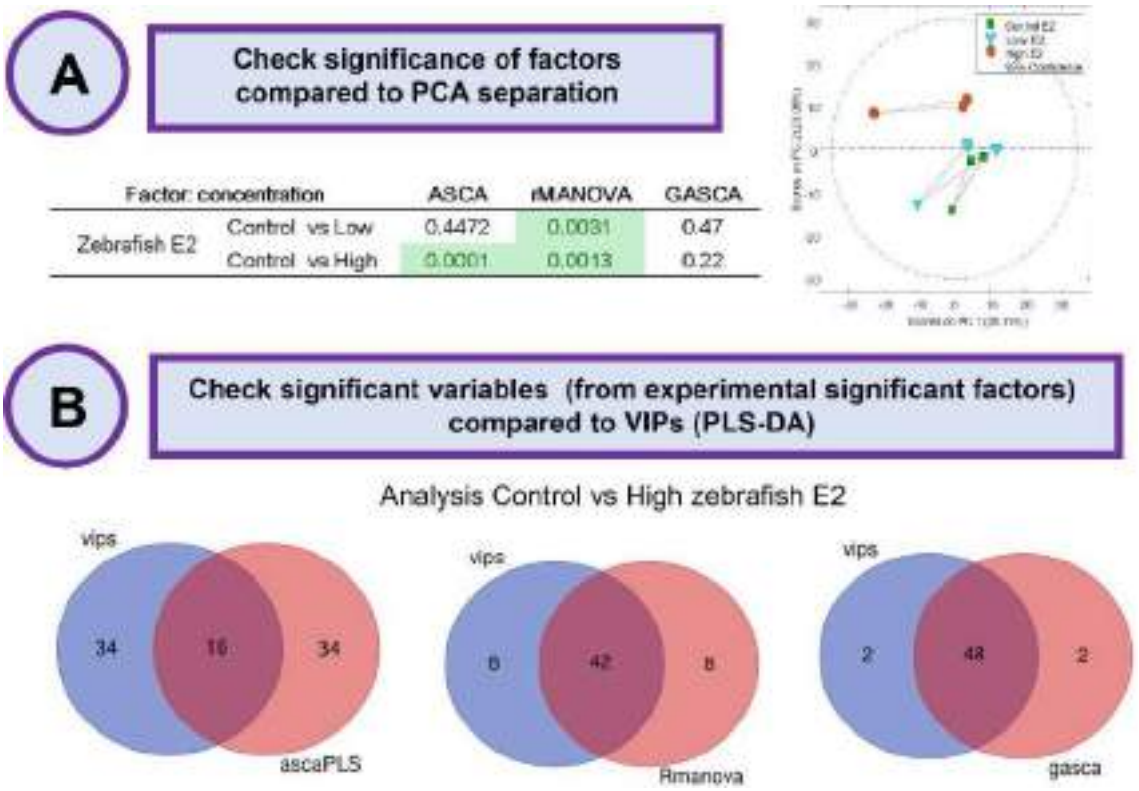


Figure 3.5. Comparison of the three multivariate ANOVA-based methods from two perspectives: A) the significance of the experimental factors and, B) the selection of the most significant variables associated with these factors.

Figure 3.6 shows the main pros, caveats, and opportunities of these three multivariate ANOVA methods, according to the results obtained in **scientific publication IV**. Overall, ASCA is the most frequently used method in metabolomic studies, and it seems to be the most suitable for statistical assessment, but the less reliable for variable selection. GASCA is too strict in the evaluation of the statistical significance of the experimental factors, but it gives more reliable results in the

selection of the more important variables in agreement with the VIPs from PLS-DA model. Lastly, rMANOVA is an intermediate option that could profit from a compromise between the significance of the factors and the choice of the relevant features. A deeper study of these initial results with larger and more complex datasets should be performed to confirm them.

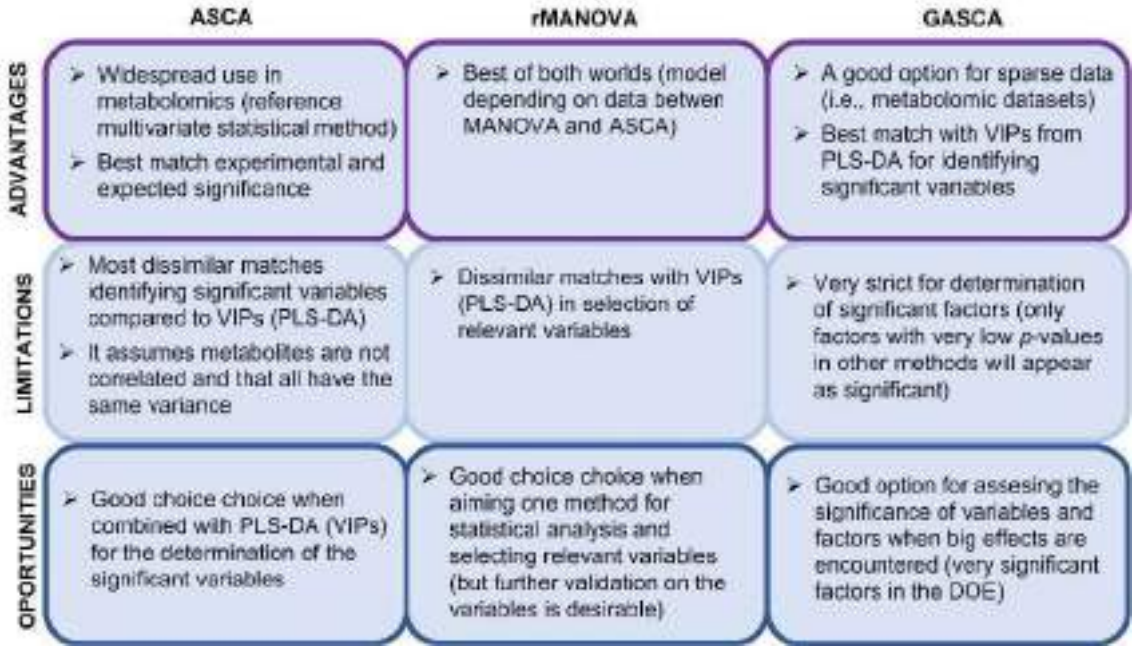


Figure 3.6. Summary of the strengths, weaknesses, and opportunities of the three multivariate statistical methods tested on **scientific publication IV:** ASCA, rMANOVA and GASCA. *Figure extracted from* ***scientific publication IV****.*

The final workflow used in this PhD Thesis consisted of the combination of ASCA and the VIPs from PLS-DA for statistical assessment and variable selection, respectively. Both are reference methods for the mentioned purposes in metabolomic studies and due to the advantage of their ease of use, especially for instance when they are used in the PLS Toolbox (MATLAB) or online platforms specific for metabolomic studies such as Metaboanalyst [42,43], where these two methods are currently integrated. However, further work is still needed to implement rMANOVA or GASCA for routine metabolomic studies.

3.3.4 Future prospects in the data analysis workflow proposed for metabolomic

The chemometric workflow described in this Chapter for metabolomic studies includes the following methods:

1) The application of the ROI strategy for pre-processing mass spectrometry data.

2) The application of MCR-ALS for the resolution of one and two-dimensional chromatographic datasets.

3) The combination of ASCA and PLS-DA for statistical assessment and selection of potential markers.

Future improvements in the ROI step through the MSroi GUI would be related to the integration strategy for LC×LC datasets. Until now, peak areas have been calculated with MSroi GUI by summing the individual peaks from subsequent modulations corresponding to the same compound (i.e., associated with the same ROI). However, an integration of the pixels from the 2D plot associated with each ROI could be implemented, analogously to the integration performed by similar software for mass spectrometry imaging datasets [5]. Besides, until now, the targeted search with the MSroi GUI has been based only on looking for a list of the *m/z* values of interest (already created by the user before importing it to the MSroi GUI). Another useful functionality to incorporate into this GUI (for targeted analysis) would be the ability to search the most common adduct forms. The MSroi GUI could, for instance, calculate the *m/z* values of the adducts of interest (which could be selected by clicking in the MSroi main menu for each analysis, according to the mobile phase composition and ionization modes) by introducing the list of exact mass values of the compounds of interest. Therefore, a more automatic and straightforward workflow for targeted analysis could be implemented.

Regarding the trilinearity of LC×LC datasets, deviations from an ideal trilinear behavior have been reported with the most common detectors, MS and UV. Consequently, the use of PARAFAC seems not recommended. However, the new possibilities of PARAFAC2 imposing a non-negativity constraint in the three modes proposed by Van Benthem et al. [32] are especially interesting for LC×LC. Current work is being pursued to evaluate the potential applications of this new version of

PARAFAC2 as well as of trilinear MCR-ALS (with a shift correction constraint) for the analysis of chromatographic data.

In addition, data fusion of LC×LC-UV-MS datasets was validated for a simple case of one sample with a mixture of already known compounds. This fusion strategy could also be useful for more complex studies involving unknown compounds (i.e., untargeted analysis), as the complementary information from both spectral modes would significantly solve potential ambiguities in the compound annotation.

Lastly, concerning the statistical evaluation of metabolomic datasets, further work is required to establish only one method able to provide significance of the experimental factors and associated the variables. On one side, it would be very interesting to look into the conditions of GASCA when determining whether a certain experimental factor is significant or not. On the other side, the reliability of rMANOVA for selecting potential markers needs to be further validated.

3.4 Conclusions

This section includes only the specific conclusions that are drawn throughout this Chapter about the chemometric strategies evaluated:

Concerning pre-processing methods for filtering and compression:

- The ROI approach allows the compression and filtering of any MS data, without any loss of spectral accuracy and without the need for any prior profile alignment.
- ROI unfolds the LC×LC-MS three-way data cube in a column-wise augmented data matrix, keeping the *m/z* axis in the columns shared among all simultaneously analyzed samples (one data matrix for each of them). In contrast, retention times are set in the rows with the subsequent modulations concatenated vertically.
- Binning can be only recommended when a UV detector is employed (e.g., LC×LC-UV datasets). However, this binning step is usually not needed (UV data are not as big as, for instance, high-resolution mass spectrometry data).

Concerning the resolution methods for LC × LC data:

- LC × LC-MS and LC × LC-UV data showed deviations from the trilinear model.
- This unfolding strategy using the ROIMCR approach (see above) leads to a better estimation of the chemical rank (mathematical rank in the absence of noise) of the analyzed system and, therefore, to the correct number of components related to different chemical constituents in the analyzed samples.
- In general, bilinear MCR-ALS is adequate for the analysis of LC × LC data.
- MCR-ALS is able to analyze independently or simultaneously LC × LC-MS, LC × LC-UV and LC × LC-UV-MS (fused) datasets without needing any prior chromatographic peak alignment or peak shape modelling steps, even in the presence of strong coelutions.
- Data fusion of multiple detectors can provide a more powerful resolution of the sample constituents and an easier identification of the compounds associated with each MCR-ALS component.

Concerning the multivariate statistical methods for metabolomic data:

- Considering the three evaluated multivariate ANOVA-based methods, ASCA results agreed the most with the expected outcomes, whereas GASCA was too strict, and only some of the factors with lower *p*-values in ASCA were considered significant.
- However, the identification of potential makers by GASCA was the most congruent with the VIPs from PLS-DA.
- rMANOVA provided intermediate results between the good statistical assessment of ASCA and the appropriate number of significant variables identified by GASCA.
- The best standard workflow for the simultaneous statistical evaluation of the factors and the appropriate variable selection is proposed to be the combination of ASCA and PLS-DA results, which is also convenient due to their widespread and ease of use.

References

[1] T.S. Bos, W.C. Knol, S.R.A. Molenaar, L.E. Niezen, P.J. Schoenmakers, G.W. Somsen, B.W.J. Pirok, Recent applications of chemometrics in one- and two-dimensional chromatography, Journal of Separation Science. 43 (2020) 1678–1727. https://doi.org/10.1002/jssc.202000011.

[2] E. Gorrochategui, J. Jaumot, S. Lacorte, R. Tauler, Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow, TrAC - Trends in Analytical Chemistry. 82 (2016) 425–442. https://doi.org/10.1016/j.trac.2016.07.004.

[3] E. Gorrochategui, J. Jaumot, R. Tauler, ROIMCR: A powerful analysis strategy for LC-MS metabolomic datasets, BMC Bioinformatics. 20 (2019) 1–17. https://doi.org/10.1186/s12859-019-2848-8.

[4] D.W. Cook, S.C. Rutan, Chemometrics for the analysis of chromatographic data in metabolomics investigations, Journal of Chemometrics. 28 (2014) 681–687. https://doi.org/10.1002/cem.2624.

[5] M. Pérez-Cova, C. Bedia, D.R. Stoll, R. Tauler, J. Jaumot, MSroi: A pre-processing tool for mass spectrometry-based studies, Chemometrics and Intelligent Laboratory Systems. 215 (2021). https://doi.org/10.1016/j.chemolab.2021.104333.

[6] A. de Juan, J. Jaumot, R. Tauler, Multivariate Curve Resolution (MCR). Solving the mixture analysis problem, Analytical Methods. 6 (2014) 4964–4976. https://doi.org/10.1039/c4ay00571f.

[7] R. Bro, PARAFAC. Tutorial and applications, Chemometrics and Intelligent Laboratory Systems. 38 (1997) 149–171. https://doi.org/10.1016/S0169-7439(97)00032-4.

[8] R. Bro, C.A. Andersson, H.A.L. Kiers, PARAFAC2 - Part II. Modeling chromatographic data with retention time shifts, Journal of Chemometrics. 13 (1999) 295–309. https://doi.org/10.1002/(SICI)1099-128X(199905/08)13:3/4<295::AID-CEM547>3.0.CO;2-Y.

[9] M. Navarro-Reig, J. Jaumot, T.A. van Beek, G. Vivó-Truyols, R. Tauler, Chemometric analysis of comprehensive LC×LC-MS data: Resolution of triacylglycerol structural isomers in corn oil, Talanta. 160 (2016) 624–635. https://doi.org/10.1016/j.talanta.2016.08.005.

[10] A.C. Olivieri, G.M. Escandar, A.M. de la Peña, Second-order and higher-order multivariate calibration methods applied to non-multilinear data using different algorithms, TrAC - Trends in Analytical Chemistry. 30 (2011) 607–617. https://doi.org/10.1016/j.trac.2010.11.018.

[11] R. Tauler, Multivariate curve resolution of multiway data using the multilinearity constraint, Journal of Chemometrics. (2020) 1–24. https://doi.org/10.1002/cem.3279.

[12] X. Zhang, R. Tauler, Flexible Implementation of the Trilinearity Constraint in Multivariate Curve Resolution Alternating Least Squares (MCR-ALS) of Chromatographic and Other Type of Data, Molecules 2022, Vol. 27, Page 2338. 27 (2022) 2338. https://doi.org/10.3390/MOLECULES27072338.

[13] C.B. Zachariassen, J. Larsen, F. van den Berg, R. Bro, A. de Juan, R. Tauler, Comparison of PARAFAC2 and MCR-ALS for resolution of an analytical liquid dilution system, Chemometrics and Intelligent Laboratory Systems. 83 (2006) 13–25. https://doi.org/10.1016/J.CHEMOLAB.2005.12.010.

[14] S. Ren, A.A. Hinzman, E.L. Kang, R.D. Szczesniak, L.J. Lu, Computational and statistical analysis of metabolomics data, Metabolomics. 11 (2015) 1492–1513. https://doi.org/10.1007/S11306-015-0823-6.

[15] J. Bartel, J. Krumsiek, F.J. Theis, Statistical methods for the analysis of high-throughput metabolomics data, Computational and Structural Biotechnology Journal. 4 (2013) e201301009. https://doi.org/10.5936/CSBJ.201301009.

[16] A. Checa, C. Bedia, J. Jaumot, Lipidomic data analysis: Tutorial, practical guidelines and applications, Analytica Chimica Acta. 885 (2015) 1–16. https://doi.org/10.1016/j.aca.2015.02.068.

[17] J.G. Scheiner, S.M., MANOVA: multiple response variables and multispecies interactions. Design and Analysis of Ecological Experiments, in: C. Press (Ed.), Design and Analysis of Ecological Experiments, 1993: pp. 94–112.

[18] A.K. Smilde, J.J. Jansen, H.C.J. Hoefsloot, R.J.A.N. Lamers, J. van der Greef, M.E. Timmerman, ANOVA-simultaneous component analysis (ASCA): A new tool for analyzing designed metabolomics data, Bioinformatics. 21 (2005) 3043–3048. https://doi.org/10.1093/bioinformatics/bti476.

[19] J. Engel, L. Blanchet, B. Bloemen, L.P. van den Heuvel, U.H.F. Engelke, R.A. Wevers, L.M.C. Buydens, Regularized MANOVA (rMANOVA) in untargeted metabolomics, Analytica Chimica Acta. 899 (2015) 1–12. https://doi.org/10.1016/j.aca.2015.06.042.

[20] E. Saccenti, A.K. Smilde, J. Camacho, Group-wise ANOVA simultaneous component analysis for designed omics experiments, Metabolomics. 14 (2018) 1–18. https://doi.org/10.1007/s11306-018-1369-1.

[21] N.J. Nielsen, G. Tomasi, R.J.N. Frandsen, M.B. Kristensen, J. Nielsen, H. Giese, J.H. Christensen, A pre-processing strategy for liquid chromatography time-of-flight mass spectrometry metabolic fingerprinting data, Metabolomics. 6 (2010) 341–352. https://doi.org/10.1007/s11306-010-0211-1.
[22] M. Navarro-Reig, J. Jaumot, A. Baglai, G. Vivó-Truyols, P.J. Schoenmakers, R. Tauler, Untargeted Comprehensive Two-Dimensional Liquid Chromatography Coupled with High-Resolution Mass Spectrometry Analysis of Rice Metabolome Using Multivariate Curve Resolution, Analytical Chemistry. 89 (2017) 7675–7683. https://doi.org/10.1021/acs.analchem.7b01648.
[23] A.W. Dowsey, J.A. English, F. Lisacek, J.S. Morris, G.Z. Yang, M.J. Dunn, Image analysis tools and emerging algorithms for expression proteomics, Proteomics. 10 (2010) 4226–4257. https://doi.org/10.1002/pmic.200900635.
[24] R. Tautenhahn, C. Bottcher, S. Neumann, Highly sensitive feature detection for high resolution LC/MS, BMC Bioinformatics. 9 (2008) 1–16. https://doi.org/10.1186/1471-2105-9-504.
[25] and M.A. Hiroshi Tsugawa, Tomas Cajka, Tobias Kind, Yan Ma, Brendan Higgins, Kazutaka Ikeda, Mitsuhiro Kanazawa, Jean VanderGheynst, Oliver Fiehn, MS-DIAL: Data Independent MS/MS Deconvolution for Comprehensive, Nat Methods. 12 (2015) 523–526. https://doi.org/10.1038/nmeth.3393.MS-DIAL.
[26] M. Navarro-Reig, J. Jaumot, R. Tauler, An untargeted lipidomic strategy combining comprehensive two-dimensional liquid chromatography and chemometric analysis, Journal of Chromatography A. 1568 (2018) 80–90. https://doi.org/10.1016/j.chroma.2018.07.017.
[27] R.C. Allen, S.C. Rutan, Semi-automated alignment and quantification of peaks using parallel factor analysis for comprehensive two-dimensional liquid chromatography–diode array detector data sets, Analytica Chimica Acta. 723 (2012) 7–17. https://doi.org/10.1016/J.ACA.2012.02.019.
[28] C. Tistaert, H.P. Bailey, R.C. Allen, Y. vander Heyden, S.C. Rutan, Resolution of spectrally rank-deficient multivariate curve resolution: Alternating least squares components in comprehensive two-dimensional liquid chromatographic analysis, Journal of Chemometrics. 26 (2012) 474–486. https://doi.org/10.1002/cem.2434.
[29] H.P. Bailey, S.C. Rutan, Chemometric resolution and quantification of four-way data arising from comprehensive 2D-LC-DAD analysis of human urine, Chemometrics and Intelligent Laboratory Systems. 106 (2011) 131–141. https://doi.org/10.1016/j.chemolab.2010.07.008.
[30] M.B. Anzardi, J.A. Arancibia, A.C. Olivieri, Interpretation of matrix chromatographic-spectral data modeling with parallel factor analysis 2 and multivariate curve resolution, Journal of Chromatography A. 1604 (2019). https://doi.org/10.1016/j.chroma.2019.460502.
[31] S.A. Bortolato, A.C. Olivieri, Chemometric processing of second-order liquid chromatographic data with UV-vis and fluorescence detection. A comparison of multivariate curve resolution and parallel factor analysis 2, Analytica Chimica Acta. 842 (2014) 11–19. https://doi.org/10.1016/j.aca.2014.07.007.
[32] M.H. van Benthem, T.J. Keller, G.D. Gillispie, S.A. DeJong, Getting to the Core of PARAFAC2, A Nonnegative Approach, Chemometrics and Intelligent Laboratory Systems. 206 (2020). https://doi.org/10.1016/J.CHEMOLAB.2020.104127.
[33] R. Tauler, I. Marqués, E. Casassas, Multivariate curve resolution applied to three-way trilinear data: Study of a spectrofluorimetric acid-base titration of salicylic acid at three excitation wavelengths, Journal of Chemometrics. 12 (1998) 55–75. https://doi.org/10.1002/(sici)1099-128x(199801/02)12:1<55::aid-cem501>3.0.co;2-%23.
[34] R. Bro, H.A.L. Kiers, A new efficient method for determining the number of components in PARAFAC models, Journal of Chemometrics. 17 (2003) 274–286. https://doi.org/10.1002/cem.801.
[35] A. Malik, R. Tauler, Performance and validation of MCR-ALS with quadrilinear constraint in the analysis of noisy datasets, Chemometrics and Intelligent Laboratory Systems. 135 (2014) 223–234. https://doi.org/10.1016/j.chemolab.2014.04.002.
[36] R. Tauler, Multivariate curve resolution applied to second order data, Chemometrics and Intelligent Laboratory Systems. 30 (1995) 133–146. https://doi.org/10.1016/0169-7439(95)00047-X.
[37] D.W. Cook, S.C. Rutan, D.R. Stoll, P.W. Carr, Two dimensional assisted liquid chromatography - a chemometric approach to improve accuracy and precision of quantitation in liquid chromatography using 2D separation, dual detectors, and multivariate curve resolution, Analytica Chimica Acta. 859 (2015) 87–95. https://doi.org/10.1016/j.aca.2014.12.009.
[38] D.W. Cook, M.L. Burnham, D.C. Harmes, D.R. Stoll, S.C. Rutan, Comparison of multivariate curve resolution strategies in quantitative LCxLC: Application to the quantification of furanocoumarins in apiaceous vegetables, Analytica Chimica Acta. (2017). https://doi.org/10.1016/j.aca.2017.01.047.
[39] H.P. Bailey, S.C. Rutan, P.W. Carr, Factors that affect quantification of diode array data in comprehensive two-dimensional liquid chromatography using chemometric data analysis, Journal of Chromatography A. 1218 (2011) 8411–8422. https://doi.org/10.1016/j.chroma.2011.09.057.

[40] M. Pérez-Cova, J. Jaumot, R. Tauler, Untangling comprehensive two-dimensional liquid chromatography data sets using regions of interest and multivariate curve resolution approaches, TrAC - Trends in Analytical Chemistry. 137 (2021). https://doi.org/10.1016/j.trac.2021.116207.

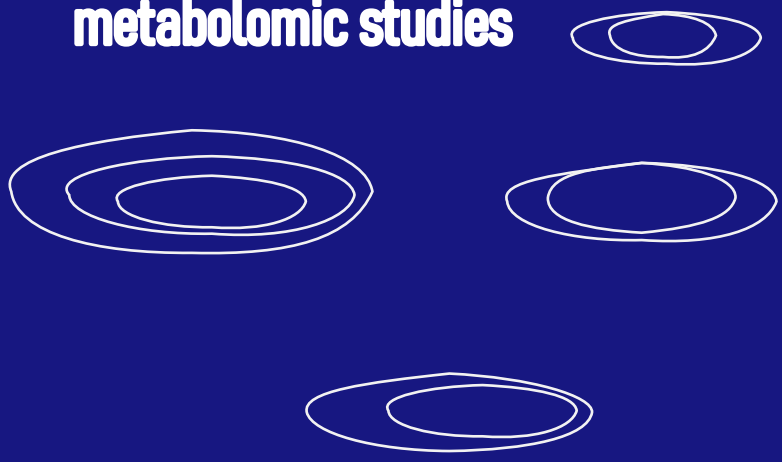
[41] P.S. Gromski, H. Muhamadali, D.I. Ellis, Y. Xu, E. Correa, M.L. Turner, R. Goodacre, A tutorial review: Metabolomics and partial least squares-discriminant analysis - a marriage of convenience or a shotgun wedding, Analytica Chimica Acta. 879 (2015) 10–23. https://doi.org/10.1016/j.aca.2015.02.012.

[42] J. Xia, I. v. Sinelnikov, B. Han, D.S. Wishart, MetaboAnalyst 3.0-making metabolomics more meaningful, Nucleic Acids Research. 43 (2015) W251–W257. https://doi.org/10.1093/nar/gkv380.

[43] Z. Pang, J. Chong, G. Zhou, D.A. de Lima Morais, L. Chang, M. Barrette, C. Gauthier, P.É. Jacques, S. Li, J. Xia, MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights, Nucleic Acids Research. 49 (2021) W388–W396. https://doi.org/10.1093/NAR/GKAB382.

Chapter four

Development and applications of LC×LC methodology for metabolomic studies



4.1 Introduction

Liquid chromatography coupled to mass spectrometry (LC-MS) arises as the leading analytical technique in metabolomics thanks to its versatility and sensitivity. Nevertheless, in untargeted metabolomics, the number of analytes is significantly high, and many similar compounds are found (e.g., enantiomers or structural isomers) when biological matrices are involved. In these cases, the chromatographic separation provided by LC, even in ultra-high performance (UHPLC) mode, may not be enough. Hence, multidimensional analytical platforms have emerged in recent years as a potential solution to expand metabolome coverage thanks to the increased resolving power and selectivity [1,2]. Besides, the peak capacity considerably augments when coupled to another orthogonal separation, which is especially useful in the case of complex samples, as exemplified in **Figure 4.1**. For instance, LC has been commonly combined in metabolomic studies with supercritical fluid chromatography (SFC) [3], ion mobility (IM) [4,5] or another LC separation with different chromatographic conditions or retention mechanisms, i.e., two-dimensional liquid chromatography (2DLC) [6].

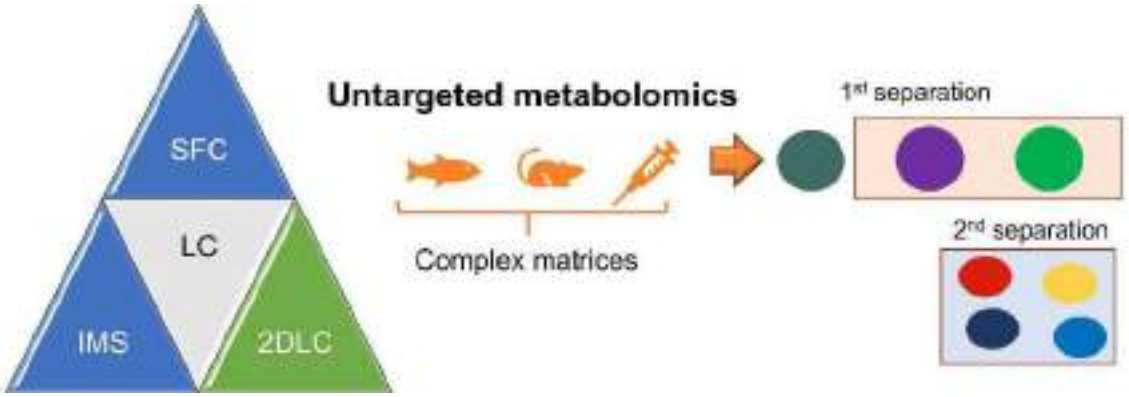


Figure 4.1. Visual exemplification of how multidimensional separations can increase metabolome coverage. 2DLC: two-dimensional liquid chromatography; IM: ion mobility; LC: liquid chromatography; SFC: supercritical fluid chromatography.

2DLC offers many possible combinations (see **Figure 4.2**) due to the wide variety of instrumental set-ups, retention mechanisms available, and parameters that can be optimized (e.g., stationary phases, the composition of mobile phases, salts and organic modifiers content, temperature, pH) [7]. This Chapter discussion aims to shed some light on the main challenges of the development and application of 2DLC methodology in metabolomics studies. More specifically, this Chapter focuses on the online comprehensive mode (LC × LC) because it is especially appealing for untargeted analysis, where global metabolome profiling is pursued, and unknown compounds can be discovered. Reversed phase (RP) and hydrophilic interaction chromatography (HILIC) have been combined for the analysis of lipids and metabolites using RP × HILIC and HILIC × RP modes, respectively (see **Figure 4.2**). The suitability of each order depends on the analytes, as discussed later in this Chapter, together with the main analytical and instrumental difficulties encountered (e.g., solvent incompatibility between the two dimensions, reduced sensitivity due to the dilution in the ^{2}D or long analysis time). In addition, active solvent modulation (ASM) [8] is suggested as a modulation interface for minimizing the cited problems and enhancing the separation.

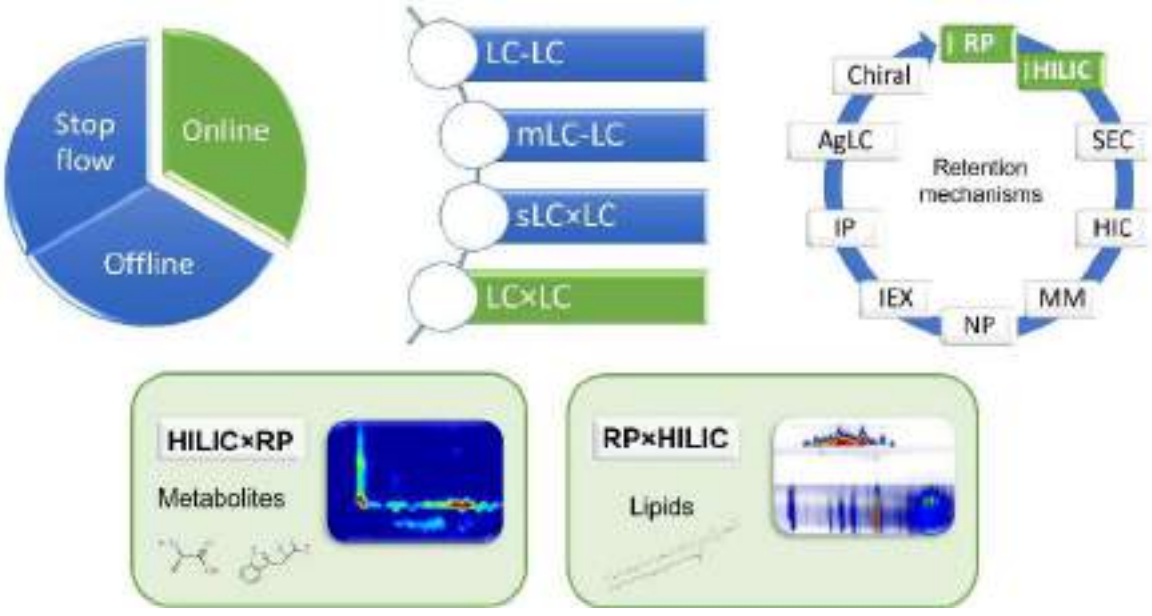


Figure 4.2. Summary of the instrumental set-ups available in 2DLC and retention mechanisms (the selected in this PhD Thesis are marked in green). LC-LC: heart-cutting; mLC-LC: multiple heart-cutting; sLC×LC: selective comprehensive, LC×LC: comprehensive; RP: reversed phase; HILIC: hydrophilic interaction; SEC: size exclusion; HIC: hydrophobic interaction; MM: mixed mode; IEX: ion exchange; IP: ion paring; AgLC: argentation.

Chemometric strategies that enhance the data analysis steps in a flexible and user-friendly manner are also discussed. The pre-processing of LC × LC datasets is performed by the regions of interest (ROI) approach implemented through a GUI interface [9] (see discussion section in the previous Chapter for more details). The main advantage of the ROI procedure for untargeted studies is that it allows a considerable reduction in the size of the datasets (e.g., in the case of LC × LC-HRMS data from tens of GB to only hundreds of MB). In the case of targeted analysis, the ROI approach (implemented via the MSROI GUI) allows selecting specific *m/z* values above the intensity evaluation. Besides, a strategy is suggested in this Chapter for the quantification in the presence of strong coelutions and unknown interferences, with and without the application of an area correlation constraint in the multivariate curve resolution alternating least squares (MCR-ALS) procedure. Some benefits of using this constraint are that they may help reducing rotation ambiguities in the case of the presence of unresolved strong chromatographic coelutions and that real concentration units can be derived directly from the iterative optimization results [10]. Besides, quantification can also be performed in the presence of unknown interferences in the samples (which were not present in the calibration mixtures), taking benefit from the second-order advantage [11].

4.2 Scientific publications

This section includes **scientific publications V and VI**, with a brief summary of each of them:

SCIENTIFIC PUBLICATION V

Global production of the endocrine disruptor bisphenol A (BPA) is still increasing nowadays, albeit this compound has been banned or strictly regulated in Europe. Consequently, unravelling its mode of action becomes a major issue. BPA is well-known for its estrogenic effect, but there is a need to characterize the obesogenic effect of this endocrine disruptor chemical. Hence, lipidomics is a powerful approach to evaluate the changes in the lipidome produced by this compound. In this study, BPA exposure is assessed on zebrafish eleutheroembryos by using an optimized LC × LC method for the untargeted analysis of their lipidome. The final method allowed the detection of changes in the lipidome caused by bisphenol A (BPA) and their comparison with the changes produced by an estrogenic control, the natural hormone 17-ß-estradiol (E2)). The design of the experiment included:

a) Two sampling days at critical stages of lipid absorption in eleutheroembryos,
b) Three doses of exposure plus Control samples for BPA and E2,
c) Two extraction protocols.

The estrogenic and obesogenic effects of BPA in the lipidome are studied.

SCIENTIFIC PUBLICATION VI

Nowadays, there is a need for standardized quantification protocols for LC × LC-MS datasets that do not require specific vendor software. This work aims to compare different quantification strategies for LC × LC-MS datasets. Three approaches based on the ROIMCR method are considered:

1) A calibration curve based on the areas for each *m/z* value from the ROI approach.
2) A classic calibration curve from areas of the resolved MCR-ALS elution profiles.
3) A calibration curve using the area correlation constraint during the iterative ALS optimization.

Prior to approaches 2 and 3, the ROI intensity matrix is analyzed by MCR-ALS. The results from the three strategies were evaluated as an alternative workflow for the quantification of LC × LC-MS datasets.

V. SCIENTIFIC PUBLICATION V

Title: Untargeted lipidomics of zebrafish (*Danio rerio*) eleutheroembryos exposed to Endocrine-disrupting chemicals using comprehensive two-dimensional liquid chromatography and advanced chemometrics

Authors: Miriam Pérez-Cova, Laia Navarro-Martin, Gabriel Leme, Romà Tauler, Benjamin Piña, Joaquim Jaumot, Dwight R. Stoll

In preparation

Title:

Untargeted lipidomics of zebrafish (*Danio rerio*) eleutheroembryos exposed to endocrine disrupting chemicals using comprehensive two-dimensional liquid chromatography and advanced chemometrics

Authors:

Miriam Pérez-Cova[a,b,c], Laia Navarro-Martin[a], Gabriel Leme[c], Romà Tauler[a], Benjamin Piña[a], Joaquim Jaumot[a*], Dwight R. Stoll[c]

[a]Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, 08034 Barcelona, Spain

[b]Department of Chemical Engineering and Analytical Chemistry, University of Barcelona, Diagonal 647, 08028, Barcelona, Spain.

[c]Department of Chemistry, Gustavus Adolphus College, Saint Peter, Minnesota 56082, United States

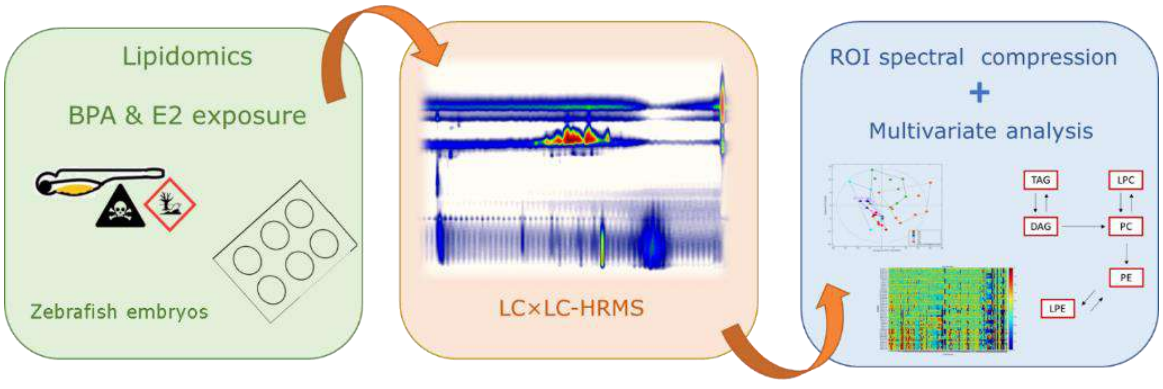
* Correspondence: joaquim.jaumot@idaea.csic.es

0. Abstract

Global production of the endocrine disruptor bisphenol A (BPA) is still increasing nowadays, albeit this compound has been banned or strictly regulated in Europe. Consequently, unravelling its mode of action becomes a major issue. BPA is well-known for its estrogenic effect, but there is a need for the characterization of the obesogenic effect of this endocrine disruptor chemical. Hence, lipidomics is a powerful approach to evaluating the changes in the lipidome produced by this compound. In this study, BPA exposure is assessed on zebrafish eleutheroembryos using an optimized two-dimensional liquid chromatography coupled to high resolution mass spectrometry (LC×LC-HRMS) method for untargeted analysis of their lipidome. The use of LC×LC provided a greater separation power than one-dimensional liquid chromatography, which allowed a better characterization of the zebrafish lipidome. In addition, the use of active solvent modulation (ASM) as interface between the two chromatographic dimensions enhanced the sensitivity of the LC×LC method as well as solvent compatibility between both chromatographic dimensions. The final method allowed the detection of changes in the lipidome caused by bisphenol A (BPA) and their comparison with the changes produced by an estrogenic control, the natural hormone 17-ß-estradiol (E2)). The altered lipids for both EDCs seemed to be linked to the estrogenic effect of BPA. However, an obesogenic effect was also found for E2 exposure, which difficulted the characterization of the non-estrogenic effect caused by BPA.

Keywords: LC × LC-HRMS, zebrafish, lipidomics, endocrine disrupting chemicals, chemometrics, bisphenol A

Graphical Abstract



1. Introduction

Bisphenol A (BPA) is a synthetic chemical compound used mainly in the manufacturing of polycarbonates (65%), epoxy resins (28%), and flame retardants (7%) (Abraham and Chakraborty, 2019). Its environmental releases usually come from industrial effluents or plastic

wastes that end up in rivers when BPA is not fully removed in wastewater plants (Chakraborty et al., 2021; Sun et al., 2017; Wang et al., 2019). BPA is also present in daily life products such as thermal paper, or metal and plastic drinking bottles Banaderakhshan et al., 2022; Farooq et al., 2021; Kovačič et al., 2020) and in indoor dust and air from daily frequented places (e.g., houses, offices, laboratories) (Caban and Stepnowski, 2020; Lee et al., 2021; Vasiljevic and Harner, 2021). Although its use has been restricted in Europe and even banned, for example for its use in materials for baby bottles, sippy cups, and infant formula packaging ("COMMISSION REGULATION (EU) 2018/213"), global BPA production is expected to increase in the next ten years a 4.6% regarding the 2013-2019 period ("Bisphenol A (BPA) Market Size, Share, Industry Report 2030 | ChemAnalyst,"). Besides, the Environmental Protection Agency (EPA) ("EPA - Risk Management for Bisphenol A (BPA)," states that BPA releases into the environment can exceed one million pounds per year. Therefore, BPA has been found in a variety of samples with concentrations up to 30 μg g^{-1} dry weight in fish tissues (Tao et al., 2021). In children, BPA levels have been reported with concentrations up to 2 μg L-1 in urine (Tkalec et al., 2021).

BPA is a well-known endocrine disruptor chemical (EDC) ("EPA - Risk Management for Bisphenol A (BPA)"). This plastic additive can bioaccumulate and biomagnify through the food chain, causing severe damage to aquatic ecosystems and human populations (Pop et al., 2021; Wang et al., 2020). BPA produces an estrogenic effect (Ben-Jonathan and Steinmetz, 1998; Chen et al., 2002; Eramo et al., 2010; Heindel and Blumberg, 2019; Moon, 2019; Paris et al., 2002) because it mainly acts as agonist of estrogen receptors (Heindel and Blumberg, 2019; Mu et al., 2018) but is also known for interacting with other receptors such as retinoid (Martínez et al., 2018), estrogen-related gamma (Tohmé et al., 2014) and peroxisome proliferatoractivated gamma (Martínez et al., 2018). Apart from its well-known estrogenic properties, BPA has been shown to act as an endocrine disruptor chemical (EDC) by interfering with other signaling pathways of the endocrine system. For example, this compound is also classified as obesogenic (Santangeli et al., 2018) since can be linked to obesity, adipogenesis, diabetes and cardiovascular diseases (Longo et al., 2020; Pérez-Bermejo et al., 2021; Silva et al., 2021). In addition, BPA also impacts on the immune system (Gear and Belcher, 2017; Sawai et al., 2003; Tkalec et al., 2021; Ishido et al., 2004; Yamaguchi et al., 2006). Besides, other studies demonstrated that most of the effects caused by BPA exposure in development are not mimicked by an estrogenic control (Gould et al., 1998; Martínez et al., 2020c). Hence, BPA has multiple targets and effects, apart from its estrogenicity, capable of influencing multiple endocrine-related pathways. Altogether, despite the fact that the estrogenic activity of BPA has been widely studied (Cano-Nicolau et al., 2016; Chen et al., 2002; Eramo et al., 2010) more research is needed to characterize the non-estrogenic effect of this compound.

Lipidomics is a useful tool to study the changes in the lipidome caused by certain stressors and is especially appealing to characterize the obesogenic effect produced by BPA. The alterations in the lipidome due to BPA exposure have been previously assessed in liver cells (Marqueño et al., 2021), aquatic invertebrates (Fuertes et al., 2018), or rodents (Nguyen et al., 2021). In this work, aquatic vertebrate zebrafish eleutheroembryos (*Danio rerio)* have been selected to unravel BPA exposure. Other omic studies have been carried out in zebrafish for this EDC before, such as metabolomics and transcriptomics (Huang et al., 2020; Martínez et al., 2020c; Ortiz-Villanueva et al., 2018, 2017; Tian et al., 2021; Nguyen et al., 2021). In this work, aquatic vertebrate zebrafish eleutheroembryos (*Danio rerio*) have been selected to unravel BPA exposure. Other omic studies have been carried out in zebrafish for this EDC before, such as metabolomics and transcriptomics (Huang et al., 2020; Martínez et al., 2020c; Ortiz-Villanueva et al., 2018, 2017; Tian et al., 2021). A previous lipidomic study focused on the BPA exposure in zebrafish eleutheroembryos at the same concentration level for several days post-fertilization (Martínez et al., 2020a). The study presented here characterizes the BPA effects with a simultaneous comparison with an estrogenic control on early development stages of zebrafish and at various concentration levels of exposure. The aim is to untangle the mode of action of this EDC, from both the obesogenic and estrogenic points of view.

In this work, a cutting-edge analytical methodology has been employed to increase the lipidome coverage (i.e., separate and identify isobaric lipids as well as increase the resolving power of current methods). A comprehensive two-dimensional liquid chromatography coupled to high-resolution mass spectrometry (LC×LC-HRMS) set-up has been optimized for untargeted lipid analysis, employing RP in the first dimension (^{1}D) and HILIC in the second dimension (^{2}D). The main advantage over previous LC×LC methods is the use of Active Solvent Modulation (ASM) as a valve-based interface between both dimensions (Stoll et al., 2017), applied for the first time for lipid analysis, which enhances analytical sensitivity and solvent compatibility between both dimensions, while reducing the total analysis time. Besides, a data analysis strategy for LC×LC datasets, usually the bottleneck of this type of analysis, is also proposed. The approach joins a first step of spectral compression by the regions of interest (ROI) approach for reducing the dimensionality of the data, and a second step of combined univariate and multivariate analyses approaches.

The goal of this study is to characterize the effect of BPA exposure by comparing the changes in the lipidome caused by this EDC with an estrogenic control, the natural hormone 17-ß-estradiol (E2). To do so, our study exposed zebrafish eleutheroembryos from 2 to 6 days post-fertilization (dpf) to BPA and E2 in a dose-response manner and collected samples at 4 and 6 dpf for further lipidomic analysis. Two extraction protocols were employed for a broader lipid coverage (a

general lipid extraction and a sphingolipid-based extraction). Finally, the optimized RP×HILIC method with ASM as modulation interface, and the data analysis strategy for multidimensional datasets were used to determine changes in the lipidomic profiles of exposed animals and to shed some light on the mechanisms of action of BPA.

2. Materials and methods

2.1 Chemicals and reagents

Bisphenol A (BPA, ≥99.0%), 17β-estradiol (E2, ≥98.0%, dimethyl sulfoxide (DMSO, for molecular biology, ≥99.9%), ammonium acetate (NH_4Ac, ≥99.0%), acetic acid (HAc, ≥95.0%), calcium sulfate dihydrate ($CaSO_4 \cdot 2H_2O$) and dibutylhydroxytoluene (BHT) were purchased from Sigma-Aldrich (St. Louis, USA). Chloroform ($CHCl_3$ ≥99.0%) was provided by Carlo-Erba reagents, and formic acid (FAc, 50%) was purchased from Honeywell Fluka. HPLC grade water, HPLC grade methanol (MeOH) and HPLC grade acetonitrile (AcN) were supplied by Merck KGaA (Darmstadt, Germany). HPLC grade water used in HPLC analysis was obtained from a Merck-Millipore Milli-Q® system (Burlington, United States) ultra-filtration system.

Seven lipid standards from different families were used in this study, grouped into two mixes for extraction purposes. Phospholipids mix was composed by 17:0-17:1-17:0 D5 triglyceride, 16:0 D31-18:1 phosphatidylethanolamine, 17:1 lysophosphatidylethanolamine, 17:1 lysophosphatidylglycerol, Sphingolipids mix included 18:1-12:0 N-lauroyl-D-erythro-sphingosine, 18:1-12:0 D-glucosyl-ß-1,1'-N-lauroyl-D-erythro-sphingosine, 17:1 D-erythro-sphingosine. All these lipid standards were purchased from Avanti Polar Lipids (Alabaster, AL, US). For identification purposes, all seven lipids were jointed in a lipid standard mix and measured in the same conditions as the samples.

2.2 Animal maintenance and rearing conditions

Adult wild-type zebrafish (*Danio rerio*) were maintained under standard conditions: 28 ± 1 °C, with 12 Light:12 Dark photoperiods, fed twice a day with dried flakes (TetraMin, Tetra, Germany). They were kept in fish water, composed of reverse-osmosis purified water, which contained 90 $\mu g \cdot mL^{-1}$ of Instant Ocean (Aquarium Systems, Sarrebourg, France) and $CaSO_4 \cdot 2H_2O$ (100 $\mu g \cdot mL^{-1}$). Eleutheroembryos from zebrafish were obtained by natural mating by placing five females and three males on 4-L breeding tanks, kept separated from adults through a bottom mesh. Eggs were collected and rinsed at 2 hours post-fertilization (hpf). At 24 hpf, the fertilized eggs were distributed randomly in 6-well multi-plates at a density of 3.3 embryos per

mL (10 individuals per well, in 3.0 mL) in clean fish water until the start of the exposures. All experiments were conducted in accordance with the institutional guidelines under a license from the local government (DAMM 7669, 7964) and approved by the Institutional Animal Care and Use Committees at the Research and Development Centre of the Spanish National Research Council (CID-CSIC).

2.3 Zebrafish eleutheroembryos exposures and sample collection

A concentrated stock solution for each compound and intermediate diluted solutions for each concentration level were prepared at the beginning of the experiment in DMSO and kept at 4 ˚C. All working solutions were prepared by diluting the intermediate stocks and changed every day to ensure continuous exposure to the chemicals until embryo collection. In these solutions, DMSO was used as a vehicle at a final concentration of 0.2% in fish water (the percentage of DMSO that was also added to the control groups). Exposures started at 2 dpf (days post-fertilization) and were carried out until 6 dpf. Although the chemical stability of the working solutions for the two compounds was found stable for at least 48 h in the absence of any degradation agent (Jordão et al., 2016; Jürgens et al., 2002), water renewal was performed every day to ensure proper exposure.

The present study aimed to use concentrations under sub-lethal and low phenotypic effects at the morphological level. Based on previous studies (Martínez et al., 2020a, 2019; Ortiz-Villanueva et al., 2017), a preliminary range-finding test was performed in which tested concentrations ranged from 26 to 43.8 μM for BPA and 8 to 12 μM for E2. A stereomicroscope was used to oversee eleutheroembryos development during exposure. Different parameters were controlled daily until its collection at 6 days post-fertilization (144 hpf), with an especial interest in the following rates: mortality/survival (from 24 hpf), hatching (from 72 hpf) and swim bladder inflation (from 96 hpf).

Based on these results and to perform the lipidomic study at concentrations that avoided possible molecular events related to cellular/organisms death processes, the highest levels of exposure were set to 26 μM for BPA and 8 μM and E2, and the rest selected to be comparable to previous lipidomic (Martínez et al., 2020a) and metabolomic (Ortiz-Villanueva et al., 2018) studies. Hence, the nominal concentrations were: 4, 18 and 26 μM for BPA, and 1, 4 and 8 μM for E2. Real concentrations (summarized in **Table 1**) were determined by triplicate preparing mock solutions and measured using liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS), as explained in **Supplementary Material A Section 2**. **Table 1** also expresses the % that represents the real concentrations regarding the nominal ones. Collection days were se 4 and 6 dpf because they are critical stages of yolk sac reabsorption (Martínez et al., 2020a). Pools

of 10 eleuteroembryos per replicate were gathered for each biological replicate (n= 4), and samples were frozen at -80 ˚C until extraction.

Table 1. Real concentrations of exposure determined by LC-MS/MS.

Concentrations	BPA			E2		
Nominal (µM)	4	18	26	1	4	8
Real (µM)	3.50 ± 0.07	22.3 ± 0.1	30.2 ± 0.1	0.83 ± 0.07	4.8 ± 0.1	7.4 ± 0.1
%	87.5	123.9	116.2	83.0	120.0	92.5

2.4 Lipid extraction

2.4.1General lipid extraction

The detailed protocol of the general lipid extraction can be found elsewhere (Martínez et al., 2020a). Briefly, it consists of a $CHCl_3$:MeOH (2:1) extraction using BHT to avoid lipid oxidation. The procedure starts with the addition of 750 µL of a $CHCl_3$:MeOH (2:1) solution with 1% BHT to the frozen sample. Phospholipids extraction standards mix (10 µL at 20 µM, per sample) and 2 stainless steel beads (7 mm diameter) were also added. Samples were homogenized in the tissuelyzer LT Qiagen (Hilden, Germany) at 50Hz for 4 min. Other 75 µL of a $CHCl_3$:MeOH (2:1) solution (without BHT) were added, and vortexed. The extract was transferred to a new Eppendorf with a glass pipette. Afterwards, the addition of 375 µL of saline solution (0.88% KCl) allowed a two-phases separation (organic phase at the bottom, and aqueous phase at the top), which were vortexed and centrifuged for 4 min at 14500 rpm and 4˚C. The organic phase was evaporated until dryness under nitrogen steam, resuspended in another Eppendorf in 800 µL of $CHCl_3$:MeOH (2:1), centrifuged under the same conditions, and re-evaporated. Finally, extracts were resuspended in 200 µL of $CHCl_3$:MeOH (2:1), centrifuged and transferred to glass chromatographic vials. They were evaporated again until dryness and kept in an argon atmosphere at -80 ˚C until analysis. Prior to analysis, samples were resuspended in 100 µL of AcN and the sphingolipids instrumental standards mix (5 µL at 20 µM, per sample) was added. Quality control (QC) samples were generated by pooling 50 µL of the fourth replicate of the extracted samples for each experimental condition.

2.4.2Sphingolipids-based extraction

The sphingolipid extraction was based on a combination of previously published protocols (Dalmau et al., 2015; Martínez et al., 2020a) with minor modifications. Briefly, frozen samples were homogenized in the tissuelyzer LT Qiagen (Hilden, Germany) for 4 min at 50Hz, after adding 60 µL of a $CHCl_3$:MeOH (1:2) solution containing 1% BHT, the sphingolipids standards

mix (10 µL at 20 µM, per sample) and 2 stainless steel beads (7 mm diameter). Then, 60 µL of a $CHCl_3$:MeOH (1:2) solution (without BHT) were added. After vortexing, the mixture was transferred to a glass vial using a glass pipette. The saponification step started with the addition of 50 µL of KOH 1M. Samples were sonicated for 15 min and then put in the oven at 37 ˚C for 2 hours. Afterwards, 150 µL of CH_3COOH 1 M were added to stop the saponification. Then, 200 µL of $CHCl_3$ and 375 µL of 0.88% KCl were added. Two layers were formed: the organic phase at the bottom, and the aqueous phase at the top. Samples were vortexed and centrifuged for 4 min at 14500 rpm and 4˚C. The organic phase was evaporated until dryness under nitrogen steam and resuspended in 8 µL of $CHCl_3$:MeOH (1:2) in another Eppendorf. The new tubes were centrifuged under the same conditions, and re-evaporated. Another resuspension was performed in 200 µL of $CHCl_3$:MeOH (1:2), followed by a centrifugation and transfer to glass chromatographic vials. Finally, extracts were evaporated again until dryness and kept in an argon atmosphere at -80 ˚C until analysis. Prior to analysis, samples were resuspended in 50 µL and the phospholipid instrumental standards mix (10 µL at 20 µM, per sample) was added. Quality control (QC) samples were generated by pooling 25 µL of the fourth replicate of the extracted samples for each experimental condition.

2.5 LC×LC-HRMS analysis

Three biological replicates for each dose (Control, Low, Medium and High concentrations of exposure), day (4 or 6 dpf), treatment (BPA or E2 treatments) and extraction type (general or sphingolipid-based) were analyzed by LC×LC-HRMS. In addition, QC samples and blanks were incorporated along the sequence.

An RP×HILIC setting was employed for the LC×LC analyses. An Agilent Poroshell HPH C18 (150 mm x 2.1 mm i.d., 1.9 µm) was selected for the first-dimension separation, at a temperature of 50 ºC. Mobile phases of ^{1}D were: A) 30 mM ammonium formate (pH 4.5), and B) ACN/IPA 33.3:66.6 (v/v). Gradient composition and flow rate were established as follows (%B, flow rate mL/min): 0 min (70, 0.04), 60 min (90, 0.04), 75 min (100, 0.04), 110 min (100, 0.04), 111 min (70, 0.12), 118 min (70, 0.12), 119 min (70, 0.04), 120 min (70, 0.04) (110-120 min re-equilibration step). The total chromatographic run was 120 min per sample.

A HILIC column was prepared in-house for the second-dimension separation, by slurry packing unmodified bare Zorbax silica (3.5 µm, 80 Å pore size) into a small column (20 mm x 2.1 mm i.d.). ^{2}D column was held at 40˚C. A flow rate of 2 mL min^{-1} was employed, and a passive flow splitting using a simple T-piece before detection with a split ratio of 1:2 (1 part to MS, 2 parts to waste). Mobile phases of ^{2}D were: A) 30 mM ammonium formate (pH 4.5), and B) pure AcN. Gradient composition expressed as % B was: 0 min (100), 0.22 min (100), 0.8 min (65),

1.00 min (100). Modulation time was set to 1 min, and the first 0.22 min corresponded to the ASM step (Stoll et al., 2017); the actual gradient time was 0.8 min. ASM dilution factor was 5.

A 6545XT AdvanceBio LC/Q-TOF (Agilent Technologies, Santa Clara, CA) mass spectrometer with an Agilent JetStream (AJS) electrospray ionization source in positive mode was employed. Full scan spectra were acquired from 100 to 1500 Da. The acquisition frequency was 125 ms/spectrum. Auto MS/MS was set for obtaining iterative fragmentations MS/MS of the QCs. The collision energy was 25 eV, the acquisition rate was 8 spectra/s. Precursors were sorted by abundancy only, and the scan speed varied with the precursor abundance (target 25000 counts/spectrum).

2.6 Data analysis

A first compression step was applied to LC×LC-HRMS datasets, followed by multivariate analysis aiming to unravel the most significant effects of BPA and E2 in the lipidome of the zebrafish eleutheroembryos.

2.6.1 Survival, hatching and swim bladder inflation rates

Survival, hatching and swim bladder inflation rates were determined in 6 and 16 replicates per condition (ranging test and exposures for lipidomics, respectively) with a total of 10 eleutheroembryos per replicate. Kruskal-Wallis followed by pairwise comparisons using the Wilcoxon rank sum test were performed to assess statistically significant differences (considering p-values < 0.05). Both statistical analysis and associated graphs were performed by *tidyr* (Hadley Wickham, Maximilian Girlich, 2016), *stats* (R Foundation for Statistical Computing, Vienna, 2020), *ggplot2* (Wickham, 2017), *ggpubr* (Kassambara, 2020) and *grid* (R Foundation for Statistical Computing, Vienna, 2020) packages in the R environment v.4.0.3 (R Foundation for Statistical Computing, Vienna, 2021), using RStudio (RStudio, Inc. and MA, 2018).

2.6.2 Lipidomic analysis

Data compression, filtering and normalization

Raw vendor LC×LC-HRMS data were converted to (.mzXML*)* format with the MS Convert GUI (Palo Alto, CA, USA) from the Proteowizard open-source software (Chambers et al., 2012). First, a signal threshold prefilter of absolute intensity higher than 100 was applied during this conversion, to reduce the number of low intensity *m/z* values associated with each retention time allowing a size reduction from 13 GB to approximately 1 GB per file. The MSroi procedure was selected for further data compression and filtering, implemented through the Msroi app (Pérez-Cova et al., 2021). The regions of interest approach (ROI) applied to mass spectrometry data allows a spectral compression without any loss of spectral accuracy (Gorrochategui et al., 2019).

This strategy selects the most intense *m/z* values above an intensity threshold set by the user. The signals above this threshold are kept for further analysis, whereas the others are considered noise and, therefore, omitted. Other parameters are also required, for instance, related to the maximum spectral resolution of the mass spectrometer (i.e., mass error tolerance) or the minimum points needed to define a chromatographic peak within all samples (i.e., minimum occurrences). More information about the MSroi procedure and parameters is included in **Supplementary Material A Section 3**. Briefly, an *m/z* range of 400-1500 was selected, with a signal threshold of 6000, a signal factor of 4, a mass error tolerance of 0.1 Da, a minimum of 5 occurrences and the final *m/z* values for each ROI was calculated by the median of all values detected for this ROI. MSroi approach also provided the areas of each *m/z* (feature) as output. A total of eight datasets were compressed separately, according to EDC (BPA or E2), extraction (general or sphingolipid-based) and exposure time (4 and 6 days post-fertilization).

Approximately 450 features (300 for the general extraction and 150 for the sphingolipid) were obtained for each day and EDC type. Days, EDC and extraction sets were then grouped into a single matrix with up to 567 unique features. This matrix was imported into R studio for the post-processing analysis first steps, including outliers removal (in both dimensions, samples and features), replacing zeros with the minimum value of the corresponding feature and normalization. Areas were normalized according to the number of the zebrafish eleutheroembryos per sample and the measured surrogates and internal standards area values, with the aim of correcting possible instrumental drifts or extraction losses. Finally, the PQN mathematical normalization was applied.

Statistical assessment, exploratory analysis, and discovery of potential markers of the exposure

Once the normalized area matrix was obtained, post-processing strategies were applied to obtain an exploratory overview of the data and to identify features (lipids) affected by the exposures, being either common for both compounds or specific to each one. Univariate statistical assessment was performed with SPSS 27.0.1.0 (©Copyright IBM Corporation), and further multivariate analysis was carried out in MATLAB environment (Release 2020b, The Mathworks Inc, Natick, MA, US), using the PLS Toolbox 8.9.1 (Eigenvector Research Inc, Wenatchee, WA, US). First, a principal component analysis (PCA) was used for unsupervised exploratory analysis (Vidal et al., 2016). In the PCA scores plot, clusters of the biological replicates were expected and control samples should cluster separately from the higher doses of exposure. In addition, differences between collection days and even between control samples are also expected, due to natural biological changes in the lipidome at those stages of growth (i.e., more lipids from the yolk sac are absorbed at 6 dpf than 4 dpf). PCA was applied after a PQN and autoscaling normalization.

Next, ANOVA-simultaneous component analysis (ASCA) was employed for an initial multivariate statistical assessment. ASCA combines the multivariate analysis of variance method (for multiple variables) and the simultaneous component analysis (SCA) with the aim of modelling the effect of different factors defined in the design of the experiment (DoE). More details on ASCA principles and applications can be found elsewhere (Bertinetto et al., 2020; Smilde et al., 2005). ASCA was applied in this study, after a PQN and mean-centered normalization, to evaluate the significance of the concentration levels, days of exposure (4 and 6 dpf) and EDC (BPA and E2). The statistical assessment was carried out via a permutation test in which 10,000 replicates were considered.

The selection of the relevant variables was performed through ANOVA tests, considering four sets of samples: BPA-4dpf, BPA-6dpf, E2-4dpf, E2-6pdf. Each data set was composed of the three exposure doses for each EDC and control samples. A second filtering step was performed on the normalized matrix, keeping only the significant values from at least one of the four sets (p-values < 0.05).

Hierarchical clustering analysis (HCA) combined with dendrograms and partition around medoids (PAM) clustering analyses were carried out on the matrix containing the relevant features, with the aim of identifying clusters of lipids presenting a similar behavior according to the different EDC exposure conditions. PAM and clustering analyses were both performed using the packages *gplots* (Gregory R. Warnes, Ben Bolker, Lodewijk Bonebakker, Robert Gentleman et al., 2020), *fpc* (Hennig, 2020) and *cluster* (Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., 2021) packages in the R environment v.4.0.3 (R Foundation for Statistical Computing, Vienna, 2021) using RStudio (RStudio, Inc. and MA, 2018).

Lipid identification

Lipid identification was focused on the *m/z* values associated with significant changes between treated and control samples, according to the ANOVA results.

First, QCs with MS/MS information were loaded into the MS-DIAL software (Tsugawa et al., 2020). More information on the MS-DIAL parameters can be found in **Supplementary Material A Section 4.** Lipids were annotated based on MS/MS matches with MS-DIAL MS/MS spectral libraries (Tsugawa et al., 2020). Besides, retention time windows were determined for both dimensions according to the relative position of the lipid standard mix measured in the same conditions. Next, the proposed annotation was validated by comparison with available lipid libraries from the literature on zebrafish eleutheroembryos (Fraher et al., 2016; Martínez et al., 2020a; Zhao et al., 2019) and LIPIDMAPS database (Fahy et al., 2009, 2007).

3. Results and discussion

3.1 Zebrafish eleutheroembryos exposures

Dose-ranging tests were carried out to set the highest concentration of the exposure ensuring working at sublethal doses with low morphological effects. The results from the different doses impact (see **Supplementary Material Section A Figure 1**) showed statistical differences in survival and hatching rates only for the highest BPA dose (44 µM) when compared to controls (Wilcoxon, *p*-values<0.05). Moreover, swim bladder inflation (SBI) rates were decreased at all tested doses for both BPA and E2 at 4 dpf, and only individuals exposed to 26 µM and 8 µM recovered to control rates at 6 dpf. Therefore, the lowest observed effect concentrations (LOEC) were set to 26 µM and 8 µM for BPA and E2, respectively. These results are in agreement with the previously reported by Martínez et al. (Martínez et al., 2020a, 2019, 2018). Hence, these LOECs were established as the highest doses selected for the lipidomic study.

In the final exposure study, no statistical differences in survival and hatching rates were detected between controls and any exposed groups to BPA or E2. Conversely, a significant decrease in the swim bladder inflation rates was observed in all tested doses for both BPA and E2 at 4 dpf, that was only recovered at 6 dpf in the individuals exposed to 4 and 18 µM of BPA and 1 µM of E2 (see **Figure 1**).

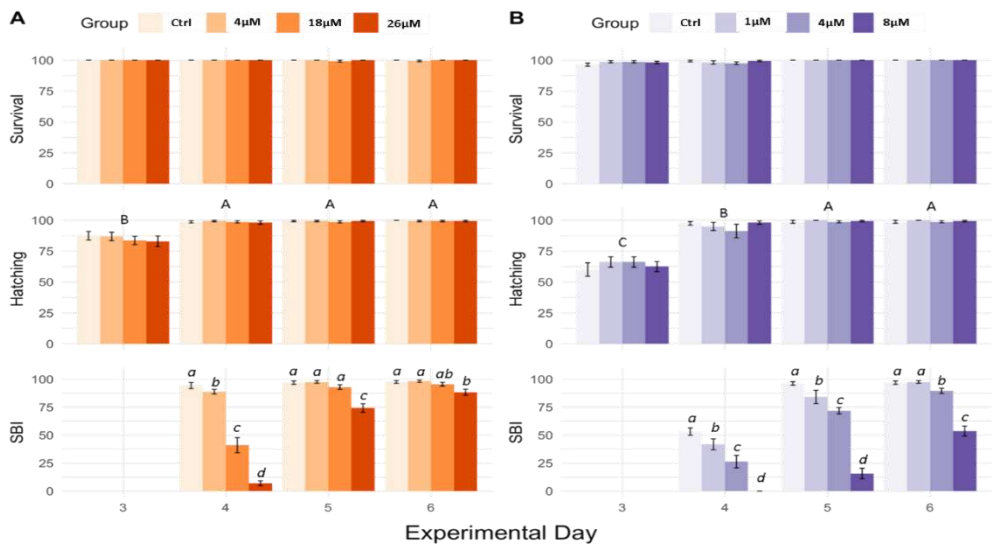


Figure 1. Effects of increasing concentrations of **(A)** BPA (4-26 µM) and **(B)** E2 (1-8 µM) on survival, hatching and swim bladder inflation (SBI) rates at 3, 4, 5 and 6 dpf. The mean value ± SD (standard deviation) is shown for each group (n=16 groups of 10 individuals each). A non-parametric test (Kruskal-Wallis with Wilcoxon rank sum test for multiple comparisons, *p*-values < 0.05) was performed. Capital letters denote significant differences between days of development regardless of the treatment. Italic letters denote differences between concentrations within each developmental day.

3.2 Development of the RP×HILIC-HRMS method for lipid analysis

The main advantage of an LC×LC method for lipid analysis in comparison with one-dimensional liquid chromatography is that isobaric compounds can be separated according to different retention mechanisms. For instance, in RP, lipids are separated by their hydrophobicity (i.e., by the length of their chains and the number and positions of the double bonds), whereas HILIC separates by their hydrophilicity (i.e., polar head groups, characteristic of each lipid family) (Cífková et al., 2016; Holčapek et al., 2015a). Hence, lipid resolution is considerably increased in an LC×LC set-up. Previous works have pointed out the use of an RP×HILIC set-up (with RP in the ^{1}D and HILIC in the ^{2}D), which has been preferred due to its higher efficiency in the separation (Holčapek et al., 2015a; Xu et al., 2020). This is because, in RP×HILIC, the stress of the separation is on the hydrophobic part of the lipids, provided by the RP dimension, whereas HILIC provides a quick screening discriminating by lipid families. In the opposite set-up, a short RP separation as ^{2}D may not be powerful enough to differentiate between similar compounds, at least in a comprehensive LC×LC mode. Hence, the configuration RP×HILIC has been selected in this work.

One of the major improvements of the LC×LC method developed for this study regarding the already existing literature for lipid analysis is the use of ASM. ASM is a valve-based approach designed for improving solvent compatibility between both dimensions while enhancing global sensitivity (Stoll et al., 2017). More information on the followed ASM strategy can be found in **Section 5 of Supplementary Material A**. Thus, ASM allowed higher fractions of the ^{2}D effluent to access the MS, with a split ratio of 1:2 (1 part to MS, 2 parts to waste) and the use of a bigger loop size (40 μL loops) to store ^{1}D fractions before entering in the ^{2}D column. Besides, the effluent from the ^{1}D was diluted with ^{2}D mobile phase composition before reaching the ^{2}D column, which enhanced the retention in the ^{2}D separation (the water content at the beginning of the ^{2}D separation was drastically reduced, improving the retention of the most polar compounds in the HILIC column). Consequently, the total sensitivity was considerably increased, and the total analysis time was reduced, in comparison with other RP×HILIC approaches from the literature (Baglai et al., 2017; Holčapek et al., 2015b; Navarro-Reig et al., 2018; Xu et al., 2020).

Another benefit of the proposed method developed regarding lipid identification compared to the previous study on BPA exposure in zebrafish embryos (Martínez et al., 2020a) is the use of MS/MS which allows more confidence in the identification step. Most lipids were annotated at level 2 (MS/MS information, exact mass and retention time from both dimensions) according to the confidence level of compound annotation re-defined in the Compound Identification workgroup of the Metabolomics Society in 2017 (Blaženović et al., 2018).

Figure 2 shows two LC×LC chromatograms represented by 2D plots obtained with the optimized method. **Figure 2.A** displays an example of a mixture of nine lipid standards. As it is shown, the main lipid families are well distributed in the 2D space. Three regions of compounds are appreciated in the HILIC separation, corresponding to three main groups of lipids according to their polarity. The first group can be associated with barely retained compounds in HILIC ^{2}D, including ceramides, glucosylceramides, triacylglycerides, and diacylglycerides. The second group comprises phosphatidylethanolamines, and the third group is composed of phosphatidylcholines, sphingomyelins and lyso forms of the main glycerophospholipids (e.g., lysophosphatidylcholines, lysophosphatidylethanolamines, lysophosphoglycerols). **Figure 2.B** shows a 2D plot of a chromatogram measured for a control sample at 6 dpf. In the sample, there are very intense signals related to less retained compounds in HILIC (e.g., ceramides and triacylglycerides), but also from the third group according to **Figure 2.A** (e.g., phosphatidylcholines and sphingomyelins). Other less intense signals can be associated to the second group (e.g., phosphatidylethanolamines), or with lyso forms, at the very beginning of the ^{1}D separation (e.g., lysophosphatidylcholines).

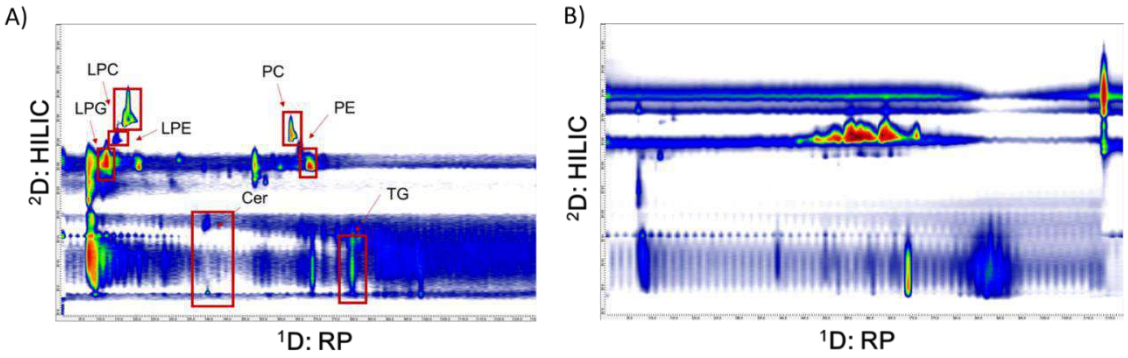


Figure 2. Examples of LC×LC chromatograms obtained for **A)** a lipid mixture of standards, and **B)** for a control sample at 6 dpf as collection day. The mixture of lipid standards (A) contained: 17:0 monoacylglycerol (MG), 17:0 lysophosphatidic acid (LPA), 17:1 phosphatidylethanolamine (LPE), 17:1 lysophosphoglycerol (LPG), 17:1 lysophosphatidylserine (LPS), 17:0 lysophosphatidylcholine (LPC), 1,3-17:0 D5 diacylglycerides (DG), 17:0 cholesteryl ester (CE), 16:0 D31-18:1 phosphatidylethanolamine (PE), 16:0 D31-18:1 phosphoglycerol (PG), 16:0 D31-18:1 phosphatidylcholine (PC), 16:0 D31-18:1 phosphatidylserine (PS) and 1,2,3-17:0 triacylglyceride (TG).

3.3 Multivariate statistical assessment and exploratory analysis

A first multivariate assessment of the different DOE factors was performed with an ASCA analysis. Three factors were evaluated: concentration of exposure (Control, 4, 18 and 26 μM for BPA; Control, 1, 4 and 8 μM for E2)), collection day (4 or 6 dpf) and EDC (BPA or E2). Both collection day and EDC factors were significant individually. Although the concentration was not significant considering the different doses of both EDC simultaneously, when analyzed separately, most of the combinations against control samples resulted significant (except 1 and 8 μM, and the interaction between all doses for E2). These results are summarized in **Table 2**. Regarding multiple factors at a time, no interaction was considered significant.

Table 2. ASCA results of statistical factors from the experimental design: concentration of exposure (Control, 4, 18 and 26 μM for BPA; Control, 1, 4 and 8 μM for E2)), collection day (4 or 6 dpf) and EDC (BPA or E2).

Comparison by Conc	**BPA 4 dpf**	**BPA 6 dpf**	**By factor**	
C - 26	0.0001	0.0001	Day	0.0001
C - 18	0.0001	0.0001	EDC	0.0001
C - 4	0.0001	0.0001	Conc	0.1554
C - 4 - 18 - 26	0.0007	0.0015	**By pairs of factors**	
Comparison by Conc	**E2 4 dpf**	**E2 6 dpf**	(Day) x (EDC)	1
C - 8	0.5981	0.0001	(Day) x (Conc)	1
C - 4	0.0001	0.0001	(EDC) x (Conc)	1
C - 1	0.306	0.0001		
C - 1 - 4 - 8	0.0934	0.0001		

An initial exploratory analysis of all sets and all features (no filtering applied based on their univariate statistical significance) was performed using PCA. When the four sets were plotted simultaneously (**see Figure 3.I**), the separation of the samples provided by the first principal component (23% of variance) was performed regarding the EDC (BPA samples on one side, and E2 samples on the opposite side), whereas the second principal component separated the samples according to the collection day (19% of variance). Therefore, the four sets clustered separately. However, a closer look at each set individually gave more information on the distribution by exposure concentration level for each EDC. **Figure 3.II** shows the PCA scores plot for each of the sets: **A)** BPA-4dpf, **B)** BPA-6dpf, **C)** E2-4dpf, **D)** E2-6dpf.

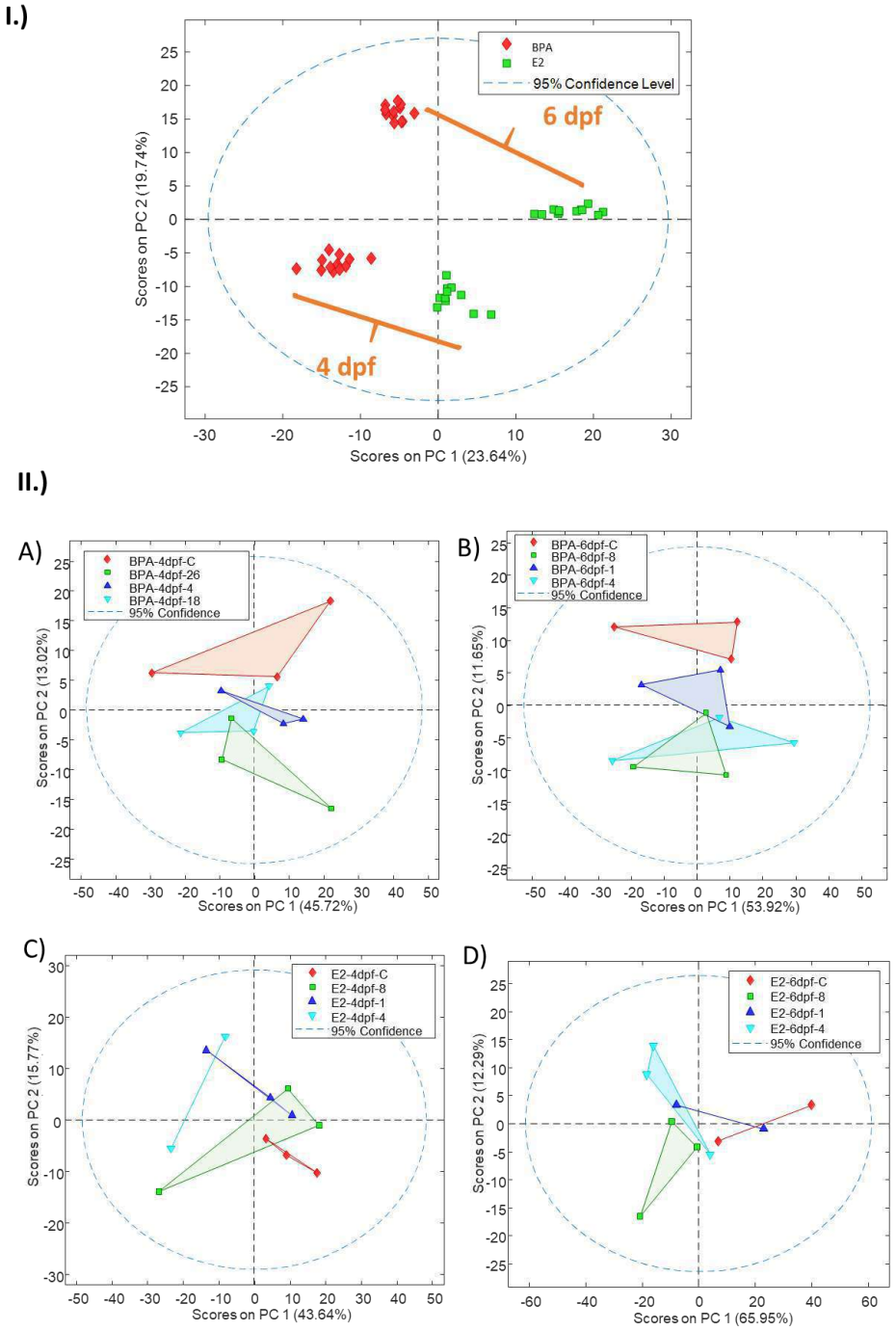


Figure 3. PCA score plots for **I.**) the four sets analyzed together, and **II.**) all sets separately, corresponding to A) BPA-4pdf, B) BPA-6dpf, C) E2-4 dpf, D) E2-6dpf.

In all cases, the explained variance of the first two principal components exceeded 50%. For BPA on both days, the second component separated the concentration levels in increasing order (C – 4 – 18 – 26 μM). A similar trend was also found for the E2-6 dpf set (C – 1 – 4 – 8 μM), but from the first component. However, in the case of E2-4 dpf the best separation was found for C vs 4 μM, instead of 8 μM. Therefore, the highest effect of E2 at 4 dpf was assigned to the medium

concentration, instead of the high dose, as could be expected. These results agreed with the previously obtained with ASCA. The previous trend was only observed for C – 1 – 4 μM, as the 8 μM clustered very close to control samples.

3.4 Selection of relevant features, lipid annotation and clustering analysis

ANOVA analysis resulted in 143 unique relevant features from the four datasets (BPA-4dpf, BPA-6dpf, E2-4dpf, E2-6pdf). These compounds were then annotated (using MS-DIAL spectral lipid library) with a confidence level 2 regarding annotation guidelines from the Metabolomics Society from 2017 (Blaženović et al., 2018), as information about MS/MS spectra, RT and accurate mass were included. Some exceptions were only annotated with RT and accurate mass due to a lack of MS/MS reference spectra (level 3 annotation). Information on lipid annotation and fold-changes can be found in **Supplementary Material**.

PAM analysis was performed on the relevant feature matrix. PAM provided five main clusters of lipids, as displayed in **Figure 4**. The first cluster (28 lipids) showed clear differences associated with the collection day regardless of the EDC. Lipids from this cluster were generally present at higher abundances at 4 dpf than at 6dpf. It is important to notice that natural differences between zebrafish eleutheroembryos occur at the two stages of growth studied. At 4 dpf lipids from yolk sac were not completely absorbed yet in non-exposed individuals, whereas this absorption was completed at 6 dpf (Martínez et al., 2020b). In addition to this tendency, this group also contained lipids for which the EDC exposure accelerated or slowed down the natural absorption, regarding control samples. The second, third, fourth and fifth clusters (44, 15, 33, 23 lipids, respectively) included lipids whose abundances were higher due to the following exposures respectively: BPA-6pdf, E2-6dpf, E2-4dpf and BPA-4dpf. Hence, lipids from clusters 3 and 4 were related to estrogenic alterations, whereas lipids from 2 and 5 could be associated with specific effects from BPA exposure at different stages of growth. **Table 3** shows the fold-changes values of the identified lipids at level 2 (with MS/MS matches), organized by cluster and by lipid family. **Table 3** shows that both EDC altered energy-related lipids (e.g., triacylglicerides (TG), lysophosphatidylcholines (LPC) and phosphatidylcholines (PC)), as previously reported (Martínez et al., 2020b). In addition, from this table it can also be observed that there are some lipids affected similarly for both EDCs. For instance, for both BPA and E2 exposures, some of the TGs from cluster 1 were more present in controls than in treated samples, with especial emphasis in the highest doses of exposure (i.e., these lipids were more consumed because of the treatment). This shared trend suggest that E2 may share some of the obesogenic effect of BPA, which difficults the characterization of the obesogenic effects of BPA.

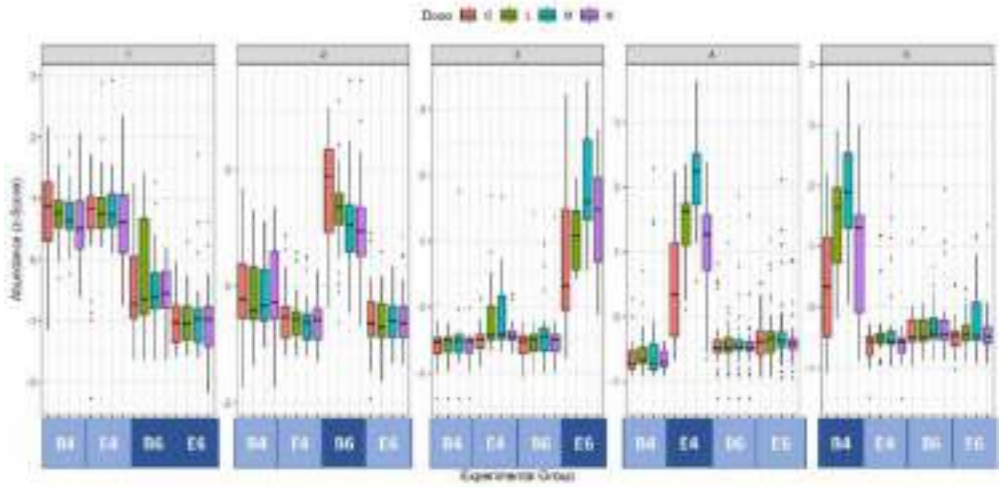


Figure 4. PAM analysis including five main cluster of lipids. A root-mean-scale transformation and an autoscaling normalization was performed. B4: BPA-4dpf, E4: E2-4dpf, B6: BPA-6dpf, E6: E2-6dpf.

Table 3. Fold-changes values of the significant lipids identified at level 2. Color bar: blue indicates higher values in control samples than in treated, whereas red is associated with lipids more present in treated samples than in controls. B4: BPA-4dpf, E4: E2-4dpf, B6: BPA-6dpf, E6: E2-6dpf, LPC: lysophosphatidylethanolamine, PC: phosphatidylcholine, PE: phosphatidylethanolamine, Cer: ceramide, TG: triacylglicerides, B4: BPA-4dpf, E4: E2-4dpf, B6: BPA-6dpf, E6: E2-6dpf, L: Low dose, M: Medium dose, H:High dose.

Cluster	Family	Lipid	B4_L	B4_M	B4_H	E4_L	E4_M	E4_H	B6_L	B6_M	B6_H	E6_L	E6_M	E6_H
1	Cer	Cer(34:1)	0.93	1.11	1.08	4.40	3.64	3.56	1.08	1.12	0.87	1.23	2.90	1.78
		LPC(16:0)	0.52	0.53	0.72	0.79	0.67	1.17	0.83	0.85	0.85	0.99	0.93	0.92
	PC	PC(36:6)	1.00	1.05	1.11	0.79	0.87	0.88	1.52	3.72	5.60	2.53	9.11	13.18
		PC(41:6)	1.28	1.35	1.87	2.08	3.82	0.96	1.00	1.00	1.00	0.99	0.93	0.92
		PC(42:6)	1.14	1.10	1.10	0.71	0.84	0.83	1.08	1.81	3.99	1.01	1.49	1.58
	TG	TG(53:2)	0.78	0.66	0.60	0.90	0.97	0.96	1.00	1.81	1.00	0.99	0.93	0.92
		TG(58:12)	0.86	0.73	0.70	0.77	0.75	0.75	1.00	1.00	1.00	0.99	0.93	0.92
		TG(58:6)	0.87	0.61	0.57	0.99	0.93	0.84	1.12	0.49	0.55	0.99	0.93	0.92
		TG(60:10)	0.88	0.74	0.67	0.93	0.90	0.75	1.20	0.64	0.75	0.45	0.35	0.00
		TG(62:12)	0.86	0.74	0.66	0.97	0.95	0.81	1.18	0.75	0.85	0.90	0.74	0.32
2	Cer	Cer(42:2)	1.05	0.96	1.03	0.91	0.84	0.75	0.71	0.54	0.65	0.99	0.93	0.92
	PC	LPC(18:2)	1.09	1.12	1.14	1.57	0.93	0.96	0.44	0.73	0.83	0.99	0.93	0.92
		LPC(18:0)	1.05	0.96	1.03	0.97	0.93	0.96	2.10	8.66	24.07	0.99	0.93	0.92
	TG	TG(62:13)	0.87	0.74	0.65	0.97	0.93	0.96	1.13	0.73	0.82	0.35	0.15	0.33
3	PE	PE(O-40:8)	1.05	0.96	1.03	4.78	3.69	0.96	1.00	1.00	1.00	0.94	0.86	1.14
	PC	PC(35:1)	1.05	0.96	1.03	6.53	6.49	2.26	1.00	1.00	1.00	1.08	1.30	1.36
		PC(36:3)	1.05	0.96	1.03	0.97	0.93	0.96	1.00	1.00	1.00	2.83	4.18	2.37
		PC(38:2)	1.05	0.96	1.03	0.97	2.16	0.96	1.00	1.00	1.00	2.23	2.90	2.83
		PC(42:8)	1.05	0.96	1.03	0.97	0.93	0.96	1.00	1.00	1.00	1.21	2.68	3.08
	TG	TG(57:7)	0.73	0.47	0.44	0.98	0.93	0.83	1.00	1.00	1.00	0.64	0.76	0.78
4	PC	PC(36:2)	1.05	0.96	1.03	2.86	5.02	2.62	1.00	1.00	1.00	0.99	0.93	0.92
		PC(32:1)	1.72	3.17	1.03	2.28	2.43	0.96	1.00	1.00	1.00	0.99	0.93	0.92
5	PC	PC(38:5)	4.50	4.35	4.17	0.97	0.93	0.96	1.00	1.00	1.00	0.99	0.93	0.92
	TG	TG(50:2)	0.99	0.94	1.04	1.27	2.09	0.96	1.00	1.00	1.00	0.99	0.93	0.92

4. Conclusions

The BPA impact was characterized from a lipidomic point of view by comparing the caused alterations with an estrogenic control, E2. An RP×HILIC-HRMS method was developed for the analysis of the zebrafish eleutheroembyros lipidome at early growth stages. The use of ASM enhanced sensitivity and solvent compatibility between mobile phases while reducing the total analysis time. Besides, the MS/MS information obtained contributed to a more confident lipid annotation.

Regarding the EDC exposure, the lowest observed effect concentrations (LOEC) were 26 μM and 8 μM for BPA and E2, respectively, because no statistical significance was found for survival and hatching during the whole stage of growth (up to 6 dpf). The statistical assessment of the different factors from the experimental design (dose of exposure, collection day and EDC) showed that all tested doses were significant for both EDC at 6 dpf, but only for BPA at 4 dpf. These results were in agreement with exploratory analyses, where a clear differentiation of control and treated samples was observed for BPA at both days, and for E2 at 6dpf, but not at 4 dpf.

The significant lipids obtained from a multiple dose comparisons (for each day and EDC) were selected for further analysis and biological interpretation. These lipids were clustered, each of them associated with the exposure of an EDC on a certain day, plus an extra cluster that collects the differences between both collection days. The altered lipids for both EDCs seemed to be linked to the estrogenic effect of BPA. However, an obesogenic effect was also found for E2 exposure, which difficulted the characterization of the non-estrogenic effect caused by BPA.

Acknowledgements

The authors would like to thank Rubén Martínez, Marta Casado, Claudia Sanz Lanzas, Olga Burgos and Marica Erminia Schiano for the help with the zebrafish experiments. In addition, the 1290 LC system and 6545 XT QTOF instrumentation were provided to DS as gifts by Agilent Technologies through their Thought Leader program.

Funding

The research leading to these results has received funding from grants CTQ2017-82598-P and CEX2018-000794-S funded by the Spanish Ministry of Science and Innovation (MCIN/AEI/10.13039/501100011033). The authors also want to grant support from the Catalan Agency for Management of University and Research Grants (AGAUR, Grant 2017SGR753).

MPC acknowledges a predoctoral FPU 16/02640 scholarship from the Spanish Ministry of Education and Vocational Training (MEFP). LNM was supported by grant RyC2019-026426-I, funded by the Spanish Ministry of Science and Innovation (MCIN/AEI/10.13039/501100011033).

Conflict of interests

The authors declare no conflict of interest.

References

Abraham, A., Chakraborty, P., 2019. A review on sources and health impacts of bisphenol A. Reviews on Environmental Health 35, 201–210. https://doi.org/10.1515/reveh-2019-0034

Akhbarizadeh, R., Russo, G., Rossi, S., Golianova, K., Moore, F., Guida, M., de Falco, M., Grumetto, L., 2021. Emerging endocrine disruptors in two edible fish from the Persian Gulf: Occurrence, congener profile, and human health risk assessment. Marine Pollution Bulletin 166, 112241. https://doi.org/10.1016/j.marpolbul.2021.112241

Baglai, A., Gargano, A.F.G., Jordens, J., Mengerink, Y., Honing, M., van der Wal, S., Schoenmakers, P.J., 2017. Comprehensive lipidomic analysis of human plasma using multidimensional liquid- and gas-phase separations: Two-dimensional liquid chromatography–mass spectrometry vs. liquid chromatography–trapped-ion-mobility–mass spectrometry. Journal of Chromatography A 1530, 90–103. https://doi.org/10.1016/j.chroma.2017.11.014

Banaderakhshan, R., Kemp, P., Breul, L., Steinbichl, P., Hartmann, C., Fürhacker, M., 2022. Bisphenol A and its alternatives in Austrian thermal paper receipts, and the migration from reusable plastic drinking bottles into water and artificial saliva using UHPLC-MS/MS. Chemosphere 286, 131842. https://doi.org/10.1016/j.chemosphere.2021.131842

Ben-Jonathan, N., Steinmetz, R., 1998. Xenoestrogens: The Emerging Story of Bisphenol A. Trends in Endocrinology & Metabolism 9, 124–128. https://doi.org/10.1016/S1043-2760(98)00029-0

Bertinetto, C., Engel, J., Jansen, J., 2020. ANOVA simultaneous component analysis: A tutorial review. Analytica Chimica Acta: X 6, 100061. https://doi.org/10.1016/j.acax.2020.100061

Bisphenol A (BPA) Market Size, Share, Industry Report 2030 | ChemAnalyst [WWW Document], n.d. URL https://www.chemanalyst.com/industry-report/bisphenol-a-market-57 (accessed 4.26.22).

Blaženović, I., Kind, T., Ji, J., Fiehn, O., 2018. Software tools and approaches for compound identification of LC-MS/MS data in metabolomics. Metabolites 8. https://doi.org/10.3390/metabo8020031

Caban, M., Stepnowski, P., 2020. Determination of bisphenol A in size fractions of indoor dust from several microenvironments. Microchemical Journal 153, 104392. https://doi.org/10.1016/j.microc.2019.104392

Cajka, T., Fiehn, O., 2014. Comprehensive analysis of lipids in biological systems by liquid chromatography-mass spectrometry. TrAC - Trends in Analytical Chemistry 61, 192–206. https://doi.org/10.1016/j.trac.2014.04.017

Cano-Nicolau, J., Vaillant, C., Pellegrini, E., Charlier, T.D., Kah, O., Coumailleau, P., 2016. Estrogenic effects of several BPA analogs in the developing zebrafish brain. Frontiers in Neuroscience 10, 1–14. https://doi.org/10.3389/fnins.2016.00112

Chakraborty, P., Shappell, N.W., Mukhopadhyay, M., Onanong, S., Rex, K.R., Snow, D., 2021. Surveillance of plasticizers, bisphenol A, steroids and caffeine in surface water of River Ganga and Sundarban wetland along the Bay of Bengal: occurrence, sources, estrogenicity screening and

ecotoxicological risk assessment. Water Research 190, 116668. https://doi.org/10.1016/J.WATRES.2020.116668

Chambers, M.C., Maclean, B., Burke, R., Amode, D., Ruderman, D.L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T.A., Brusniak, M.-Y., Paulse, C., Creasy, D., Flashner, L., Kani, K., Moulding, C., Seymour, S.L., Nuwaysir, L.M., Lefebvre, B., Kuhlmann, F., Roark, J., Rainer, P., Detlev, S., Hemenway, T., Huhmer, A., Langridge, J., Connolly, B., Chadick, T., Holly, K., Eckels, J., Deutsch, E.W., Moritz, R.L., Katz, J.E., Agus, D.B., MacCoss, M., Tabb, D.L., Mallick, P., 2012. A cross-platform toolkit for mass spectrometry and proteomics. Nature Biotechnology. https://doi.org/10.1038/nbt.2377

Chen, M.Y., Ike, M., Fujita, M., 2002. Acute toxicity, mutagenicity, and estrogenicity of bisphenol-A and other bisphenols. Environmental Toxicology 17, 80–86. https://doi.org/10.1002/tox.10035

Cífková, E., Hájek, R., Lísa, M., Holčapek, M., 2016. Hydrophilic interaction liquid chromatography-mass spectrometry of (lyso)phosphatidic acids, (lyso)phosphatidylserines and other lipid classes. Journal of Chromatography A 1439, 65–73. https://doi.org/10.1016/j.chroma.2016.01.064

COMMISSION REGULATION (EU) 2018/213 [WWW Document], n.d. URL https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018R0213&from=EL (accessed 4.13.22).

Dalmau, N., Jaumot, J., Tauler, R., Bedia, C., 2015. Epithelial-to-mesenchymal transition involves triacylglycerol accumulation in DU145 prostate cancer cells. Mol. BioSyst. 11, 3397–3406. https://doi.org/10.1039/C5MB00413F

ECHA - BPA [WWW Document], n.d. URL https://echa.europa.eu/es/substance-information/-/substanceinfo/100.001.133 (accessed 4.13.22).

EPA - Risk Management for Bisphenol A (BPA) [WWW Document], n.d. URL https://www.epa.gov/assessing-and-managing-chemicals-under-tsca/risk-management-bisphenol-bpa (accessed 4.13.22).

Eramo, S., Urbani, G., Sfasciotti, G.L., Brugnoletti, O., Bossù, M., Polimeni, A., 2010. Estrogenicity of bisphenol A released from sealants and composites: a review of the literature. Ann Stomatol (Roma) 1, 14–21.

Fahy, E., Subramaniam, S., Murphy, R.C., Nishijima, M., Raetz, C.R.H., Shimizu, T., Spener, F., van Meer, G., Wakelam, M.J.O., Dennis, E.A., 2009. Update of the LIPID MAPS comprehensive classification system for lipids. Journal of Lipid Research 50, 9–14. https://doi.org/10.1194/jlr.R800095-JLR200

Fahy, E., Sud, M., Cotter, D., Subramaniam, S., 2007. LIPID MAPS online tools for lipid research. Nucleic Acids Research 35. https://doi.org/10.1093/nar/gkm324

Farooq, M.U., Jalees, M.I., Qurat-ul-Ain, Hussain, G., Anis, M., Islam, U., 2021. Health risk assessment of endocrine disruptor bisphenol A leaching from plastic bottles of milk and soft drinks. Environmental Science and Pollution Research 28, 57090–57098. https://doi.org/10.1007/s11356-021-14653-4

Fraher, D., Sanigorski, A., Mellett, N.A., Meikle, P.J., Sinclair, A.J., Gibert, Y., 2016. Zebrafish Embryonic Lipidomic Analysis Reveals that the Yolk Cell Is Metabolically Active in Processing Lipid. Cell Reports 14, 1317–1329. https://doi.org/10.1016/j.celrep.2016.01.016

Fuertes, I., Jordão, R., Casas, F., Barata, C., 2018. Allocation of glycerolipids and glycerophospholipids from adults to eggs in Daphnia magna: Perturbations by compounds that enhance lipid droplet accumulation. Environmental Pollution 242, 1702–1710. https://doi.org/10.1016/j.envpol.2018.07.102

Gear, R.B., Belcher, S.M., 2017. Impacts of Bisphenol A and Ethinyl Estradiol on Male and Female CD-1 Mouse Spleen. Scientific Reports 7, 1–12. https://doi.org/10.1038/s41598-017-00961-8

Gorrochategui, E., Jaumot, J., Tauler, R., 2019. ROIMCR: A powerful analysis strategy for LC-MS metabolomic datasets. BMC Bioinformatics 20, 1–17. https://doi.org/10.1186/s12859-019-2848-8

Gould, J.C., Leonard, L.S., Maness, S.C., Wagner, B.L., Conner, K., Zacharewski, T., Safe, S., McDonnell, D.P., Gaido, K.W., 1998. Bisphenol A interacts with the estrogen receptor α in a distinct manner from estradiol. Molecular and Cellular Endocrinology 142, 203–214. https://doi.org/10.1016/S0303-7207(98)00084-7

Gregory R. Warnes, Ben Bolker, Lodewijk Bonebakker, Robert Gentleman, W., Huber, Andy Liaw, Thomas Lumley, Martin Maechler, Arni Magnusson, S.M., Venables, M.S. and B., 2020. gplots: Various R Programming Tools for Plotting Data. R package version 3.1.1. [WWW Document]. URL https://cran.r-project.org/package=gplots

Hadley Wickham, Maximilian Girlich, Rs., n.d. tidyr: Tidy Messy Data. R package version 1.1.3. [WWW Document]. 2016. URL https://cran.r-project.org/package=tidyr (accessed 4.24.22).

Heindel, J.J., Blumberg, B., 2019. Environmental obesogens: Mechanisms and controversies. Annual Review of Pharmacology and Toxicology 59, 89–106. https://doi.org/10.1146/annurev-pharmtox-010818-021304

Hennig, C., 2020. fpc: Flexible Procedures for Clustering. R package version 2.2-9. [WWW Document]. URL https://cran.r-project.org/package=fpc

Holčapek, M., Ovčačíková, M., Lísa, M., Cífková, E., Hájek, T., 2015a. Continuous comprehensive two-dimensional liquid chromatography-electrospray ionization mass spectrometry of complex lipidomic samples. Anal Bioanal Chem 407, 5033–5043. https://doi.org/10.1007/s00216-015-8528-2

Holčapek, M., Ovčačíková, M., Lísa, M., Cífková, E., Hájek, T., 2015b. Continuous comprehensive two-dimensional liquid chromatography-electrospray ionization mass spectrometry of complex lipidomic samples. Anal Bioanal Chem 407, 5033–5043. https://doi.org/10.1007/s00216-015-8528-2

Huang, W., Zheng, S., Wang, X., Cai, Z., Xiao, J., Liu, C., Wu, K., 2020. A transcriptomics-based analysis of toxicity mechanisms of zebrafish embryos and larvae following parental Bisphenol A exposure. Ecotoxicology and Environmental Safety 205, 111165. https://doi.org/10.1016/j.ecoenv.2020.111165

Ishido, M., Masuo, Y., Kunimoto, M., Oka, S., Morita, M., 2004. Bisphenol A Causes Hyperactivity in the Rat Concomitantly with Impairment of Tyrosine Hydroxylase Immunoreactivity. Journal of Neuroscience Research 76, 423–433. https://doi.org/10.1002/JNR.20050

Jordão, R., Garreta, E., Campos, B., Lemos, M.F.L., Soares, A.M.V.M., Tauler, R., Barata, C., 2016. Compounds altering fat storage in Daphnia magna. Sci Total Environ 545–546, 127–136. https://doi.org/10.1016/J.SCITOTENV.2015.12.097

Jürgens, M.D., Holthaus, K.I.E., Johnson, A.C., Smith, J.J.L., Hetheridge, M., Williams, R.J., 2002. The potential for estradiol and ethinylestradiol degradation in English rivers. Environmental Toxicology and Chemistry 21, 480–488. https://doi.org/10.1002/etc.5620210302

Kassambara, A., n.d. ggpubr: ggplot2 Based Publication Ready Plots [WWW Document]. 2020. URL https://rpkgs.datanovia.com/ggpubr/ (accessed 4.26.22).

Kovačič, A., Gys, C., Gulin, M.R., Kosjek, T., Heath, D., Covaci, A., Heath, E., 2020. The migration of bisphenols from beverage cans and reusable sports bottles. Food Chemistry 331, 127326. https://doi.org/10.1016/J.FOODCHEM.2020.127326

Lee, Byoung cheun, Yoon, H., Lee, Byeongwoo, Kim, P., Moon, H.B., Kim, Y., 2021. Occurrence of bisphenols and phthalates in indoor dust collected from Korean homes. Journal of Industrial and Engineering Chemistry 99, 68–73. https://doi.org/10.1016/j.jiec.2021.03.051

Longo, M., Zatterale, F., Naderi, J., Nigro, C., Oriente, F., Formisano, P., Miele, C., Beguinot, F., 2020. Low-dose bisphenol-a promotes epigenetic changes at pparγ promoter in adipose precursor cells. Nutrients 12, 1–23. https://doi.org/10.3390/nu12113498

Lv, W., Shi, X., Wang, S., Xu, G., 2019. Multidimensional liquid chromatography-mass spectrometry for metabolomic and lipidomic analyses. TrAC - Trends in Analytical Chemistry 120, 115302. https://doi.org/10.1016/j.trac.2018.11.001

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., H., 2021. Cluster Analysis Basics and Extensions. R package version 2.1.2. [WWW Document]. URL https://cran.r-project.org/package=cluster. (accessed 4.26.22).

Marqueño, A., Pérez-Albaladejo, E., Denslow, N.D., Bowden, J.A., Porte, C., 2021. Untargeted lipidomics reveals the toxicity of bisphenol A bis(3-chloro-2- hydroxypropyl) ether and bisphenols A and F in

zebrafish liver cells. Ecotoxicology and Environmental Safety 219. https://doi.org/10.1016/j.ecoenv.2021.112311

Martínez, R., Esteve-Codina, A., Herrero-Nogareda, L., Ortiz-Villanueva, E., Barata, C., Tauler, R., Raldúa, D., Piña, B., Navarro-Martín, L., 2018. Dose-dependent transcriptomic responses of zebrafish eleutheroembryos to Bisphenol A. Environmental Pollution 243, 988–997. https://doi.org/10.1016/j.envpol.2018.09.043

Martínez, R., Herrero-Nogareda, L., van Antro, M., Campos, M.P., Casado, M., Barata, C., Piña, B., Navarro-Martín, L., 2019. Morphometric signatures of exposure to endocrine disrupting chemicals in zebrafish eleutheroembryos. Aquatic Toxicology 214, 105232. https://doi.org/10.1016/j.aquatox.2019.105232

Martínez, R., Navarro-Martín, L., van Antro, M., Fuertes, I., Casado, M., Barata, C., Piña, B., 2020a. Changes in lipid profiles induced by bisphenol A (BPA) in zebrafish eleutheroembryos during the yolk sac absorption stage. Chemosphere 246. https://doi.org/10.1016/j.chemosphere.2019.125704

Martínez, R., Navarro-Martín, L., van Antro, M., Fuertes, I., Casado, M., Barata, C., Piña, B., 2020b. Changes in lipid profiles induced by bisphenol A (BPA) in zebrafish eleutheroembryos during the yolk sac absorption stage. Chemosphere 246. https://doi.org/10.1016/j.chemosphere.2019.125704

Martínez, R., Tu, W., Eng, T., Allaire-Leung, M., Piña, B., Navarro-Martín, L., Mennigen, J.A., 2020c. Acute and long-term metabolic consequences of early developmental Bisphenol A exposure in zebrafish (Danio rerio). Chemosphere 256. https://doi.org/10.1016/j.chemosphere.2020.127080

Moon, M.K., 2019. Moon M.K., Concern about the safety of bisphenol A substitutes, Diabetes and Metabolism Journal, 2019, 43: 46-48 46–48.

Mu, X., Huang, Y., Li, Xuxing, Lei, Y., Teng, M., Li, Xuefeng, Wang, C., Li, Y., 2018. Developmental Effects and Estrogenicity of Bisphenol A Alternatives in a Zebrafish Embryo Model. Environmental Science and Technology 52, 3222–3231. https://doi.org/10.1021/acs.est.7b06255

Navarro-Reig, M., Jaumot, J., Tauler, R., 2018. An untargeted lipidomic strategy combining comprehensive two-dimensional liquid chromatography and chemometric analysis. Journal of Chromatography A 1568, 80–90. https://doi.org/10.1016/j.chroma.2018.07.017

Nguyen, H.T., Li, L., Eguchi, A., Kannan, K., Kim, E.Y., Iwata, H., 2021. Effects on the liver lipidome of rat offspring prenatally exposed to bisphenol A. Science of the Total Environment 759, 143466. https://doi.org/10.1016/j.scitotenv.2020.143466

Ortiz-Villanueva, E., Jaumot, J., Martínez, R., Navarro-Martín, L., Piña, B., Tauler, R., 2018. Assessment of endocrine disruptors effects on zebrafish (Danio rerio) embryos by untargeted LC-HRMS metabolomic analysis. Science of the Total Environment 635, 156–166. https://doi.org/10.1016/j.scitotenv.2018.03.369

Ortiz-Villanueva, E., Navarro-Martín, L., Jaumot, J., Benavente, F., Sanz-Nebot, V., Piña, B., Tauler, R., 2017. Metabolic disruption of zebrafish (Danio rerio) embryos by bisphenol A. An integrated metabolomic and transcriptomic approach. Environmental Pollution 231, 22–36. https://doi.org/10.1016/j.envpol.2017.07.095

Paglia, G., Smith, A.J., Astarita, G., 2021. Ion mobility mass spectrometry in the omics era: Challenges and opportunities for metabolomics and lipidomics. Mass Spectrometry Reviews. https://doi.org/10.1002/mas.21686

Paris, F., Balaguer, P., Térouanne, B., Servant, N., Lacoste, C., Cravedi, J.P., Nicolas, J.C., Sultan, C., 2002. Phenylphenols, biphenols, bisphenol-A and 4-tert-octylphenol exhibit α and β estrogen activities and antiandrogen activity in reporter cell lines. Molecular and Cellular Endocrinology 193, 43–49. https://doi.org/10.1016/S0303-7207(02)00094-1

Pérez-Bermejo, M., Mas-Pérez, I., Murillo-Llorente, M.T., 2021. The role of the bisphenol a in diabetes and obesity. Biomedicines 9, 1–17. https://doi.org/10.3390/biomedicines9060666

Pérez-Cova, M., Bedia, C., Stoll, D.R., Tauler, R., Jaumot, J., 2021. MSroi: A pre-processing tool for mass spectrometry-based studies. Chemometrics and Intelligent Laboratory Systems 215. https://doi.org/10.1016/j.chemolab.2021.104333

Pop, C.E., Draga, S., Măciucă, R., Niță, R., Crăciun, N., Wolff, R., 2021. Bisphenol A effects in aqueous environment on lemna minor. Processes 9. https://doi.org/10.3390/pr9091512

R Foundation for Statistical Computing, Vienna, A., n.d. R Core Team (2020). R: A language and environment for statistical computing. [WWW Document]. 2020. URL (accessed 4.26.22).

RStudio, Inc., B., MA, n.d. RStudio Team (2018). RStudio: Integrated Development for R. [WWW Document]. 2018. URL http://www.rstudio.com/. (accessed 4.26.22).

Santangeli, S., Notarstefano, V., Maradonna, F., Giorgini, E., Gioacchini, G., Forner-Piquer, I., Habibi, H.R., Carnevali, O., 2018. Effects of diethylene glycol dibenzoate and Bisphenol A on the lipid metabolism of Danio rerio. Science of the Total Environment 636, 641–655. https://doi.org/10.1016/j.scitotenv.2018.04.291

Sawai, C., Anderson, K., Walser-Kuntz, D., 2003. Effect of bisphenol A on murine immune function: modulation of interferon-gamma, IgG2a, and disease symptoms in NZB X NZW F1 mice. Environmental Health Perspectives 111, 1883–1887. https://doi.org/10.1289/EHP.6359

Silva, C.C.V., Jaddoe, V.W.V., Sol, C.M., el Marroun, H., Martinez-Moral, M.P., Kannan, K., Trasande, L., Santos, S., 2021. Phthalate and Bisphenol Urinary Concentrations, Body Fat Measures, and Cardiovascular Risk Factors in Dutch School-Age Children. Obesity 29, 409–417. https://doi.org/10.1002/oby.23082

Smilde, A.K., Jansen, J.J., Hoefsloot, H.C.J., Lamers, R.J.A.N., van der Greef, J., Timmerman, M.E., 2005. ANOVA-simultaneous component analysis (ASCA): A new tool for analyzing designed metabolomics data. Bioinformatics 21, 3043–3048. https://doi.org/10.1093/bioinformatics/bti476

Stoll, D.R., Shoykhet, K., Petersson, P., Buckenmaier, S., 2017. Active Solvent Modulation: A Valve-Based Approach to Improve Separation Compatibility in Two-Dimensional Liquid Chromatography. Analytical Chemistry 89, 9260–9267. https://doi.org/10.1021/acs.analchem.7b02046

Sun, Q., Wang, Y., Li, Y., Ashfaq, M., Dai, L., Xie, X., Yu, C.P., 2017. Fate and mass balance of bisphenol analogues in wastewater treatment plants in Xiamen City, China. Environmental Pollution 225, 542–549. https://doi.org/10.1016/J.ENVPOL.2017.03.018

Tao, H. yu, Zhang, J., Shi, J., Guo, W., Liu, X., Zhang, M., Ge, H., Li, X. yan, 2021. Occurrence and emission of phthalates, bisphenol A, and oestrogenic compounds in concentrated animal feeding operations in Southern China. Ecotoxicology and Environmental Safety 207, 111521. https://doi.org/10.1016/j.ecoenv.2020.111521

Tian, S., Yan, S., Meng, Z., Huang, S., Sun, W., Jia, M., Teng, M., Zhou, Z., Zhu, W., 2021. New insights into bisphenols induced obesity in zebrafish (Danio rerio): Activation of cannabinoid receptor CB1. Journal of Hazardous Materials 418, 126100. https://doi.org/10.1016/j.jhazmat.2021.126100

Tkalec, Ž., Kosjek, T., Snoj Tratnik, J., Stajnko, A., Runkel, A.A., Sykiotou, M., Mazej, D., Horvat, M., 2021. Exposure of Slovenian children and adolescents to bisphenols, parabens and triclosan: Urinary levels, exposure patterns, determinants of exposure and susceptibility. Environment International 146. https://doi.org/10.1016/j.envint.2020.106172

Tohmé, M., Prud'Homme, S.M., Boulahtouf, A., Samarut, E., Brunet, F., Bernard, L., Bourguet, W., Gibert, Y., Balaguer, P., Laudet, V., 2014. Estrogen-related receptor γ is an in vivo receptor of bisphenol A. FASEB Journal 28, 3124–3133. https://doi.org/10.1096/fj.13-240465

Tsugawa, H., Ikeda, K., Takahashi, M., Satoh, A., Mori, Y., Uchino, H., Okahashi, N., Yamada, Y., Tada, I., Bonini, P., Higashi, Y., Okazaki, Y., Zhou, Z., Zhu, Z.J., Koelmel, J., Cajka, T., Fiehn, O., Saito, K., Arita, Masanori, Arita, Makoto, 2020. A lipidome atlas in MS-DIAL 4. Nature Biotechnology 38, 1159–1163. https://doi.org/10.1038/s41587-020-0531-2

Vasiljevic, T., Harner, T., 2021. Bisphenol A and its analogues in outdoor and indoor air: Properties, sources and global levels. Science of the Total Environment 789, 148013. https://doi.org/10.1016/j.scitotenv.2021.148013

Vidal, R., Ma, Y., Sastry, S.S., 2016. Principal component analysis. Interdisciplinary Applied Mathematics 40, 25–62. https://doi.org/10.1007/978-0-387-87811-9_2

Wang, H., Liu, Z. hua, Zhang, J., Huang, R. ping, Yin, H., Dang, Z., Wu, P. xiao, Liu, Y., 2019. Insights into removal mechanisms of bisphenol A and its analogues in municipal wastewater treatment plants. Science of the Total Environment 692, 107–116. https://doi.org/10.1016/j.scitotenv.2019.07.134

Wang, H., Liu, Z. hua, Zhang, J., Huang, R.P., Yin, H., Dang, Z., 2020. Human exposure of bisphenol A and its analogues: understandings from human urinary excretion data and wastewater-based epidemiology. Environmental Science and Pollution Research 27, 3247–3256. https://doi.org/10.1007/s11356-019-07111-9

Wickham, H., 2017. ggplot2 - Elegant Graphics for Data Analysis (2nd Edition). Journal of Statistical Software 77, XVI–260.

Xu, M., Legradi, J., Leonards, P., 2020. Evaluation of LC-MS and LC×LC-MS in analysis of zebrafish embryo samples for comprehensive lipid profiling. Analytical and Bioanalytical Chemistry 412, 4313–4325. https://doi.org/10.1007/s00216-020-02661-1

Yamaguchi, H., Zhu, J., Yu, T., Sasaki, K., Umetsu, H., Kidachi, Y., Ryoyama, K., 2006. Low-level bisphenol A increases production of glial fibrillary acidic protein in differentiating astrocyte progenitor cells through excessive STAT3 and Smad1 activation. Toxicology 226, 131–142. https://doi.org/10.1016/J.TOX.2006.06.011

Zhao, X., Chen, J., Zhang, W., Yang, C., Ma, X., Zhang, S., Zhang, X., 2019. Lipid Alterations during Zebrafish Embryogenesis Revealed by Dynamic Mass Spectrometry Profiling with C=C Specificity. J Am Soc Mass Spectrom 30, 2646–2654. https://doi.org/10.1007/s13361-019-02334-z

Supplementary Material A

Untargeted lipidomics of zebrafish (*Danio rerio*) eleutheroembryos exposed to endocrine disrupting chemicals using comprehensive two-dimensional liquid chromatography and advanced chemometrics

Miriam Pérez-Cova[a,b,c], Laia Navarro-Martin[a], Gabriel Leme[c], Romà Tauler[a], Benjamin Piña[a], Joaquim Jaumot[a], Dwight R. Stoll[c]*

[a]Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, 08034 Barcelona, Spain

[b]Department of Chemical Engineering and Analytical Chemistry, University of Barcelona, Diagonal 647, 08028, Barcelona, Spain.

[c]Department of Chemistry, Gustavus Adolphus College, Saint Peter, Minnesota 56082, United States

* Correspondence: joaquim.jaumot@idaea.csic.es

1. Zebrafish eleutheroembryos ranging tests for BPA and E2 exposure

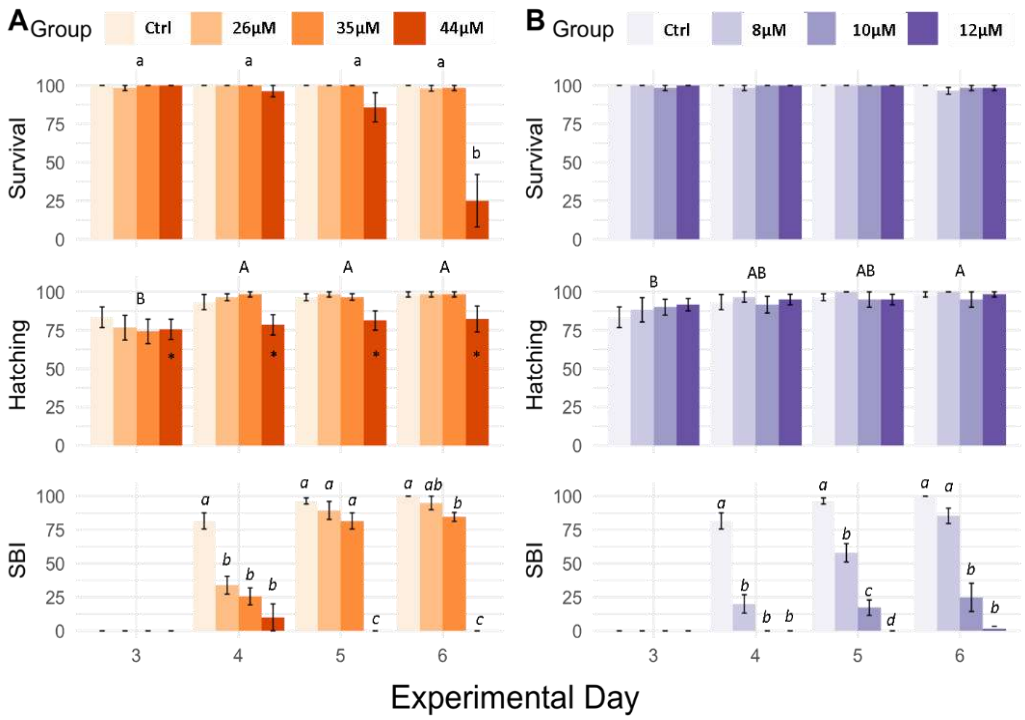


Figure S1. Effects of increasing concentrations of **(A)** BPA (26-44 µM) and **(B)** E2 (8-12 µM) on survival, hatching and swim bladder inflation (SBI) rates at 3, 4, 5 and 6 dpf. The mean value ± SD (standard deviation) is shown for each group (n=6 groups of 10 individuals each). Non-parametric test (Kruskal-Wallis with Wilcoxon rank sum test for multiple comparisons, *p*-value < 0.05) was performed. Lower cap letters denote significant differences between all possible comparisons (within day and dose). Capital letters denote significant differences between days of development regardless of the treatment. Italic letters denote differences between concentrations within each developmental day.

2. Determination of real concentrations of exposure

BPA and E2 real concentrations of exposure were determined by a LC-MS/MS method (Silcock et al., 2013) with minor modifications. A calibration curve for each endocrine disrupting chemical was prepared with the proper dilutions of the concentrated stocks (BPA: 0.44-70 µM; E2: 0.1-22 µM).

Chromatographic separations were carried on an Acquity UHPLC system (Waters, Milford, MA, US), using an Acquity BEH C18 (2.1 x 50 mm; 1.7 µM) from Waters (Milford, MA, US), at 40 ˚C. Mobile phases employed were: A) MeOH, B) 0.05% NH_4OH (aqueous). Elution gradient started at 35% of A, increased until 65% A in 2.5 min, reached 70% at 3.5 min, returned to initial

conditions at 3.6 min, and held until 5 min. Flow rate was set at 0.4 mL/min, Injection volume was 10 µL, and the temperature of the autosampler was 10 ˚C.

A Xevo TQS, Acquity Waters (Mildford, USA) was employed as detector for LC-MS/MS analysis, in ESI negative mode. MS conditions were capillary voltage 2 kV, desolvation temperature 500 ˚C, desolvation gas flow 1000 L/hour, cone gas flow 150 L/ hour. Full scan mass range was set from 50 to 300 Da, with a scan time of 0.3. MS/MS analysis was performed in MRM mode, with the following parameters:

Compound	Nominal mass	Precursor ion	Product ion	Cone (V)	Collision energy (eV)
BPA	228	227.1	133	31	25
	228	227.1	212	31	17
E2	272	271.2	183.1	51	35
	272	271.2	145.1	51	40

3. Regions of interest (ROI) strategy for spectral compression

ROI is a spectral compression strategy based on the selection of *m/z* values with intensities higher than a certain signal-to-noise ratio (SN threshold) among the different chromatograms of each dataset. ROI also considers a mass error tolerance, related to the mass accuracy of the mass spectrometer, and a minimum number of occurrences, required for defining a chromatographic peak. A factor can also be set to establish an intensity threshold low, but only considering the features whose intensities are a multiple of this factor (e.g. min max 2, means features kept have intensities at least twice the SN threshold). ROI *m/z* values are searched for each retention time, and the final value will be the mean (or the median) of all the values corresponding to the same chromatographic peak. If an *m/z* value is detected for some samples but others no, then non-present ROIs will be set to a low random intensity value at the noise level. With this strategy, the original *m/z* vector is reduced for all samples simultaneously; the new vector is composed of discrete *m/z* values. In this work, ROI strategy was applied through the MSroi app (Pérez-Cova et al., 2021a). ROI parameters were set to 400-1500 (*m/z* range), 6000 (threshold), 0.1 Da/e (mass accuracy), 5 (minimum occurrences) and *m/z* final values calculated by the median of the values for each chromatographic peak.

The main outputs of the ROI procedure are: A) a column-wise data matrix containing the intensities of the *m/z* values selected for each of the retention times from both dimensions, and B) a vector with the actual *m/z* values of relevance. The ROI augmented data matrices dimensions were, in the rows, the total number of elution times considered in the whole set of samples and

for both dimensions, and in the columns, the total number of relevant *m/z* values of each ROI. Due to the huge size of the LC×LC datasets (1Gb per sample), ROI analysis was performed separately to eight different datasets, as shown in the following table:

Dataset	EDC	Day post-fertilization	Extraction	Concentration levels
1	BPA	4	General	C-L-M-H
2	BPA	4	Sphingolipids	C-L-M-H
3	BPA	6	General	C-L-M-H
4	BPA	6	Sphingolipids	C-L-M-H
5	E2	4	General	C-L-M-H
6	E2	4	Sphingolipids	C-L-M-H
7	E2	6	General	C-L-M-H
8	E2	6	Sphingolipids	C-L-M-H

For more information about ROI strategy, see (Gorrochategui et al., 2019; Pérez-Cova et al., 2021a), and more specifically about 2DLC data on the review (Pérez-Cova et al., 2021b).

4. MS-DIAL parameters for lipid identification

In the following table all parameters employed in MS-DIAL are listed:

Start up a project	**RP×HILIC-HRMS method**
Ionization type	Soft ionization
Separation type	Chromatography (LC)
Method type	Data dependent MS/MS
Data type (MS1)	Centroid (centroided in Proteowizard)
Data type (MS/MS)	Centroid (centroided in Proteowizard)
Ion mode	Positive ion mode
Target omics	Lipidomics
Data collection	
MS1 tolerance	0.01
MS2 tolerance	0.01
Retention time begin	0
Retention time end	120
Mass range begin	100
Mass range end	1500
Maximum charged number	2
Consider Cl and Br elements	Unchecked
Number of threads	20
Execute retention time corrections	Unchecked
Peak detection	

Minimum peak height	10000
Mass slice width	0.1
Smoothing method	Linear weighted moving average
Smoothing level	3
Minimum peak width	5
Exclusion mass list (tolerance: 0.01Da)	922.0098
MS2Dec	
Sigma window value	0.5
MS2Dec amplitude cut off	100
Exclude after precursor	Checked
Keep isotope until	0.5
Keep the isotopic ion w/o MS2Dec	Unchecked
Identification	
Retention time tolerance	100
Accurate mass tolerance (MS1)	0.01
Accurate mass tolerance (MS2)	0.01
Identification score cut off	80
Use retention time for scoring	Unchecked
Use retention time for filtering	Unchecked
Postidentification	Not used
Adduct	
Molecular species	[M+H]+, [M+NH_4]+, [M+Na]+, [M+CH_3OH+H]+, [M+H-H_2O]+, [M+2Na-H]+, [2M+H]+
Alignment	
Retention time tolerance	0.5
MS1 tolerance	0.015
Retention time factor	0.2
MS1 factor	0.8
Peak count filter	5
N% detected in at least one group	5
Remove feature based on blank information	Unchecked
Sample average / blank average	5
Keep "reference matched" metabolite features	Checked
Keep "suggested (w/o MS2)" metabolite features	Unchecked
Keep removable features and assign the tag	Checked
Gap filling by compulsion	Checked
Isotope tracking	
	Not used

5. Active Solvent Modulation

ASM is a valve-based approach recently developed by Stoll et al. (Stoll et al., 2017) that uses an 8-port interface with a 4-position design, modified with a bypass capillary. When the bypass

path is isolated, the valve acts as a normal 8- or 10-port valve with 2 positions. This means that one of the loops is being refilled with ^{1}D effluent, while the other loop is being discharged into the ^{2}D column with ^{2}D mobile phase. However, when the bypass is on, the ^{1}D effluent from the loop is being displaced and diluted with ^{2}D initial mobile phase composition. This dilution step (also called ASM step) depends on the flow rate and loop size and takes place at the very beginning of each modulation. The dilution is performed according to split ratios (i.e., in this work, a $^{1}/_{5}$ dilution was employed, meaning that 1 part goes through loop and 5 parts go through bypass). Hence, ASM uses a bypass capillary that dilute the fractions coming out from the ^{1}D column before reaching the ^{2}D column, which improves solvent compatibility between the two separations, while enhancing sensitivity and decreasing the total analysis time.

References

Gorrochategui, E., Jaumot, J., Tauler, R., 2019. ROIMCR: A powerful analysis strategy for LC-MS metabolomic datasets. BMC Bioinformatics 20, 1–17. https://doi.org/10.1186/s12859-019-2848-8

Pérez-Cova, M., Bedia, C., Stoll, D.R., Tauler, R., Jaumot, J., 2021a. MSroi: A pre-processing tool for mass spectrometry-based studies. Chemometrics and Intelligent Laboratory Systems 215. https://doi.org/10.1016/j.chemolab.2021.104333

Pérez-Cova, M., Jaumot, J., Tauler, R., 2021b. Untangling comprehensive two-dimensional liquid chromatography data sets using regions of interest and multivariate curve resolution approaches. TrAC - Trends in Analytical Chemistry 137. https://doi.org/10.1016/j.trac.2021.116207

Silcock, P., Wainwright, A., Hunter, C., 2013. Advancing Endocrine Disrupting Compound Analysis Through Integrated Technology and Workflow Solutions 1–10.

Stoll, D.R., Shoykhet, K., Petersson, P., Buckenmaier, S., 2017. Active Solvent Modulation: A Valve-Based Approach to Improve Separation Compatibility in Two-Dimensional Liquid Chromatography. Analytical Chemistry 89, 9260–9267. https://doi.org/10.1021/acs.analchem.7b02046

VI. SCIENTIFIC PUBLICATION VI

Title: Quantification strategies for two-dimensional liquid chromatography datasets using regions of interest and multivariate curve resolution approaches

Authors: Miriam Pérez-Cova, Stefan Platikanov, Romà Tauler, Joaquim Jaumot

Citation reference: Talanta 247 (2022) 123586.

DOI: 10.1016/j.talanta.2022.123586

Talanta 247 (2022) 123586

Contents lists available at ScienceDirect

Talanta

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

Quantification strategies for two-dimensional liquid chromatography datasets using regions of interest and multivariate curve resolution approaches

Miriam Pérez-Cova [a,b,*], Stefan Platikanov [a], Romà Tauler [a], Joaquim Jaumot [a,**]

[a] Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, 508034 Barcelona, Spain
[b] Department of Chemical Engineering and Analytical Chemistry, University of Barcelona, Diagonal 647, E08028, Barcelona, Spain

ARTICLE INFO

Keywords:
LC×LC
MCR-ALS
RODMCR
Chemometrics
Data analysis
Quantification

ABSTRACT

In this work, three chemometrics-based approaches are compared for quantification purposes when using two-dimensional liquid chromatography (LC×LC-MS), taking as a study case the quantification of amino acids in commercial drug mixtures. Although the approaches have been already used for one-dimensional gas or liquid chromatography, the main novelty of this work is the demonstration of their applicability to LC×LC-MS datasets. Besides, steps such as peak alignment and modelling, commonly applied in this type of data analysis, are not required with the approaches proposed here. In a first step, regions of interest (ROI) strategy is used for the spectral compression of the LC×LC-MS datasets. Then the first strategy consists of building a calibration curve from the areas obtained in this ROI compression step. Alternatively, the ROI intensity matrices can be used as input for a second analysis step employing the multivariate curve resolution alternating least squares (MCR-ALS) method. The main benefit of MCR-ALS is the resolution of elution and spectral profiles for each of the analytes in the mixture, even in the case of strong coelutions and high signal overlapping. Classical MCR-ALS based calibration curve from the peak areas resolved only applying non-negativity constraints (second strategy) is compared to the results obtained when an area correlation constraint is imposed during the *ALS* optimization (third strategy). All in all, similar quantification results were achieved by the three approaches but, especially in prediction studies, the more accurate quantification is obtained when the calibration curve is built from the peak areas obtained with MCR-ALS when the area correlation constraint is imposed.

1. Introduction

The use of comprehensive two-dimensional liquid chromatography (LC×LC) has considerably grown in the last decade [1]. Instrumental developments (for instance, modulation interfaces [2]) have facilitated the implementation of multidimensional liquid chromatography in various research fields, such as food [3,4], environmental [5,6], traditional Chinese medicine [7], pharmaceutical [8,9], proteomic [10,11], or metabolomic [12,13] analysis, among many others. However, regardless of the latest improvements in LC×LC, data analysis remains a major bottleneck due to the high complexity and huge size of the generated datasets. More specifically, quantification in LC×LC is more complex than in one-dimension liquid chromatography (1DLC), because each chromatographic peak from the first dimension (^{1}D), corresponding to a single compound, is split into several peaks in the second dimension (^{2}D). Therefore, peak integration in LC×LC requires the sum of individual peaks from the same compound or the graphical integration of its corresponding spot in a 2D plot. Consequently, the integration step becomes more difficult in the case of complex samples (e.g., biological matrices), where overlapping peaks are commonly found. Besides, most of the LC×LC visualization and quantification is performed directly on vendor software or specific software designed for multiple vendor data (e.g., GC Image LC×LC Edition Software from GC Image™, AnalyzerPro® XD from SpectralWorks, and ChromSquare from Shimadzu). These tools are user-friendly. Nevertheless, the main drawbacks are their relatively high cost and lack of flexibility and control over the results, being sometimes perceived as a "black box". For this reason, in this work, three different chemometric-based quantification

* Corresponding author. Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, E08034 Barcelona, Spain.
** Corresponding author.
E-mail address: miriampco7@gmail.com (M. Pérez-Cova).

https://doi.org/10.1016/j.talanta.2022.123586
Received 31 March 2022; Received in revised form 23 May 2022; Accepted 24 May 2022
Available online 27 May 2022
0039-9140/© 2022 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

strategies are proposed and compared. Here, all data analyses are performed in the MATLAB environment with the correspondent toolboxes and through GUI interfaces. However, the same strategies can be developed in common programming languages such as Python and R software.

The first strategy is based on the regions of interest (ROI) approach for spectral compression [14]. A preliminary filtering step is highly recommended in LC×LC, especially when employed by mass spectrometry (MS). Thus, the ROI procedure implements a spectral compression algorithm without the loss of MS spectral accuracy by keeping the most relevant m/z values with intensities above a pre-established signal-to-noise ratio threshold, according to a mass error tolerance previously defined (related to the MS instrument accuracy). Among other outputs, ROI procedure provides a features area matrix with the information for each sample and feature characterized by its m/z value. Therefore, calibration curves can be built based directly on this ROI areas output.

The second and third quantification strategies proposed in this work are based on the ROIMCR procedure, which means that after the ROI compression step, Multivariate Curve Resolution (MCR) is applied. MCR is a bilinear factor analysis method that allows the resolution of the elution and spectral profiles [15]. MCR has been widely employed for quantification purposes [16] and successfully implemented for LC×LC data when coupled to a diode array detector (DAD), as demonstrated by the work of Rutan et al. [17–19]. The main advantage of applying MCR to LC×LC datasets is that no prior chromatographic alignment or peak shape modelling is required (steps often required with other LC×LC data analysis pipelines due to chromatographic shifts or slight changes in peak shapes between samples). MCR allows direct analysis of the chromatograms, simplifying the pre-processing step, as well as the resolution of overlapping peaks.

Alternating Least Squares (ALS) has been selected for optimizing the initial guesses (i.e., estimates) of elution or spectra profiles [15]. During this optimization, mathematical constraints are applied to give physical (i.e., chemical) significance to the pure numerical solutions, and decrease the rotational ambiguities inherent to bilinear models [16,20]. The second quantification strategy proposed in this work is a classic calibration curve obtained with the areas from the MCR-ALS resolved elution profiles. This approach has been previously validated for the quantitative analysis of one-dimensional liquid chromatography coupled to mass spectrometry (LC-MS) and gas chromatography coupled to mass spectrometry (GC-MS) datasets [21,22]. Lastly, the third procedure implies using the area correlation constraint [23–25] during the MCR-ALS optimization. This constraint is specially recommended for the analysis of second-order multivariate calibration data, where unknown interferences can be present in the real samples (and not in the calibration set) without affecting the quantification [15,23].

This work aims to compare these quantification strategies for LC×LC-MS datasets, without the need for specific vendor software or extensive prior chromatographic pre-processing. First, a calibration curve is built only with the areas provided by the ROI approach. Then, the ROI intensity matrix is analyzed by MCR-ALS for the resolution and quantification of the LC×LC data. Two other strategies are then considered: a classic calibration curve from the resolved MCR elution profiles and a calibration curve using the area correlation constraint during the iterative ALS optimization. These three approaches have been tested for the quantification of several amino acids in commercial drug mixtures.

2. Materials and methods

2.1. Reagents, standards and samples

Ammonium formate (NH_4Form, ≥99%), formic acid (HForm, ≥95.0%) and acetic acid (HAc, ≥95.0%) were purchased from Sigma-Aldrich (Merck, Darmstadt, Germany). HPLC grade water and acetonitrile (AcN), and the amino acid standards mix, physiological acids neutrals, and basics were supplied by Merck KGaA (Darmstadt, Germany). The amino acid mix contained: L-alanine, ammonium chloride, L-arginine, L-aspartic acid, L-cystine, L-glutamic acid, glycine, L-histidine, L-isoleucine, L-leucine, L-lysine, L-methionine, L-phenylalanine, L-proline, L-serine, L-threonine, L-tyrosine, L-valine, β-alanine, L-α-amino-n-butyric acid, γ-amino-n-butyric acid, DL-β-aminoisobutyric acid, L-anserine, L-carnosine, L-citrulline, L-creatinine, cystathionine, ethanolamine, L-homocystine, δ-hydroxylysine, hydroxy-L-proline, 1-methyl-L-histidine, 3-methyl-L-histidine, L-ornithine, sarcosine, taurine, L-tryptophan, and urea.

Two commercial products composed of several amino acids were employed as experimental samples. Sample A: Aminoven 10% for hospital use in solution format, including lysine acetate, L-valine, L-serine, L-tryptophan, L-alanine, L-arginine, taurine, L-phenylalanine, glycine, L-histidine, L-isoleucine, L-leucine, L-methionine, L-proline and L-tyrosine; and Sample B) Aminosäuren komplex from Vitamaze GmbH (Heidelberg, Germany) in pill format, including eight essential amino acids: L-leucine, L-lysine, L-valine, L-isoleucine, L-phenylalanine, L-threonine, L-methionine and L-tryptophan. From now on, the two samples will be referred to as **Sample A** (perfusion solution of Aminoven 10%) and **Sample B** (pill extract of amino acids).

2.2. Standards mix dilution and sample treatments

In the standard mixture, each amino acid was at a concentration of 0.5 μmol mL^{-1}, which coincides with the highest concentration value of the calibration curve. The amino acids mix stock solution was diluted in a mixture of AcN:H_2O (1:1) to obtain the following final concentrations 0.1, 0.2, 0.3 and 0.4 μmol mL^{-1}.

Sample A was diluted 250 times in AcN:H_2O (1:1) so that the final amino acid concentrations were within the calibration range (0.1–0.5 μmol mL^{-1}) and three replicates of the diluted solution were analyzed. **Sample B** was initially triturated and homogenized. A quarter of each tablet was dissolved in a mixture of 1:4 HAc:H_2O and filtered with a 0.45 μm pore size and 25 mm diameter Millex-LCR hydrophilic PTFE membrane supplied by Merck KGaA (Darmstadt, Germany). Then, a 1:1 dilution was performed with a mixture of AcN:H_2O (1:1), and the diluted solution was analyzed in triplicate. Those amino acids with concentration levels under 0.1 μmol mL^{-1} were considered out of the calibration range, which was delimited by the sensibility of the mass spectrometer and the dilution effect of the two-dimensional separation. Besides, only the amino acids with an m/z value higher than 80 were studied. The mass range (80–400 Da) was established to avoid possible noisy signals from solvents or impurities.

2.3. LC×LC-MS method

LC×LC-MS required two chromatographic separations. The second chromatographic separation was performed in an Acquity UHPLC system (Waters, Milford, MA, US) equipped with a quaternary pump and an autosampler, whereas the first chromatographic separation employed an auxiliary Waters 1525 binary HPLC pump. An Acquity UPLC Column Manager (Waters, Milford, MA, US), equipped with two 6-port two-position valves and two 30 μL loops, was used as interface [26,27].

The chromatographic column used in the first dimension was a BEH Waters HILIC Acquity UPLC (100 × 2.1 mm; 1.7 μm), running at 50 °C, and at a flow rate of 0.03 mL min^{-1}, with an injection volume of 20 μL; mobile phases A) 10 mM NH_4Form and 0.2% HForm in 90:10 AcN:H_2O, B) 10 mM NH_4Form and 0.2% HForm in 50:50 AcN:H_2O, eluting according to the following gradient: 0 min, 0.1% B; 130 min, 90% B; 131 min, 0.1% B and kept until 160 min at 0.1% B for re-equilibration. The total analysis time per chromatogram was 160 min.

The second dimension column was an Xbridge BEH Shield C18 Waters (50 × 4.6 mm; 3.5 μm), running at 50 °C, and at a flow rate of 1.00 mL min^{-1} with a split ratio of 1:1 (1 part to waste, 1 part to MS), mobile

phases A) 0.1% HForm in AcN and B) 0.1% HForm in H_2O, and following the gradient: 0 min, 2% B; 1 min, 20% B; 1.1 min, 2% B; and kept until 1.5 min at 2% B for re-equilibration. The total modulation time was 1.5 min.

A triple quadrupole detector (TQD, Waters, Milford, MA, US) was employed as mass spectrometer, acquiring positive and negative ionization modes with an electrospray (ESI) as ionization source. Desolvation gas (nitrogen, >99.98%) flow rate was set at 900 L h^{-1}, desolvation temperature at 450 °C and cone voltage at 50 V. Full scan mode was employed, with a mass acquisition range from 80 to 400 Da.

2.4. Data management and analysis

2.4.1. Raw data conversion and import

LC×LC-MS raw files (.raw) were converted to Common Data Format files (.cdf) for each sample using the DataBridge file converter (Waters MassLynx® software). Then, the MSroi software [14] was employed for importing the generated (.cdf) files into the MATLAB environment using the apps: *ImportMS* for single file importation and *ImportMSmulti* for batch importation of multiple files [14].

2.4.2. Regions of interest algorithm

The ROI strategy targeted the selection of the most interesting mass traces [14,28] by applying several filters as filtering parameters. First, a selection of mass traces with intensity signals higher than a fixed signal-to-noise ratio threshold (SNRThr) which had to be encountered a minimum number of times in the chromatographic mode. In this study, the ROI parameters were fixed to an SNRThr set at 0.1% of the observed maximum mass signal intensity. However, some mass traces required larger SNRThr values (up to 0.5%) to avoid background noise and be filtered appropciately. The mass error tolerance of the spectrometer was set at 0.5Da/e for the TQD analyzer used for this MS analysis. The minimum number of occurrences to be considered as a chromatographic peak was set at 25. However, for some m/z this number had to be elevated to 120, when the chromatographic peak elapsed for too long time. Detailed information about the ROI approach and its applicability for mass spectrometry-based studies can be found in Refs. [14,28].

An LC×LC-MS single sample is a three-dimensional data cube, as shown in Fig. 1. Two modes for the ^{1}D and the ^{2}D retention times (chromatographic information) and the third mode represents the mass spectral information. This data cube is unfolded during the ROI approach, keeping in common the spectral dimension. The output is an augmented two-dimensional matrix containing all measured retention times (considering both chromatographic dimensions in the rows) in the first mode and the filtered spectral m/z data as columns in the second mode. In the case of multiple LC×LC-MS samples, the ROI approach builds a higher-order structure based on these augmented individual matrices (i.e., by stacking one individual matrix above the others) to generate a super augmented two-dimensional data matrix.

In this study, the ROI approach generated two super augmented data matrices $\mathbf{D}_{\mathbf{aug},1}$ and $\mathbf{D}_{\mathbf{aug},2}$, corresponding to **Sample A** and **Sample B**, respectively. Each one of these super augmented matrices was built by concatenating the five matrices with amino acids standards mixtures data at five concentrations ($\mathbf{D}_{\mathbf{std},1\text{-}5}$), in addition to the data matrices obtained for the three replicates for the two commercial products with mixtures of amino acids ($\mathbf{D}_{\mathbf{san}1,1\text{-}3}$ and $\mathbf{D}_{\mathbf{san}2,1\text{-}3}$). Fig. S1 exemplifies the generic augmentation step for $\mathbf{D}_{\mathbf{aug},\mathbf{x}}$ (applicable to both samples). The quantitative information for the resolved amino acids in the first five matrices (used as calibration samples) was used to predict the "unknown" concentrations of the corresponded amino acids resolved in the commercial mixtures. It is worth mentioning that the concatenation of these eight matrices was done keeping in common the spectral m/z data dimension. Since $\mathbf{D}_{\mathbf{std},1\text{-}5}$, $\mathbf{D}_{\mathbf{san}1,1\text{-}3}$ and $\mathbf{D}_{\mathbf{san}2,1\text{-}3}$ can contain different m/z traces after the ROI approach, a new alignment step by m/z correspondence was required [14,28,29]. All amino acids were detected in the ROI step except L-glycine and taurine from **Sample A**, and L-lysine

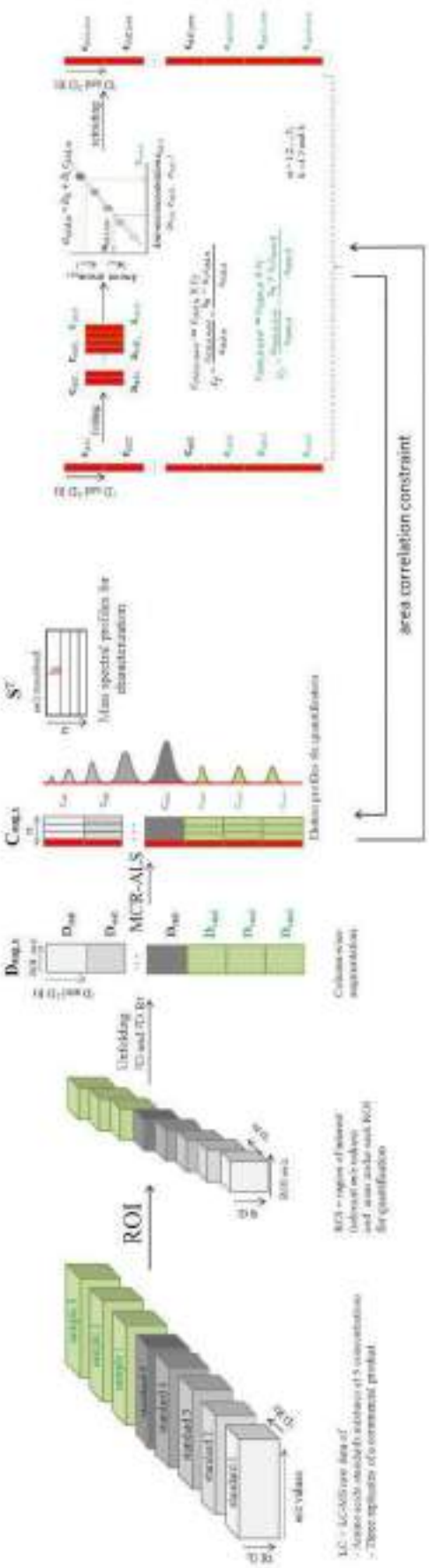


Fig. 1. Schematic implementation of MSROI and the MCR-ALS analysis of multiple LC×LC–MS data matrices using a bilinear decomposition model.

from **Sample B**. L-leucine and L-isoleucine were detected in the same ROI, and therefore, considered as the same compound.

Before further analysis, all these ROI data matrices are normalized to correct for instrumental intensity drifts among different sample injections. This normalization was done by dividing the measured mass intensity values of each data matrix by the chromatographic area of a selected *m/z* value (from an unknown compound) persistent in all chromatographic runs.

2.4.3. MCR-ALS analysis

Next, the super augmented matrices were subjected to further analysis using the Multivariate Curve Resolution (MCR-ALS) method. In this study, MCR-ALS is proposed for the resolution and quantitative analysis of complex mixture of amino acids. MCR-ALS relies on a bilinear data decomposition to resolve the mixture analysis problem in multicomponent systems by decomposing the experimental data matrix, **D**, into the product of the pure component response profiles, **C** and $\mathbf{S}^T$, according to the multivariate Beer-Lambert law:

$$\mathbf{D} = \mathbf{C}\mathbf{S}^T + \mathbf{E} \quad \text{Eq.1}$$

Here, **D** contains the chromatographic information, **C** is a matrix associated with chromatographic elution profiles of the resolved components (quantitative information) and $\mathbf{S}^T$ is a matrix related to the mass spectra of the resolved components (qualitative information useful, for instance, for compound identification). In addition, **E** represents the residuals not explained by the model using the considered number of components.

In this study, MCR-ALS was extended to the simultaneous analysis of multiple data sets or data matrices coming from multiple LC×LC-MS experiments. In this case, data was arranged in two super augmented column-wise data matrices, $\mathbf{D}_{aug,1}$ and $\mathbf{D}_{aug,2}$ (one super augmented matrix per commercial product) and the MCR bilinear model, generalized for both cases ($\mathbf{D}_{aug,x}$ either for **Sample A** or **Sample B**), was applied as shown:

$$\mathbf{D}_{aug,x} = \begin{pmatrix} \mathbf{D}_{std1} \\ \mathbf{D}_{std2} \\ \dots \\ \mathbf{D}_{std5} \\ \mathbf{D}_{unk1} \\ \mathbf{D}_{unk2} \\ \mathbf{D}_{unk3} \end{pmatrix} = \begin{pmatrix} \mathbf{C}_{std1} \\ \mathbf{C}_{std2} \\ \dots \\ \mathbf{C}_{std5} \\ \mathbf{C}_{unk1} \\ \mathbf{C}_{unk2} \\ \mathbf{C}_{unk3} \end{pmatrix} \mathbf{S}^T + \begin{pmatrix} \mathbf{E}_{std1} \\ \mathbf{E}_{std2} \\ \dots \\ \mathbf{E}_{std5} \\ \mathbf{E}_{unk1} \\ \mathbf{E}_{unk2} \\ \mathbf{E}_{unk3} \end{pmatrix} = \mathbf{C}_{aug,x}\mathbf{S}^T + \mathbf{E}_{aug,x} \quad \text{Eq.2}$$

A $\mathbf{D}_{aug,x}$ matrix has dimensions of ^{1}D x ^{2}D Rt x 8 matrices as number of rows (number of retention times in the first and second dimensions for the 8 matrices – five standards and three samples) and as columns – the number of selected ROI *m/z* values. Accordingly, the $\mathbf{C}_{aug,x}$ dimensions are ^{1}D x ^{2}D Rt x 8 matrices (i.e., the same rows as $\mathbf{D}_{aug,x}$) as rows and the resolved *n* number of components as columns. The $\mathbf{S}^T$ matrix dimensions correspond to the *n* number of components as rows and the number of selected ROI *m/z* values as columns. Finally, the $\mathbf{E}_{aug,x}$ matrix is the matrix of residuals not explained by the MCR model and has the same dimensions as matrix $\mathbf{D}_{aug,x}$. Using the proposed number of components, an initial estimate of $\mathbf{C}_{aug,x}$ or $\mathbf{S}^T$ is determined by a pure variable detection method to start the alternating least squares (ALS) iterative optimization. Then, the algorithm calculates $\mathbf{C}_{aug,x}$ and $\mathbf{S}^T$ matrices iteratively. During this ALS optimization, the non-negativity constraint was imposed to give physical meaning to the pure mathematical solutions, facilitate and speed up the convergence of the iterative process and minimize ambiguities inherent to factor analysis methods.

$\mathbf{C}_{aug,x}$ is the augmented concentration matrix giving the concentration profiles of every component in every individual data matrix $\mathbf{D}_{std,1\text{-}5}$ and $\mathbf{D}_{sam,1\text{-}3}$ (again either for **Sample A** or **Sample B**), and describes the concentration changes of the resolved components in each related **D** matrices. This flexibility allows the usage of this model in chromatographic studies. This resolved concentration profiles can be associated with chromatographic elution profiles, allowing component changes in position or shape from one sub-matrix to another.

Bayat et al. [30] have recently used MCR-ALS with non-negativity (nn) and the area correlation constraints to perform a second-order data quantitative analysis as a calibration step during the ALS optimization. The area correlation constraint step was implemented by regressing the peak areas (or heights) of the analytes resolved concentration profiles in the calibration samples against their known analyte concentrations. The estimated offset and slope of the calibration equation were used during the ALS optimization for the calculation of the analytes concentration in unknown samples. In this study, we applied the same strategy using the amino acids standards mixtures known five concentrations to achieve a calibration equation in order to predict the unknown concentrations of the common amino acids in the three replicates of **Sample A** or **Sample B**. This quantification approach based on the regression of the resolved areas on known concentrations is also applicable in a similar way for the case of MSROI resolved areas under each selected *m/z* as a chromatographic peak [14].

In these previous works [23,24], the authors concluded that the optimal recovery of the quantitative information was obtained by applying the additional constraint of area correlation during the ALS optimization, MCR-ALS (nn + acc). This new procedure led to better results, optimal from a least-squares criterion. Moreover, this area correlation constraint allowed for the concentration profiles of the analytes to be recovered in their proper quantitative units.

The area correlation constraint is imposed at each iteration step of the ALS process and the sum of the concentration from the analyte profiles are regressed to known analyte concentrations. This strategy is schematically shown in Fig. 1 right panel. The area (or height) of **C** profiles of each analyte are calibrated to the known concentration in submatrices of **C** (c) with linear line (Eq. (3)):

$$\mathbf{a}_{std,n} = \mathbf{b}_0 + \mathbf{b}_1\mathbf{c}_{std,n} \quad \text{Eq.3}$$

where $\mathbf{a}_{std,n}$ represents the area of the **C** elution profiles, $\mathbf{c}_{std,n}$ is the known nominal analyte concentration of the known samples (in this study, $\mathbf{D}_{std1}$ to $\mathbf{D}_{std5}$ matrices). $\mathbf{b}_0$ and $\mathbf{b}_1$ are the intercept and slope of the regression line for calibration samples, respectively. Then, the obtained slope and intercept of the linear model should be used to rescale the elements of the analyte profiles (c vectors) in all the calibration samples using a correlation factor (Eqs. (4) and (5)):

$$\mathbf{c}_{std,n,new} = \mathbf{c}_{std,n} \times \mathbf{c}_f \quad \text{Eq.4}$$

$$\mathbf{c}_f = \frac{\mathbf{a}_{std,n,new}}{\mathbf{a}_{std,n}} = \frac{\mathbf{b}_0 + \mathbf{b}_1\mathbf{C}_{std,n}}{\mathbf{a}_{std,n}} \quad \text{Eq.5}$$

where $\mathbf{c}_{std,n,new}$ and $\mathbf{c}_{std,n}$ are the vectors representing analyte profile after and before rescaling, respectively. $\mathbf{a}_{std,n,new}$ and $\mathbf{a}_{std,n}$ are the sum area (or height) of c profiles. $\mathbf{c}_n$ is the known concentration of the nth calibration simple. For the "unknown" k-samples ($\mathbf{D}_{sam,1}$, $\mathbf{D}_{sam,2}$ and $\mathbf{D}_{sam,3}$), which are not used in the calibration step, the analyte concentration profile is also scaled according to parameters of the regression line mentioned before (Eqs. (6) and (7)):

$$\mathbf{c}_{unk,new} = \mathbf{c}_{unk} \times \mathbf{c}_f \quad \text{Eq.6}$$

$$\mathbf{c}_f = \frac{\mathbf{a}_{unk,new}}{\mathbf{a}_{unk}} = \frac{\mathbf{b}_0 + \mathbf{b}_1\mathbf{C}_{unk}}{\mathbf{a}_{unk}} \quad \text{Eq.7}$$

where $\mathbf{a}_{unk}$ is the resolved analyte concentration profile area in "unknown" commercial product samples, $\mathbf{c}_{unk}$ is the concentration of "unknown" samples, obtained by the calibration model of known samples built previously $\left(\mathbf{c}_{unk} = \frac{a_{unk} - b_0}{b_1}\right)$.

As shown in Fig. 1, this procedure is implemented in every ALS step. The iterative estimation of analyte concentration is obtained, when the desired convergence is achieved, while concentration and spectral profiles are optimal [23,24].

3. Results and discussion

3.1. Visualization of LC×LC chromatograms

2D plots of the Total Ion Chromatograms (TICs) obtained are shown in Fig. 2. Few compounds (the most apolar, eluting at approximately 20 min) were barely retained in the ^{1}D column and appeared at three different retention times of the ^{2}D separation (RP), according to their hydrophobicity. However, most of the compounds were well retained in the ^{1}D HILIC column, and two regions could be differentiated along the ^{2}D separation (especially between 60 and 80 min). Although the standard mix compounds were distributed in the 2D space, there were certain overlapping regions (marked with a dashed line) with a higher density of compounds eluting. This scenario was even clearer for the analyzed samples (Samples A and B), where fewer compounds were detected. These coeluting regions were especially interesting for testing

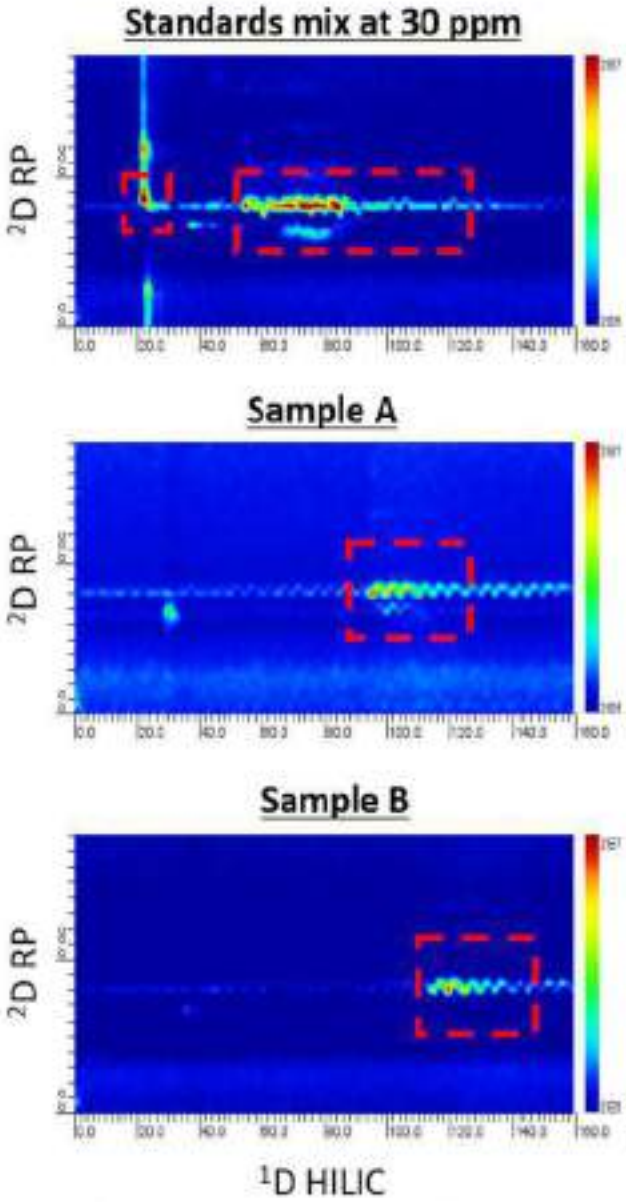


Fig. 2. 2D plots of the chromatograms obtained with the LC×LC Edition Software from GCImage, LLC (Lincoln, Nebraska), marked with dashed lines the regions with important coelution of compounds. The chromatograms correspond to the standard mix of amino acids at 30 ppm (top), and the Samples A (middle) and B (bottom).

the resolution capabilities of the ROI/MCR approach, as it is developed in the following sections.

3.2. ROI and MCR-ALS based quantification strategies

The first part of this study included an ROI approach analysis, performed on different data matrices: 1) the augmented data matrix $\mathbf{D}_{aug,1\text{-}5}$ of the amino acids standards mixtures at five concentrations; 2) the augmented data matrix $\mathbf{D}_{aug,1\text{-}3}$ for the three replicates of **Sample A**; and 3) the augmented data matrix $\mathbf{D}_{aug,1\text{-}3}$ of the three replicates of **Sample B**. This step would allow identifying all possible amino acids from the experimental LC×LC-MS before further MCR-ALS analysis.

The second section included the MCR-ALS analysis of the ROI outputs, arranged, and aligned by m/z ROI concordance into new augmented matrices, $\mathbf{D}_{aug,1}$ and $\mathbf{D}_{aug,2}$. Here, the MCR-ALS resolution will be focused on amino acids common for the standard mixtures at five different concentrations and the three replicates from one of the commercial products.

The last part included the quantification of resolved amino acids using three alternative approaches: 1) MSROI area quantification; 2) MCR-ALS with non-negativity (nn) constraint and external quantification; and 3) MCR-ALS with non-negativity and area correlation constraints (nn + acc) for quantification.

The first two approaches were based on the use of external calibration curves. This strategy employed a linear regression to relate the known concentrations of the recovered amino acids resolved in the five mixtures samples and the commercial products with the corresponding peak areas of their elution profiles obtained from the MSROI or MCR-ALS (nn) analyses. The third approach was based on the quantification of the recovered amino acids using the proposed area correlation constraint approach implemented during the MCR-ALS modelling [23, 30].

3.2.1. ROI approach analysis of the amino acids standards mixtures and drug samples

The application of the ROI approach for m/z selection resulted in a significant reduction of the large number of m/z values acquired by the MS instrument. First, the ROI approach was applied on LC×LC-MS data sets of individual concentrations of amino acids standards mixtures, $\mathbf{D}_{aug,1\text{-}5}$. The ROI approach resulted in a data matrix with dimensions of 23,305 rows as the total number of retention time channels for the ^{1}D and ^{2}D elution times (5 concentrations (chromatographic runs) x 466 elution times measures in each LC×LC-MS run), and 208 columns of finally selected m/z ROI values of interest. Fig. 3a shows the obtained LC×LC-MS ROI chromatograms for the five concentrations of amino acids mixtures, analyzed in positive mode. The peak clusters became denser with the increase of the concentration. Fig. 3b figures out that some m/z ROIs (for example m/z 105; m/z 110; m/z 113.9; m/z 158; m/z 214 shown on Fig. 3b present very high ion abundance, in contrast to the majority of the other selected m/z ROIs. However, these m/z ROIs were persistent background signals in all chromatographic runs. All m/z ROIs of particular interest have lower abundance, thus reinforcing the MCR-ALS application as the next step in this study.

The ROI approach, applied on the LC×LC-MS data sets of the three replicates for the two commercial products, resulted in two new augmented data matrices with dimensions of 13,983 rows (3 × 4661, defined as above) as the total number of retention time channels for the ^{1}D and ^{2}D elution times, and 99 columns of finally selected m/z ROI values for the **Sample A** (see Fig. 3c and 3d) and 54 columns of finally selected m/z ROI values for the **Sample B** (see Fig. 3e and 3f).

In contrast with the chromatograms observed for the five amino acids standards mixtures, the elution profiles of the selected m/z ROIs for these two commercial products pointed out that one m/z ROI in a single sample split clearly into several chromatographic peaks corresponding to its different modulations in the second column (i.e., the process of collecting the effluent from the first chromatographic

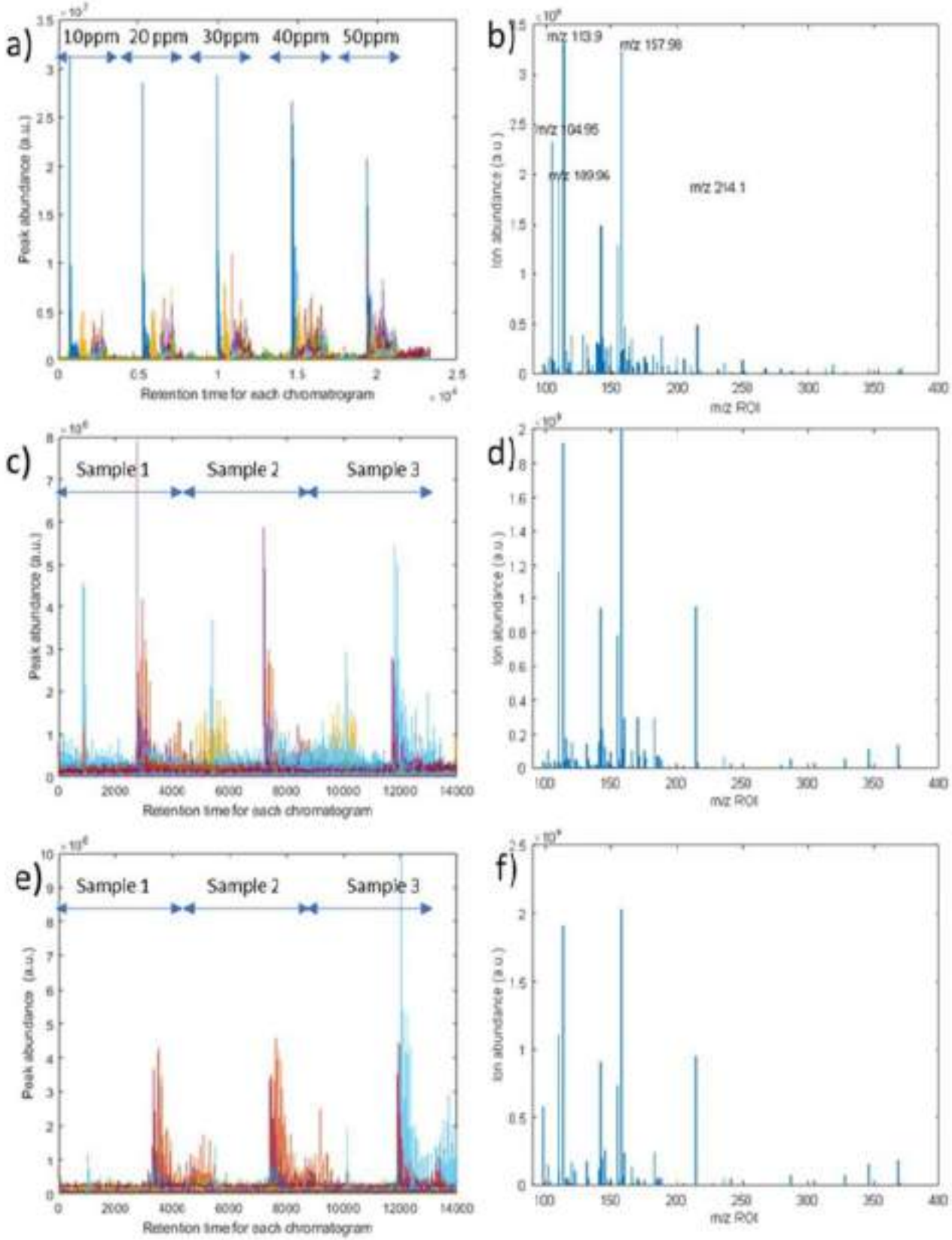


Fig. 3. ROI analysis of concatenated $\mathbf{D}_{ROI\,3}$ matrices showing the chromatographic data of the amino acids standards mixtures at five concentrations: a) elution profiles and b) mass spectra. ROI analysis of concatenated $\mathbf{D}_{term1\text{-}3}$ matrices showing the chromatographic data of the amino acids for the three replicates of **Sample A**: c) elution profiles and d) mass spectra. ROI analysis of concatenated $\mathbf{D}_{term1\text{-}3}$ matrices showing the chromatographic data of the amino acids for the three replicates of **Sample B**: e) elution profiles and f) mass spectra.

dimension and injecting these fractions into the second chromatographic dimension). Peak clusters in the three replicates were at a relatively equal density.

ROI approach allowed the investigation of the selected m/z ROIs one-by-one to confirm as many as possible amino acids in the five standard mixture samples and the two commercial products before proceeding with the MCR-ALS data analysis.

For example, Fig. 4 presents a selection of m/z ROI values, which corresponded to the L-threonine and was identified in the a) in the amino acids standards mixtures samples, b) in the three replicates of the **Sample A**, and c) in the three replicates of the **Sample B**.

The evaluation of the selected three m/z ROIs (see Fig. 4) in the data sets revealed homogeneous dispersion of the averaged m/z values around 119.95. In addition, the chromatographic separation for this particular m/z ROI, considering the concatenated elution profiles (**a.I.** panel) showed gradually thickened clusters of peak ions following the increase of the concentrations of the amino acid standard in these five samples. On the contrary, relatively equal thickness peak clusters were observed on the panels showing the three replicates for each commercial product (**b.I.** for **Sample A**, and **c.I.** for **Sample B** panels). All sample elution profiles were overlapped (**II.** panels), indicating the similarity of the elution profiles of these selected ROIs and using as x-axis the same retention time scale for all concatenated samples. Moreover, Fig. 4 also shows the areas under each chromatographic peak (see **a.II.**, **b.II.**, **c.II.** panels), giving an initial quantitative information about the relative concentration of this particular feature with m/z 119.95. This compound was identified as L-threonine amino acid in the considered samples. The five recovered areas in Fig. 4a.II. (the amino acids standard mixtures at five concentrations) followed a smooth gradient of increase, indicating strong linearity and thus, a possibility to build a linear equation

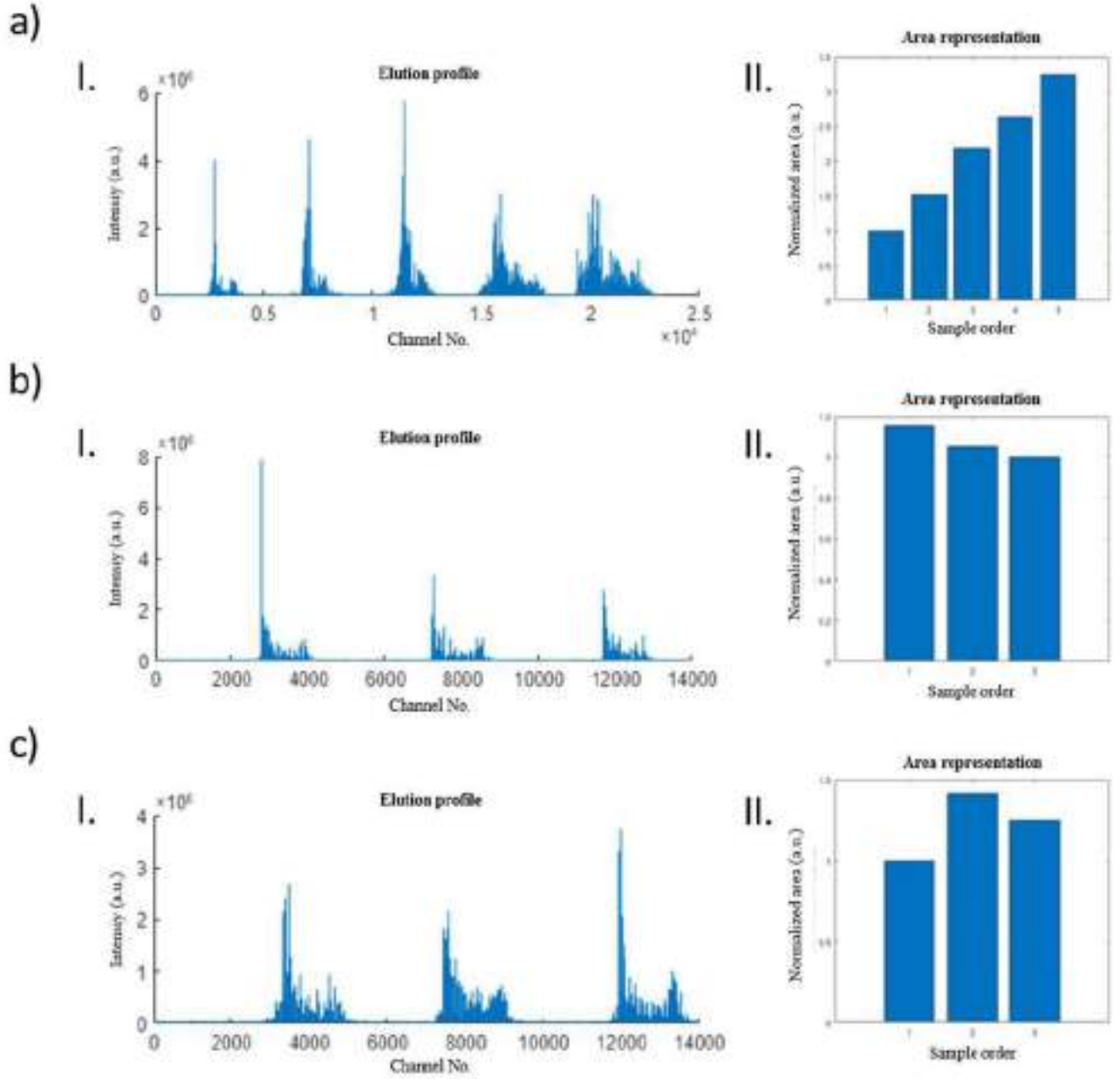


Fig. 4. The graphical output provided by the MSROI GUI for the evaluation of m/z ROI with value 119.9507 corresponding to L-threonine ($M + H^+$), considering elution profiles (I. panels) and the quantitative information (normalized areas using the minimum obtained value for easier comparison) related to the selected m/z ROI (II. panels), for: a) LC×LC-MS data of the amino acids standards mixtures at five concentrations; b) LC×LC-MS data of the three replicates samples for **Sample A**; c) LC×LC-MS data of the three replicates samples for **Sample B**.

externally to be used with increasing concentration, for the quantitative calibration of this amino acid. Areas observed for the three replicates of each sample (in Fig. 4b.II. and 4c.II.) presented similar values for L-threonine amino acid, which allowed its subsequent quantification using the different chemometric-based strategies.

Since ROI augmented matrices had different m/z values, a procedure for their arrangement by m/z correspondence was performed to achieve the final column-wise augmented data matrices with the same number of columns (m/z values) [28], $\mathbf{D}_{aug}$. These augmented matrices new dimensions were 37,288 rows (i.e., the total number of retention time channels for the ^{1}D and ^{2}D elution times), 95 columns of finally selected m/z ROI values for the combination of amino acids standards mixtures and the three **Sample A** replicates, and 49 columns of finally selected m/z ROI values for the combination of amino acids standards mixtures and the three **Sample B** replicates.

Despite the significant reduction of m/z values, a subsequent MCR-ALS chemometric analysis was still required to identify the amino acids standards. A large number of ROIs was observed due to the presence of many unknown compounds in the samples with relatively intense MS signals.

3.2.2. MCR-ALS analysis of D_{aug} matrices merging amino acids standards mixtures and drug samples

These two $\mathbf{D}_{aug}$ data matrices were then subjected to MCR-ALS analysis to resolve the elution profiles (in **C**) and pure MS spectra (in $\mathbf{S}^T$) of all amino acids present in the five mixtures samples concatenated, either with the three replicates of **Sample A**, or with the replicates of **Sample B**.

The first MCR-ALS analysis was done on $\mathbf{D}_{aug}$ (37,288 rows x 95 columns) with concatenated data for the five mixtures and the three replicates of **Sample A**. Fifty MCR-ALS components were used after an inspection of the sizes of the singular values of the data matrix. Constraints applied during the MCR-ALS analysis were non-negativity and spectral normalization (equal height). A list of resolved amino acids, common for the mixture samples and the three replicates of **Sample A**, correctly detected, is shown in Table 1.

Fig. 5a shows the elution profiles of all the resolved amino acids constituents (C) in the eight chromatographic runs. Changes in peak areas and heights of the elution profiles of the amino acids at the five different concentration levels and the three replicates are distinguished clearly. In addition, Fig. 5b displays the corresponding pure mass spectra of the amino acid constituents resolved by the MCR-ALS (in $\mathbf{S}^T$) procedure. Fig. 5c shows the resolved elution profile of the L-threonine standard in the different samples, and Fig. 5d depicts the resolved mass spectrum of the L-threonine amino acid standard (m/z 119.9507).

In a similar way, MCR-ALS analysis was done on $\mathbf{D}_{aug}$ (37,288 rows x 49 columns) with concatenated data for the five amino acids standard mixtures and the three replicates of **Sample B**. In this case, 26 MCR-ALS components were used after an inspection of the sizes of the singular values of the data matrix. The detected and resolved amino acids are listed in Table 1. The same spectral and elution profiles representations than in Fig. 5 but for **Sample B** are included in Fig. S2. Again, it is possible to distinguish the gradual changes in the peak areas and their heights of the elution profiles of the amino acids at the five different

Table 1
Figures of merit of the calibration and prediction of several amino acids concentrations using three quantification strategies.

Sample A		Amino acids calibration[a]			Amino acids prediction in the three replicates of Sample A[b]		
Amino acids	Method	r	RMSEC	Rel. Error %	Actual conc. µg/ml	RMSEP	Rel. Error %
L-proline	MS ROI[c]	0.991	2.12	5.5	44.8	4.72	11.7
	MCRALS[d]	0.991	2.12	5.6		4.71	11.7
	MCRALS[e]	0.990	2.19	5.7		2.80	6.0
L-valine	MS ROI[c]	0.990	2.36	6.0	24.8	5.34	21.2
	MCRALS[d]	0.990	2.29	5.9		5.41	20.4
	MCRALS[e]	0.993	2.24	5.6		5.37	20.4
L-threonine	MS ROI[c]	0.993	1.02	2.6	17.6	3.71	21.1
	MCRALS[d]	0.994	1.76	4.5		3.44	19.5
	MCRALS[e]	0.992	1.94	4.9		6.17	35.2
L-leucine/L-isoleucine	MS ROI[c]	0.993	2.12	4.9	29.6	2.27	7.2
	MCRALS[d]	0.993	2.11	4.9		2.53	7.4
	MCRALS[e]	0.993	2.14	4.9		2.25	7.2
L-phenylalanine	MS ROI[c]	0.989	3.44	6.3	20.4	4.24	18.5
	MCRALS[d]	0.988	3.60	6.6		3.83	16.9
	MCRALS[e]	0.980	4.72	8.6		2.86	14.8
L-tryptophan	MS ROI[c]	0.998	1.89	2.8	8.0	1.53	19.8
	MCRALS[d]	0.996	2.48	3.7		2.86	25.0
	MCRALS[e]	0.998	1.91	2.8		1.80	22.5
Sample B		**Amino acids calibration[a]**			**Amino acids prediction in the three replicates of Sample B[b]**		
Amino acids	Method	r	RMSEC	Rel. Error %	Actual conc. µg/ml	RMSEP	Rel. Error %
L-threonine	MS ROI[c]	0.999	0.90	2.3	27.1	4.28	15.8
	MCRALS[d]	0.998	1.01	2.5		4.45	16.4
	MCRALS[e]	0.997	1.24	3.1		3.31	12.2
L-leucine/L-isoleucine	MS ROI[c]	0.993	2.12	4.9	48.9	9.74	20.3
	MCRALS[d]	0.994	1.97	4.5		12.10	25.2
	MCRALS[e]	0.992	2.25	5.2		11.83	24.6
L-methionine	MS ROI[c]	0.974	4.74	9.6	17.1	4.12	24.1
	MCRALS[d]	0.978	4.41	8.9		5.56	32.5
	MCRALS[e]	0.973	5.03	10.1		5.22	30.5
L-phenylalanine	MS ROI[c]	0.989	3.35	6.3	31.3	4.20	13.3
	MCRALS[d]	0.989	3.52	6.4		4.17	13.3
	MCRALS[e]	0.988	3.65	6.7		3.31	10.8

[a] AAs calibration using the five different concentrations of the amino acids standards mixtures.
[b] AAs prediction in the three replicates, either of the **Sample A** or the **Sample B**.
[c] MSROI area as method for quantification.
[d] MCR-ALS with non-negativity constraint as method for quantification.
[e] MCR-ALS with non-negativity and area correlation constraints as method for quantification.

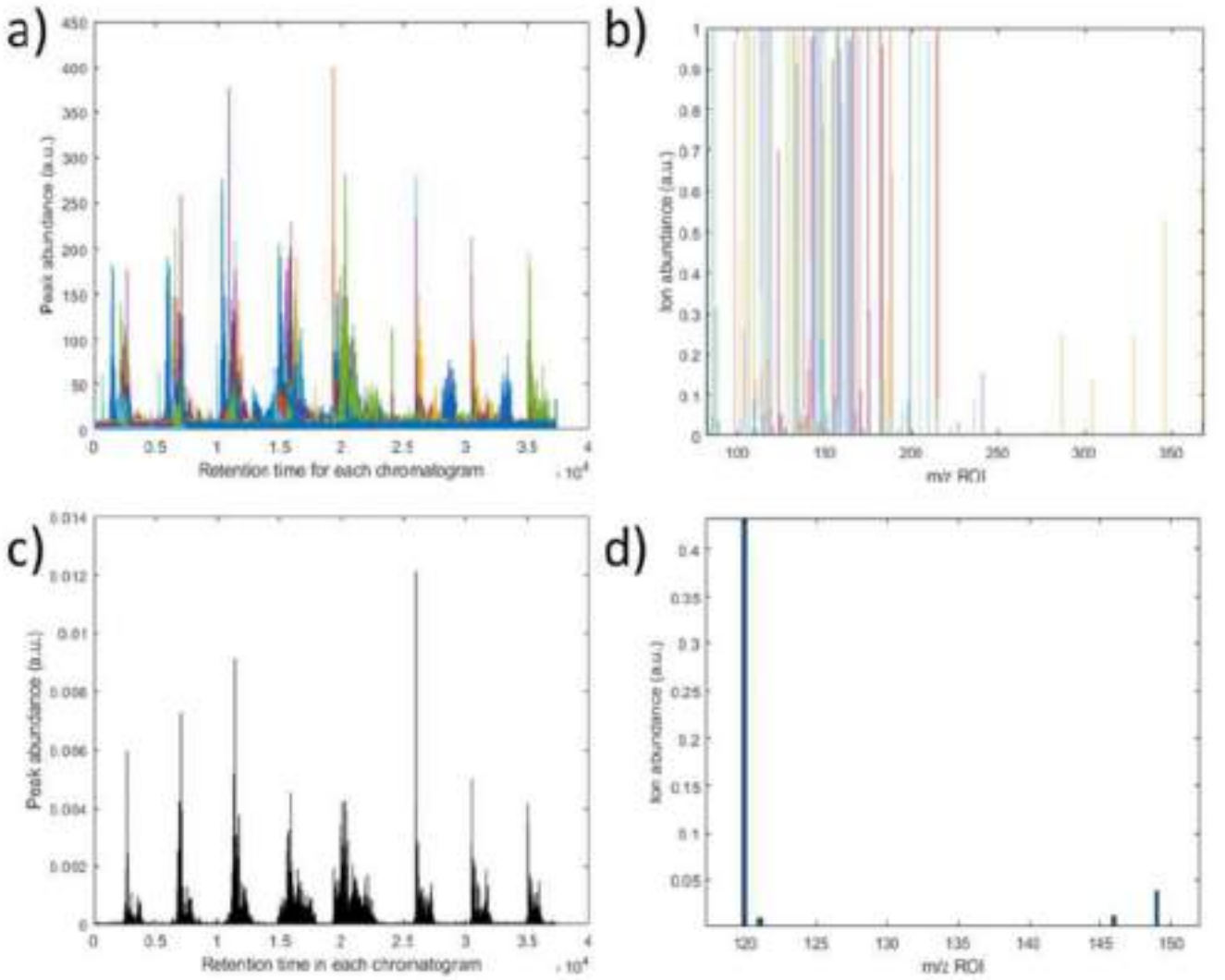


Fig. 5. MCR-ALS analysis of concatenated $\mathbf{D}_{aug}$ matrix with chromatographic data of the amino acids standards mixtures at five concentrations and the three replicates of **Sample A**: a) elution profiles, (**C**); b) mass spectral ROIs resolved components, ($\mathbf{s}^T$); c) elution profile, (**c**), and pure mass spectral ROI resolved component, ($\mathbf{s}^T$) corresponding to L-threonine standard (119.9507 m/z).

concentration levels and the three replicates.

3.2.3. *MSROI and MCR-ALS based quantification strategies*

The area correlation constraint in MCR-ALS was used in previous works for the quantitative determination of analyte mixtures [23,24]. In our study, its practical application is compared with MSROI and the MCR-ALS classical areas approaches for the quantification of amino acids in the five standard mixture samples and the three replicates for the two commercial products using their corresponding LC×LC-MS data.

MSROI peak areas (see Fig. S3a.I) and the resolved by MCR-ALS (nn) peak areas in the elution profiles of the amino acids for the five standards mixtures (see Fig. S3b and S3c, chromatograms 1 to 5) were used to build calibration curves by their regression against their known concentrations in the standards mixtures. Once these calibration curves were built, the amino acids concentration in the three replicates of **Sample A** (see Fig. S3a.II and S3b chromatograms 6 to 8) or **Sample B** (see Fig. S3a.III and S3c chromatograms 6 to 8) were estimated from their peak areas, and compared with their actual known concentration values. These predicted concentrations were plotted against the known amino acids concentrations in the third replicates and the regression parameters were calculated. For instance, Fig. S4 shows the calibration curves and the predictions for the L-threonine (119.9507 m/z) amino acid in the three replicates of **Sample A** and **Sample B** obtained with the three quantification approaches applied in this study: MSROI peak areas, MCR-ALS with non-negativity (nn) and MCR-ALS with non-negativity and area correlation coefficient (nn + ac). Merit figures, calculated from these curves, are presented in Table 1. As it can be observed for the calibration of the L-threonine, a good linearity and merit figures were obtained with values close to one for the correlation coefficient and small RMSEC and relative errors (around 5% or below). In addition, the prediction results for the L-threonine amino acid in the three replicates of **Sample A** or **Sample B** commercial products were also satisfactory, reporting low RMSEP values (slightly larger than RMSEC) and relative prediction errors around 20%. The magnitude of this error was relatively large but can be explained by the inherent variability associated with the two-dimensional chromatographic system. If the quantification strategies were considered, the three approaches provided similar results (Table 1). In the case of **Sample A**, the MSROI peak areas approach seemed to provide the best results in the calibration, whilst prediction results were comparable between the three approaches. However, in the case of **Sample B**, the MCR-ALS with the correlation constraint provided the best results both for predicting L-threonine.

When considering a representative set of amino acids detected for each sample, the same trend as in L-threonine can be observed. In general, the results using the three approaches were similar especially in the

case of calibration. In contrast, MCR-ALS with non-negativity and correlation constraints provided, in most cases, slightly better results (i.e., lower relative errors values) when considering the prediction of the amino acids in the real samples. The results obtained were similar when considering the other two approaches (based on MSROI areas and classical MCR-ALS resolution). Therefore, calibration and prediction error values obtained for the MCR-ALS (nn + ac) could be considered satisfactory, taking into account the current inherent variability in the multidimensional chromatography analyses. In addition, these error values for a single compound were similar in almost all cases, which provided greater confidence in the MCR-ALS models for prediction. The obtained results were acceptable for the calibration studies when evaluating the mean of the individual relative errors (in percentage). Still, in the prediction case, the obtained errors were relatively high in some cases. These large values could be explained because an optimal resolution of elution and spectral profiles could not be obtained. For instance, L-methionine for **Sample B** provided large relative errors of prediction (around 25%) in agreement with a poor calibration model (relative errors of calibration of around 10%).

Regarding previous validation of the ROIMCR strategy for absolute quantification purposes for both LC-MS and GC-MS datasets, similar values to the obtained in the present work were achieved (i.e., relative errors lower than 10% for most of the calibration mixtures and lower than 20% for drug samples) [21,22]. It is important to highlight that these results were obtained with a classical MCR-ALS calibration, and only non-negativity and spectral normalization (i.e., equal height) were selected as constraints. On the other side, the comparison performed by Bayat et al. [24,30] between MCR-ALS calibration (nn) and MCR-ALS (nn and ac) showed slight improvements in the relative errors in the area correlation approach for the majority of the compounds (in some specific cases, a reduction of up to a 10% in the relative values was appreciated). This slight enhancement in the quantification accuracy was also found in the present work as a global trend, although especially encountered for the prediction of amino acids in unknown samples (i.e., this tendency was not observed in the calibration studies). Most of the analyzed compounds presented relative errors lower than 10% for calibration mixtures and lower than 20% in the case of drug samples, which are in agreement with the reported by Bayat et al. [24].

Besides, although in this work only the linear concentration range between 10 and 50 ppm was used (due to the low sensitivity of the mass detector employed), other previous applications of MCR-ALS with and without the correlation constraint have demonstrated that it can be used in other applications and concentration ranges, for instance with calibration curves in the ppb range (e.g., from 1 to 400 ppb [22], or from 5 to 25 ppb [24]). This broad applicability is because the detection limit is conditioned by the sensitivity of the detector, rather than by the chemometric approach itself. Hence, the versatility of the MCR-ALS method for quantification is also very appealing in a wide variety of applications where the qualitative implementation of this approach has already been demonstrated, such as in environmental analyses (e.g., pharmaceutical compounds [31–33], in the determination of organic matter [34], organic compounds [35] or proteins [36], both in water bodies and wastewater treatment plants, and also metabolomic studies [37–39]. The ROIMCR procedure has also been validated for qualitative analysis and relative quantification of LC×LC-MS datasets, in metabolomics and pharmaceutical analyses [26,27]. In addition, there are still many other fields to be explored, such as foodomics [3,4] or environmental analysis [5,6].

One final comment regarding the results obtained in the comparison of the three calibration procedures is that the different approaches provided a similar quantification, which could be expected in the case of good chromatographic resolution and mass spectrometry detection. This good analytical performance favors the simpler methods such as the only based in the ROI-obtained areas. In contrast, MCR-based methods could show a much better performance in the case of complex scenarios such as overlapping chromatographic peaks or biological samples with multiple compounds, some of them being, for instance, isobaric. A pair of resolution examples are included in Fig. S5 to illustrate the main benefits of the MCR-ALS approach for quantification when applying area correlation constraint. Two compounds (L-leucine and L-methionine) from the highly dense chromatographic area in Fig. 2 are represented considering both the Extracted Ion Chromatograms (EICs) and resolved elution and spectral profiles obtained with MCR-ALS when applying the area correlation constraint. Although not isobaric, these compounds eluted in a very narrow chromatographic region were perfectly resolved and quantified with the strategy proposed. Therefore, these examples demonstrated for a simple case that MCR-ALS with the area correlation constraint can deal with coeluting compounds from LC×LC. In addition, the obtained good chromatographic resolution also minimized the presence of rotation ambiguities in MCR solutions which could harm the results obtained using only natural MCR constraints. For these reasons, the results obtained using the area correlation constraint did not improve much the results already obtained by using only non-negativity constraints or directly from the ROI signals. Again, it is only when exists a strong overlapping of the elution profiles of the standards used for calibration that the area correlation constraint can decrease the effect of rotation ambiguities and improve the calibration results. In the results obtained in this work, the situation is intermediate, with some slight improvements in the results when the area correlation constraint was applied.

4. Conclusions

Different strategies for the quantification of two-dimensional liquid chromatography datasets have been compared in this manuscript because a standardized approach to carry out analyte's quantification (alternative to vendor software approaches) is still missing. Besides, an alignment or a peak modelling step (often needed when analyzing LC×LC datasets) are not required in any of the three strategies tested in this work, simplifying the data analysis workflow. Here, we propose a pipeline composed of two steps. First, compression and filtering performed with the ROI method, and second, the resolution of analytes (elution and spectra profiles) using MCR-ALS with an area correlation constraint. This approach generates calibration models able to predict the analytes in unknown samples, even in the presence of unknown interferences regarding the calibration mixtures. The use of this area correlation constraint globally provided slightly lower prediction errors (i.e., a more accurate absolute quantification of the analytes). However, the other considered approaches based on an external calibration using the areas data from the ROI compression or the classical MCR-ALS approach (without the area correlation constraint) could also be acceptable options for preliminary studies. For instance, ROI calibration may be an interesting option in the absence of strong coelutions (i.e., if the analytes are well distributed in the 2D space, and there are no overlapping signals) due to its simplicity and their straightforward application. Otherwise, the combination of ROI with MCR-ALS is highly recommended. Even if a good separation was already achieved by LC×LC-MS analysis of the mixture of standards, there are some advantages of applying the ROIMCR method. First, this approach allows the same single-shot simultaneous analysis of all calibration and quantification steps. Second, when unknown real samples are simultaneously analyzed, the ROIMCR method allows the direct quantification of the standard compounds in the unknown samples even in the presence of unknown overlapping interferences and matrix effects. This good performance of the ROIMCR based approach opens the possibility of being used in different research fields in which LC×LC-MS has a promising future, such as pharmaceutical, food, metabolomics or environmental analyses, where complex matrices with multiple isobaric compounds are often encountered.

Funding

The research leading to these results have received funding from grants CTQ2017-82596-P and CEX2018-000794-S funded by MCIN/AEI. The authors also want to grant support from the Catalan Agency for Management of University and Research Grants (AGAUR, Grant 2017SGR753). MPC acknowledges a predoctoral FPU 16/02640 scholarship from the Spanish Ministry of Education and Vocational Training (MEFP).

Author contributions

Joaquim Jaumot and Miriam Pérez-Cova: Conceptualization. Miriam Pérez-Cova: Methodology. Stefan Platikanov: Formal analysis. Miriam Pérez-Cova and Stefan Platikanov: Data curation. Miriam Pérez-Cova, Stefan Platikanov, and Joaquim Jaumot: Writing – original draft. Miriam Pérez-Cova, Stefan Platikanov, Romà Tauler and Joaquim Jaumot: Writing – review & editing. Romà Tauler and Joaquim Jaumot: Supervision. Joaquim Jaumot: Funding acquisition. All authors have read and agreed to the published version of the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.talanta.2022.123586.

References

[1] B.W.J. Pirok, D.R. Stoll, P.J. Schoenmakers, Recent developments in two-dimensional liquid chromatography: fundamental improvements for practical applications, Anal. Chem. 91 (2019) 240–263, https://doi.org/10.1021/acs.analchem.8b04841.

[2] Y. Chen, L. Montero, O.J. Schmitz, Advance in on-line two-dimensional liquid chromatography modulation technology, Trends Anal. Chem. 120 (2019), https://doi.org/10.1016/J.TRAC.2018.115647.

[3] F. Cacciola, P. Rigano, P. Dugo, L. Mondello, Comprehensive two-dimensional liquid chromatography as a powerful tool for the analysis of food and food products, Trends Anal. Chem. 127 (2020) 115894, https://doi.org/10.1016/J.TRAC.2020.115894.

[4] L. Montero, M. Herrero, Two-dimensional liquid chromatography approaches in Foodomics – a review, Anal. Chim. Acta 1063 (2019) 1–18, https://doi.org/10.1016/J.ACA.2019.02.006.

[5] U. Jost, F. Hubeduth, Two-dimensional hydrophilic interaction and reversed phase liquid chromatography easily extracted pesticides and polar pesticides multi-residue method-A concept, J. Chromatogr. A 1621 (2020) 461040, https://doi.org/10.1016/J.CHROMA.2020.461040.

[6] C. Armutcu, E. Özgür, T. Kurnaz, E. Bayram, L. Uzun, M.E. Çormen, Rapid analysis of polycyclic aromatic hydrocarbons in water samples using an automated on-line two-dimensional liquid chromatography, Water Air Soil Pollut. 230 (2019) 249, https://doi.org/10.1007/s11270-019-4306-7.

[7] W. Zhou, Y. Liu, J. Wang, Z. Guo, A. Shen, Y. Liu, X. Liang, Application of two-dimensional liquid chromatography in the separation of traditional Chinese medicine, J. Separ. Sci. 43 (2020) 87–104, https://doi.org/10.1002/jssc.201900765.

[8] M. Iguiniz, S. Heinisch, Two-dimensional liquid chromatography in pharmaceutical analysis. Instrumental aspects, trends and applications, J. Pharm. Biomed. Anal. 145 (2017) 483–503, https://doi.org/10.1016/J.JPBA.2017.07.019.

[9] K. Wicht, M. Boert, A. Kojani, S. Schipperges, N. von Doehren, G. Desmet, A. de Villiers, P. Lynen, Pharmaceutical impurity analysis by comprehensive two-dimensional temperature responsive × reversed phase liquid chromatography, J. Chromatogr., A 1630 (2020) 461561, https://doi.org/10.1016/J.CHROMA.2020.461561.

[10] X. Wang, G. Yao, H. Zheng, N. Gao, X. Zhang, Strategy for high-throughput identification of protein complexes by array-based multi-dimensional liquid chromatography-mass spectrometry, J. Chromatogr. A 1652 (2021) 462351, https://doi.org/10.1016/J.CHROMA.2021.462351.

[11] H.C.H. Law, R.P.W. Kong, S.S.W. Sneo, Y. Zhao, Z. Zhang, Y. Wang, G. Li, Q. Quan, S.M.Y. Lee, H.C. Lam, I.K. Chu, A versatile reversed phase-strong cation exchange-reversed phase (RP-SCX-RP) multidimensional liquid chromatography platform for qualitative and quantitative shotgun proteomics, Analyst 140 (2015) 1237, https://doi.org/10.1039/c4an01193a.

[12] M. Pérez-Cova, R. Tauler, J. Jaumot, Two-dimensional liquid chromatography in metabolomics and lipidomics, Neuromethods 159 (2021) 25–47, https://doi.org/10.1007/978-1-0716-0864-7_3.

[13] W. Lv, X. Shi, S. Wang, G. Xu, Multidimensional liquid chromatography-mass spectrometry for metabolomic and lipidomic analyses, Trends Anal. Chem. 120 (2019) 115302, https://doi.org/10.1016/J.TRAC.2018.11.001.

[14] M. Pérez-Cova, C. Bedia, D.R. Stoll, R. Tauler, J. Jaumot, MSroi a preprocessing tool for mass spectrometry-based studies, Chemometr. Intell. Lab. Syst. 215 (2021), https://doi.org/10.1016/j.chemolab.2021.104333.

[15] A. de Juan, J. Jaumot, R. Tauler, Multivariate Curve Resolution (MCR). Solving the mixture analysis problem, Anal. Methods 6 (2014) 4964–4976, https://doi.org/10.1039/c4ay00571f.

[16] R. Tauler, A. de Juan, Chapter 5 - multivariate curve resolution for quantitative analysis, in: A.M. de la Pena, H.C. Goicoechea, G.M. Escandar, A.C. Olivieri (Eds.), Fundamentals and Analytical Applications of Multiway Calibration, Elsevier, 2015, pp. 247–292, https://doi.org/10.1016/B978-0-444-63527-3.00005-6.

[17] D.W. Cook, M.L. Burnham, D.C. Hamon, D.R. Stoll, S.C. Rutan, Comparison of multivariate curve resolution strategies in quantitative LCxLC: application to the quantification of furanocoumarins in apiaceous vegetables, Anal. Chim. Acta 961 (2017) 49–58, https://doi.org/10.1016/j.aca.2017.01.047.

[18] H.P. Bailey, S.C. Rutan, P.W. Carr, Factors that affect quantification of diode array data in comprehensive two-dimensional liquid chromatography using chemometric data analysis, J. Chromatogr. A 1218 (2011) 8411–8422, https://doi.org/10.1016/j.chroma.2011.09.057.

[19] H.P. Bailey, S.C. Rutan, Chemometric resolution and quantification of four-way data arising from comprehensive 2D-LC-DAD analysis of human urine, Chemometr. Intell. Lab. Syst. 106 (2011) 131–141, https://doi.org/10.1016/j.chemolab.2010.07.008.

[20] G. Ahmadi, R. Tauler, H. Abdollahi, Multivariate calibration of first-order data with the correlation constrained MCR-ALS method, Chemometr. Intell. Lab. Syst. 142 (2015) 143–150, https://doi.org/10.1016/J.CHEMOLAB.2014.11.010.

[21] N. Dalmau, C. Bedia, R. Tauler, Validation of the regions of interest multivariate curve resolution (ROIMCR) procedure for untargeted LC-MS lipidomic analysis, Anal. Chim. Acta 1025 (2018) 80–91, https://doi.org/10.1016/j.aca.2018.04.023.

[22] R.S.M. Pourzail, J. Cristale, S. Lacorte, R. Tauler, Non-targeted gas chromatography orbitrap mass spectrometry qualitative and quantitative analysis of semi-volatile organic compounds in indoor dust using the regions of interest multivariate curve resolution chemometrics procedure, J. Chromatogr. A (2022) 462907, https://doi.org/10.1016/J.CHROMA.2022.462907.

[23] A.C. de Oliveira Neves, R. Tauler, E.M.G. de Lima, Area correlation constraint for the MCR–ALS quantification of cholesterol using EEM fluorescence data: a new approach, Anal. Chim. Acta 937 (2016) 21–28, https://doi.org/10.1016/j.aca.2016.08.011.

[24] M. Bayot, M. Marín-García, J.B. Ghasemi, R. Tauler, Application of the area correlation constraint in the MCR-ALS quantitative analysis of complex mixture samples, Anal. Chim. Acta 1113 (2020) 52–65, https://doi.org/10.1016/j.aca.2020.03.057.

[25] G. Ahmadi, R. Tauler, H. Abdollahi, Multivariate calibration of first-order data with the correlation constrained MCR-ALS method, Chemometr. Intell. Lab. Syst. 142 (2015) 143–150, https://doi.org/10.1016/j.chemolab.2014.11.010.

[26] M. Navarro-Reig, J. Jaumot, R. Tauler, An untargeted lipidomic strategy combining comprehensive two-dimensional liquid chromatography and chemometric analysis, J. Chromatogr. A 1568 (2018) 80–90, https://doi.org/10.1016/j.chroma.2018.07.017.

[27] M. Pérez-Cova, R. Tauler, J. Jaumot, Chemometrics in comprehensive two-dimensional liquid chromatography: a study of the data structure and its multilinear behavior, Chemometr. Intell. Lab. Syst. 201 (2020) 104009, https://doi.org/10.1016/j.chemolab.2020.104009.

[28] E. Gorrochategui, J. Jaumot, R. Tauler, ROIMCR: a powerful analysis strategy for LC-MS metabolomic datasets, BMC Bioinf. 20 (2019) 1–17, https://doi.org/10.1186/s12859-019-2849-8.

[29] M. Pérez-Cova, J. Jaumot, R. Tauler, Untangling comprehensive two-dimensional liquid chromatography data sets using regions of interest and multivariate curve resolution approaches, Trends Anal. Chem. 137 (2021), https://doi.org/10.1016/j.trac.2021.116207.

[30] M. Bayot, M. Marín-García, J.B. Ghasemi, R. Tauler, Application of the area correlation constraint in the MCR-ALS quantitative analysis of complex mixture samples, Anal. Chim. Acta 1113 (2020) 52–65, https://doi.org/10.1016/j.aca.2020.03.057.

[31] A. Mostafa, H. Ibrahim, Quantitative analysis and resolution of pharmaceuticals in the environment using multivariate curve resolution-alternating least squares (MCR-ALS), Acta Pharm. 69 (2019) 217–231, https://doi.org/10.2478/ACPH-2019-0011.

[32] M. Marín-García, M. de Luca, G. Ragno, R. Tauler, Coupling of spectrometric, chromatographic, and chemometric analysis in the investigation of the photodegradation of sulfamethoxazole, Talanta 239 (2022) 122953, https://doi.org/10.1016/J.TALANTA.2021.122953.

[33] M. Marín-García, G. Ioele, H. Fraquet-Grisell, S. Lacorte, G. Ragno, R. Tauler, Investigation of the photodegradation profile of tamoxifen using spectroscopic and chromatographic analysis and multivariate curve resolution, Chemometr. Intell. Lab. Syst. 174 (2018) 128–141, https://doi.org/10.1016/j.chemolab.2018.01.011.

[34] M. Marín-García, R. Tauler, Chemometrics characterization of the Ebro river dissolved organic matter, Chemometr. Intell. Lab. Syst. 201 (2020), https://doi.org/10.1016/j.chemolab.2020.104010.

[35] S. Mínbas, S. Bieber, T. Letzel, Spotlight on mass spectrometric non-target screening analysis: advanced data processing methods recently communicated for extracting, prioritizing and quantifying features, Anal. Sci. Adv. 3 (2022) 103–112, https://doi.org/10.1002/ANSA.202200001.

[36] C. Perez-Lopez, A. Glieberds, M. Comacoel, D. Barceló, J. Abian, R. Tauler, Non-target protein analysis of samples from wastewater treatment plants using the regions of interest multivariate curve resolution (ROIMCR) chemometrics method, J. Environ. Chem. Eng. 9 (2021) 105752, https://doi.org/10.1016/J.JECE.2021.105752.

[37] M. Navarro-Reig, J. Jaumot, A. Baglai, G. Vivó-Truyols, P.J. Schoenmakers, R. Tauler, Untargeted comprehensive two-dimensional liquid chromatography coupled with high-resolution mass spectrometry analysis of rice metabolome using multivariate curve resolution, Anal. Chem. 89 (2017) 7675–7683, https://doi.org/10.1021/acs.analchem.7b01646.

[38] E. Ortiz-Villanueva, J. Jaumot, R. Martínez, L. Navarro-Martín, B. Piña, R. Tauler, Assessment of endocrine disruptor effects on zebrafish (Danio rerio) embryos by untargeted LC-HRMS metabolomic analysis, Sci. Total Environ. 635 (2018) 156–166, https://doi.org/10.1016/J.SCITOTENV.2018.03.369.

[39] F. Puig-Castellví, C. Bedia, I. Alfonso, B. Piña, R. Tauler, Deciphering the underlying metabolomic and lipidomic patterns linked to thermal acclimation in saccharomyces cerevisiae, J. Proteome Res. 17 (2018) 2034–2044, https://doi.org/10.1021/acs.jproteome.7b00921.

Supplementary Material

Quantification strategies for two-dimensional liquid chromatography datasets using regions of interest and multivariate curve resolution approaches

Miriam Pérez-Cova[1,2*], Stefan Platikanov[1], Romà Tauler[1], Joaquim Jaumot[1]

[1]Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, E08034 Barcelona, Spain

[2]Department of Chemical Engineering and Analytical Chemistry, University of Barcelona, Diagonal 647, Barcelona, E08028, Barcelona, Spain

* Correspondence: miriampeco7@gmail.com

1 Regions of interest algorithm – augmentation step

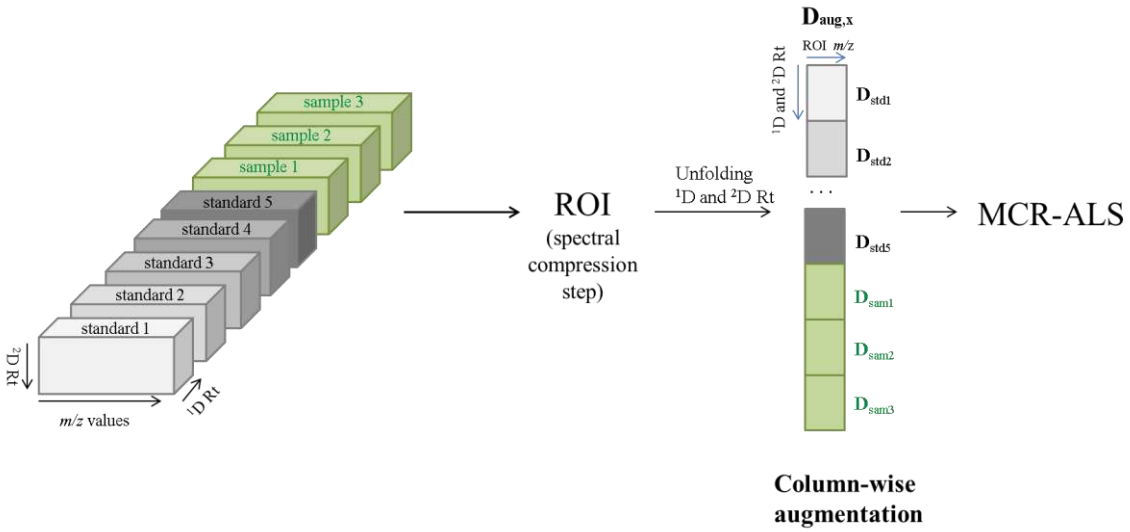


Figure S1. Scheme of the augmentation step from ROI procedure, where both standards and samples are concatenated in a column-wise manner.

2 MCR-ALS analysis of D_{aug} matrices merging amino acids standards mixtures and drug samples

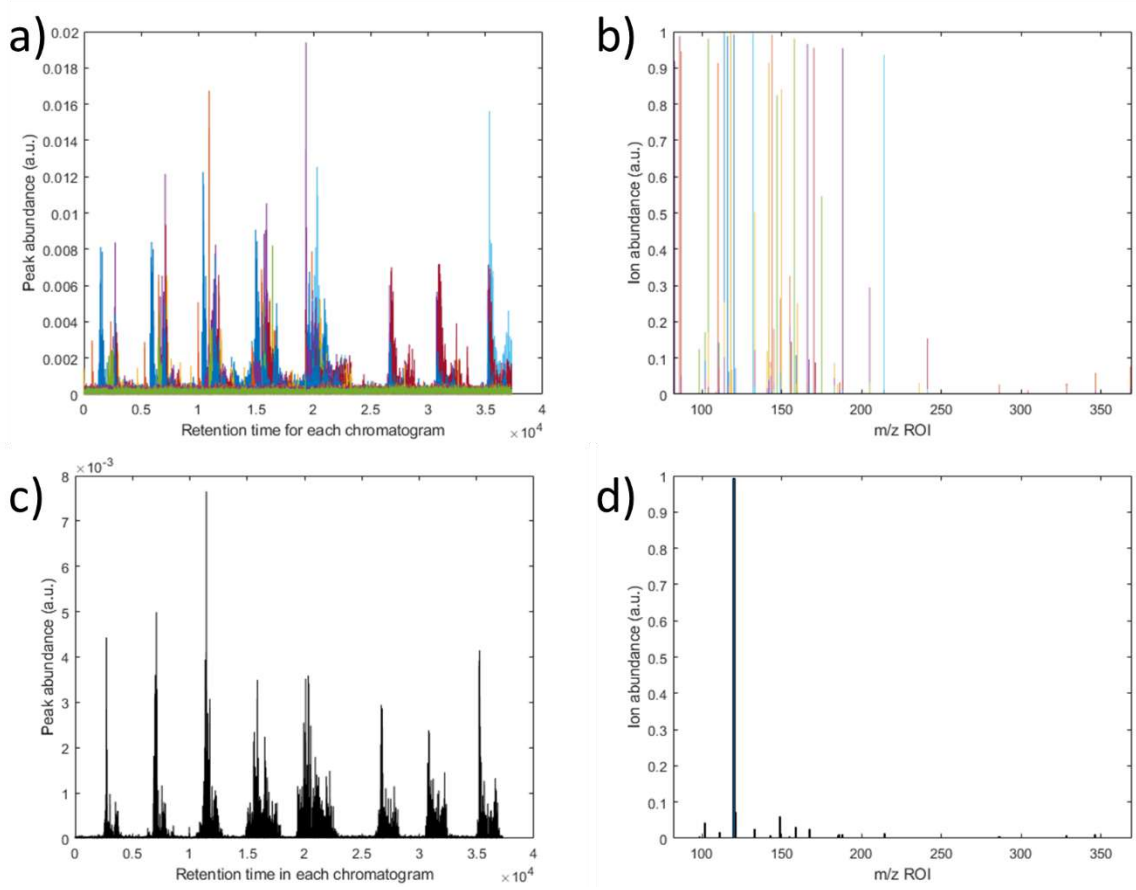


Figure S2. MCR-ALS analysis of concatenated $\mathbf{D_{aug}}$ matrix with chromatographic data of the amino acids standards mixtures at five concentrations and the three replicates of **Sample B**: **a)** elution profiles (**C**); **b)** mass spectral ROIs resolved components, ($\mathbf{S^T}$); **c)** elution profile (**c**) and pure mass spectral ROI resolved component ($\mathbf{s^t}$) corresponding to L-threonine standard (119.9507 *m/z*).

3 MSROI and MCR-ALS based quantification strategies

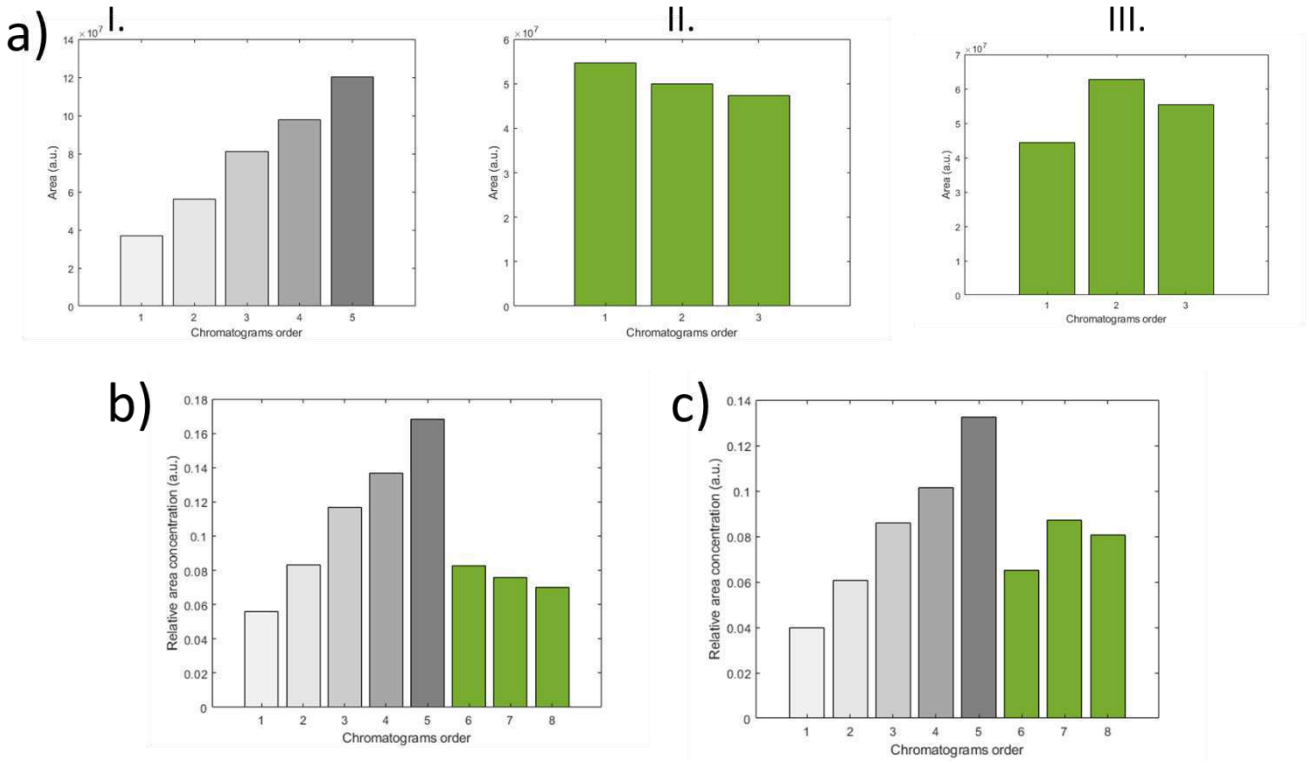


Figure S3. Peak areas of L-threonine standard (119.9507 *m/z*) for: a.I) MSROI analysis of the five amino acids standards mixtures LC×LC-MS data, a.II) the three replicates of **Sample A** LC×LC-MS data, and a.III) the three replicates of **Sample B**; b) MCR-ALS (nn) area analysis of concatenated five amino acids standards mixtures (chromatograms 1 to 5) and the three replicates of **Sample** A (chromatograms 6 to 8) LC×LC-MS data; c) MCR-ALS (nn) area analysis of concatenated five amino acids standards mixtures (chromatograms 1 to 5) and the three replicates of **Sample B** (chromatograms 6 to 8) LC×LC-MS data.

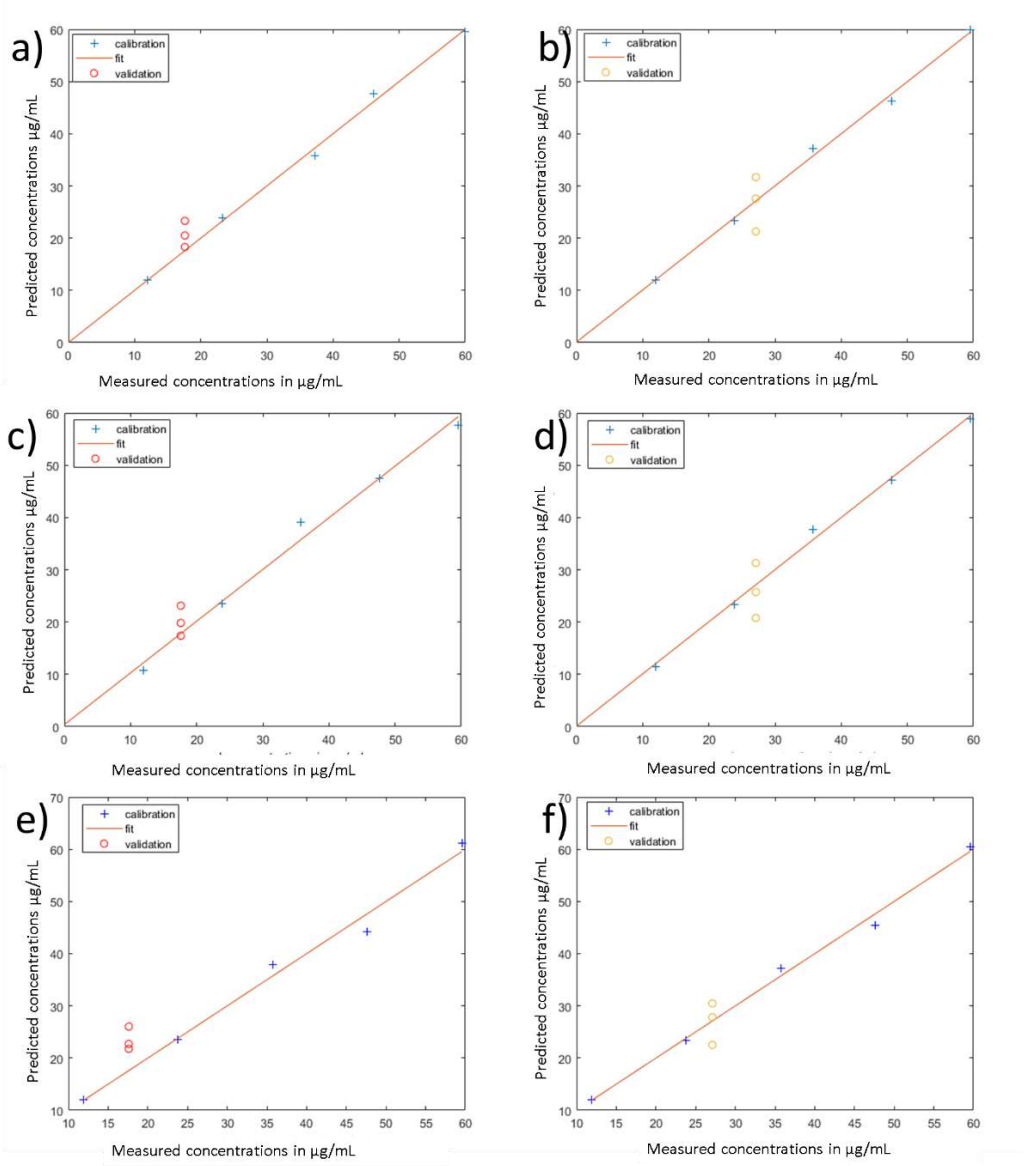


Figure S4. a) Plot of the MSROI area predicted L-threonine (119.9507 *m/z*) amino acid versus its measured (nominal) concentration for the **Sample A**; b) Plot of the MSROI area predicted L-threonine (119.9507 *m/z*) amino acid versus its measured (nominal) concentration for the **Sample B**; c) Plot of the MCR-ALS (nn) area predicted L-threonine (119.9507 *m/z*) amino acid versus its measured (nominal) concentration for the **Sample A**; d) Plot of the MCR-ALS (nn) area predicted L-threonine (119.9507 *m/z*) amino acid versus its measured (nominal) concentration for the **Sample B**; e) Plot of the MCR-ALS (nn + acc) area predicted L-threonine (119.9507 *m/z*) amino acid versus its measured (nominal) concentration for the **Sample A**; f) Plot of the MCR-ALS (nn + acc) area predicted L-threonine (119.9507 *m/z*) amino acid versus its measured (nominal) concentration for the **Sample B**.

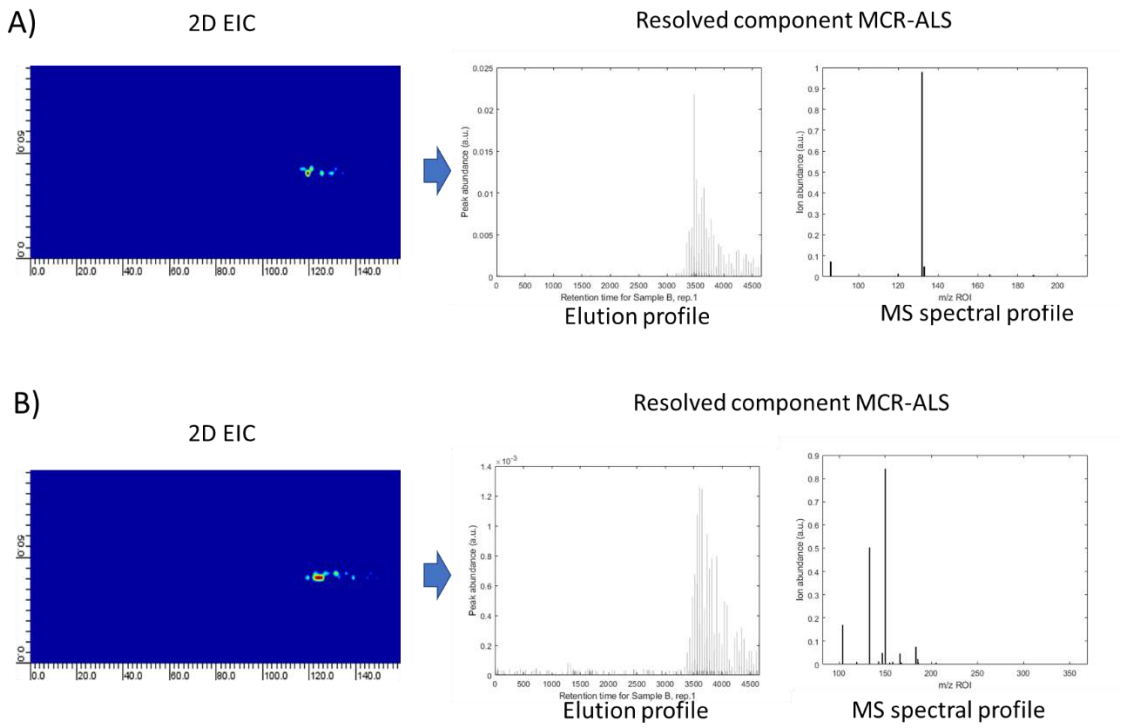


Figure S5. Two-dimensional Extracted Ion Chromatograms (EICs) obtained with the LC×LC Edition Software from GCImage, LLC (Lincoln, Nebraska), and resolved MCR-ALS elution and spectral profiles (after application of the area correlation constraint) for a) L-leucine, and b) L-methionine.

4.3 Discussion

This section discusses the results obtained in **scientific publications V and VI**. On the one hand, the discussion focuses on the analytical challenges faced in developing an LC × LC method for the untargeted analysis of different metabolites (including lipids). On the other hand, the challenges associated with the quantification of LC × LC datasets are also discussed.

4.3.1 LC × LC method development

LC × LC challenges have been extensively reported in recent reviews [7,8,12–14]. Firstly, samples suffer an extra dilution when a fraction from the ^{1}D separation accesses the ^{2}D column and is mixed with the ^{2}D mobile phase. Sensitivity is reduced compared with 1DLC, (where the fractions from the only separation process enter into the detector straightforwardly). Nevertheless, if the number of fractions is decreased in LC × LC (i.e., to reduce sensitivity loss), the ^{2}D column can be overloaded (e.g., fraction volumes higher than the ^{2}D column dead volume), and poor peak shapes can be obtained.

Secondly, the solvent compatibility between both dimensions should be considered. If too much of a strong solvent coming from the ^{1}D separation is introduced in the ^{2}D column (i.e., the strongest eluent from the mobile phase, in this case, regarding the ^{2}D separation), the less retentive analytes may not be enough retained, producing breakthrough and peak distortion.

Thirdly, the frequency of the collection of ^{1}D fractions needs to be sufficient because peaks already separated in the ^{1}D column cannot be joined again (i.e., undersampling). However, increasing the collection frequency implies that the total analysis time also augments, consequently. One important drawback associated with LC×LC separations is the acquisition of long chromatograms (that could be up to two hours in the case of complex samples, depending on the analytes and conditions tested). Therefore, a compromise in the collection frequency and total analysis time needs to be achieved.

Different analytical chromatographic strategies have been developed for overcoming these issues, mainly based on improvements in how fractions are transferred from the ^{1}D column to the ^{2}D column [7,8,15]. Among them, Active Solvent Modulation (ASM) has been tested in this PhD Thesis for metabolomic studies. The

ASM procedure is based on a single valve-based approach that aims to deal with the effect that the ^{1}D effluent has on the ^{2}D separation [8]. The flow from the ^{2}D pump is split into a pre-established ratio at the beginning of each modulation (i.e., ASM step), in accordance with the dilution needed to enhance solvent compatibility between both dimensions. While a part of the flow goes to the ^{2}D column, the other part goes through a bypass capillary, accesses the loop and mixes with the ^{1}D effluent. Consequently, each ^{1}D fraction is diluted with the ^{2}D mobile phase before entering the ^{2}D column, reducing solvent mismatch and improving retention and peak shapes. For instance, for the RP×HILIC set-up, the polar fraction in the ^{1}D (e.g., water content) is reduced before reaching the ^{2}D separation, whose strong eluent is the polar fraction (in this example, water). Furthermore, as the mismatch between mobile phases is reduced with ASM, larger sampling volumes can be injected into the ^{2}D without risking peak shapes (i.e., bigger loops can be employed), enhancing the sensitivity of the analysis and decreasing run time. Besides, after the ASM step, the ^{2}D flow enters fully into the ^{2}D column, avoiding baseline disturbances (compared to previous approaches such as in the Fixed Solvent Modulation) [8]. In this PhD Thesis, ASM was tested to analyse lipids and metabolites, in RP×HILIC and HILIC×RP approaches, respectively. Its usefulness in method development and improvements in peak separation are discussed below.

The choice of RP×HILIC for lipidomic studies

In this PhD Thesis, two possible set-ups were employed to analyse of lipids, RP×HILIC and HILIC×RP, using ASM as modulation interface between both dimensions. This study was carried out during a three-month research stay in 2019 at the Gustavus Adolphus College (St. Peter, Minnesota, USA), under the supervision of Prof. Dwight R. Stoll, who is one of the leaders worldwide in two-dimensional liquid chromatography.

RP and HILIC are the most common stationary mechanisms selected for lipidomic studies [16–18]. On the one hand, RP has been the classic stationary phase for the analysis of lipids due to its high separation efficiency for non-polar compounds, reproducibility and robustness. In RP, lipids are separated by the hydrophobicity of their fatty acyl chains, and their different number and positions of the double bonds. On the other hand, HILIC is a relatively new approach to normal phase chromatography which benefits from using water-miscible solvents, usually

the same employed in RP, but in the opposite proportions. HILIC separates lipids by families, according to the hydrophilic part of their molecules (i.e., the polar head groups). In RP × HILIC, the stress in the separation is on the hydrophobic part of the lipids, provided by the RP dimension, whereas HILIC provides a quick screening by lipid families. In the opposite scenario, HILIC × RP, lipids are thoroughly grouped by families, which facilitates their identification.

Several examples from the literature have selected both set-ups for lipid analysis in different matrices. RP × HILIC has been chosen for the study of lipids extracts from human plasma [19,20], porcine brain [20], rice [21] and zebrafish embryos samples [22]. In contrast, HILIC × RP has been used for studying lipid profiling in mussels [23], zebrafish embryos [22], and mouse brain [24]. An HILIC × RP set-up followed by an ozonolysis step was employed for characterizing egg yolk [25] and rat liver phospholipid extracts [26]. Both set-ups present advantages and disadvantages, as it will be discussed.

In LC × LC, the ^{1}D separation is normally the longest but the slowest. It usually operates at very low flow rates (in the µL mL^{-1} range instead of mL min^{-1}) determined by the whole duration of the ^{2}D separation and the loop size. In contrast, the ^{2}D separation is very short, usually about 1-2 minutes long. The capability of the ^{2}D separation to fully analyze each fraction from the ^{1}D separation (before the subsequent fraction reaches the ^{2}D column) limits the whole duration of the LC × LC method. Therefore, high flow rates are highly desirable in the ^{2}D separation. However, in the case of MS detection, a flow split after the column and prior to the detector may be required to enhance ionization. IOn the other side, a short RP separation may not be powerful enough to discriminate between similar compounds. Hence, for lipid analysis, the combination of a first HILIC and then RP fits better than with other approaches like heart-cutting or stop-flow, where the length of the ^{2}D separation is not as critical as in comprehensive mode [20]. An exhaustive comparison between C18 × HILIC, HILIC × C18 and HILIC × PFP (where PFP is a pentafluorophenyl column, a specific type of RP stationary phase) demonstrated the superiority of the C18 × HILIC set-up for the analysis of the zebrafish embryos lipidome [22]. In the mentioned study, the C18 × HILIC configuration provided a higher effective separation power, although the total analysis time was considerably higher (170 min in C18 × HILIC *versus* only 100 min for the other two combinations).

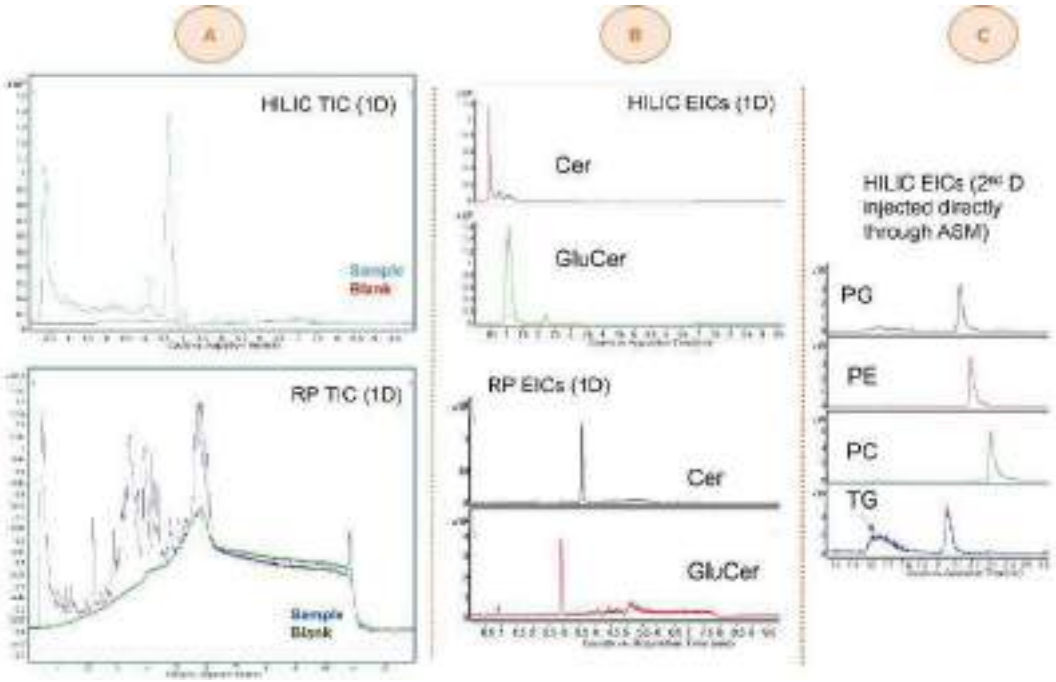


Figure 4.3. A) Comparison between TICs obtained with a bare silica column (top) and a C18 column (bottom) separations of zebrafish embryos phospholipid extract, performed in 1DLC conditions. Preliminary tests indicate that the resolving power of C18 is higher in the analysis of the studied lipids. **B)** Extracted ion chromatograms (EICs) obtained for Cer and GlucCer standards injected in the 1DLC methods corresponding to the TICs in the left. RP separation is better than HILIC for these compounds. **C)** Lipid separation by families provided by HILIC as the second dimension. The results are EICs for PG, PE, PC and TG standards. Data not published. More details about the experimental conditions in which these chromatograms are in **Annex 1**. TIC: total ion chromatogram; EIC: extracted ion chromatograms; RP: reversed phase; HILIC: hydrophilic interaction; Cer: ceramide; GlucCer: glucosylceramide; PG: phosphatidylglycerol; PE: phosphatidylethanolamine; PC: phosphatidylcholine; TG: triacylglycerol.

The results we obtained in the preliminary tests carried out during the 2019 research stay agree with the results obtained by Xu *et al.* [22], as well as by Holčapek et al. [20]. **Figure 4.3.A** shows the total ion chromatograms (TICs) collected with a commercial C18 column *versus* a bare silica column (packed manually in-house) for a similar analysis time and the same sample (zebrafish embryos phospholipid extract). Under the conditions tested (see **Annex 1** for more details), C18 separation presents a higher resolution power. Similar results were found in the analysis of sphingolipid extracts (data not shown). Besides, a detailed comparison between different lipid families in both conditions was performed. As a general conclusion, lipid peak shapes were narrower and intensities were higher in the RP mode, except in the case of *lyso* families. These *lyso* forms of the main phospholipids were globally

more intense in HILIC and showed better retention than in RP. The reason is a diminution of their hydrophobicity compared to phospholipids (i.e., lyso forms contain only one fatty acid chain instead of two).

Another problem with the tested HILIC conditions was that some sphingolipids were barely retained. **Figure 4.3.B** exemplifies this issue for the extracted ion chromatograms (EICs) of ceramides (Cer) and glucosylceramide (GlucCer) standards, injected in the same conditions as TICs in **Figure 4.3.A.** Only sphingomyelins (SM) presented an acceptable retention behavior in these HILIC conditions. Other studied lipids, such as triacylglycerides (TAG) and diacylglycerides (DAG), exhibited the same behavior as the shown sphingolipids. In contrast, RP separation of these compounds was generally more appropriate as ^{1}D mechanism, in terms of higher retention and better peak shapes and widths. On the other side, HILIC seems to be a good alternative as the ^{2}D, due to the extra selectivity provided by the differentiation of isomers by their polar head groups. **Figure 4.3.C** shows how HILIC is capable of separating different lipids according to their families in a very short period of time. The EICs presented in this PhD Thesis (an example of a phosphatidylglycerol (PG), a phosphatidylethanolamine (PE); a phosphatidylcholine (PC); a triacylglycerol (TG)) were obtained by injecting the standards directly through the ASM valve and in the solvent proportions of the ^{1}D effluent, simulating ^{2}D conditions (see **Annex 1** for more details in the employed chromatographic conditions).

The improvements due to the use of ASM in RP × HILIC for lipidomic studies

One of the major improvements due to the use of ASM in the developed RP × HILIC method is the increased sensitivity, compared to previous methods from the literature. An indicator of this enhancement is the amount of ^{2}D effluent that finally accesses the detector, in this case, the MS. In this work, the split ratio was of one effluent part to MS, and two effluent parts to waste (1:2). Consequently, a much higher fraction of the ^{2}D effluent accesses the MS than in previous examples using the same RP × HILIC configuration (e.g., 2:8 [22], 8:100 [27], 1:3 [19]. Another advantage of the method proposed in the **scientific publication V** is the considerable reduction in the total analysis time achieved, with only 120 min per sample (other examples studies required 170 min [22], 150 min [20], 190 [19], 130 min [21]. This diminution in the total run time was possible due to the use of 40 μL loops, identical

to the work by Xu [22]. However, bigger sampling volumes were collected and injected into the ^{2}D compared to the works by Holčapek [28], and Baglai., where 20 μL loops were preferred [19]. In the case of the work by Navarro-Reig *et al.* [21], the loop size was set to 70 μL. Nevertheless, higher fraction volumes than those safely established by the optimization with the ASM, produce undersampling risks, and may lead to mixing again the peaks already separated in the ^{1}D separation. In addition, further tests are required to fully validate the developed RP×HILIC method and provide a quantitative comparison of the different methods in terms of resolving power.

The final separation achieved with the RP×HILIC method using ASM at a dilution factor of 5 and 40 μL loops is depicted in **Figure 4.4,** which includes a TIC showing the spatial distribution of lipid standards from different families in the 2D space. The experimental conditions in which this chromatogram was acquired are included in Pérez-Cova *et al.* [9], **Annex 2**, and the **scientific publication V** where this RP×HILIC method was used with minor modifications in the re-equilibration step.

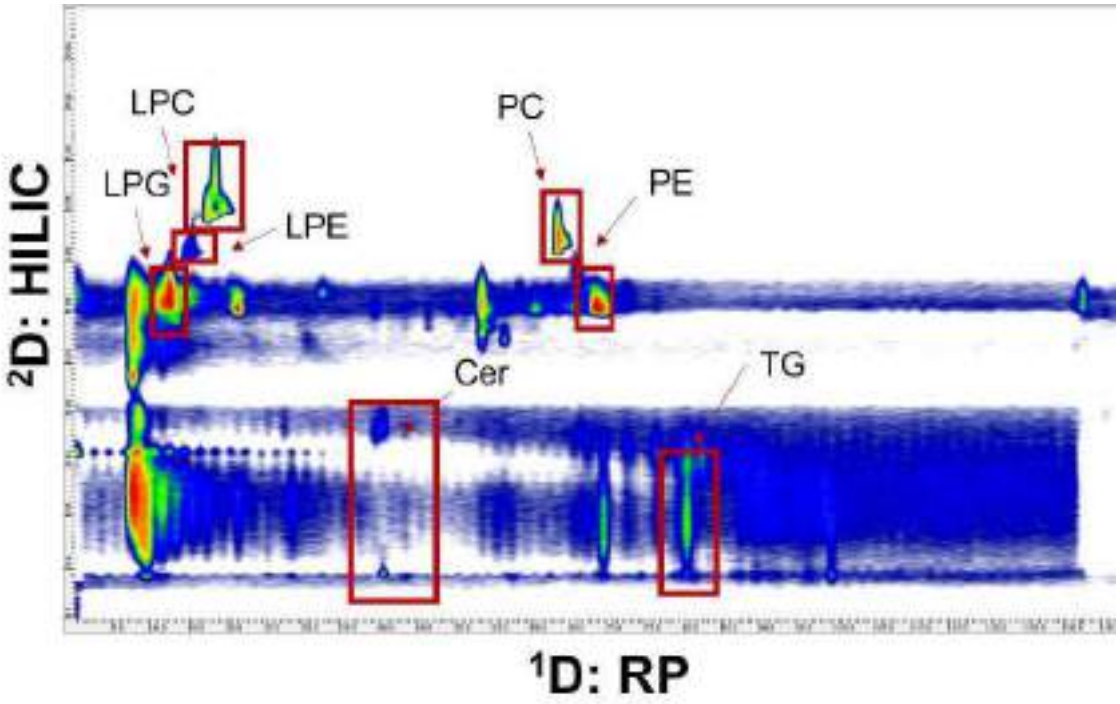


Figure 4.4. 2D representation of the final separation conditions selected for the RP×HILIC method for the untargeted analysis of zebrafish embryos lipidome. Experimental conditions of acquisition in **Annex 2**. RP: reversed phase; HILIC: hydrophilic interaction; Cer: ceramide; LPG: lysophosphatidylglycerol; PG: phosphatidylglycerol; LPE: lysophosphatidylethanolamine; PE: phosphatidylethanolamine; LPC: lysophosphatidylcholine; PC: phosphatidylcholine; TG: triacylglycerol.

A HILIC × RP separation for metabolomic studies

Both HILIC and RP have been widely employed to analyse polar and medium/non-polar metabolites, respectively [17,29,30]. Exhaustive comparisons have been performed between different analytical platforms (e.g., 1DLC employing HILIC, RP and other stationary phases). However, only one retention mechanism does not provide full coverage of the metabolome. Hence, a possible solution would be a multi-platform approach for the analysis of a broader number of compounds with different properties [31–34]. The proposed LC×LC method in this PhD Thesis seems an appealing alternative to this multi-platform strategy, because the information obtained with the multiple conditions would be ideally acquired simultaneously in the same analytical run.

The RP × RP configuration using complementary separations is commonly selected for the metabolite profiling of medium and non-polar metabolites [35–37] or the analysis of specific non-polar metabolites such as anabolic-steroids [38]. Nevertheless, polar metabolites may not be retained and can be lost during these separations. Hence, the combination of HILIC and RP could provide a broader metabolome coverage. Analogously to lipids, the retention order will depend on the target compounds. If the stress is in separating first by polarity rather than hydrophobicity, HILIC is recommended as the ^{1}D. In this case, the set-up HILIC × RP was preferred.

Some examples of applications of HILIC × RP to metabolomics include the untargeted analysis of rice [39] and licorice extracts [40], and the specific analysis of polyphenols [41], phlorotannins [42] and anthocyanins [43]. The major challenge in using this set-up for the analysis of metabolites is solvent compatibility. HILIC separation often starts with a high percentage of organic solvent (i.e., nearly 100%), that is the strongest solvent in RP. Thus, this solvent mismatch is usually more critical than in the reverse set-up configuration (at least for lipid analysis). One of the goals of the 2019 research stay was to test ASM in a HILIC × RP configuration and determine whether it could help to improve solvent compatibility. From some preliminary trials and under the tested conditions, it was found that the maximum dilution ratio provided by ASM (ASM factor 5) may not be enough to ensure good retention in RP (as ^{2}D) for the analytes tested (i.e., polar metabolites). **Figure 4.5.A** presents a TIC of zebrafish embryos metabolome extract using a HILIC×RP method (using a carbon column in the ^{2}D and with 40 μL loops; see **Annex 3** for more

details). As seen in the Figure, most metabolites eluted during ASM step (before starting the actual gradient). **Figure 4.5.B** compares the retention between a carbon column and a C18 column under similar conditions. EICs were acquired by direct injection into the ^{2}D through the ASM valve. For the tested conditions (see **Annex 3**), the carbon column provided a better separation (e.g., peaks more retained, sharper and narrower) than C18. This is the reason why a carbon column was selected for the 2D tests (e.g., **Figure 4.5.A**). However, more studies are required to evaluate the suitability of ASM and the best RP retention mechanism in metabolomic studies.

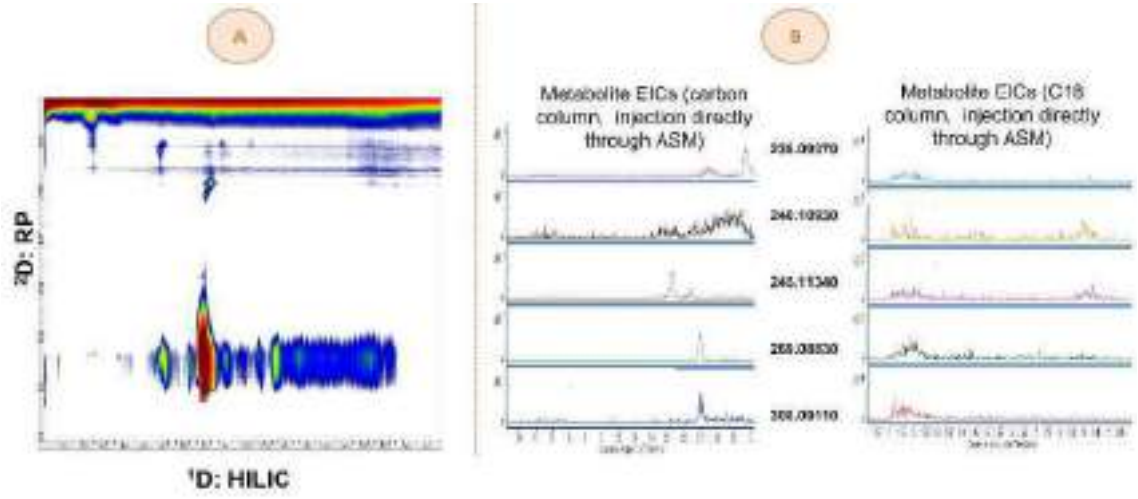


Figure 4.5. A) Preliminary test of a HILIC×RP method applied to the analysis of zebrafish embryos metabolome extracts. Most of metabolites present in the samples were not sufficiently retained in the ^{2}D and eluted during the ASM step. **B)** Comparison of the retention obtained with a carbon column (left) and a C18 column (right) for the same metabolites (numbers in the middle of the EICs correspond to the *m/z* values of the compounds shown). Sharper and more retained peaks were achieved with the carbon column.

Other options for improving solvent compatibility would be different active modulation strategies such as “at-column dilution” [44] or “trapping columns” [38]. Alternatively, it is possible to add water with an extra third pump at a constant flow rate for diluting the organic content from the ^{1}D before accessing the ^{2}D column [39]. In the methodology employed in **scientific publication VI,** a reduced loop size (30 μL) was adopted to reduce the fractioning of the sample from the ^{1}D column when it is introduced into the ^{2}D column. Consequently, the total analysis time increased considerably (160 min per run, including a re-equilibration step of half an hour). Nevertheless, it is important to remark that the objective of the mentioned publication was not the development of a novel LC×LC methodology with increased solvent compatibility, but rather to explore a new chemometric strategy that allowed

the quantification of multiple analytes in the presence of interferences and overlapping signals. This approach will be discussed in the following section.

4.3.2 Chemometric developments for LC×LC

This section focuses on pre-processing and quantification steps of LC×LC data. First, the regions of interest (ROI) approach is suggested for spectral compression and signal filtering. This strategy can be employed for both untargeted and targeted analysis, as it has been demonstrated in **scientific publications V and VI**, respectively. Second, a combination of regions of interest with multivariate curve resolution least squares (with and without an area correlation constraint) is proposed for quantification purposes, taking the analysis of amino acids samples as a case of study.

Pre-processing strategy for LC×LC-HRMS datasets based on the regions of interest approach

ROI approach, discussed in detail in **Chapter 3**, was applied to the datasets obtained by analyzing zebrafish embryos lipidome extracts with the RP×HILIC-HRMS optimized method. The entire experimental design is detailed in **scientific publication V**. Briefly, zebrafish embryos were exposed to two EDCs (BPA and E2), at three concentration levels of exposure each (Low, Medium and High doses, plus Control samples, without EDC), collected at two days (4 and 6 days post-fertilization) and extracted following two protocols (general extraction and sphingolipid-based extraction).

As already described in **scientific publication V**, the ROI strategy could not be directly applied on *.mzXML* transformed files due to their big size (12-13 GB per each raw file; higher than 20 GB when converted directly to *.mzXML* format using peak picking). Thus, a filtering step was performed during their transformation with the MSconvertGUI tool of the Proteowizard suite, to reduce noise (prefilter of absolute intensity signals higher than 100). Then, these filtered *.mzXML* files (0.7-1 GB each) were imported into MATLAB and compressed simultaneously with the MSroi GUI app [9] by sets of samples (example of one set: general extraction batch of bisphenol A exposure all concentrations levels at day 6; 18 samples per set, including QCs; 8 sets in total). For each set (15-20 GB per set), the importing time was less than 2 hours.

In a previous study from our research group [39], LC × LC-HRMS datasets were first compressed in the spectral dimension with the ROI approach and further reduced in the time dimension using wavelets, followed by a time-windowing step that divided the chromatograms into three regions that were analyzed separately (with some overlapping between the regions to avoid any loss of information derived from the cutting). Then, MCR-ALS was applied by triplicate, according to each 'superaugmented' data matrix for each of the three selected time windows. This MCR-ALS step aims to resolve the chromatographic and spectral profiles, to get the qualitative (i.e., through mass spectra for identification purposes) and quantitative (i.e., through an area matrix) information about the samples. Several components are resolved, corresponding to the potential sample constituents. Ideally, all the MS adducts and isotopic forms from the same molecule are joined into the same MCR component. This resolution process can be called, at present as, 'componentization' [45]. A number of potential features can be lost between the ROI step (the ROI intensity matrix is the input for MCR-ALS) and the output of MCR-ALS (a matrix containing the areas of each component, plus their elution and spectra profiles). These discarded features are considered noisy signals or minor instrumental contributions (e.g., from background or solvent) [39], and they are moved to residuals. This may happen, for instance [21,39], when the intensity cut off is very low, obtaining up to 1000 features per set, but only 50-100 components are considered in the MCR-ALS calculations. Careful examination of the original ROI data and MCR-ALS residuals is therefore recommended.

The detailed selection of the relevant variables (i.e., potential markers) for the specific study is performed based on MCR-ALS components from. The aim of **scientific publication V** was to assess the reliability of this ROI strategy, without including the MCR-ALS resolution step afterwards, looking for the number of potential markers of the studied exposure.

In **scientific publication V**, optimization of the ROI parameters was investigated, trying to select a manageable number of features and a reasonable analysis time for the compression of the data. Different conditions are tested at the beginning of the ROI procedure each time new datasets are analyzed. Although some of the parameters can be used routinely when the same type of data acquisition is employed (i.e., mass error), others are data dependent (i.e., intensity threshold). Lipids standards added to the samples (for ensuring data quality in terms of

extraction and instrumental response) can help to optimize ROI parameters. The final ROI conditions need to include the *m/z* values corresponding to the standards, and the intensity thresholds are adjusted consequentially in the first trials. These standards are also useful for testing mass error factor if a new instrument is used. It is important to notice that selecting an appropriate signal-to-noise (S/N) ratio (established by intensity threshold and other factor parameters in the ROI step), especially in untargeted analysis. S/N will limit the whole pre-processing step duration and the number of potential features. Lower threshold values require longer analysis times (e.g., more than one day for large LC × LC-HRMS datasets), whereas too high threshold values can lose relevant features. When optimizing ROI parameters, a compromise between an adequate S/N ratio (including as many relevant features as possible but avoiding noisy contributions) and an acceptable pre-processing time should be achieved. For this purpose, a *minmax* signal factor parameter can be very handy. It is a multiplication factor that allows locating the intensity threshold very low, but only considering those ROIs (i.e., potential features) whose intensities are above the product of the threshold and this factor (e.g., intensity threshold x 5). The parameters selected in **scientific publication V,** (untargeted analysis) produced more than 500 features in the simultaneous analysis of all the data sets (i.e., potential features) with an ROI step shorter than two hours per set. Hence, in **scientific publication V**, only the ROI step was used to achieve an acceptable reduction in size for each dataset. In this case, peak areas were obtained directly from the ROI features, instead of from the MCR-ALS resolved components, with the aim of investigating and validating the use of ROI for providing relative quantitative information directly.

A preliminary study of the datasets obtained with zebrafish embryos was performed using a pseudo-targeted strategy. A list of the *m/z* values of the lipids already detected in positive ionization mode in zebrafish embryos in a previous work by Martínez *et al.* [46] were selected and introduced using the targeted mode of the MSroi GUI for their analysis. Only the batches corresponding to 'general extraction' were considered because the extraction procedure employed was the same as in the work by Martínez. In this way, their comparison was possible. The proposed pseudo-targeted analysis validated the ROI procedure for LC×LC-HRMS using those lipids already annotated in the LC-HRMS analysis of the same samples [46]. It was also ensured the adequate data normalization during the untargeted analysis, as the trends in pseudo-targeted sets should match the untargeted ones (i.e., both are the

same data). Besides, this preliminary approach facilitated the compound annotation in the final stage of the analysis (i.e., untargeted mode) of the lipids obtained in the 'general extraction batches'.

A comparison of the pseudo-targeted and the fully untargeted results is summarized in **Figure 4.6**. Principal component analysis (PCA) score plots show the same trends considering both approaches. When all samples are analyzed together, the separation is based on the collection day and the considered EDC (**PCA 1**). Four clusters are distinguished (BPA-4dpf, BPA-6dpf, E2-4dpf, E2-6dpf). Regarding the exposure concentration, the samples are ordered from Control to High in both approaches (an example is shown on BPA exposure at 4dpf; **PCA 2**). ANOVA Simultaneous Component Analysis (**ASCA**) results were also coincident. For instance, at 6 dpf, both EDC presented *p*-values lower than 0.05 regarding the pairs Control vs High dose or Control vs Medium dose. An overlook of the relevant variables (VIPs > 1) obtained through partial least square discriminant analysis (**PLS-DA**) models indicated an increase in the specific lipids related to the exposure, which could allow better differentiation between treatments. An example is also shown for Control vs High dose at 4dpf for BPA exposure, expressing the matching lipids through Venn diagrams [47] (https://bioinformatics.psb.ugent.be/webtools/Venn/). Lipid annotation in the untargeted analysis was performed in two steps: 1) using the information about the pseudo-annotated compounds in the targeted approach, and 2) based on matches between experimental MS/MS spectra and the theoretical MS/MS spectra from the MS-DIAL library [48]. Some examples of the lipids with an important role in BPA exposure detected by both data analysis strategies (pseudo-targeted and fully untargeted) were LPC(18:0), PC(32:0), PC(34:4), PC(34:2), PC(36:6) and TG(56:3). The use of the fully untargeted strategy also allowed the identification of the relevant lipids from other lipid classes, such as ceramides (e.g., Cer(34:1), Cer(42:2), Cer(34:0)), or from the same classes as those considered in the pseudo-targeted analysis but not included in the targeted list (e.g., PC(42:8), TG(58:6), SM(32:2), PC(33:4)).

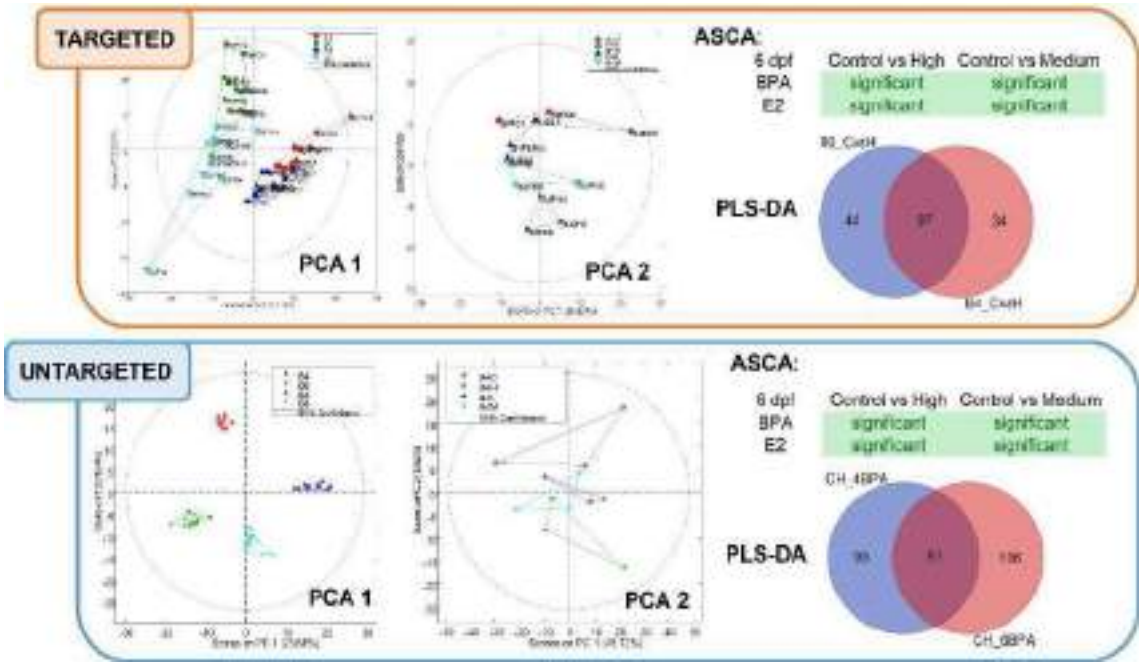


Figure 4.6. Comparison of results obtained with the pseudo-targeted and the fully untargeted data analysis approaches. Both used ROI as compression step, but in the first case, there was a prior selection of *m/z* values associated with lipids already reported in zebrafish embryos, whereas in the second case, the selection of the variables was performed based mainly on their intensity, without any *a priori* assumptions. **PCA 1**: BPA and E2 exposure separated by EDC and by day. **PCA 2**: BPA exposure day 4 separated by concentration level from Control to High. **ASCA**: Both doses High and Medium are significant for both EDC at 6 dpf. **PLS-DA**: An increase in the number of specific lipids for each treatment with the untargeted approach.

In conclusion, the use of the ROI strategy was validated for both pseudo-targeted and untargeted analysis of LC×LC-HRMS datasets, allowing both filtering and compression without requiring any signal or peak alignment. Post-processing (e.g., exploratory, classification or statistical analysis) was performed directly on the areas obtained from ROI, as no resolution step was applied in this case. Special attention is needed in order to discard different adducts or isotopic forms for the same compound, since a large number of potential features can be obtained from each of them. Lipid annotation is still a major challenge, but combining the information of retention times from both dimensions and the accurate mass and MS/MS spectra, the annotation step (see **Figure 4.7**) becomes more reliable, especially for isobaric compounds discrimination. MS-DIAL software [48] is a useful tool to match theoretical spectra with deconvoluted experimental MS/MS spectra thanks to its

incorporated large MS/MS library. However, it is important to keep in mind that when analyzing data from LC × LC analysis, the same compound can result in several hits between experimental and theoretical spectra due to different retention times associated with the subsequent modulations, which are perceived as different compounds by the software. Besides, when using MS-DIAL software to get the quantitative information from the peak areas, peak integration is performed on a basis of one-dimensional chromatography. Therefore, each ^{1}D fraction is considered as an individual peak. Thus, peak integration is more reliable with the ROI strategy, in which the subsequent modulations of the same compound are simultaneously considered in the same peak ROI. However, a novel strategy based on demodulating the list of features obtained using MS-DIAL on the LC × LC chromatograms has recently been developed [49]. This approach provides the sum of the areas of the different modulated points for a certain feature at the retention time where the maximum intensity was found for each *m/z* value, overcoming the current limitations of MS-DIAL for LC × LC analysis.

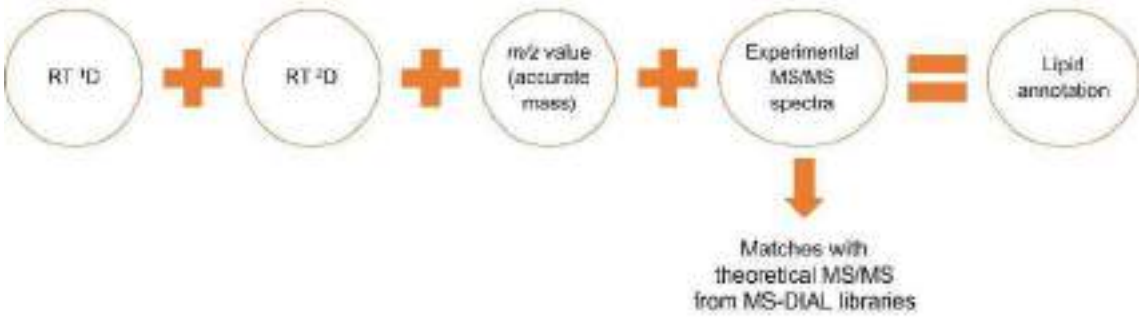


Figure 4.7. Information provided by the LC × LC-HRMS pipeline including MS/MS analysis, which contributes to an easier annotation step.

Quantification approaches for LC × LC targeted analysis based on ROIMCR

Three quantification strategies were evaluated in **scientific publication VI**. All three strategies present the particularity that no vendor software is required and the advantage that no prior chromatographic alignment nor peak shape modelling steps are needed (contrary to most multidimensional chromatography data analysis strategies). The areas used for building the calibration curves were obtained at three different stages of the ROIMCR pipeline, and derived quantification results were

compared. The complexity of the quantification procedures proposed increases with each step added, but the reliability and accuracy of the results are also improved (see **Figure 4.8**). The first approach is based only on the direct application of the ROI approach. In **scientific publication V**, the ROI approach was evaluated for the first time for relative quantification purposes in untargeted analysis). In contrast, in **scientific publication VI,** the capabilities of the ROI procedure for absolute quantification in targeted analysis without the use of MCR were assessed for the first time. Thus, the areas of each ROI feature were obtained in a straightforward manner from the simultaneous integration of all the chromatographic peaks from the modulations subsequently associated with the same compound (i.e., related to the same ROI). Regression correlation coefficients of the calibration curves ranged from 0.974 to 0.999 for the amino acids in the test mixture. Relative errors (RE) were lower than 10% in the calibration mixture and lower than 25% in tested samples. Therefore, the quantitative information recovered using this direct ROI strategy was considered acceptable.

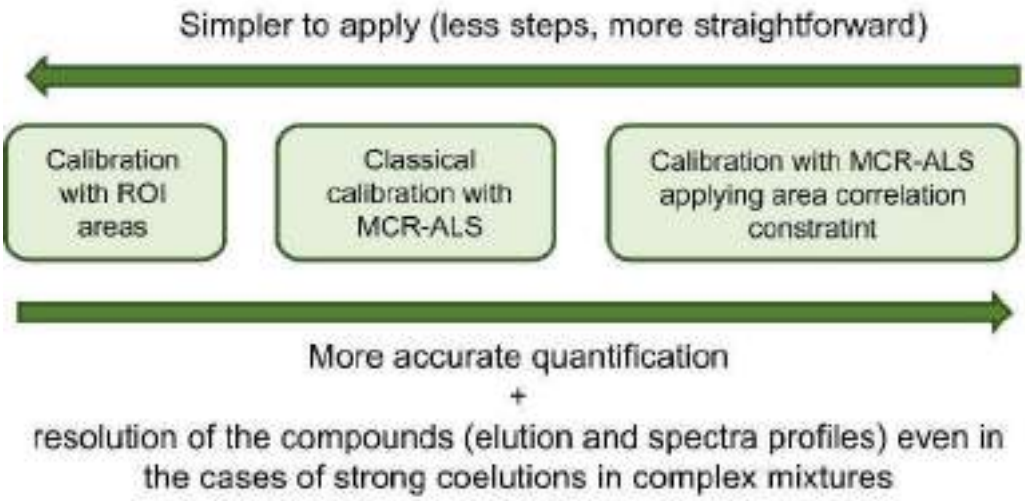


Figure 4.8. Summary of quantification strategies based on ROIMCR pipeline.

In chromatographic scenarios where strong coelutions between two or more sample compounds exist with even identical or very close *m/z* values, these compounds may be integrated into the same ROI, and their independent analysis may not be feasible. Although 2DLC resolution power is considerably increased in comparison with 1DLC, all compounds may not be completely resolved, especially

when very complex samples are analyzed. One of the main advantages of MCR-ALS quantification strategies is that compounds with strongly overlapping elution profiles and very similar mass spectra can be unambiguously resolved [50]. The ROIMCR quantitative applications to LC-MS datasets have already been demonstrated and validated in the previous work by Dalmau *et al.* [50] and to gas chromatography coupled to mass spectrometry (GC-MS) datasets by Pourasil et al [51]. In Dalmau's work, the strategy was first evaluated in a mixture of lipid standards with a targeted analysis, from both identification and quantification points of view. The calibration curves were built using the peak areas of the elution profiles of the lipids resolved by MCR-ALS and the known concentrations of the standards as reference values. This calibration curve was then applied to the quantification of lipids in biological samples, also resolved by the ROIMCR method. RE values were lower than 20% in all cases, and regression coefficients ranged from 0.9908 to 0.9997 for lipid standard mixtures, and from 0.9723 to 0.9978 for cell samples. The quantification strategy employed by Dalmau was also used in **scientific publication VI**, referred to as "classical MCR-ALS quantification". LC×LC results were similar to those previously achieved for LC-MS datasets (i.e., regression coefficients from 0.978 to 0.998, and RE again lower than 10% in the calibration mixture, and lower than 25% in both samples tested).

During the iterative ALS optimization step in MCR, the use of mathematical constraints provides chemical significance to the pure numerical solutions, while the rotational ambiguities inherent to bilinear models are reduced [52]. For instance, non-negativity constraints are used in both LC-MS and LC×LC-MS because elution and spectra profiles have only positive values. Another possible constraint (i.e., area correlation constraint) has been suggested for the analysis of second-order multivariate calibration data [53,54]. This constraint presents some advantages compared to classical MCR-ALS calibration. It improves the quantification in the presence of strong unknown interferences (not found in the calibration sets) in the real samples, which is especially interesting in the analysis of samples with complex matrices [11,55]. Besides, the application of this constraint may reduce rotation ambiguities and may improve the accuracy of the quantification [53]. An exhaustive comparison of results obtained only with non-negativity constraint and with area correlation constraint was performed by Bayat [53]. Higher regression coefficient values and lower RE values were reported for all analyzed compounds when area correlation constraint was applied (only non-negativity: 0.938-0.990 and 4.5-20.8%; area correlation constraint: 0.975-0.999 and 4.5-10.9, for regression coefficients and

RE % ranges, respectively). Although the improvement was not so high in **scientific publication VI** in the compound-by-compound comparison, the general trend was in agreement, with better regression coefficients and lower RE values as well (area correlation constraint regression values ranging from 0.973 to 0.998, and RE values than 10% in the calibration mixture and lower than 25% in both samples tested). All in all, a more accurate quantification was achieved with the area correlation constraint, especially for the prediction of amino acids in unknown samples. These improvements would be higher in cases where overlapping elution profiles are present as in lower performance fast chromatographic approaches.

As a general conclusion, it is worth noticing that the three tested strategies provided acceptable quantitative results and can be all used for quantification purposes in LC×LC targeted analysis. Whenever a good calibration model is achieved, then the combination of MCR-ALS with the area correlation constraint seems to be the more suitable strategy because the accuracy in the quantification is improved, and stronger coelutions can be resolved. A further step in the validation of the proposed approaches for LC×LC datasets would be their application in more complex scenarios, such as for datasets acquired in metabolomic studies in the presence of strong biological interferences. In these cases, a significant improvement in the accuracy of the quantitative determinations may be expected when applying the area correlation constraint.

4.3.3 Future perspectives on the use of LC × LC for metabolomic studies

This section aims to discuss briefly the present of multidimensional liquid chromatography applied to metabolomic studies, what improvements are needed, which challenges are we still facing, and to shed light on possible solutions that could lead the way in the future of this field. This discussion is organized into three main topics: LC×LC for lipids, LC×LC for small polar metabolites, and LC×LC data analysis software.

LC × LC analysis of lipids: what is next?

The usefulness of RP×HILIC in lipidomic studies has been widely demonstrated, not only in scientific **publication V**, but also in previous works from the literature

[19–22]. Besides, the set-up proposed here benefits active solvent modulation strategies, especially ASM, which allow increasing sensitivity and solvent compatibility while diminishing run times. Although minor improvements are still needed, such as for instance, the reduction of the total analysis time (still two hours long), and the optimization of the ^{2}D separations to increase retention on less polar lipids (e.g., ceramides or triacylglycerides), current ASM methods are already very powerful strategies for untargeted analysis.

In fact, the choice of different versions of the already employed as stationary phases for RP and HILIC may not be very powerful yet. For instance, some lipids are not enough retained with shorter linear alkylsilane phases (e.g., C4), leading to important losses in lipid variety, which was the goal of establishing untargeted methods for their analysis. These results were obtained from preliminary tests during the 2019 research stay, when four types of RP stationary phases were tested: C4, C8, C18 and carbon columns (data not published). For the last column type, no peaks were obtained (lipids may be too much retained). Separations provided by C18 were superior by far, and other options were not worthy even for reducing the isopropanol content in the ^{1}D mobile phases (which increases the polar solvent percentage that enters the ^{2}D column, potentially worsening the ^{2}D separation). In addition, most of the HILIC separations (in RP×HILIC set-ups) are based on unbonded silica columns. In contrast to C18-based columns, multiple options could be tested in HILIC approaches, using different bondings to hydrophilic functional groups (e.g., cyano, amide, amine, zwitterionic). Unfortunately, only the bare silica type of C18 could be tested in this preliminary study, although no significant improvements in column behavior with other variants of stationary phases are expected, at least for the current technologies.

On the contrary, new combinations of stationary phases are yet to be explored. Particularly, the use of recent advances in active modulation and combinations that have been discarded until now could be employed due to the enhanced solvent compatibility. An example would be mixed mode (MM) chromatography, which is characterized by combining more than one retention mechanism in the same column (i.e., ion exchange and reversed phase, or reversed phase and HILIC). A system with only one column (i.e., a mixed-mode C18-Diol) has already been developed for both polar and non-polar compounds [56]. Alternatively, a 2DLC system with only one of the dimensions replaced by MM, could also be set-up. For instance, if ^{2}D separation

is replaced by a MM, where RP and HILIC are mixed, separation of compounds that were not retained under HILIC conditions (e.g., ceramides or triacylglycerides) may be then improved. Nevertheless, special attention would be required when using MM to ensure orthogonality, which could be jeopardized due to the use of similar retention mechanisms. In addition to MM, chiral columns also present great potential in 2DLC lipid separations. They have been widely employed for the separation of diverse lipid isomers by LC-MS/MS [57–59]. Besides, the combination of RP and chiral columns for the analysis of triacylglyceride isomers are also found in the literature [60,61]. Although isomers from different lipid classes are separated in HILIC (as ^{2}D), identifying isomers from the same lipid class with the same chain length is still a major challenge with RP×HILIC, even with the extra information provided by MS/MS. Thus, as chiral columns are a useful tool in isomer discrimination, the RP×Chiral set-up could be a powerful alternative for isomer separation.

Indeed, the combination of RP and chiral columns in a comprehensive mode is appealing, but instead of LC in both dimensions, one of them is replaced by supercritical fluid chromatography. Some advantages of SFC for lipid analysis are reported [62,63], with the following advantages:

- Reduced consumption of solvent (i.e., greener separations)
- A larger diversity of both stationary phases and mobile phase mixtures
- Broader lipid coverage (i.e., different polarities can be analyzed simultaneously)
- Potentially more confidence in annotation
- An increased global orthogonality when coupled to LC in contrast with LC×LC

Nevertheless, there are some important drawbacks as well. For instance, some of the main disadvantages of the coupling with SFC are the undesired injection effects leading to peak distortion and broadening in the ^{2}D (in LC × SFC mode), the complexity of the interfaces and the lack of commercial instrumentation [63]. SFC×SFC configurations are still to be tested [3]. Some pioneer examples in the recent literature about lipid analysis combine SFC as ^{1}D with RP as ^{2}D in online [64,65] and offline formats [66].

Alternatively, another powerful tool that, combined with LC, allows the discrimination between lipid isomers (i.e., both structural and stereoisomers) is ion

mobility spectrometry (IMS) [66]. The collision cross-section (CCS) ion mobility is an orthogonal parameter that facilitates compound identification (in addition to retention time, *m/z* value and fragmentation patterns, if MS/MS is employed afterwards), and helps to increase confidence in their annotation. Besides, ion mobility capabilities have been improved considerably in recent years [67] and CCS libraries are continuously growing [68]. For instance, the use of trapped ion mobility spectrometry coupled to time-of-flight mass spectrometry (TIMS-TOF-MS) in combination with liquid chromatography has risen recently in metabolomic (and lipidomics) applications [68–71]. Indeed, LC × LC × IM-MS has already been suggested for the analysis of very complex samples [72]. In the future, this strategy could also be employed in lipid analysis, merging all benefits from LC × LC with an extra dimension able to differentiate isomers.

LC × LC for the analysis of small and polar molecules: what are the challenges?

As mentioned before in this Chapter, some modulation improvements are still required for HILIC×RP configurations for the analysis of polar metabolites. In contrast to the analysis of large molecules (lipids, polymers or peptides), where active modulation (e.g., ASM) has led to significant advances in dealing with solvent mismatches [73–75], there are still big challenges remaining for the most polar compounds of the metabolome spectrum. Until now, the best solutions have been the dilution of the ^{1}D effluent with water from a third external pump [39], actively removing it completely before it reaches the ^{2}D column, using cartridges instead of loops [38] or an evaporative system at the interface [76]. Alternatively, it would be interesting to test the at-column dilution [77] procedure for this type of applications, as it has been proved when RP×HILIC columns are combined for the analysis of the chemical composition of medicinal herbs [44].

In addition, other possible retention mechanism combinations may give a greater benefit when the ASM approach is used. For instance, when MM is used in the ^{2}D, as suggested for lipid analysis. Another example would be the HILIC×HILIC method. Historically, this retention mechanism has been associated with long re-equilibration times caused by slow desorption and reformation of the aqueous layer [78]. Nevertheless, a recent study by Seidl *et al.* demonstrated that acceptable repeatability can be achieved even in the absence of a full re-equilibration [79]. These results validated the use of HILIC not only as ^{1}D but also as ^{2}D. In an

HILIC×HILIC set-up, a major threat would be to accomplish a sufficient degree of orthogonality. However, the work from Wang *et al.* shows not only an orthogonal method, but also the ability to separate isomeric species [80]. Despite the success of these isolated examples and in contrast to RP×RP separations, HILIC×HILIC are less popular by far. Current advances in modulation interfaces combined with a HILIC×HILIC configuration using complementary stationary phases may contribute to relevant advances in metabolomic separations. Besides, improved orthogonality and solvent compatibility strategies in RP×RP set-ups can be implemented as well in HILIC×HILIC. For instance, the use of segment gradients [37], or the more recent proposal of parallel gradients [81], may be useful in the cases where separation mechanisms employed in both dimensions are correlated.

Analogously to lipid analysis, SFC and IMS also provide an extra dimension with great potential in untargeted metabolomics, for instance, in the separation and identification of known and unknown isomeric compounds [5]. Thus, the combinations of these techniques with LC (e.g., an LC × LC × IMS-MS method) seem appealing in this case.

2DLC data analysis software: what is needed?

The multidimensional chromatography community agrees that one of the more urgent needs is the development of alternative software packages complementary to the vendor options available nowadays (e.g., freeware) and more flexible workflows for data analysis. The validation of these new software tools is also a key aspect to consider, and the use of benchmark datasets seems appealing for this purpose. In this last section of this Chapter, the currently available software and community needs regarding data analysis are reported based on the feedback from the discussions in the Multidimensional Chromatography Workshops carried out in 2021 and 2022 (http://www.multidimensionalchromatography.com/)

One of the main interests throughout this PhD Thesis has been the development of new approaches for data pre-processing, resolution and post-processing of 2DLC datasets. An important effort has been put into adapting previously developed tools for the analysis of LC-MS metabolomic studies (i.e., based on ROI, MCR-ALS and other multivariate analysis methods) to LC×LC-MS datasets. The strategies used in this PhD Thesis encompass 2DLC plot visualization, data compression and filtering,

compound resolution and annotation, and both absolute and relative quantification (i.e., for targeted and untargeted analysis, respectively). Apart from the work in this direction of our research group, it is worth highlighting the work by Molenaar *et al* [82,83]. Two types of software have recently been released by the Chemometrics and Advanced Separations Team (CAST, https://cast-amsterdam.org/). First is the MOREPEAKS or Multivariate Optimization and Refinement Program for Efficient Analysis of Key Separations [84], which allows the easy visualization of both LC × LC or GC × GC raw data, as well as the calculation of quality descriptors such as orthogonality or peak capacity. Besides, MOREPEAKS facilitates the optimization of chromatographic separations due to its ability to model and simulate analyte retention in different conditions. This software also has peak-tracking tools to extract information in an easy way. Quantification is one of the future goals that this software will have. Second, the MOREDISTRIBUTIONS or Multivariate and Otherwise Rapid and Efficient Determination and Identification Software for Thorough Representation and Interpretation By Unveiling Traits Informing On Novel Synthetics, [85], is a data analysis software specifically designed for synthetic polymers, which does not require computational skills for the application of a bunch of chemometric tools. In addition to their wide accessibility, both the approaches in this PhD Thesis and the work from Molenaar *et al*. also have in common the choice of the MATLAB environment for software development. Some of the main advantages of the use of MATLAB for data analysis are: the extensive documentation available (including official resources and technical support), the facility to develop user-friendly interfaces, and the large amount of built-in algorithms, toolboxes, and diverse functions accessible (e.g., PLS Toolbox 8.9.1 from Eigenvector Research Inc). Although MATLAB is proprietary software (meaning that a commercial license is needed to use it) when the MATLAB runtime compiler is used, users without a MATLAB license can also access the software tools developed for this computer environment, as it is the case for instance of the work from Molenaar.In this PhD Thesis the MSroi and MCR-ALS GUIs have been used, released in MATLAB for free (http://mcrals.info/), although they are planned to be implemented in other common programming languages (e.g., Python and R software).

Although MATLAB has historically led the chemometrics field, other open-source languages are increasing their popularity for the analysis of multidimensional chromatographic data. For instance, some R packages have been directly developed for the processing of GC × GC data [86–88], which could be potentially used for LC ×

LC datasets as well. Other programming languages, such as Python or Java, are less frequently employed in multidimensional chromatography, but some examples can also be found for GC × GC data analysis [89]. A hybrid option would be interfacing open source and commercial software, as already proposed by Wilde et al. [90]. The major advantage of this proposal would be a customize and more automatized workflow for data processing. Thus, this combined pipeline will benefit from the main pros of commercial software (standardized workflows, use of GUIs) but at the same time overcome their main limitations. For instance, the user will gain more control over the results and more flexibility upon the methods applied in the different data analysis steps. This intermediate option is especially interesting in the case of large-scale studies with multiple users (e.g., metabolomic cohorts).

Lastly, a query that arises when developing new software or new data analysis strategies is how to validate them. It is crucial to evaluate these new algorithms to ensure reproducible and reliable results compared with the existing tools. Besides, findings should be the same regardless of the approach employed (e.g., in metabolomic studies, the same biomarkers should be obtained). This issue has already been addressed for GC × GC data [91,92]. In these works, benchmark datasets that enable comparison between multiple software is also suggested. Following this initiative, the chromatogram shown in **Figure 4.4** earlier in this Chapter, obtained with the optimized method achieved during the 2019 research stay, was used in the validation of the MSroi GUI for 2DLC [9]) (which can be freely downloaded in the mcrals.info website). Furthermore, previous works from our research group have validated the use of MCR-ALS for LC-MS metabolomic studies compared to XCMS platforms [50,93], and their quantification capabilities have also been assessed compared with the peak areas provided by other vendors software [50]. Some examples of LC-MS benchmark datasets are also found on mcrals.info website. However, LC × LC benchmark datasets are still lacking, because of the still too recent development of LC × LC data analysis strategies compared to commercial software. What requirements should have these benchmark datasets are still an open question, as too many parameters are needed to be considered. For instance, is high-resolution mass spectrometry a must? Should data come from multiple instrument vendors? and in which format (e.g., .netCDF, .csv)? Are different experimental conditions required (different stationary phases, modulation strategies, etc)? Hence, this is still a need for the LC × LC community which should be addressed soon.

4.4 Conclusions

This section summarizes the specific conclusions drawn throughout this Chapter from different perspectives such as the proposed analytical improvements in LC×LC and the advantages in the developed data analysis strategies, especially related to the data pre-processing and the quantification steps.

Concerning LC × LC developments for the analysis of lipids:

- RP×HILIC configuration provides a higher resolution power compared to HILIC×RP for the analysis of lipids.
- The use of ASM in the analysis of lipids with an RP×HILIC set-up, increases sensitivity, solvent compatibility and decreases total run time.
- An optimized RP×HILIC method for untargeted lipid analysis is proposed in lipidomic studies, which has been employed to assess environmental effects caused by bisphenol A exposure in zebrafish embryos.

Concerning chemometric strategies suggested for the analysis of LC × LC datasets:

- In the case of very big LC × LC datasets, an intensity prefilter is recommended in the conversion process prior to the ROI step.
- The proposed ROI approach (i.e., filtering and spectral compression) has sufficiently reduced the dimensionality of LC × LC-HRMS datasets. Other compression strategies (i.e., in the time dimension) and peak alignment steps were not required, significantly simplifying the whole pre-processing workflow.
- ROI strategy demonstrated its usefulness for targeted, pseudo-targeted and untargeted approaches (i.e., providing absolute and relative quantification information).
- From the different quantification approaches tested (ROI, classic calibration with MCR-ALS, and MCR-ALS with area correlation constraint), the best results were obtained when the ROIMCR procedure with the area correlation constraint (applied during the ALS optimization) were combined. However, acceptable quantification results can be obtained with all three strategies.

References

[1] P. Miggiels, B. Wouters, G.J.P. van Westen, A.C. Dubbelman, T. Hankemeier, Novel technologies for metabolomics: More for less, TrAC - Trends in Analytical Chemistry. 120 (2019) 115323. https://doi.org/10.1016/j.trac.2018.11.021.

[2] W. Lv, X. Shi, S. Wang, G. Xu, Multidimensional liquid chromatography-mass spectrometry for metabolomic and lipidomic analyses, TrAC - Trends in Analytical Chemistry. 120 (2019) 115302. https://doi.org/10.1016/j.trac.2018.11.001.

[3] A.S. Kaplitz, M.E. Mostafa, S.A. Calvez, J.L. Edwards, J.P. Grinias, Two-dimensional separation techniques using supercritical fluid chromatography, Journal of Separation Science. 44 (2021) 426–437. https://doi.org/10.1002/jssc.202000823.

[4] A. Delvaux, E. Rathahao-Paris, S. Alves, Different ion mobility-mass spectrometry coupling techniques to promote metabolomics, Mass Spectrometry Reviews. (2021). https://doi.org/10.1002/mas.21685.

[5] M. du Luo, Z.W. Zhou, Z.J. Zhu, The Application of Ion Mobility-Mass Spectrometry in Untargeted Metabolomics: from Separation to Identification, Journal of Analysis and Testing. 4 (2020) 163–174. https://doi.org/10.1007/S41664-020-00133-0/FIGURES/3.

[6] M. Grübner, A. Dunkel, F. Steiner, T. Hofmann, Systematic Evaluation of Liquid Chromatography (LC) Column Combinations for Application in Two-Dimensional LC Metabolomic Studies, Analytical Chemistry. 93 (2021) 12565–12573. https://doi.org/10.1021/acs.analchem.1c01857.

[7] B.W.J. Pirok, D.R. Stoll, P.J. Schoenmakers, Recent Developments in Two-Dimensional Liquid Chromatography: Fundamental Improvements for Practical Applications, Analytical Chemistry. 91 (2019) 240–263. https://doi.org/10.1021/acs.analchem.8b04841.

[8] D.R. Stoll, K. Shoykhet, P. Petersson, S. Buckenmaier, Active Solvent Modulation: A Valve-Based Approach to Improve Separation Compatibility in Two-Dimensional Liquid Chromatography, Analytical Chemistry. 89 (2017) 9260–9267. https://doi.org/10.1021/acs.analchem.7b02046.

[9] M. Pérez-Cova, C. Bedia, D.R. Stoll, R. Tauler, J. Jaumot, MSroi: A pre-processing tool for mass spectrometry-based studies, Chemometrics and Intelligent Laboratory Systems. 215 (2021). https://doi.org/10.1016/j.chemolab.2021.104333.

[10] J. Jaumot, B. Igne, C.A. Anderson, J.K. Drennen, A. de Juan, Blending process modeling and control by multivariate curve resolution, Talanta. 117 (2013) 492–504. https://doi.org/10.1016/j.talanta.2013.09.037.

[11] A.C. de O. Neves, R. Tauler, K.M.G. de Lima, Area correlation constraint for the MCR – ALS quantification of cholesterol using EEM fluorescence data: A new approach, Analytica Chimica Acta. 937 (2016) 21–28. https://doi.org/10.1016/j.aca.2016.08.011.

[12] D.R. Stoll, P.W. Carr, Two-Dimensional Liquid Chromatography: A State of the Art Tutorial, Analytical Chemistry. 89 (2017) 519–531. https://doi.org/10.1021/acs.analchem.6b03506.

[13] B.W.J. Pirok, A.F.G. Gargano, P.J. Schoenmakers, Optimizing separations in online comprehensive two-dimensional liquid chromatography, Journal of Separation Science. 41 (2018) 68–98. https://doi.org/10.1002/jssc.201700863.

[14] M. Pérez-Cova, J. Jaumot, R. Tauler, Untangling comprehensive two-dimensional liquid chromatography data sets using regions of interest and multivariate curve resolution approaches, TrAC - Trends in Analytical Chemistry. 137 (2021). https://doi.org/10.1016/j.trac.2021.116207.

[15] Y. Chen, L. Montero, O.J. Schmitz, Advance in on-line two-dimensional liquid chromatography modulation technology, TrAC - Trends in Analytical Chemistry. 120 (2019) 115647. https://doi.org/10.1016/j.trac.2019.115647.

[16] T. Cajka, O. Fiehn, Comprehensive analysis of lipids in biological systems by liquid chromatography-mass spectrometry, TrAC - Trends in Analytical Chemistry. 61 (2014) 192–206. https://doi.org/10.1016/j.trac.2014.04.017.

[17] M. Roca, M.I. Alcoriza, J.C. Garcia-Cañaveras, A. Lahoz, Reviewing the metabolome coverage provided by LC-MS: Focus on sample preparation and chromatography-A tutorial, Analytica Chimica Acta. 1147 (2021) 38–55. https://doi.org/10.1016/j.aca.2020.12.025.

[18] T. Xu, C. Hu, Q. Xuan, G. Xu, Recent advances in analytical strategies for mass spectrometry-based lipidomics, Analytica Chimica Acta. 1137 (2020) 156–169. https://doi.org/10.1016/j.aca.2020.09.060.

[19] A. Baglai, A.F.G. Gargano, J. Jordens, Y. Mengerink, M. Honing, S. van der Wal, P.J. Schoenmakers, Comprehensive lipidomic analysis of human plasma using multidimensional liquid- and gas-phase separations: Two-dimensional liquid chromatography–mass spectrometry vs. liquid chromatography–trapped-ion-mobility–mass spectrometry, Journal of Chromatography A. 1530 (2017) 90–103. https://doi.org/10.1016/j.chroma.2017.11.014.

[20] M. Holčapek, M. Ovčačíková, M. Lísa, E. Cífková, T. Hájek, Continuous comprehensive two-dimensional liquid chromatography-electrospray ionization mass spectrometry of complex lipidomic samples, Anal Bioanal Chem. 407 (2015) 5033–5043. https://doi.org/10.1007/s00216-015-8528-2.

[21] M. Navarro-Reig, J. Jaumot, R. Tauler, An untargeted lipidomic strategy combining comprehensive two-dimensional liquid chromatography and chemometric analysis, Journal of Chromatography A. 1568 (2018) 80–90. https://doi.org/10.1016/j.chroma.2018.07.017.

[22] M. Xu, J. Legradi, P. Leonards, Evaluation of LC-MS and LC × LC-MS in analysis of zebrafish embryo samples for comprehensive lipid profiling, Analytical and Bioanalytical Chemistry. 412 (2020) 4313–4325. https://doi.org/10.1007/s00216-020-02661-1.

[23] P. Donato, G. Micalizzi, M. Oteri, F. Rigano, D. Sciarrone, P. Dugo, L. Mondello, Comprehensive lipid profiling in the Mediterranean mussel (Mytilus galloprovincialis) using hyphenated and multidimensional chromatography techniques coupled to mass spectrometry detection, Analytical and Bioanalytical Chemistry. 410 (2018) 3297–3313. https://doi.org/10.1007/s00216-018-1045-3.

[24] R. Berkecz, F. Tömösi, T. Körmöczi, V. Szegedi, J. Horváth, T. Janáky, Comprehensive phospholipid and sphingomyelin profiling of different brain regions in mouse model of anxiety disorder using online two-dimensional (HILIC/RP)-LC/MS method, Journal of Pharmaceutical and Biomedical Analysis. 149 (2018) 308–317. https://doi.org/10.1016/j.jpba.2017.10.043.

[25] C. Sun, Y.Y. Zhao, J.M. Curtis, Characterization of phospholipids by two-dimensional liquid chromatography coupled to in-line ozonolysis-mass spectrometry, Journal of Agricultural and Food Chemistry. 63 (2015) 1442–1451. https://doi.org/10.1021/jf5049595.

[26] C. Sun, Y.Y. Zhao, J.M. Curtis, Elucidation of phosphatidylcholine isomers using two dimensional liquid chromatography coupled in-line with ozonolysis mass spectrometry, Journal of Chromatography A. 1351 (2014) 37–45. https://doi.org/10.1016/j.chroma.2014.04.069.

[27] M. Holčapek, M. Ovčačíková, M. Lísa, E. Cífková, T. Hájek, Continuous comprehensive two-dimensional liquid chromatography-electrospray ionization mass spectrometry of complex lipidomic samples, Anal Bioanal Chem. 407 (2015) 5033–5043. https://doi.org/10.1007/s00216-015-8528-2.

[28] M. Holčapek, M. Ovčačíková, M. Lísa, E. Cífková, T. Hájek, Continuous comprehensive two-dimensional liquid chromatography-electrospray ionization mass spectrometry of complex lipidomic samples, Anal Bioanal Chem. 407 (2015) 5033–5043. https://doi.org/10.1007/s00216-015-8528-2.

[29] Ö.C. Zeki, C.C. Eylem, T. Reçber, S. Kır, E. Nemutlu, Integration of GC–MS and LC–MS for untargeted metabolomics profiling, Journal of Pharmaceutical and Biomedical Analysis. 190 (2020). https://doi.org/10.1016/j.jpba.2020.113509.

[30] E.M. Harrieder, F. Kretschmer, S. Böcker, M. Witting, Current state-of-the-art of separation methods used in LC-MS based metabolomics and lipidomics, Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences. 1188 (2022). https://doi.org/10.1016/j.jchromb.2021.123069.

[31] A. Klåvus, M. Kokla, S. Noerman, V.M. Koistinen, M. Tuomainen, I. Zarei, T. Meuronen, M.R. Häkkinen, S. Rummukainen, A.F. Babu, T. Sallinen, O. Kärkkäinen, J. Paananen, D. Broadhurst, C. Brunius, K. Hanhineva, “Notame”: Workflow for non-targeted LC-MS metabolic profiling, Metabolites. 10 (2020) 1–35. https://doi.org/10.3390/metabo10040135.

[32] M. Xu, J. Legradi, P. Leonards, Cross platform solutions to improve the zebrafish polar metabolome coverage using LC-QTOF MS: Optimization of separation mechanisms, solvent additives, and resuspension solvents, Talanta. 234 (2021) 122688. https://doi.org/10.1016/j.talanta.2021.122688.

[33] C. Martias, N. Baroukh, S. Mavel, H. Blasco, A. Lefèvre, L. Roch, F. Montigny, J. Gatien, L. Schibler, D. Dufour-Rainfray, L. Nadal-Desbarats, P. Emond, Optimization of sample preparation for metabolomics exploration of urine, feces, blood and saliva in humans using combined nmr and uhplc-hrms platforms, Molecules. 26 (2021). https://doi.org/10.3390/molecules26144111.

[34] J. Pezzatti, V. González-Ruiz, S. Codesido, Y. Gagnebin, A. Joshi, D. Guillarme, J. Schappler, D. Picard, J. Boccard, S. Rudaz, A scoring approach for multi-platform acquisition in metabolomics, Journal of Chromatography A. 1592 (2019) 47–54. https://doi.org/10.1016/j.chroma.2019.01.023.

[35] K. Arena, F. Cacciola, L. Dugo, P. Dugo, L. Mondello, Determination of the metabolite content of Brassica juncea cultivars using comprehensive two-dimensional liquid chromatography coupled with a photodiode array and mass spectrometry detection, Molecules. 25 (2020) 1–12. https://doi.org/10.3390/molecules25051235.

[36] A. Corgier, M. Sarrut, G. Crétier, S. Heinisch, Potential of Online Comprehensive Two-Dimensional Liquid Chromatography For Micro-Preparative Separations of Simple Samples, Chromatographia. 79 (2016) 255–260. https://doi.org/10.1007/s10337-015-3012-x.

[37] Y.F. Wong, F. Cacciola, S. Fermas, S. Riga, D. James, V. Manzin, B. Bonnet, P.J. Marriott, P. Dugo, L. Mondello, Untargeted profiling of Glycyrrhiza glabra extract with comprehensive two-dimensional liquid chromatography-mass spectrometry using multi-segmented shift gradients in the second dimension: Expanding the metabolic coverage, Electrophoresis. 39 (2018) 1993–2000. https://doi.org/10.1002/elps.201700469.

[38] A. Baglai, M.H. Blokland, H.G.J. Mol, A.F.G. Gargano, S. van der Wal, P.J. Schoenmakers, Enhancing detectability of anabolic-steroid residues in bovine urine by actively modulated online comprehensive two-dimensional liquid chromatography – high-resolution mass spectrometry, Analytica Chimica Acta. 1013 (2018) 87–97. https://doi.org/10.1016/j.aca.2017.12.043.

[39] M. Navarro-Reig, J. Jaumot, A. Baglai, G. Vivó-Truyols, P.J. Schoenmakers, R. Tauler, Untargeted Comprehensive Two-Dimensional Liquid Chromatography Coupled with High-Resolution Mass Spectrometry Analysis of Rice Metabolome Using Multivariate Curve Resolution, Analytical Chemistry. 89 (2017) 7675–7683. https://doi.org/10.1021/acs.analchem.7b01648.

[40] L. Montero, E. Ibáñez, M. Russo, R. di Sanzo, L. Rastrelli, A.L. Piccinelli, R. Celano, A. Cifuentes, M. Herrero, Metabolite profiling of licorice (Glycyrrhiza glabra) from different locations using comprehensive two-dimensional liquid chromatography coupled to diode array and tandem mass spectrometry detection, Analytica Chimica Acta. 913 (2016) 145–159. https://doi.org/10.1016/j.aca.2016.01.040.

[41] E. Sommella, O.H. Ismail, F. Pagano, G. Pepe, C. Ostacolo, G. Mazzoccanti, M. Russo, E. Novellino, F. Gasparrini, P. Campiglia, Development of an improved online comprehensive hydrophilic interaction chromatography × reversed-phase ultra-high-pressure liquid chromatography platform for complex multiclass polyphenolic sample analysis, Journal of Separation Science. 40 (2017) 2188–2197. https://doi.org/10.1002/jssc.201700134.

[42] L. Montero, A.P. Sánchez-Camargo, V. García-Cañas, A. Tanniou, V. Stiger-Pouvreau, M. Russo, L. Rastrelli, A. Cifuentes, M. Herrero, E. Ibáñez, Anti-proliferative activity and chemical characterization by comprehensive two-dimensional liquid chromatography coupled to mass spectrometry of phlorotannins from the brown macroalga Sargassum muticum collected on North-Atlantic coasts, Journal of Chromatography A. 1428 (2016) 115–125. https://doi.org/10.1016/j.chroma.2015.07.053.

[43] C.M. Willemse, M.A. Stander, J. Vestner, A.G.J. Tredoux, A. de Villiers, Comprehensive Two-Dimensional Hydrophilic Interaction Chromatography (HILIC) × Reversed-Phase Liquid Chromatography Coupled to High-Resolution Mass Spectrometry (RP-LC-UV-MS) Analysis of Anthocyanins and Derived Pigments in Red Wine, Analytical Chemistry. 87 (2015) 12006–12015. https://doi.org/10.1021/acs.analchem.5b03615.

[44] Y. Chen, L. Montero, J. Luo, J. Li, O.J. Schmitz, Application of the new at-column dilution (ACD) modulator for the two-dimensional RP×HILIC analysis of Buddleja davidii, Analytical and Bioanalytical Chemistry. 412 (2020) 1483–1495. https://doi.org/10.1007/s00216-020-02392-3.

[45] L.L. Hohrenk, M. Vosough, T.C. Schmidt, Implementation of Chemometric Tools to Improve Data Mining and Prioritization in LC-HRMS for Nontarget Screening of Organic Micropollutants in Complex Water Matrixes, Analytical Chemistry. 91 (2019) 9213–9220. https://doi.org/10.1021/ACS.ANALCHEM.9B01984/SUPPL_FILE/AC9B01984_SI_001.PDF.

[46] R. Martínez, L. Navarro-Martín, M. van Antro, I. Fuertes, M. Casado, C. Barata, B. Piña, Changes in lipid profiles induced by bisphenol A (BPA) in zebrafish eleutheroembryos during the yolk sac absorption stage, Chemosphere. 246 (2020) 125704. https://doi.org/10.1016/J.CHEMOSPHERE.2019.125704.

[47] A. Jia, L. Xu, Y. Wang, Venn diagrams in bioinformatics, Briefings in Bioinformatics. 22 (2021). https://doi.org/10.1093/BIB/BBAB108.

[48] H. Tsugawa, K. Ikeda, M. Takahashi, A. Satoh, Y. Mori, H. Uchino, N. Okahashi, Y. Yamada, I. Tada, P. Bonini, Y. Higashi, Y. Okazaki, Z. Zhou, Z.-J. Zhu, J. Koelmel, T. Cajka, O. Fiehn, K. Saito, M. Arita, M. Arita, A lipidome atlas in MS-DIAL 4, Nature Biotechnology. (n.d.). https://doi.org/10.1038/s41587-020-0531-2.

[49] L. Montero, S.W. Meckelmann, H. Kim, J.F. Ayala-Cabrera, O.J. Schmitz, Differentiation of industrial hemp strains by their cannabinoid and phenolic compounds using LC × LC-HRMS, Analytical and Bioanalytical Chemistry. (2022). https://doi.org/10.1007/S00216-022-03925-8.

[50] N. Dalmau, C. Bedia, R. Tauler, Validation of the Regions of Interest Multivariate Curve Resolution (ROIMCR) procedure for untargeted LC-MS lipidomic analysis, Analytica Chimica Acta. 1025 (2018) 80–91. https://doi.org/10.1016/j.aca.2018.04.003.

[51] R.S.M. Pourasil, J. Cristale, S. Lacorte, R. Tauler, Non-targeted Gas Chromatography Orbitrap Mass Spectrometry qualitative and quantitative analysis of semi-volatile organic compounds in indoor dust using the Regions of Interest Multivariate Cuarve Resolution chemometrics procedure, Journal of Chromatography A. (2022) 462907. https://doi.org/10.1016/J.CHROMA.2022.462907.

[52] G. Ahmadi, R. Tauler, H. Abdollahi, Multivariate calibration of first-order data with the correlation constrained MCR-ALS method, Chemometrics and Intelligent Laboratory Systems. 142 (2015) 143–150. https://doi.org/10.1016/j.chemolab.2014.11.010.

[53] M. Bayat, M. Marín-García, J.B. Ghasemi, R. Tauler, Application of the area correlation constraint in the MCR-ALS quantitative analysis of complex mixture samples, Analytica Chimica Acta. 1113 (2020) 52–65. https://doi.org/10.1016/j.aca.2020.03.057.

[54] A.C. de O. Neves, R. Tauler, K.M.G. de Lima, Area correlation constraint for the MCR－ALS quantification of cholesterol using EEM fluorescence data: A new approach, Analytica Chimica Acta. 937 (2016) 21–28. https://doi.org/10.1016/j.aca.2016.08.011.

[55] A. de Juan, J. Jaumot, R. Tauler, Multivariate Curve Resolution (MCR). Solving the mixture analysis problem, Analytical Methods. 6 (2014) 4964–4976. https://doi.org/10.1039/c4ay00571f.

[56] Q. Wang, M. Ye, L. Xu, Z. guo Shi, A reversed-phase/hydrophilic interaction mixed-mode C18-Diol stationary phase for multiple applications, Anal Chim Acta. 888 (2015) 182–190. https://doi.org/10.1016/J.ACA.2015.06.058.

[57] J. Ito, N. Shimizu, E. Kobayashi, Y. Hanzawa, Y. Otoki, S. Kato, T. Hirokawa, S. Kuwahara, T. Miyazawa, K. Nakagawa, A novel chiral stationary phase LC-MS/MS method to evaluate oxidation mechanisms of edible oils, Scientific Reports. 7 (2017) 1–10. https://doi.org/10.1038/s41598-017-10536-2.

[58] T. Řezanka, K. Sigler, Separation of enantiomeric triacylglycerols by chiral-phase HPLC, Lipids. 49 (2014) 1251–1260. https://doi.org/10.1007/s11745-014-3959-7.

[59] S.A. Brose, B.T. Thuen, M.Y. Golovko, LC/MS/MS method for analysis of E2 series prostaglandins and isoprostanes, Journal of Lipid Research. 52 (2011) 850–859. https://doi.org/10.1194/jlr.D013441.

[60] T. Řezanka, I. Kolouchová, A. Čejková, T. Cajthaml, K. Sigler, Identification of regioisomers and enantiomers of triacylglycerols in different yeasts using reversed- and chiral-phase LC-MS, Journal of Separation Science. 36 (2013) 3310–3320. https://doi.org/10.1002/jssc.201300657.

[61] T. Řezanka, L. Nedbalová, K. Sigler, Enantiomeric separation of triacylglycerols containing polyunsaturated fatty acids with 18 carbon atoms, Journal of Chromatography A. 1467 (2016) 261–269. https://doi.org/10.1016/j.chroma.2016.07.006.

[62] S. Song, H. Liu, Y. Bai, Supercritical Fluid Chromatography and Its Application in Lipid Isomer Separation, Journal of Analysis and Testing. 1 (n.d.). https://doi.org/10.1007/s41664-017-0031-7.

[63] M. Burlet-Parendel, K. Faure, Opportunities and challenges of liquid chromatography coupled to supercritical fluid chromatography, TrAC Trends in Analytical Chemistry. 144 (2021) 116422. https://doi.org/10.1016/J.TRAC.2021.116422.

[64] L. Feng, L. Wu, Y. Guo, N. Hamada, Y. Hashi, X. Li, L. Cao, Determination of vitamin D3 in daily oily supplements by a two-dimensional supercritical fluid chromatography-liquid chromatography-mass spectrometry system, J Chromatogr A. 1629 (2020). https://doi.org/10.1016/J.CHROMA.2020.461510.

[65] L. Yang, H. Nie, F. Zhao, S. Song, Y. Meng, Y. Bai, H. Liu, A novel online two-dimensional supercritical fluid chromatography/reversed phase liquid chromatography–mass spectrometry method for lipid profiling, Analytical and Bioanalytical Chemistry. 412 (2020) 2225–2235. https://doi.org/10.1007/s00216-019-02242-x.

[66] J.E. Kyle, X. Zhang, K.K. Weitz, M.E. Monroe, Y.M. Ibrahim, R.J. Moore, J. Cha, X. Sun, E.S. Lovelace, J. Wagoner, S.J. Polyak, T.O. Metz, S.K. Dey, R.D. Smith, K.E. Burnum-Johnson, E.S. Baker, Uncovering biologically significant lipid isomers with liquid chromatography, ion mobility spectrometry and mass spectrometry, Analyst. 141 (2016) 1649–1659. https://doi.org/10.1039/c5an02062j.

[67] C. Gonzaíez-Riano, D. Dudzik, A. Garcia, A. Gil-De-La-Fuente, A. Gradillas, J. Godzien, A.´ Ngeles Lópezlópez-Gonzaívez, F. Rey-Stolle, D. Rojo, F.J. Ruperez, J. Saiz, C. Barbas, Recent Developments along the Analytical Process for Metabolomics Workflows, (2019). https://doi.org/10.1021/acs.analchem.9b04553.

[68] M. Schroeder, S.W. Meyer, H.M. Heyman, A. Barsch, L.W. Sumner, Generation of a collision cross section library for multi-dimensional plant metabolomics using UHPLC-trapped ion mobility-MS/MS, Metabolites. 10 (2020). https://doi.org/10.3390/metabo10010013.

[69] M. Chen, Y. Hao, S. Chen, A protocol for investigating lipidomic dysregulation and discovering lipid biomarkers from human serums, STAR Protocols. 3 (2022) 101125. https://doi.org/10.1016/j.xpro.2022.101125.

[70] B. Spanier, A. Laurençon, A. Weiser, N. Pujol, S. Omi, A. Barsch, A. Korf, S.W. Meyer, J.J. Ewbank, F. Paladino, S. Garvis, · Hugo Aguilaniu, Comparison of lipidome profiles of Caenorhabditis elegans-results from an inter-laboratory ring trial, Metabolomics. 1 (123AD) 25. https://doi.org/10.1007/s11306-021-01775-6.

[71] C. di Poto, X. Tian, X. Peng, H.M. Heyman, M. Szesny, S. Hess, L.H. Cazares, Metabolomic Profiling of Human Urine Samples Using LC-TIMS-QTOF Mass Spectrometry, J Am Soc Mass Spectrom. 32 (2021) 2072–2080. https://doi.org/10.1021/jasms.0c00467.

[72] P. Venter, M. Muller, J. Vestner, M.A. Stander, A.G.J. Tredoux, H. Pasch, A. de Villiers, Comprehensive Three-Dimensional LC × LC × Ion Mobility Spectrometry Separation Combined with High-Resolution MS for the Analysis of Complex Samples, Analytical Chemistry. 90 (2018) 11643–11650. https://doi.org/10.1021/acs.analchem.8b03234.

[73] D.R. Stoll, H.R. Lhotka, D.C. Harmes, B. Madigan, J.J. Hsiao, G.O. Staples, High resolution two-dimensional liquid chromatography coupled with mass spectrometry for robust and sensitive characterization of therapeutic antibodies at the peptide level, Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences. 1134–1135 (2019) 121832. https://doi.org/10.1016/j.jchromb.2019.121832.

[74] P. Yang, W. Gao, T. Zhang, M. Pursch, J. Luong, W. Sattler, A. Singh, S. Backer, Two-dimensional liquid chromatography with active solvent modulation for studying monomer incorporation in copolymer dispersants, Journal of Separation Science. 42 (2019) 2805–2815. https://doi.org/10.1002/jssc.201900283.

[75] D.R. Stoll, D.C. Harmes, G.O. Staples, O.G. Potter, C.T. Dammann, D. Guillarme, A. Beck, Development of Comprehensive Online Two-Dimensional Liquid Chromatography/Mass Spectrometry Using Hydrophilic Interaction and Reversed-Phase Separations for Rapid and Deep Profiling of Therapeutic Antibodies, Analytical Chemistry. 90 (2018) 5923–5929. https://doi.org/10.1021/acs.analchem.8b00776.

[76] E. Fornells, B. Barnett, M. Bailey, E.F. Hilder, R.A. Shellie, M.C. Breadmore, Evaporative membrane modulation for comprehensive two-dimensional liquid chromatography, Analytica Chimica Acta. 1000 (2018) 303–309. https://doi.org/10.1016/j.aca.2017.11.053.

[77] Y. Chen, J. Li, O.J. Schmitz, Development of an At-Column Dilution Modulator for Flexible and Precise Control of Dilution Factors to Overcome Mobile Phase Incompatibility in Comprehensive Two-Dimensional Liquid Chromatography, Analytical Chemistry. 91 (2019) 10251–10257. https://doi.org/10.1021/acs.analchem.9b02391.

[78] P. Hemström, K. Irgum, Hydrophilic interaction chromatography, 2006. https://doi.org/10.1002/jssc.200600199.

[79] C. Seidl, D.S. Bell, D.R. Stoll, A study of the re-equilibration of hydrophilic interaction columns with a focus on viability for use in two-dimensional liquid chromatography, Journal of Chromatography A. 1604 (2019) 460484. https://doi.org/10.1016/J.CHROMA.2019.460484.

[80] Y. Wang, X. Lu, G. Xu, Development of a comprehensive two-dimensional hydrophilic interaction chromatography/quadrupole time-of-flight mass spectrometry system and its application in separation and identification of saponins from Quillaja saponaria, Journal of Chromatography A. 1181 (2008) 51–59. https://doi.org/10.1016/j.chroma.2007.12.034.

[81] A.A. Aly, M. Muller, A. de Villiers, B.W.J. Pirok, T. Górecki, Parallel gradients in comprehensive multidimensional liquid chromatography enhance utilization of the separation space and the degree of orthogonality when the separation mechanisms are correlated, Journal of Chromatography A. 1628 (2020) 461452. https://doi.org/10.1016/j.chroma.2020.461452.

[82] S.R.A. Molenaar, P.J. Schoenmakers, B.W.J. Pirok, MOREPEAKS, (2021). https://doi.org/10.5281/ZENODO.6375413.

[83] S.R.A. Molenaar, B. van de Put, B.W.J. Pirok, MOREDISTRIBUTIONS, (2021). https://doi.org/10.5281/ZENODO.5710530.

[84] & B.W.J.P. Stef R.A. Molenaar, Peter J. Schoenmakers, MOREPEAKS, (2021). https://doi.org/https://doi.org/10.5281/zenodo.5786549.

[85] & B.W.J.Pirok. Stef R.A. Molenaar, Bram van de Put, MOREDISTRIBUTIONS, (2021). https://doi.org/https://doi.org/10.5281/zenodo.5710530.

[86] C. Quiroz-Moreno, M.F. Furlan, J.R. Belinato, F. Augusto, G.L. Alexandrino, N.G.S. Mogollón, RGCxGC toolbox: An R-package for data processing in comprehensive two-dimensional gas chromatography-mass spectrometry, Microchemical Journal. 156 (2020) 104830. https://doi.org/10.1016/j.microc.2020.104830.

[87] E. Hoh, N.G. Dodder, S.J. Lehotay, K.C. Pangallo, C.M. Reddy, K.A. Maruya, Nontargeted comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry method and software for inventorying persistent and bioaccumulative contaminants in marine environments, Environmental Science and Technology. 46 (2012) 8001–8008. https://doi.org/10.1021/es301139q.

[88] S. Moayedpour, H. Parastar, RMet: An automated R based software for analyzing GC-MS and GC × GC-MS untargeted metabolomic data, Chemometrics and Intelligent Laboratory Systems. 194 (2019) 103866. https://doi.org/10.1016/j.chemolab.2019.103866.

[89] I.A. Titaley, O.M. Ogba, L. Chibwe, E. Hoh, P.H.Y. Cheong, S.L.M. Simonich, Automating data analysis for two-dimensional gas chromatography/time-of-flight mass spectrometry non - targeted analysis of comparative samples, Journal of Chromatography A. 1541 (2018) 57–62. https://doi.org/10.1016/J.CHROMA.2018.02.016.

[90] M.J. Wilde, B. Zhao, R.L. Cordell, W. Ibrahim, A. Singapuri, N.J. Greening, C.E. Brightling, S. Siddiqui, P.S. Monks, R.C. Free, Automating and Extending Comprehensive Two-Dimensional Gas Chromatography Data Processing by Interfacing Open-Source and Commercial Software, Analytical Chemistry. 92 (2020) 13953–13960. https://doi.org/10.1021/acs.analchem.0c02844.

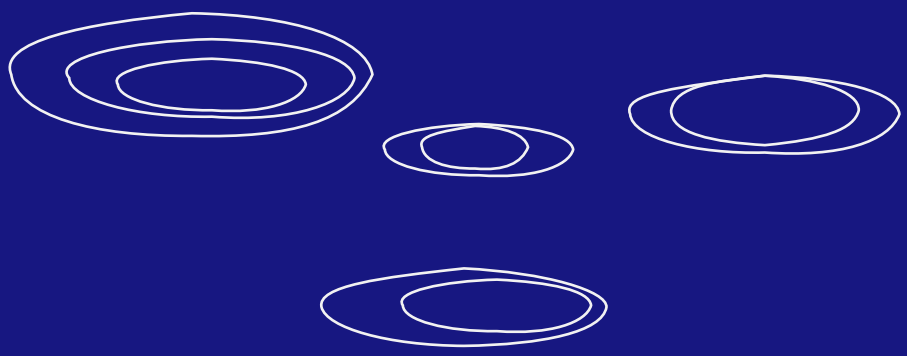
[91] B.A. Weggler, L.M. Dubois, N. Gawlitta, T. Gröger, J. Moncur, L. Mondello, S. Reichenbach, P. Tranchida, Z. Zhao, R. Zimmermann, M. Zoccali, J.F. Focant, A unique data analysis framework and open source benchmark data set for the analysis of comprehensive two-dimensional gas chromatography software, Journal of Chromatography A. 1635 (2021) 461721. https://doi.org/10.1016/j.chroma.2020.461721.

[92] N. Gawlitta, T. Gr, R. Zimmermann, J. Mass, S. Centre, H. Zentrum, New Platform-Independent Data Analysis Software with Build-in Chemometric Tools for the Processing and Statistical Analysis of Comprehensive Two-Dimensional Gas Chromatography Data Sets, (2019) 50–53.

[93] C.A. Smith, E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification, Analytical Chemistry. 78 (2006) 779–787. https://doi.org/10.1021/ac051437y.

Chapter five

Applications of metabolomic workflows for environmental assessments



5.1 Introduction

This Chapter focuses on examples of metabolomic applications used by the previously introduced chromatographic and chemometric methods. Particularly, the proposed metabolomic workflows aim to evaluate the effects of different emerging pollutants in model biosystems (i.e., arsenic exposure in rice and pharmaceutical compounds in human hepatic cells, in **scientific publications VII and VIII,** respectively). These studies combine information about the changes that the environmental stressors produce in the metabolome and lipidome of the model biosystems.

Although the term "metabolomics" refers to the analysis of small and polar compounds (i.e., metabolites) as well as other larger molecules, such as lipids (i.e., lipidomics), the analytical workflow is usually very different, depending on the analyzed compounds. For instance, extraction protocols or analytical conditions (e.g., stationary and mobile phases solvent composition in the case of liquid chromatography) can suffer modifications accordingly. The ideal situation would be to set up a single analytical workflow for molecules from high polarity to high hydrophobicity, but the broad range of physicochemical properties of the analytes difficults this scenario. Examples from the literature often collect the polar and the non-polar fractions separately during the same extraction protocol (e.g., acetonitrile/methanol layer for metabolomics, and chloroform layer for lipidomics) [1–3]. Then, the injection of these fractions into liquid chromatography coupled to mass spectrometry (LC-MS) using methods specifically designed for metabolites or lipids. Other innovative approaches run the dual analysis in parallel columns (e.g., reversed phased (RP) and hydrophilic interaction chromatography (HILIC)) and merge the obtained information into one data file [4].

Figure 5.1 summarizes the analytical workflows employed in **scientific publications VII and VIII** from the model biosystem and environmental stressor selection until the data processing strategy. In the case of studying arsenic exposure, rice is an interesting choice because its flooding conditions make this model organism especially vulnerable to inorganic arsenic species, present in the water or the soil [5]. In addition, rice is the most consumed product with a high arsenic content, becoming a threat to the human population [6]. In contrast, human hepatic cells (from the HepG2 cell line) are an appealing model biosystem for studying the effects of hepatotoxic pharmaceutical compounds released into the environment.

These compounds have been detected and quantified in wastewaters [7], but further effort is required to assess their toxicity.

Both studies have in common that the evaluation of the effects of these emerging pollutants is performed through metabolomics approaches, although the workflows followed are different. The metabolomic and lipidomic information is obtained in **scientific publication VII,** through two analytical approaches, including two extraction protocols and two untargeted LC-MS methods. In contrast, in **scientific publication VIII**, the same extraction protocol was used for both analyses (i.e., lipids and metabolites) but different LC-MS approaches, a targeted method for sphingolipids and an untargeted method for metabolites. In both studies, RP and HILIC stationary phases were employed for the analysis of lipids and metabolites, respectively, although the separation and mass spectrometry conditions were different. However, data independent acquisition (DIA) (in particular, all ion fragmentation (AIF)) was employed for untargeted metabolomics analyses in the two studies.

A more exhaustive comparison of the data analysis workflows is detailed in the discussion section of this Chapter. On the one side, the analytical and chemometric strategies from **scientific publication VII** were already optimized by the host research group at IDAEA [8–10]. Previous works from the group also merged the information of different omics (e.g., metabolomics and transcriptomics) [11] or the metabolomics and lipidomics information from diverse analytical platforms (e.g., LC-MS and nuclear magnetic resonance (NMR) [12]). In this PhD Thesis, the main novelty is the simultaneous combination of both metabolomic and lipidomic LC-MS platforms in the same environmental assessment. On the other side, the analytical workflows from **scientific publications VIII** were designed by the group of Prof. Craig Wheelock for metabolomics analysis in clinical applications [13,14], and were employed during a three-month research stay in 2021 in the Department of Medical Biochemistry and Biophysics of the Karolinska Institutet (KI, Solna, Sweden). Prof. Wheelock is the Principal Investigator of the Integrative Molecular Phenotyping laboratory, which is one of the leading groups in the development of liquid and gas chromatography coupled to mass spectrometry methods and bioinformatics approaches for the quantification of lipids and other metabolites at the population level. The research stay carried out in 2021 had a primarily formative objective, i.e.,

the familiarization with targeted analysis strategies from the analytical perspective, and also learning alternative untargeted workflows for metabolomics.

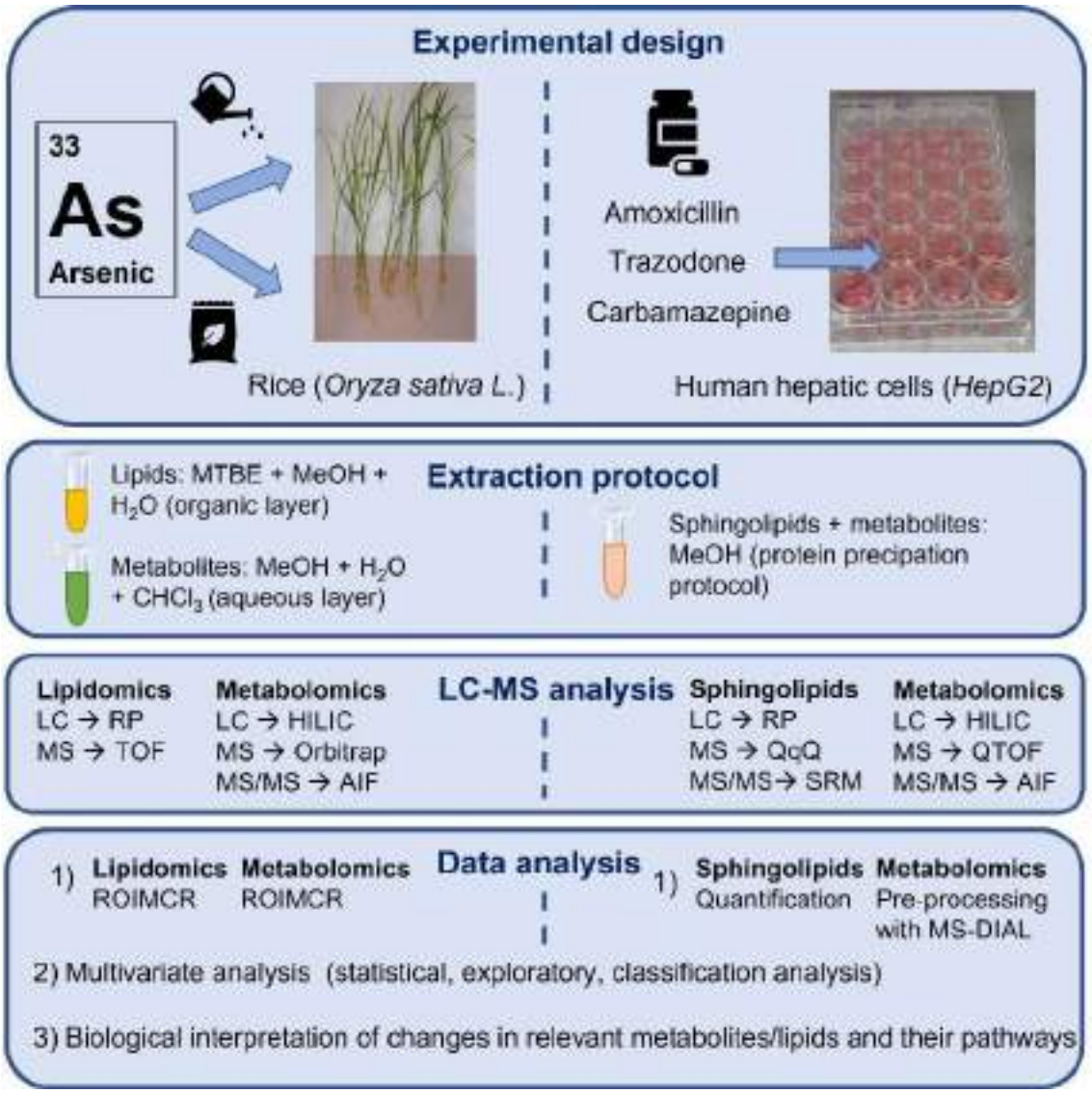


Figure 5.1. Scheme of the metabolomic/lipidomic workflows employed in **scientific publications VII and VIII.**

In summary, this Chapter is focused on discussing the metabolomics/lipidomics workflows employed in **scientific publications VII and VIII** and their main advantages and opportunities from the data analysis perspective.

5.2 Scientific publications

This section includes a brief summary of the studies presented in **scientific publications VII and VIII**:

SCIENTIFIC PUBLICATION VII

Rice crops are especially vulnerable to arsenic exposure compared to other cereal crops, mainly due to flooding conditions in which rice is grown, which facilitates its uptake. Because rice and rice-based products are among the main food staples worldwide, they are also the most consumed products with higher arsenic content, becoming a threat to the environment and human population. In this work, arsenic exposure in rice is assessed with two treatments (supplying arsenic to rice crops at early stages through irrigation or through the soil), and two concentration levels. Although considerable effort has been put in understanding how this metalloid is translocated and accumulated once it has accessed through the roots, this study aims to shed some light on the mode of action of arsenic in rice and the changes caused in the metabolome and lipidome of the crops. Untargeted metabolomics and lipidomics platforms included one-dimensional liquid chromatography coupled to mass spectrometry (LC-MS) as instrumental analytical platform and regions of interest – multivariate curve resolution (ROIMCR) as chemometric data analysis approach.

SCIENTIFIC PUBLICATION VIII

In recent years, pharmaceutical compounds have arisen as one of the main emerging contaminants (ECs) because their consumption and release into the environment have considerably increased worldwide. The goal of this study is to assess the effects caused by three widely consumed hepatoxic pharmaceutical compounds: an antibiotic (amoxicillin), an antiepileptic (carbamazepine), and an antidepressant (trazodone) at environmentally relevant concentrations. A combination of an untargeted metabolomic workflow and a targeted sphingolipid platform has been selected to unravel the metabolic alterations in human hepatic cells exposed to these ECs at three concentrations of exposure for 24 hours. Univariate and multivariate statistical methods were employed for discriminating the most affected metabolites and sphingolipids for each drug exposure. Therefore, this study allowed identifying the main metabolic pathways that suffered changes due to the exposure to the pharmaceutical compounds.

VII. SCIENTIFIC PUBLICATION VII

Title: Adverse Effects of Arsenic Uptake in Rice Metabolome and Lipidome Revealed by Untargeted Liquid Chromatography Coupled to Mass Spectrometry (LC-MS) and Regions of Interest Multivariate Curve Resolution

Authors: Miriam Pérez-Cova, Romà Tauler and Joaquim Jaumot

Citation reference: Separations 9 (2022) 79.

DOI: 10.3390/separations9030079

separations

MDPI

Article

Adverse Effects of Arsenic Uptake in Rice Metabolome and Lipidome Revealed by Untargeted Liquid Chromatography Coupled to Mass Spectrometry (LC-MS) and Regions of Interest Multivariate Curve Resolution

Miriam Pérez-Cova [1,2], Romà Tauler [1] and Joaquim Jaumot [1,*]

[1] Department of Environmental Chemistry, Instituto de Diagnóstico Ambiental y Estudios del Agua, Jordi Girona 18-26, 08034 Barcelona, Spain; mpcqam@idaea.csic.es (M.P.-C.); rtaqam@idaea.csic.es (R.T.)
[2] Department of Chemical Engineering and Analytical Chemistry, University of Barcelona, Diagonal 647, 08028 Barcelona, Spain
* Correspondence: joaquim.jaumot@idaea.csic.es

Abstract: Rice crops are especially vulnerable to arsenic exposure compared to other cereal crops because flooding growing conditions facilitates its uptake. Besides, there are still many unknown questions about arsenic's mode of action in rice. Here, we apply two untargeted approaches using liquid chromatography coupled to mass spectrometry (LC-MS) to unravel the effects on rice lipidome and metabolome in the early stages of growth. The exposure is evaluated through two different treatments, watering with arsenic-contaminated water and soil containing arsenic. The combination of regions of interest (ROI) and multivariate curve resolution (MCR) strategies in the ROIMCR data analyses workflow is proposed and complemented with other multivariate analyses such as partial least square discriminant analysis (PLS-DA) for the identification of potential markers of arsenic exposure and toxicity effects. The results of this study showed that rice metabolome (and lipidome) in root tissues seemed to be more affected by the watering and soil treatment. In contrast, aerial tissues alterations were accentuated by the arsenic dose, rather than with the watering and soil treatment itself. Up to a hundred lipids and 40 metabolites were significantly altered due to arsenic exposure. Major metabolic alterations were found in glycerophospholipids, glycerolipids, and amino acid-related pathways.

Keywords: rice; arsenic; metabolomics; lipidomics; LC-MS; ROIMCR

check for updates

Citation: Pérez-Cova, M.; Tauler, R.; Jaumot, J. Adverse Effects of Arsenic Uptake in Rice Metabolome and Lipidome Revealed by Untargeted Liquid Chromatography Coupled to Mass Spectrometry (LC-MS) and Regions of Interest Multivariate Curve Resolution. *Separations* **2022**, *9*, 79. https://doi.org/10.3390/separations9030079

Academic Editor: Szymon Bocian

Received: 19 February 2022
Accepted: 16 March 2022
Published: 18 March 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Nowadays, there is a need for a better understanding of molecular processes that take place in cereal crops (e.g., rice (*Oryza sativa* L.), maize (*Zea mays* L.), or wheat (*Triticum aestivum* L.)), which are one of the major food staples worldwide [1]. This increasing knowledge leads to an enhancement in their growth, production, and quality [2]. Omics sciences (e.g., genomics, transcriptomics, proteomics, or metabolomics) have arisen as powerful tools for crop improvements from different biological perspectives. These technologies allow, for instance, studying developmental stages and breeding to increase yield, quality, and the bioavailability of nutrients [3–6]. Besides, these omics have been employed in the assessment of the effects of biotic stresses on crops, such as viruses and bacteria [7], but also environmental stressors such as high temperatures [8], salinity [9], or emerging contaminants (e.g., microplastics [10] and heavy metals [11,12]).

Among cereals, rice (*O. sativa* L.) is the most suitable candidate to sequence its DNA and perform genetic modifications, due to its small and well-mapped genome. Hence, rice is currently considered the second-best plant model, after *Arabidopsis thaliana*. Rice metabolome has been widely studied [13–15]. However, metabolite annotation is still a major challenge, due to the huge number of metabolites present in *O. sativa* L. and the

limitations of currently available databases [16]. From the different emerging contaminants, there is a special interest in evaluating the effects of heavy metals and metalloids on rice [17,18]. The reason is that metals tend to accumulate in different parts of the plant (e.g., in the leaves, roots, or stems), becoming a concern not only for the development of the organism itself, but also for human populations through the food chain [19]. Arsenic is a metalloid included in the top ten chemicals of major public health concern by the World Health Organization (WHO) [20]. Arsenic contamination has been related to natural sources like volcanism and geothermal activity, also in addition to anthropogenic sources such as industrial and agricultural activities [21–23]. The arsenic occurrence has been reported to be up to 183 $\mu g\ L^{-1}$ in groundwaters [24] and 8 $mg\ kg^{-1}$ in agricultural soils [25], but can also be highly absorbed by composts and biochars [26]. Besides, recent changes in land-use have led to an increase in arsenic levels, compromising human health [27]. The Food and Drug Administration (FDA) from the United States stated that rice presented the second highest arsenic levels, following seafood, and was the most consumed product with high arsenic content, due to its presence in many daily products [28]. Arsenic levels in polished rice have been reported up to 0.629 and 0.055 $mg\ kg^{-1}$, for total arsenic and inorganic arsenic content, respectively [29]. From the different arsenic forms, the inorganic is the most toxic. It is highly bioavailable because roots capture it and accumulate in the edible parts using the same transport system as silicon or phosphorus [30]. Therefore, the accumulation and translocation of arsenic is a major concern in rice crops [31–33].

Untargeted metabolomics (and lipidomics) based on liquid chromatography coupled to mass spectrometry (LC-MS) seem suitable approaches for the discovery of unknown metabolites and lipids, respectively [34]. On the one hand, LC-MS is a versatile analytical technique that allows the identification and quantification of a variety of metabolites, ranging from small and polar to big and non-polar, without the need for derivatization steps [35–37]. On the other hand, the goal of untargeted approaches is to have a qualitative profile of the major changes in metabolic pathways due to, for instance, exposure to certain stressors. Therefore, the focus is not on individual metabolites or lipids, but rather a global perspective without any a priori assumptions on the effects these exposures may have on specific biological pathways [38]. This work employed untargeted LC-MS metabolomic and lipidomic workflows to characterize rice metabolome and lipidome, respectively.

Here, the objective is to shed some light on the absorption and translocation (uptake) mechanisms of arsenic in rice together, with its potential impact on their metabolome and lipidome. Previous research in our group already addressed how arsenic affects rice lipidome when supplied by watering the crop [39]. This study aims to complement the previous work with a comparison with the scenario where arsenic content comes from contaminated soil. Therefore, arsenic was supplied through two main routes: watering with contaminated water or soil containing arsenic. In addition, this new study includes metabolomic as well as lipidomic analysis, in order to have a more global overview of arsenic exposure. Two analytical platforms based on LC-MS were employed for the analysis of polar and non-polar metabolites, including lipid species, that were affected by this contaminant. First, a compression strategy based on regions of interest (ROI) was performed to filter the data feature matrices. This step was followed by the application of multivariate curve resolution alternating least squares (MCR-ALS) analysis to resolve the elution and the spectra profiles of the compounds of interest and obtain the areas of the chromatographic peaks necessary to perform subsequent multivariate analyses. The untargeted metabolomic and lipidomic workflows proposed were able to discover potential markers of arsenic exposure and facilitate the identification of the metabolic pathways affected in the different treatments.

2. Materials and Methods

2.1. Chemicals and Reagents

Sodium arsenate dibasic heptahydrate (≥98.0%), calcium carbonate ($CaCO_3$, ≥99.0%), ammonium acetate (≥99.0%), acetic acid (≥95.0%), ammonium formate (≥98.0%), and

formic acid (≥95.0%) were purchased from Sigma-Aldrich (St. Louis, MO, USA). HPLC grade water, HPLC grade acetonitrile (AcN), HPLC grade methanol (MeOH), methyl tert-butyl ether (MTBE), and chloroform ($CHCl_3$) were supplied by Merck KGaA (Darmstadt, Germany). Water used for plant irrigation and preparing arsenic solutions was filtered through a 0.22 µm nylon filter and purified with an Elix 3 Milli-Q system (Millipore, Belford, MA, USA).

For the lipidomics study, thirteen lipid standards from several lipid families were used as extraction standards: 17:0 monoacylglycerol, 1,2,3-17:0 triglyceride, 17:1 lysophosphatidylethanolamine, 17:0 lysophosphatidylcholine, 17:0 lysophosphatidic acid, 17:0 lysophosphatidylglycerol, 17:0 lysophosphatidylserine, 17:0 cholesteryl ester, 16:0D31-18:1 phosphatidic acid, 16:0D31-18:1 phosphatidylethanolamine, 16:0D31-18:1 phosphatidyl glycerol, 16:0D31-18:1 phosphatidylcholine, and 16:0 D31-18:1 phosphatidylserine. Three sphingolipids were used as internal instrumental standards: N-lauroyl-D-erythro-sphingosylphosphorylcholine, N-(dodecanoyl)-1-ß-glucosyl-sphing-4-ene, and N-(dodecanoyl)-sphing-4-enine. All these lipid standards were purchased from Avanti Polar Lipids (Alabaster, AL, US). For the metabolomics study, L-methionine sulfone and piperazine-1,4-bis (2-ethanesulfonic acid) (PIPES) were used as the extraction and internal instrumental standards, respectively, and were purchased from Sigma-Aldrich (St. Louis, MO, USA).

A stock solution of arsenic (V), from now on As (V), at 10,000 µM, was prepared from the sodium arsenate salt. For the watering treatment, solutions containing 1 and 1000 µM of As (V) were prepared weakly by diluting the initial concentrated stock. For the soil treatment, solutions of 5 and 50 mg L^{-1} were prepared directly from the sodium arsenate salt. The solution containing 0.001 µM of As (V) used for watering the soil treatment harvest was prepared daily diluting from the initial concentrated stock.

The following abbreviations have been used to describe lipid families: lysophosphatidic acid (LPA), lysophosphatidylcholines (LPC), lysophosphatidylglycerol (LPG), lysophosphatidylethanolamine (LPE), lysophosphatidylserine (LPS), phosphatidic acid (PA), phosphatidylcholines (PC), phosphatidylglycerol (PG), phosphatidylinositols (PI), phosphatidylethanolamine (PE), lysophosphatidylserine (LPS), sphingomyelin (SM), ceramides (Cer), hexosylceramide (HexCer), monogalactosyldiacylglycerol (MGDG), digalactosyldiacylglycerol (DGDG), sulfolipid sulfoquinovosyldiacylglycerol (SQDG), diacylglycerol (DG), triacylglycerol (TG), fatty acid (FA), cholesteryl ester (CE), sterol lipid (ST), and eicosanoyl-EA (NAE).

2.2. *Plant Growth, Arsenic Treatments, and Extraction Protocols*

2.2.1. General Growing Conditions and Harvesting

Plant growth and lipid extraction were performed using the procedure described elsewhere [39–41]. Briefly, rice seeds were obtained from the Centre for Research in Agricultural Genomics (CRAG, Bellaterra, Spain). Seeds were incubated for 48 h at 30 °C in an oven (J.P. Selecta) in a wet environment. After incubation, plants were grown on an Environmental Test Chamber MLE-352H (Panasonic) for 22 days, where cyclic environmental changes of temperature and light intensity were simulated, as shown in Supplementary Material A Figure S1. Soil employed for planting included a mixture of peat, vermiculite, fertilizer, and $CaCO_3$. Plates containing different samples were placed in random order inside the chamber, and re-located each watering cycle, established at three times per week.

During the harvest, roots and aerial tissues (i.e., corresponding to the part of the plant above ground) were separated, quenched with liquid nitrogen, and kept at −80 °C until extraction. Before extraction, rice samples were ground to a fine powder with a liquid nitrogen mortar and lyophilized to dryness for 24 h.

2.2.2. Watering and Soil Treatments

For the watering treatment, during the first 11 days, rice was irrigated with Milli-Q water. From that day until harvesting, plants were watered with 1 and 1000 µM of As (V) for the two concentration levels of exposure, and with Milli-Q water for control samples.

European legislation established the lowest concentration at 1 µM as it is the limit of the acceptable arsenic concentration in water [42]. The upper concentration was set at 1000 µM, a threshold established to ensure that the experiment was performed under sub-lethal arsenic concentration for the plant, based on previous studies [39].

For the soil treatment, two containers were prepared with 1 kg of soil two days before planting. Soil from the container was exposed to two arsenic concentration levels (5 and 50 mg L^{-1}). Once sowing, rice was irrigated for the whole growth period with a solution containing 0.001 µM of As (V). The lowest arsenic limit in this treatment was set at 5 mg L^{-1} as a maximum value of common arsenic leaches without toxic characteristics [43]. However, the background soil content of arsenic varies between 1 and 40 ppm, according to the US Food and Drug Administration (FDA) report [28]. The highest arsenic limit was established to 50 mg L^{-1}, as a considerably high arsenic content in the soil, slightly above the maximum frequently encountered levels. Table S1 summarizes the arsenic concentration levels selected in this work, expressed in µM for the sake of clarity. The two treatments are referred to with a W (watering) or an S (soil), followed by the concentration dose (L for low and H for high).

2.2.3. Lipid Extraction

A general lipid extraction for untargeted analysis was performed following a previous extraction protocol [39,44]. Briefly, 5 mg of the dried tissue were weighted in individual tubes for each replicate and dissolved in 1 mL of MTBE:MeOH (3:1). Extraction standards mix were added (10 µL at 20 µM, per sample), and then, the mixture was vortexed for 1 min and sonicated for 10 min. Afterwards, 0.5 mL of H_2O:MeOH (3:1) were added, vortexed for 1 min again, and centrifuged for 5 min at 14,500 rpm. The upper organic fraction was collected, whereas the lower aqueous phase was re-extracted with 0.65 mL of MTBE and 0.35 mL of MeOH:H_2O (1:0.85), vortexed for 1 min and centrifuged for 5 min at 2000× *g*. Next, combined organic phases were evaporated to dryness under nitrogen gas. Extracts were stored at −20 °C until analysis, and resuspended before injection with 250 µL of MeOH:H_2O (4:1). Finally, 10 µL of the internal standards mix at 20 µM were added to each sample.

2.2.4. Metabolite Extraction

For metabolite extraction, based on previous works from Ortiz-Villanueva et al. [45] and Navarro-Reig et al. [46], 40 mg of the dried tissue were weighted in individual tubes for each replicate and 1 mL of MeOH, and 50 µL of L-methionine sulfone (L-met) at 50 mg L^{-1} were added, acting as a surrogate. The mixture was vortexed for 1 min, and sonicated for 10 min, twice. Then, it was centrifuged at 14,500 rpm, and 750 µL of the supernatant were taken, and mixed with 500 µL of $CHCl_3$ and 400 µL H_2O. Next, it was vortex for 1 min, kept during 15 min at 4 °C, and centrifuged again for 20 min at 14,500 rpm. The aqueous fraction was collected and evaporated to dryness under nitrogen gas. Extracts were stored at −20 °C until analysis, and resuspended before injection with 450 µL of AcN:H_2O (1:1). Finally, 50 µL of 50 mg L^{-1} solution of the instrumental internal standard, PIPES, was added.

2.3. *LC-MS Analysis*

Five biological replicates were analyzed for each sample condition (control, low, and high exposure concentrations), each treatment (watering or soil treatments) and each extraction type (lipid or metabolite extractions). In total, 60 samples were analyzed in each analytical platform (lipidomics or metabolomics). In addition, quality control (QCs) pools composed of 70 (lipidomics) or 50 µL (metabolomics) of solution of each sample condition were prepared separately for different tissues (roots or aerial parts) and extraction types (lipid or metabolite extractions). QCs were repeatedly analyzed during the chromatographic batch every five samples.

2.3.1. Lipidomic Analysis

The lipidomic analysis was performed using a Waters Acquity UPLC system (Waters Corporation, MA, USA), connected to a Waters LCT Premier orthogonal accelerated time of flight mass spectrometer (Waters), operated in both positive and negative electrospray (ESI) ionization modes. Full scan spectra were acquired from 50 to 1500 Da at a scan cycle time of 0.3 s. The following parameters were set for positive ionization mode: capillary voltage, 3000.0 V; sample cone voltage, 50.0 V; desolvation temperature, 350.0 °C; source temperature, 100.0 °C; desolvation gas flow 600.0 L h^{-1}. The same parameters were also used for negative ionization mode, except capillary voltage, set to 2800.0 V instead.

The chromatographic column employed was a Kinetex C8 (100 × 2.1 mm, 1.7 µm) (Phenomenex) under the following conditions (already used in [47]): temperature at 30 °C, injection volume at 10 µL, and flow rate at 0.3 mL min^{-1}. Mobile phases selected were (A) MeOH 1 mM ammonium formate and (B) H_2O 2 mM ammonium formate, both at 0.2% formic acid. The gradient started at 80% A, increased to 90% A in 3 min, from 3 to 6 min remained at 90% A, changed to 99% A until minute 15, remained constant 1 min, and returned to initial conditions until minute 20.

2.3.2. Metabolomic Analysis

The metabolomic analysis was performed using a Waters Acquity UPLC system connected to a Q-Exactive (Thermo Fisher Scientific, Hemel Hempstead, UK) equipped with a quadrupole-Orbitrap mass analyzer. Electrospray (ESI) was used as an ionization source in both positive and negative ion modes. Full scan mass range was set from m/z 90 to 1000. The following parameters were set for positive ionization mode: electrospray voltage, 3000.0 V; sheath gas flow rate, 25 arbitrary units (a.u.); auxiliary gas flow rate, 10 a.u.; capillary temperature, 350 °C; and S-lens level, 60%. Negative ionization mode conditions were the same, except for the sheath gas flow rate, set to 40 a.u. All ion fragmentation (AIF) was performed with normalized collision energy (NCE) of 35 eV.

The column employed was a HILIC TSK gel amide-80 column (250 × 2.0 mm i.d., 5 µm) provided by Tosoh Bioscience (Tokyo, Japan), under the following experimental conditions (already employed in [45]): flow rate at 0.15 mL min^{-1}, at room temperature, and 5 µL injection volume. Mobile phases were (A) AcN and (B) 5 mM ammonium acetate, adjusted at pH 5.5 with acetic acid. The gradient employed was: starting conditions at 25% B, then increased until 30% B in 8 min; a 60% B was reached at 10 min, held for 2 min more and then back to 25% B until minute 14 min; lastly, a re-equilibration step was added and from 14 to 20 min at 25% B.

2.4. *Data Analysis*

2.4.1. Data Compression, Filtering, and Normalization

For data acquisition control and initial data preprocessing, MassLynx 4.1 (Waters Corporation, MA, USA) and Thermo Xcalibur 3.1.66.7 (Thermo Scientific, Hemel, UK) were used for lipidomics and metabolomics studies, respectively. In lipidomics analysis, LC–MS raw files (.raw) were converted to the 'common data format', cdf files using the DataBridge file converter tool available from MassLynx software suite. In the metabolomic analysis, raw files were converted into mzXML format using MS Convert GUI (Palo Alto, CA, USA) using the Proteowizard open-source software [48].

Raw data were then imported to MATLAB computer and visualization environment (Release 2020a, The Mathworks Inc, Natick, MA, US) and analyzed with the ROIMCR chemometrics strategy [49]. This approach was employed for data compression and filtering on one side, and for the resolution of the elution and mass spectra profiles of the different constituents (metabolites or lipids) present in the analyzed rice samples. More information about the ROIMCR approach can be found in Supplementary Material A Section S2. Briefly, spectral compression based on regions of interest (ROI) was performed through the MATLAB MSroi app [50]. ROI strategy allows significant data compression in the spectral dimension without losing their instrumental spectral accuracy. The approach

establishes an intensity threshold value, and MS signals below this threshold are discarded (considered noise). Two additional parameters should be defined, the mass error tolerance (related to the mass spectrometer maximum spectral resolution), and the minimum number of values (minimum number of MS signal occurrences) required to define a chromatographic peak across all the samples (which depends on the type of chromatographic column and conditions used). The main parameters used for the ROI procedure in this work are, briefly, mass error tolerance of 30 and 10 ppm (lipidomics and metabolomics analysis, respectively), a minimum signal factor of 2, a minimum occurrence value of 100, and ROIs were calculated with the median of the m/z values determined for each chromatographic peak. More details can be found in Supplementary Material A Table S2. The ROI approach provides two main outputs: a vector including the list of relevant m/z ROI values (according to the previously mentioned parameters selected for the analysis), and a data matrix with the MS intensities at the selected ROIs (for all considered retention times and samples).

Then, MCR-ALS was applied to the ROI feature data matrices obtained by the workflow described above. Details on the MCR-ALS procedure and parameters employed in the analysis of the data sets in this work are given in Supplementary Material A Section S2. Briefly, MCR-ALS is a bilinear model that decomposes the original data matrix into two-factor matrices related to the elution and spectral profile of the different components. Ideally, each component can be associated with lipid or metabolite constituents of the analyzed samples and possible contributions to the solvent and backgrounds instrument signals. The sample constituents can be identified using the information from the MCR-ALS resolved spectra profiles. MS signals from the same chemical compound, including multiple isotopic forms or adducts and possible mass and ion fragments, are merged in the same MCR-ALS component (i.e., componentization). On the other hand, quantitative information can be retrieved from the elution profiles of the resolved MCR-ALS components, by integrating the areas of their resolved chromatographic peaks. Hence, a data matrix containing the peak areas of each MCR-ALS component is one of the outputs of this method (i.e., component matrix). In this work, four peak area data matrices were obtained for each analytical platform (i.e., lipidomics and metabolomics). Each data matrix corresponded to a specific tissue of the rice plant (roots and aerial parts) and an electrospray ionization mode (positive and negative mode). Finally, these peak areas were normalized by the internal standards added before instrumental analysis, the surrogates employed to correct extraction losses and the dried weight of each replicate. QCs were used as an internal check of the data quality, obtaining similar values within each batch and between different batches. Therefore, no further normalization based on QCs was required.

2.4.2. Statistical Assessment, Exploratory Analysis, and Discovery of Markers of the Exposure

Chemometric analysis of the normalized peak areas of the different components resolved by MCR-ALS was performed with the PLS Toolbox 8.9.1 (Eigenvector Research Inc, Wenatchee, WA, US) under MATLAB (Release 2020b, The Mathworks Inc, Natick, MA, US) Different types of data analyses were applied for statistical assessment, exploratory analysis, and discrimination analysis of markers of the exposure.

The first step was the statistical assessment of the different rice sample treatments with ANOVA-simultaneous component analysis (ASCA) [51]. ASCA combines the multivariate analysis of variance, ANOVA, and simultaneous component analysis (SCA). The null hypothesis of ASCA is that the experimental factors from the experimental design have no effect on the observed results. ASCA was applied to the component matrices for both sample treatments (soil and watering), at the different concentration levels (high, low and control samples). Statistical assessment is performed by a permutation test considering 10,000 replicates.

Principal component analysis (PCA) [52] and hierarchical clustering analysis (HCA) [53] were used for the exploratory study of the effects produced by the different treatments

and conditions, on each of the MCR-ALS component peak area data matrices (related with the metabolites and lipids present in the analyzed samples). PCA describes the experimental data variation in a few components or contributions, explaining the most relevant information from the original variables. The scores plot visualizes the major trends in samples, clustering or discerning them according to their different levels of exposure compared to control samples. In this work, PCA was especially useful to analyze sample trends in an unsupervised manner (no prior information is provided about the different sample classes, i.e., type of treatment or arsenic concentration levels). Biological replicates are expected to cluster together, whereas control samples and samples at similar exposure concentration levels will hopefully cluster separately. On the other hand, HCA using a dendrogram (clustergram) representation allows visualizing trends in the different compounds (i.e., lipids/metabolites that cluster due to similar behavior) and in the samples (i.e., samples ordered by sample type). HCA was performed on data matrices with the fold changes (FC) in the logarithm scale. FC is a standard measurement in metabolomics to compare how much an original condition (control) has changed when related to another condition (exposed or treated). Thus, in untargeted type of data analysis, FCs are usually calculated as the ratio between areas of exposed samples divided by the areas of control samples. Therefore, FCs are expressed as relative abundances. In this case, the area of each component and each exposed replicate was divided by the mean value of the control samples. In this work, HCA was performed using only the more significant peak areas of the MCR-ALS resolved components from each dataset.

Finally, partial least squares discriminant analysis (PLS-DA) [54,55] was applied to the same data matrix of the peak areas of the MCR-ALS resolved components. PLS-DA is a useful and powerful approach for discriminating samples from a supervised perspective. Contrary to PCA, the model is built using information regarding the class membership of each sample (e.g., watering or soil treatment, and exposure levels). The analysis was performed by considering pairs of exposure concentrations (e.g., control samples versus lowest exposure level of the watering treatment, C vs. WL, etc.). A leave-one-out cross-validation method was applied. Variables important in projection (VIPs) of the PLS-DA models allow the identification of possible markers of arsenic exposure and unravel the uptake mechanisms by comparing the significant MCR-ALS components resulted from the different treatments, especially against control samples. Matthews correlation coefficient (MCC) was evaluated as an indicator of the quality of the binary classifications, ranging from −1 to 1 (1 represents a perfect model, −1 a wrong prediction, and 0 random predictions) [56]. In this work, the variables (peak areas of the different MCR-ALS components) associated with VIPs higher than 1.0 were considered relevant according to the various As (V) treatments. Each significant component was then associated with the most intense m/z value from their spectral profile for annotation purposes. Hence, only the compounds related to the relevant MCR-ALS components (with higher VIP values) were finally investigated and annotated.

For PCA and PLS-DA analysis, MCR-ALS resolved component peak area matrices were normalized with probabilistic quotient normalization (PQN) and autoscaled. For ASCA, the same peak areas data matrices were mean-centered. In HCA analysis, a logarithmic normalization was applied to the fold peak area changes before analysis.

2.4.3. Compound Identification

Relevant compounds (MCR-ALS components whose peak areas changed significantly between treatments) from the PLS-DA analysis were selected for identification, due to their implication in the metabolic changes caused by the different arsenic treatments. On one side, lipids were identified according to an in-house built database composed of a list of retention times (RT) associated with compounds frequently detected in plant matrices using the same LC-MS method employed in this work. Besides, LIPIDMAPS [57] and online spectral library human metabolite database [58] were also used for lipid annotation,

selecting the candidates that provided a lower error comparing m/z values of the mass spectra resolved by MCR-ALS and the theoretical one.

In addition, metabolites present in QC samples could be confirmed based on MS/MS spectral matches using public metabolite libraries from the MS-DIAL website [59]. Theoretical and experimental spectra provided by HMDB [58], Massbank [60], and Global Natural Product Social Molecular Networking (GNPS) [61] were compared with our experimental MS/MS data. Plantcyc online database [62] was employed to confirm whether the annotated compounds have been previously found in rice (*Oryza sativa* L.).

3. Results

3.1. Statistical Assessment and Exploratory Analysis of Arsenic Exposure

First, ASCA was employed to evaluate the statistical significance of the experimental treatment (watering/soil) and the As (V) concentration levels (high/low/control), as well as the interaction between these two factors. Both "treatment" factor, and the potential interaction between "treatment" and "As (V) concentration" resulted in being not significant, whereas the "As (V) concentration" factor (with all levels considered at a time) was statistically significant in all cases (i.e., in both tissues: roots and aerial parts, in both ionization modes: positive and negative, and in both platforms: lipidomics and metabolomics; for the eight datasets analyzed in total). Individual studies were also analyzed at two concentration levels (e.g., C vs. WL, etc.) or simultaneously at all concentration levels (e.g., C vs. WL vs. WH). In lipidomics datasets, all combinations (even at the lowest concentration level) were significant (with p-values between 0.003 and 0.0001), regardless of the two types of tissue (roots or aerial) and ionization (positive or negative) modes. Metabolomic datasets exhibited the same behavior (p-values ranging from 0.0346 to 0.0001), with some exceptions. Indeed, the lowest concentration level in soil treatment (C vs. SL) and, consequently, soil treatment in general (C vs. SL vs. SH) were not statistically significant, neither in aerial parts positive ionization nor in roots negative ionization for the metabolomic datasets. Besides, the lowest watering treatment (WL) was not significant in aerial parts positive ionization. In conclusion, aerial parts positive ionization metabolomic set was the least affected by the arsenic exposure in this study. The only clear significant factor for this data set was watering at the highest concentration level (with p-values C vs. WH: 0.0249, and C vs. WL vs. WH: 0.0045). All ASCA results are summarized in Table S3 in Supplementary Material A.

Second, PCA was applied to all datasets to visualize the effects of arsenic exposure. In all cases, more than 40% of all data variance was explained only with the first two components (PC1 and PC2) of the model. On one side, PCA scores plots of the lipidomic datasets showed a clear differentiation between control and exposed groups (samples were separated by PC1 in aerial tissues and by PC2 in roots), as expected and in agreement with previous ASCA results. In addition, a distinction between the two treatments (watering and soil) was found in the analysis of root samples. In contrast, aerial tissue samples were separated in PC2 accordingly to the concentration level, rather than with the treatment itself. Figure 1A,B summarize this trend for both tissues in negative ionization mode, although a similar tendency was obtained for positive mode as well (Figure S3A,B in Supplementary Material A). Hence, root lipids were affected differently according to how rice was exposed to arsenic (from soil or watering). However, this discernment was not present for lipids in aerial parts of the plant, which were more affected by the total arsenic content.

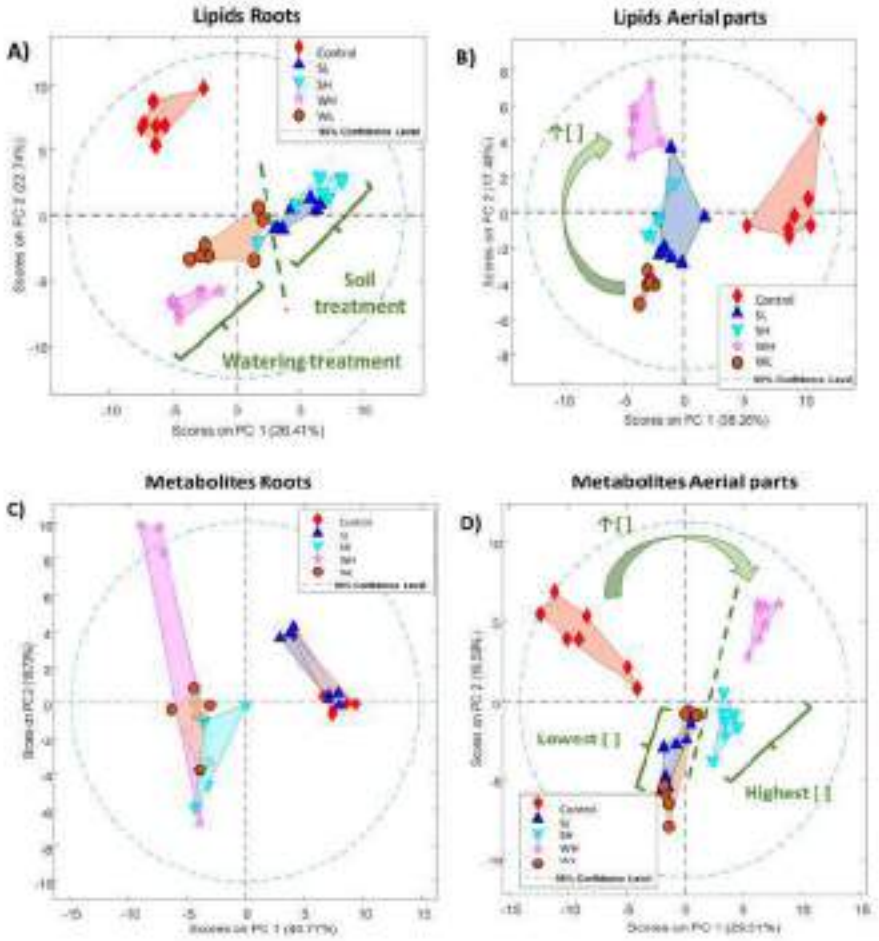


Figure 1. PCA score plots are shown for negative ionization mode obtained in the analysis of both tissues, roots, and aerial parts. (**A**,**B**) represent lipidomic analysis of roots and aerial tissues, respectively. Analogously, (**C**,**D**) refer to metabolomic analysis of roots and aerials. Both in lipidomic and metabolomic samples from root tissue are more affected by treatment rather than by concentration level, whereas aerial parts have the opposite scenario.

On the other side, similar exposure effects were observed for metabolomic datasets. Concentration levels differentiated metabolites from aerial parts, as shown in Figure 1D for negative mode (and in Figure S3D in Supplementary Material A for positive mode). Three clusters apart from controls were identified, one corresponding to the two lowest exposures (WL and SL), and the other two levels in an increasing order regarding its concentration (WH the most isolated). PCA score plots for metabolites in roots negative ionization (Figure 1C) were basically defined by the lack of differentiation between control and SL groups, separated in PC1 (also observed with ASCA). However, a closer look at PCAs by treatment (only considering C, WL, and WH) revealed that WH and WL clustered together separated from control samples in PC1, which explained 47% of the variance (data not shown). The same C-WL-WH trend was observed for roots positive ionization (Figure S3C in Supplementary Material A). Again, this fact confirmed that the treatment itself affected metabolites from roots more than by the different concentration levels.

3.2. *MCR-ALS Component Selection and Annotation*

PLS-DA models were built to identify the MCR-ALS components responsible for sample discrimination between different treatments and concentration levels (comparing control vs. treated samples by pairs). Thus, the chemical compounds associated with these components would be considered potential markers of arsenic exposure and helpful in unravelling metabolic changes caused in rice by this metalloid. Table S4 in Supplementary

Material A summarizes the total number of relevant VIPs > 1.0 and MCC values obtained for all datasets. Sample classification was excellent for all lipidomic datasets (MCC equal to 1.0 in all cases). Good discrimination was also achieved for all metabolomic datasets (MCC ranging from 0.7 to 1.0). The first 50 and 20 MCR-ALS components (for lipidomics and metabolomics analysis, respectively) with the higher VIP values for each pair of control vs. treated samples were considered significant, and therefore, contemplated for annotation.

Lipids were annotated using an in-house built retention time database in plant matrices and using the external aid of LIPIDMAPS [57] and HMDB [58]. Up to 100 significant lipids were annotated in total, considering both ionization modes. Metabolites MS/MS spectra from QCs samples were deconvoluted using MS-DIAL [63]. Up to 40 significant metabolites were annotated in total. Figure S4 in Supplementary Material A shows MS/MS spectrum match (experimental vs. theoretical) for L-tryptophan as an example. Tables S5 and S6 in Supplementary Material A list all parameters used in MS-DIAL analysis, employed exclusively for annotation purposes.

Annotation confidence corresponded to level 3 for lipids (no MS/MS information, only exact mass and retention time) and level 2 for metabolites (MS/MS, retention time, exact mass), according to the confidence level of compound annotation re-defined the Compound Identification workgroup of the Metabolomics Society in 2017 [64]. In these cases, when fragment ions were not detected under the mass range conditions of this study, the corresponding metabolites were only tentatively annotated (level 3). Supplementary Material B provides all significant annotated MCR-ALS components. Tables S7–S10 correspond to lipids grouped by tissue and ionization mode, whereas Tables S11–S14 are analogous for metabolites. In addition to compound information, each table furnishes details on which variables were significant for each of the treatments and concentration levels, fold change ratios (areas of all samples in one class divided by the mean area of control samples), and the global tendency of all replicates compared to controls (up/down).

3.3. *Lipidomic Results*

Annotated lipids, selected from those MCR-ALS resolved components whose peak areas were significant (higher VIPs from PLS-DA results, see above), mainly belonged to three lipid classes: glycerophospholipids (52%), glycerolipids (30%), and sphingolipids (12%). Figure 2A shows the proportion of significant lipids among each family in aerial parts, roots, and both tissues simultaneously. There is an increase of affected glycerolipids in roots (e.g., DGs, MGDGs, and DGDGs), in contrast with a slight increment in certain glycerophospholipids (e.g., PGs and PAs) in aerial tissues. Figure 2B depicts the number of significantly affected lipids is given for the different treatments. Most of the annotated lipids were found significant in all four treatments (WH, WL, SH, and SL) in comparison to control samples. Nevertheless, some specific compounds were related only to watering or soil exposure (e.g., four DGs suffered changes in roots when soil treatment was applied).

HCA was applied to the logarithm of the fold changes of the annotated MCR-ALS components to give a global perspective on lipid changes with the different treatments., There were primarily two clusters found for both tissues and ionization modes: upregulated (marked in red) and downregulated (marked in blue). HCA maps plus dendrograms (clustergrams) are included in Figure 3 for aerial positive ionization, and in Supplementary Figure S5 for aerial negative ionization, roots positive, and negative ionization. Besides, certain lipids in both aerial and root tissues exhibited differences regarding the treatments (watering and soil), although this effect was clearer in roots (in agreement with previous results from PCAs). Among the lipids that increased their concentration regarding controls in aerials, several LPCs, PAs, and DGs stood out, whereas PCs, PGs, and PSs generally decreased. Concerning roots, PCs have also reduced their concentrations, as well as DGDGs and MGDGs. Again, PAs abundances were also incremented due to arsenic exposure.

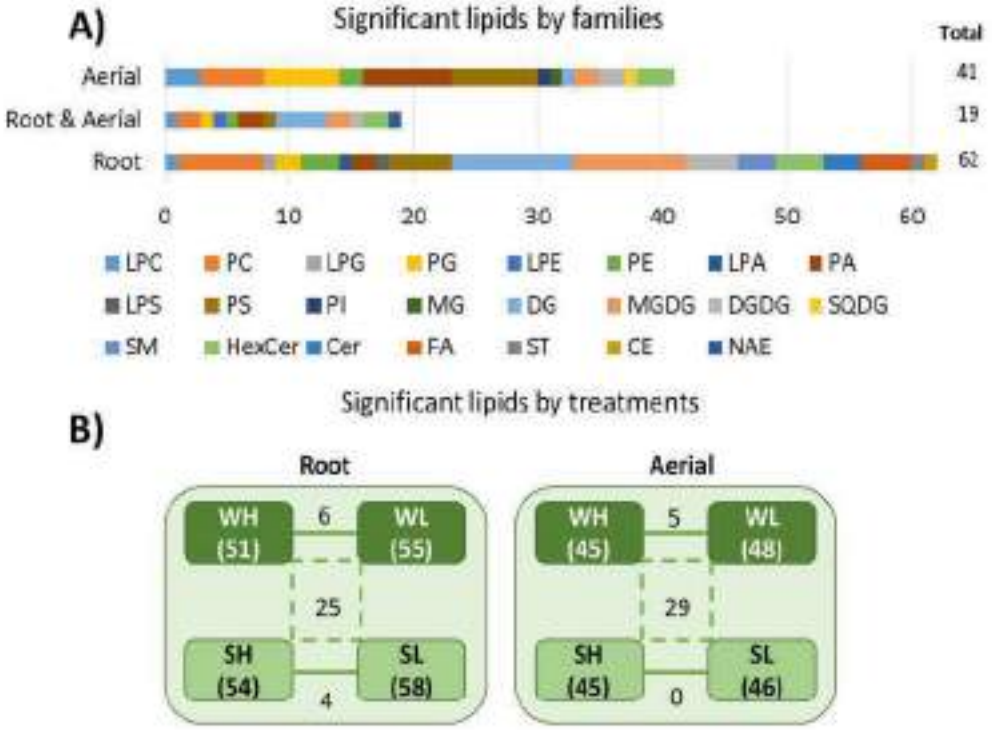


Figure 2. Representation of significant lipids annotated belonging to the different tissues, organized by (**A**) families and (**B**) specific treatments.

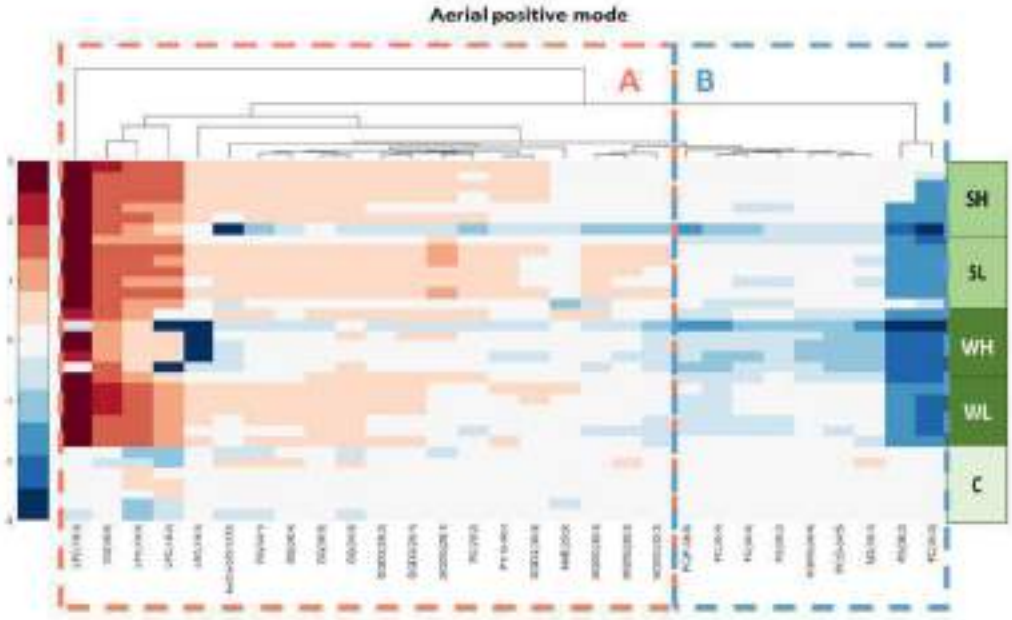


Figure 3. Hierarchical clustering heatmap applied to the logarithm of the fold changes of the annotated lipids for aerial positive ionization set. An intensity color bar is included on the left side of the figure, indicating the relative abundance of the lipid regarding control samples (higher abundance in red, lower abundance in blue). Two clusters are differentiated, with increasing abundance (**A**) and diminishing abundance (**B**).

Some potential markers of arsenic exposure had already been suggested in a previous lipidomics study [39]. In this previous work, the main aim was to develop a multidimensional chromatographic method and evaluate the effects only at the two concentration levels referred to as "watering treatment". However, the coincident lipids from the prior

work and this study did not show an exclusive behavior for watering exposure, but they were instead associated with increasing concentration levels of arsenic (e.g., PA (36:2), PG (34:1), or PC (36:6), which were also markers of soil treatment). The present work also pointed out new lipids highly affected by arsenic exposure regardless of the treatment (e.g., LPC(18:3) in aerial tissues). Besides, the present study showed new insights regarding how arsenic can access rice, and allowed detecting potential markers of the different treatments. For instance, lipids that showed the same tendency in both tissues but accentuated in one of them (e.g., PS (39:3), especially decreased with watering treatments in aerials; or LPE (18:2), more affected by soil treatments in roots). Other lipids presented other remarkable changes. For instance, LPC (16:0) augmented in aerials tissues for the lower doses, but decreased in roots at the highest doses. In other cases, relevant lipids were only annotated for a single tissue (e.g., PG (34:2), PG (32:2), or DG (34:2) diminished due to watering treatment in roots). Nevertheless, further complementary targeted studies and MS/MS confirmation are necessary to assess the effects of arsenic exposure in rice lipidome completely.

3.4. Metabolomic Results

Significant annotated metabolites in Tables S11–S14 were previously detected in *Oryza sativa L.* according to the Plantcyc database [62]. Contrarily to lipids, no clear specific effect based on treatments (watering/soil) was detected in roots. In addition, control and SL groups for this tissue in negative mode cannot be distinguished, as already observed in PCA and ASCA. Most of the annotated metabolites were found in WH, SH, and WL groups, as shown in Figure 4A, (e.g., palmitic acid, allantoin, norvaline, succinic acid, tryptophan, and isoleucine). In addition, the arsenic effect in aerial tissues was dominated by its concentration level rather than by the treatment itself, as previously seen in PCA results. MetaboAnalyst pathway analysis [65] was performed to have a closer look into the metabolic pathways that could be affected by arsenic exposure in general (without tissue differentiation). Table 1 exhibits a detailed list of these pathways, ordered by decreasing significance, including the number of significant metabolites found for each pathway, their *p*-values and False Discovery Rate (FDR) results. Moreover, Figure 4B graphically displays the obtained results, pointing out the main pathways with letters. The five principal metabolic pathways altered by arsenic exposure in this study were amino acid related, i.e., aminoacyl-tRNA biosynthesis (A); alanine, aspartate, and glutamate metabolism (B); glycine, serine, and threonine metabolism (C); phenylalanine, tyrosine, and tryptophan biosynthesis (H); arginine and proline metabolism (I); arginine biosynthesis (C) (Figure 4B). A pathway comparison based on the analyzed tissue is provided in Figure 4C. Overlapping altered pathways were found for both roots and aerials, but the individual metabolites related to these pathways were not necessarily the same. For instance, some common metabolites in both tissues were tryptophan, phenylalanine, serine, proline, glutamine, shikimic acid, allantoin, and succinic acid. On the contrary, adenosine, palmitic acid, and betaine were exclusively detected as significant for roots, and dimethylglycine, pyroglutamic acid, and benzoic acid were only significant in aerials. A more in-depth characterization of the specific metabolic routes affected (e.g., via targeted analysis) could complement these findings and confirm metabolic changes in both tissues according to the treatments applied. Besides, larger spectral databases for secondary metabolites in plants are still lacking, which in the end, still limits their potential discovery to those with an already available MS/MS spectrum.

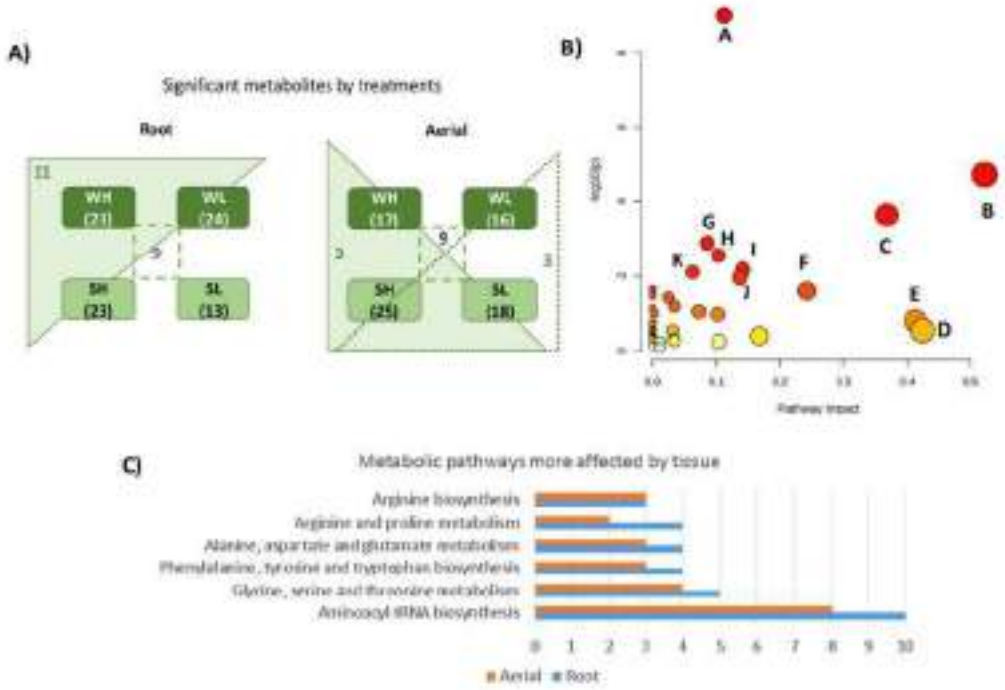


Figure 4. (**A**) Summary of the common metabolites expressed by the different treatments, with especial emphasis in the lower effect on SL group in roots. (**B**) Visualization of MetaboAnalyst results indicating the most affected metabolic pathways regarding arsenic exposure from Table 1, for the simultaneous analysis of all tissues and ionization modes. Letter code for the pathways corresponds to: (A) aminoacyl-tRNA biosynthesis; (B) alanine, aspartate, and glutamate metabolism; (C) glycine, serine, and threonine metabolism; (D) phenylalanine metabolism; (E) isoquinoline alkaloid biosynthesis; (F) tryptophan metabolism; (G) arginine biosynthesis; (H) phenylalanine, tyrosine, and tryptophan biosynthesis; (I) arginine and proline metabolism; (J) butanoate metabolism; (K) glyoxylate and dicarboxylate metabolism. (**C**) Comparison of the number of significant metabolites related to the main metabolic pathways affected by the exposure from (**B**), according to the tissue analyzed (roots vs. aerials).

Table 1. Metabolomic results from pathway analysis in Metaboanalyst online platform, for all tissues and ionization modes simultaneously. Metabolic pathways affected are ordered according to their significance.

Result from Pathway Analysis	Total	Expected	Hits	Raw p	$-\text{Log}_{10}$(p)	Holm Adjust	FDR	Impact
Aminoacyl-tRNA biosynthesis	46	1.31	12	9.86×10^{-10}	9.01	9.37×10^{-8}	9.37×10^{-8}	0.11
Alanine, aspartate, and glutamate metabolism	22	0.63	6	1.951×0^{-5}	4.71	1.85×10-3	9.26×10^{-4}	0.52
Glycine, serine, and threonine metabolism	33	0.94	6	2.29×10^{-4}	3.64	2.13×10-2	7.26×10^{-3}	0.37
Arginine biosynthesis	18	0.51	4	1.30×10^{-3}	2.89	1.20×10-1	3.09×10^{-2}	0.08
Phenylalanine, tyrosine, and tryptophan biosynthesis	22	0.63	4	2.86×10^{-3}	2.54	2.61×10-1	5.44×10^{-2}	0.10
Arginine and proline metabolism	28	0.80	4	7.08×10^{-3}	2.15	6.37×10^{-1}	1.09×10^{-1}	0.34
Glyoxylate and dicarboxylate metabolism	29	0.83	4	8.04×10^{-3}	2.09	7.16×10^{-1}	1.09×10^{-1}	0.06
Butanoate metabolism	17	0.48	3	1.11×10^{-2}	1.96	9.74×10^{-1}	1.51×10^{-1}	0.14
Valine, leucine, and isoleucine biosynthesis	22	0.63	3	2.27×10^{-2}	1.64	1.00	2.21×10^{-1}	0.00
Lysine biosynthesis	9	0.26	2	2.51×10^{-2}	1.60	1.00	2.21×10^{-1}	0.00
Tryptophan metabolism	23	0.66	3	2.76×10^{-2}	1.56	1.00	2.21×10^{-1}	0.24
Cyanoamino acid metabolism	26	0.74	3	3.54×10^{-2}	1.45	1.00	2.83×10^{-1}	0.00
Cysteine and methionine metabolism	46	1.31	4	3.90×10^{-2}	1.41	1.00	2.85×10^{-1}	0.02
Sulfur metabolism	15	0.43	2	6.79×10^{-2}	1.18	1.00	4.47×10^{-1}	0.00
Phenylpropanoid biosynthesis	35	1.00	3	7.46×10^{-2}	1.13	1.00	4.74×10^{-1}	0.00
beta-Alanine metabolism	18	0.51	2	9.10×10^{-2}	1.04	1.00	5.40×10^{-1}	0.07
Purine metabolism	63	1.80	4	1.01×10^{-1}	9.9710^{-1}	1.00	5.63×10^{-1}	0.00
Citrate cycle (TCA cycle)	20	0.57	2	1.09×10^{-1}	9.6210^{-1}	1.00	5.76×10^{-1}	0.10
Isoquinoline alkaloid biosynthesis	6	0.17	1	1.60×10^{-1}	7.9710^{-1}	1.00	7.08×10^{-1}	0.41
Galactose metabolism	27	0.77	2	1.78×10^{-1}	7.4910^{-1}	1.00	8.47×10^{-1}	0.00
Monobactam biosynthesis	8	0.23	1	2.07×10^{-1}	6.8410^{-1}	1.00	8.94×10^{-1}	0.00
Tropane, piperidine, and pyridine alkaloid biosynthesis	8	0.23	1	2.07×10^{-1}	6.8410^{-1}	1.00	8.94×10^{-1}	0.00
Valine, leucine, and isoleucine degradation	37	1.05	2	2.65×10^{-1}	5.4510^{-1}	1.00	1.00	0.00

Table 1. Cont.

Result from Pathway Analysis	Total	Expected	Hits	Raw p	−Log(p)	Holm Adjust	FDR	Impact
Nitrogen metabolism	12	0.74	1	2.94×10^{-1}	5.2310^{-1}	1.00	1.00	0.00
Phenylalanine metabolism	12	0.74	1	2.94×10^{-1}	5.2310^{-1}	1.00	1.00	0.42
Pyrimidine metabolism	38	1.08	2	2.96×10^{-1}	5.2910^{-1}	1.00	1.00	0.03
Nicotinate and nicotinamide metabolism	13	0.37	1	3.15×10^{-1}	5.0210^{-1}	1.00	1.00	0.00
Cutin, suberine, and wax biosynthesis	14	0.40	1	3.34×10^{-1}	4.76×10^{-1}	1.00	1.00	0.00
Sphingolipid metabolism	17	0.48	1	3.90×10^{-1}	4.09×10^{-1}	1.00	1.00	0.01
Ascorbate and aldarate metabolism	18	0.51	1	4.08×10^{-1}	3.90×10^{-1}	1.00	1.00	0.00
Tyrosine metabolism	19	0.51	1	4.08×10^{-1}	3.90×10^{-1}	1.00	1.00	0.17
Fructose and mannose metabolism	20	0.57	1	4.42×10^{-1}	3.55×10^{-1}	1.00	1.00	0.00
Propanoate metabolism	20	0.57	1	4.42×10^{-1}	3.55×10^{-1}	1.00	1.00	0.00
Carbon fixation in photosynthetic organisms	21	0.60	1	4.58×10^{-1}	3.39×10^{-1}	1.00	1.00	0.00
Zeatin biosynthesis	21	0.60	1	4.58×10^{-1}	3.39×10^{-1}	1.00	1.00	0.00
Biosynthesis of unsaturated fatty acids	22	0.63	1	4.73×10^{-1}	3.25×10^{-1}	1.00	1.00	0.00
Fatty acid elongation	23	0.66	1	4.88×10^{-1}	3.11×10^{-1}	1.00	1.00	0.00
Pantothenate and CoA biosynthesis	23	0.66	1	4.88×10^{-1}	3.11×10^{-1}	1.00	1.00	0.03
Phosphatidylinositol signaling system	26	0.74	1	5.32×10^{-1}	2.74×10^{-1}	1.00	1.00	0.03
Glutathione metabolism	27	0.77	1	5.45×10^{-1}	2.63×10^{-1}	1.00	1.00	0.01
Inositol phosphate metabolism	28	0.80	1	5.59×10^{-1}	2.53×10^{-1}	1.00	1.00	0.10
Ubiquinone and other terpenoid-quinone biosynthesis	35	1.00	1	6.41×10^{-1}	1.93×10^{-1}	1.00	1.00	0.00
Fatty acid degradation	37	1.05	1	6.62×10^{-1}	1.79×10^{-1}	1.00	1.00	0.00
Flavonoid biosynthesis	47	1.34	1	7.49×10^{-1}	1.25×10^{-1}	1.00	1.00	0.00
Fatty acid biosynthesis	56	1.60	1	8.08×10^{-1}	9.23×10^{-2}	1.00	1.00	0.01

4. Discussion

Previous studies in the literature have evaluated arsenic species accumulation and translocation from roots to shoots or grains [31,33,66]. For instance, specific transfer factors have been measured to understand how several arsenic species were transported from roots to other plant parts [66]. As a general conclusion, the higher the uptake, the more arsenic content translocates to the grains. Moreover, elevated concentrations of arsenic had a negative effect on plant development. These facts were in agreement with our study, where higher concentration levels of exposure caused severer changes in the phenotype (lighter colors in the aerial parts and darker in the roots, especially for the WH group), and also in the lipidome and metabolome, as already discussed in the previous sections. The work presented here also demonstrates that regardless of the treatment tested (watering and soil), arsenic alters metabolic pathways in both tissues (roots and aerials), leading to severe damage in the whole plant, including the grain. Furthermore, the arsenic tendency to translocate from roots to shoots, and from shoots to grain, poses a threat on human populations, which take in this element and biomagnifies through the food chain [32].

Since rice is cultivated in flooded conditions, where arsenic mobility is higher [33], the plant is susceptible to uptake arsenic from two main sources, i.e., contaminated groundwaters [24] and contaminated agricultural soils [25]. This study demonstrates that both scenarios of contamination threaten the development and growth of rice. The effects in the roots lipidome differed with the two tested treatments (watering and soil), whereas aerials seem more affected by the total arsenic dosage supplied. Hence, the findings in the study draw attention to the importance of arsenic sources when proposing detoxification strategies.

Untargeted metabolomics is a useful tool to assess metal and metalloid toxicity in model organisms, such as plants [18]. This omic approach provides a snapshot of what happens at the cellular level at the moment of the harvest, which means real-time information on arsenic exposure. Main lipidic changes detected in this study were related to key alterations in glycerophospholipids and glycerolipids, which are the dominant lipid families in rice [67]. The first group is a principal component of biological membranes in animals and plants, but also a major class of lipid in rice grain, often related to its quality and nutritional significance [68]. More specifically, lysoglycerophopholipids seem to be particularly vulnerable to environmental changes, which is in agreement with the accentuated alterations found in the current study for some LPCs and LPEs. There was also a link between some of our significant annotated metabolites and glycerophospholipid pathways (e.g., serine). The second group is also found in plant cell membranes [69], and has also been linked to photosynthesis, especially glyceroglycolipids (e.g., MGDG, DGDG) [70,71]. Among glycerolipids, DGs may play a crucial role in rice. Some studies suggested that DGs derived from phosphatidic acid, and that glycerophospholipid synthesis in rice may be linked to 1,2-diacylglycerol pathways [68]. Therefore, according to our study, arsenic exposure is damaging key lipidomic pathways related to important functions in plant cells and also linked to the grain's quality.

Our metabolomic study found major effects in amino acid-related pathways, such as aminoacyl-tRNA biosynthesis, alanine, aspartate, and glutamate metabolism or glycine, serine, and threonine metabolism. These molecules are known for their essential role in the development, growth, and stress responses of plants, as they have been related to their immune system [72]. A recent review from Guo et al. summarizes current knowledge of what is the function of amino acids in rice as signal molecules, how are they transported from the roots, and how these molecules regulate plant architecture and defense against abiotic stresses, specifically mentioning the role of proline, glycine, glutamate, and glutamine in stress responses [73]. Thus, our results suggest that rice defense mechanism against arsenic involves alterations in the amino acid-related pathways, as a response of their immune system against this contaminant.

Although our study was only performed at one harvesting time, further information could be obtained with time-course experiments at several sampling times. Besides,

untangling the mode of action of this metalloid leads us to potentially identify novel detoxification mechanisms. Lastly, not all arsenic species exhibit the same toxicity and translocation [31], and the cultivar tolerance is also a key aspect to consider. For instance, a previous study evaluated the role of amino acids and thiolic ligands on arsenite tolerance in rice [74]. Results were dependent on the cultivar's tolerance. Although in our work we supplied arsenic (V) solution to rice instead of arsenic (III), both studies have in common that amino acids were altered due to this contaminant regardless of the arsenic species provided. Besides, amino acids are not only involved in rice response against arsenic stress, but could also be employed for detoxification purposes, as well as other molecules such as thiol ligands [74]. Further metabolomic studies in combination with speciation analysis could shed more light on arsenic uptake mechanism and way of action.

5. Conclusions

Untargeted lipidomic and metabolomic approaches have allowed increasing our current knowledge on arsenic exposure in rice at early stages of growth (up to 22 days), through two different treatments (watering with arsenic contaminated water or growth in contaminated soil). Metabolic and lipidomic alterations caused by As (V) treatment were present in both root and aerial tissues. The application of a proper chemometrics workflow has allowed the proposal of potential markers of these metabolic disturbances.

Specifically, arsenic impact in the roots lipidome differed according to the treatment, watering, and soil contaminations, which revealed that the nature of the arsenic source produced a different type of effects on this root tissue. In addition, severe damage in the metabolome (and lipidome) was also found in aerial tissues, confirming the presence of adverse effects due to arsenic exposure throughout the whole plant (and eventually, to the rice grains). In contrast to roots, adverse arsenic effects in aerials were more related to the arsenic dose rather than the treatment itself. Some of the lipids most affected by arsenic exposure belonged to the following lipid families: LPCs, PAs, DGs, PGs, and PSs in aerial tissues, and PCs, PAs, DGDGs, MGDGs in roots. Regarding the metabolomic alterations, the comparison of significant changes in roots and aerial metabolomes showed a considerable overlapping of biochemical pathway alterations between both tissues, although the affected metabolites did not necessarily coincide in both cases. Most of the metabolic pathways disturbed were amino acid related. In rice, amino acids have been previously associated with defense mechanisms against abiotic stresses and they also play a key role in plant immune system. These changes in amino acids may be a consequence of stress response of rice to defend itself against arsenic exposure, and they could be used as an indication of arsenic detoxification.

In conclusion, untargeted analysis has proven to be a powerful tool to generate hypothesis regarding the modes of action of toxics such as arsenic, because it furnishes a wider overview of metabolic changes; in this case, of adverse effects caused by As (V) solutions in rice lipidome and metabolome. Targeted analysis of the potential markers found in previous untargeted analysis can confirm and validate the proposed discoveries. Other complementary omic analysis, such as transcriptomics, will complementarily improve the characterization and role of specific metabolites and lipids involved in arsenic exposure.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/separations9030079/s1, Figure S1: Experimental conditions employed for rice growing in the chamber MLE-352H: light intensity and temperature; Figure S2: Scheme of the ROIMCR workflow; Figure S3: PCA score plots positive ionization mode obtained for lipidomic results for roots (A) and aerial parts (B), and metabolomic results for roots (C) and aerial parts (D); Figure S4: Experimental MS/MS spectrum of L-tryptophan compared to a theoretical MS/MS spectrum for the same compound in aerial tissues with ESI (+); Figure S5: Hierarchical clustering heatmaps applied to the logarithm of the fold changes of the annotated lipids for aerial in positive mode, roots in positive and negative modes, respectively; Table S1: Summary of the As (V) concentration levels employed in this study; Table S2: ROI parameters employed and MCR components obtained for each dataset; Table S3: Statistical results from ASCA; Table S4: PLS-DA

results: no. of Variables Important in Projection > 1.0 and Matthew Correlation Coefficients; Table S5: MS-DIAL parameters used in metabolomic annotation; Table S6: Experiment file used in MS method type section from start a project window in MS-DIAL; Table S7: Lipid identification of the most relevant features responsible for the changes induced by arsenic exposure on rice in root tissues, in positive ionization mode; Table S8: Lipid identification of the most relevant features responsible for the changes induced by arsenic exposure on rice in root tissues, in negative ionization mode; Table S9: Lipid identification of the most relevant features responsible for the changes induced by arsenic exposure on rice in aerial tissues, in positive ionization mode; Table S10: Lipid identification of the most relevant features responsible for the changes induced by arsenic exposure on rice in aerial tissues, in negative ionization mode; Table S11: Metabolite identification of the most relevant features responsible for the changes induced by arsenic exposure on rice in root tissues, in positive ionization mode; Table S12: Metabolite identification of the most relevant features responsible for the changes induced by arsenic exposure on rice in root tissues, in negative ionization mode; Table S13: Metabolite identification of the most relevant features responsible for the changes induced by arsenic exposure on rice in aerial tissues, in positive ionization mode; Table S14: Metabolite identification of the most relevant features responsible for the changes induced by arsenic exposure on rice in aerial tissues, in negative ionization mode [75,76].

Author Contributions: Conceptualization, M.P.-C., R.T. and J.J.; Methodology, M.P.-C., R.T. and J.J.; Formal analysis, M.P.-C. and J.J.; Investigation, M.P.-C. and J.J.; Writing—original draft, M.P.-C.; Writing—review and editing, M.P.-C., R.T. and J.J.; Visualization, M.P.-C. and J.J.; Supervision, R.T. and J.J.; Project administration, J.J.; Funding acquisition, R.T. and J.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Grant CTQ2017-82598-P and CEX2018-000794-S funded by MCIN/AEI/ 10.13039/501100011033, and supported from the Catalan Agency for Management of University and Research Grants (AGAUR, Grant 2017SGR753). MPC was funded by a predoctoral FPU 16/02640 scholarship from the Spanish Ministry of Education and Vocational Training (MEFP).

Data Availability Statement: The data presented in this study are openly available in Digital.CSIC (https://digital.csic.es/handle/10261/261849, accessed on 17 March 2022) and Zenodo (https://zenodo.org/record/6222067, accessed on 17 March 2022).

Conflicts of Interest: The authors declare no conflict of interest. In addition, the authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Awika, J.M. Major Cereal Grains Production and Use around the World. *ACS Symp. Ser.* **2011**, *1089*, 1–13. [CrossRef]
2. Kaur, B.; Sandhu, K.S.; Kamal, R.; Kaur, K.; Singh, J.; Röder, M.S.; Muqaddasi, Q.H. Omics for the Improvement of Abiotic, Biotic, and Agronomic Traits in Major Cereal Crops: Applications, Challenges, and Prospects. *Plants* **2021**, *10*, 1989. [CrossRef]
3. Tarpley, L.; Duran, A.L.; Kebrom, T.H.; Sumner, L.W. Biomarker Metabolites Capturing the Metabolite Variance Present in a Rice Plant Developmental Period. *BMC Plant Biol.* **2005**, *5*, 8. [CrossRef] [PubMed]
4. Xue, J.; Lu, D.; Wang, S.; Lu, Z.; Liu, W.; Wang, X.; Fang, Z.; He, X. Integrated Transcriptomic and Metabolomic Analysis Provides Insight into the Regulation of Leaf Senescence in Rice. *Sci. Rep.* **2021**, *11*, 14083. [CrossRef] [PubMed]
5. Hall, R.D.; Brouwer, I.D.; Fitzgerald, M.A. Plant Metabolomics and Its Potential Application for Human Nutrition. *Physiol. Plant.* **2008**, *132*, 162–175. [CrossRef]
6. Yang, Y.; Saand, M.A.; Huang, L.; Abdelaal, W.B.; Zhang, J.; Wu, Y.; Li, J.; Sirohi, M.H.; Wang, F. Applications of Multi-Omics Technologies for Crop Improvement. *Front. Plant Sci.* **2021**, *12*, 1846. [CrossRef]
7. Thi, K.; Vo, X.; Rahman, M.; Rahman, M.; Trinh, T.; Kim, S.T.; Jeon, J.-S. Proteomics and Metabolomics Studies on the Biotic Stress Responses of Rice: An Update. *Rice* **2021**, *14*, 30. [CrossRef]
8. Glaubitz, U.; Li, X.; Schaedel, S.; Erban, A.; Sulpice, R.; Kopka, J.; Hincha, D.K.; Zuther, E. Integrated Analysis of Rice Transcriptomic and Metabolomic Responses to Elevated Night Temperatures Identifies Sensitivity- and Tolerance-Related Profiles. *Plant Cell Environ.* **2017**, *40*, 121–137. [CrossRef] [PubMed]
9. Gupta, P.; De, B. Metabolomics Analysis of Rice Responses to Salinity Stress Revealed Elevation of Serotonin, and Gentisic Acid Levels in Leaves of Tolerant Varieties. *Plant Signal. Behav.* **2017**, *12*, e1335845. [CrossRef] [PubMed]
10. Wu, X.; Hou, H.; Liu, Y.; Yin, S.; Bian, S.; Liang, S.; Wan, C.; Yuan, S.; Xiao, K.; Liu, B.; et al. Microplastics Affect Rice (*Oryza sativa* L.) Quality by Interfering Metabolite Accumulation and Energy Expenditure Pathways: A Field Study. *J. Hazard. Mater.* **2022**, *422*, 126834. [CrossRef] [PubMed]

11. Saeed, M.; Quraishi, U.M.; Malik, R.N. Arsenic Uptake and Toxicity in Wheat (*Triticum aestivum* L.): A Review of Multi-Omics Approaches to Identify Tolerance Mechanisms. *Food Chem.* **2021**, *355*, 129607. [CrossRef] [PubMed]
12. Jamla, M.; Khare, T.; Joshi, S.; Patil, S.; Penna, S.; Kumar, V. Omics Approaches for Understanding Heavy Metal Responses and Tolerance in Plants. *Curr. Plant Biol.* **2021**, *27*, 100213. [CrossRef]
13. Oikawa, A.; Matsuda, F.; Kusano, M.; Okazaki, Y.; Saito, K. Rice Metabolomics. *Rice* **2008**, *1*, 63–71. [CrossRef]
14. Kim, T.J.; Kim, S.Y.; Park, Y.J.; Lim, S.H.; Ha, S.H.; Park, S.U.; Lee, E.; Kim, J.K. Metabolite Profiling Reveals Distinct Modulation of Complex Metabolic Networks in Non-Pigmented, Black, and Red Rice (*Oryza sativa* L.) Cultivars. *Metabolites* **2021**, *11*, 367. [CrossRef] [PubMed]
15. Kusano, M.; Yang, Z.; Okazaki, Y.; Nakabayashi, R.; Fukushima, A.; Saito, K. Using Metabolomic Approaches to Explore Chemical Diversity in Rice. *Mol. Plant* **2015**, *8*, 58–67. [CrossRef] [PubMed]
16. Yang, Z.; Nakabayashi, R.; Okazaki, Y.; Mori, T.; Takamatsu, S.; Kitanaka, S.; Kikuchi, J.; Saito, K. Toward Better Annotation in Plant Metabolomics: Isolation and Structure Elucidation of 36 Specialized Metabolites from *Oryza sativa* (Rice) by Using MS/MS and NMR Analyses. *Metabolomics* **2014**, *10*, 543–555. [CrossRef]
17. Liu, C.; Lan, M.M.; He, E.K.; Yao, A.J.; Wang, G.B.; Tang, Y.T.; Qiu, R.L. Phenomic and Metabolomic Responses of Roots to Cadmium Reveal Contrasting Resistance Strategies in Two Rice Cultivars (*Oryza sativa* L.). *Soil Ecol. Lett.* **2021**, *3*, 220–229. [CrossRef]
18. Booth, S.C.; Workentine, M.L.; Weljie, A.M.; Turner, R.J. Metabolomics and Its Application to Studying Metal Toxicity. *Metallomics* **2011**, *3*, 1142–1152. [CrossRef] [PubMed]
19. Feng, Z.; Ji, S.; Ping, J.; Cui, D. Recent Advances in Metabolomics for Studying Heavy Metal Stress in Plants. *TrAC—Trends Anal. Chem.* **2021**, *143*, 116402. [CrossRef]
20. World Health Organization Web Page. Available online: https://www.who.int/news-room/fact-sheets/detail/arsenic (accessed on 14 February 2022).
21. Shankar, S.; Shanker, U. Shikha Arsenic Contamination of Groundwater: A Review of Sources, Prevalence, Health Risks, and Strategies for Mitigation. *Sci. World J.* **2014**, *2014*, 304524. [CrossRef] [PubMed]
22. Bundschuh, J.; Schneider, J.; Alam, M.A.; Niazi, N.K.; Herath, I.; Parvez, F.; Tomaszewska, B.; Guilherme, L.R.G.; Maity, J.P.; López, D.L.; et al. Seven Potential Sources of Arsenic Pollution in Latin America and Their Environmental and Health Impacts. *Sci. Total Environ.* **2021**, *780*, 146274. [CrossRef] [PubMed]
23. Ren, S.; Song, C.; Ye, S.; Cheng, C.; Gao, P. The Spatiotemporal Variation in Heavy Metals in China's Farmland Soil over the Past 20 years: A Meta-Analysis. *Sci. Total Environ.* **2022**, *806*, 150322. [CrossRef] [PubMed]
24. Nilkamjanakul, W.; Watchalayann, P.; Chotpantarat, S. Spatial Distribution and Health Risk Assessment of As and Pb Contamination in the Groundwater of Rayong Province, Thailand. *Environ. Res.* **2022**, *204*, 111838. [CrossRef]
25. Varol, M.; Gündüz, K.; Sünbül, M.R. Pollution Status, Potential Sources and Health Risk Assessment of Arsenic and Trace Metals in Agricultural Soils: A Case Study in Malatya Province, Turkey. *Environ. Res.* **2021**, *202*, 111806. [CrossRef]
26. Lima, J.Z.; Ferreira da Silva, E.; Patinha, C.; Durães, N.; Vieira, E.M.; Rodrigues, V.G.S. Sorption of Arsenic by Composts and Biochars Derived from the Organic Fraction of Municipal Solid Wastes: Kinetic, Isotherm and Oral Bioaccessibility Study. *Environ. Res.* **2022**, *204*, 111988. [CrossRef]
27. Li, Y.; Bi, Y.; Mi, W.; Xie, S.; Ji, L. Land-Use Change Caused by Anthropogenic Activities Increase Fluoride and Arsenic Pollution in Groundwater and Human Health Risk. *J. Hazard. Mater.* **2021**, *406*, 124337. [CrossRef] [PubMed]
28. US Food and Drug Administration. *Arsenic in Rice and Rice Products Risk Assessment Report*; Center for Food Safety and Applied Nutrition of the Food and Drug Administration, 2016; Volume 1, pp. 1–284. Available online: http://www.fda.gov/Food/FoodScienceResearch/RiskSafetyAssessment/default.htm (accessed on 15 March 2022).
29. Roel, A.; Campos, F.; Verger, M.; Huertas, R.; Carracelas, G. Regional Variability of Arsenic Content in Uruguayan Polished Rice. *Chemosphere* **2022**, *288*, 132426. [CrossRef] [PubMed]
30. Zhao, F.J.; McGrath, S.P.; Meharg, A.A. Arsenic as a Food Chain Contaminant: Mechanisms of Plant Uptake and Metabolism and Mitigation Strategies. *Annu. Rev. Plant Biol.* **2010**, *61*, 535–559. [CrossRef] [PubMed]
31. Mishra, S.; Mattusch, J.; Wennrich, R. Accumulation and Transformation of Inorganic and Organic Arsenic in Rice and Role of Thiol-Complexation to Restrict Their Translocation to Shoot. *Sci. Rep.* **2017**, *7*, 40522. [CrossRef] [PubMed]
32. Kumar, S.; Dubey, R.S.; Tripathi, R.D.; Chakrabarty, D.; Trivedi, P.K. Omics and Biotechnology of Arsenic Stress and Detoxification in Plants: Current Updates and Prospective. *Environ. Int.* **2015**, *74*, 221–230. [CrossRef] [PubMed]
33. Tuli, R.; Chakrabarty, D.; Trivedi, P.K.; Tripathi, R.D. Recent Advances in Arsenic Accumulation and Metabolism in Rice. *Mol. Breed.* **2010**, *26*, 307–323. [CrossRef]
34. Perez de Souza, L.; Alseekh, S.; Naake, T.; Fernie, A. Mass Spectrometry-Based Untargeted Plant Metabolomics. *Curr. Protoc. Plant Biol.* **2019**, *4*, e20100. [CrossRef]
35. Akram, M.I.; Vincent, I.M.; Siddiqui, A.J.; Musharraf, S.G. Polymeric Hydrophilic Interaction Liquid Chromatography Coupled with Orbitrap Mass Spectrometry and Chemometric Analysis for Untargeted Metabolite Profiling of Natural Rice Variants. *J. Cereal Sci.* **2017**, *73*, 165–173. [CrossRef]
36. Xiao, R.; Ma, Y.; Zhang, D.; Qian, L. Discrimination of Conventional and Organic Rice Using Untargeted LC-MS-Based Metabolomics. *J. Cereal Sci.* **2018**, *82*, 73–81. [CrossRef]

37. Concepcion, J.C.T.; Calingacion, M.; Garson, M.J.; Fitzgerald, M.A. Lipidomics Reveals Associations between Rice Quality Traits. *Metabolomics* **2020**, *16*, 54. [CrossRef]
38. Navas-Iglesias, N.; Carrasco-Pancorbo, A.; Cuadros-Rodríguez, L. From Lipids Analysis towards Lipidomics, a New Challenge for the Analytical Chemistry of the 21st Century. Part II: Analytical Lipidomics. *TrAC—Trends Anal. Chem.* **2009**, *28*, 393–403. [CrossRef]
39. Navarro-Reig, M.; Jaumot, J.; Tauler, R. An Untargeted Lipidomic Strategy Combining Comprehensive Two-Dimensional Liquid Chromatography and Chemometric Analysis. *J. Chromatogr. A* **2018**, *1568*, 80–90. [CrossRef] [PubMed]
40. Matyash, V.; Liebisch, G.; Kurzchalia, T.V.; Shevchenko, A.; Schwudke, D. Lipid Extraction by Methyl-Tert-Butyl Ether for High-Throughput Lipidomics. *J Lipid Res.* **2008**, *49*, 1137–1146. [CrossRef]
41. Navarro-Reig, M.; Jaumot, J.; Piña, B.; Moyano, E.; Galceran, M.T.; Tauler, R. Metabolomic Analysis of the Effects of Cadmium and Copper Treatment in: *Oryza sativa* L. Using Untargeted Liquid Chromatography Coupled to High Resolution Mass Spectrometry and All-Ion Fragmentation. *Metallomics* **2017**, *9*, 660–675. [CrossRef]
42. *2006/118/EC*. Directive 2006/118/EC of the European Parliment and of the Council of 12 December 2006 on the Protection of Groundwater against Pollution and Deterioration. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32006L0118 (accessed on 15 March 2022).
43. Akhter, H.; Cartledge, F.K.; Miller, J.; McLearn, M. Treatment of Arsenic-Contaminated Soils. I: Soil Characterization. *J. Environ. Eng.* **2000**, *126*, 999–1003. [CrossRef]
44. Menéndez-Pedriza, A.; Jaumot, J.; Bedia, C. Lipidomic analysis of single and combined effects of polyethylene microplastics and polychlorinated biphenyls on human hepatoma cells. *J. Hazard. Mater.* **2022**, *421*, 126777. [CrossRef]
45. Ortiz-Villanueva, E.; Jaumot, J.; Martínez, R.; Navarro-Martín, L.; Piña, B.; Tauler, R. Assessment of Endocrine Disruptors Effects on Zebrafish (Danio Rerio) Embryos by Untargeted LC-HRMS Metabolomic Analysis. *Sci. Total Environ.* **2018**, *635*, 156–166. [CrossRef] [PubMed]
46. Navarro-Reig, M.; Jaumot, J.; García-Reiriz, A.; Tauler, R. Evaluation of Changes Induced in Rice Metabolome by Cd and Cu Exposure Using LC-MS with XCMS and MCR-ALS Data Analysis Strategies. *Anal. Bioanal. Chem.* **2015**, *407*, 8835–8847. [CrossRef] [PubMed]
47. Puig-Castellví, F.; Bedia, C.; Alfonso, I.; Piña, B.; Tauler, R. Deciphering the Underlying Metabolomic and Lipidomic Patterns Linked to Thermal Acclimation in Saccharomyces Cerevisiae. *J. Proteome Res.* **2018**, *17*, 2034–2044. [CrossRef] [PubMed]
48. Chambers, M.C.; Maclean, B.; Burke, R.; Amode, D.; Ruderman, D.L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; et al. A Cross-Platform Toolkit for Mass Spectrometry and Proteomics. *Nat. Biotechnol.* **2012**, *30*, 918–920. [CrossRef]
49. Gorrochategui, E.; Jaumot, J.; Tauler, R. ROIMCR: A Powerful Analysis Strategy for LC-MS Metabolomic Datasets. *BMC Bioinform.* **2019**, *20*, 256. [CrossRef] [PubMed]
50. Pérez-Cova, M.; Bedia, C.; Stoll, D.R.; Tauler, R.; Jaumot, J. MSroi: A Pre-Processing Tool for Mass Spectrometry-Based Studies. *Chemom. Intell. Lab. Syst.* **2021**, *215*, 104333. [CrossRef]
51. Smilde, A.K.; Jansen, J.J.; Hoefsloot, H.C.J.; Lamers, R.J.A.N.; van der Greef, J.; Timmerman, M.E. ANOVA-Simultaneous Component Analysis (ASCA): A New Tool for Analyzing Designed Metabolomics Data. *Bioinformatics* **2005**, *21*, 3043–3048. [CrossRef]
52. Joliffe, I.T.; Morgan, B. Principal Component Analysis and Exploratory Factor Analysis. *Stat. Methods Med. Res.* **1992**, *1*, 69–95. [CrossRef] [PubMed]
53. Mangiameli, P.; Chen, S.K.; West, D. A Comparison of SOM Neural Network and Hierarchical Clustering Methods. *Eur. J. Oper. Res.* **1996**, *93*, 402–417. [CrossRef]
54. Gromski, P.S.; Muhamadali, H.; Ellis, D.I.; Xu, Y.; Correa, E.; Turner, M.L.; Goodacre, R. A Tutorial Review: Metabolomics and Partial Least Squares-Discriminant Analysis—A Marriage of Convenience or a Shotgun Wedding. *Anal. Chim. Acta* **2015**, *879*, 10–23. [CrossRef]
55. Lee, L.C.; Liong, C.Y.; Jemain, A.A. Partial Least Squares-Discriminant Analysis (PLS-DA) for Classification of High-Dimensional (HD) Data: A Review of Contemporary Practice Strategies and Knowledge Gaps. *Analyst* **2018**, *143*, 3526–3539. [CrossRef] [PubMed]
56. Chicco, D.; Jurman, G. The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genom.* **2020**, *21*, 6. [CrossRef] [PubMed]
57. Fahy, E.; Sud, M.; Cotter, D.; Subramaniam, S. LIPID MAPS Online Tools for Lipid Research. *Nucleic Acids Res.* **2007**, *35*, W606–W612. [CrossRef] [PubMed]
58. Wishart, D.S.; Guo, A.C.; Oler, E.; Wang, F.; Anjum, A.; Peters, H.; Dizon, R.; Sayeeda, Z.; Tian, S.; Lee, B.L.; et al. HMDB 5.0: The Human Metabolome Database for 2022. *Nucleic Acids Res.* **2022**, *50*, D622–D631. [CrossRef]
59. MS-DIAL Web Page. Available online: http://Prime.Psc.Riken.Jp/Compms/MS-DIAL/Main.Html (accessed on 14 February 2022).
60. Schulze, T.; Meier, R.; Alygizakis, N.; Schymanski, E.; Bach, E.; LI, D.H.; lauperbe; raalizadeh; Tanaka, S.; Witting, M. *MassBank/MassBank-Data: Release Version 2021.12*. 2021. Available online: https://doi.org/10.5281/ZENODO.5775684 (accessed on 17 March 2022).

61. Wang, M.; Carver, J.J.; Phelan, V.V.; Sanchez, L.M.; Garg, N.; Peng, Y.; Nguyen, D.D.; Watrous, J.; Kapono, C.A.; Luzzatto-Knaan, T.; et al. Sharing and Community Curation of Mass Spectrometry Data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **2016**, *34*, 828–837. [CrossRef] [PubMed]
62. Hawkins, C.; Ginzburg, D.; Zhao, K.; Dwyer, W.; Xue, B.; Xu, A.; Rice, S.; Cole, B.; Paley, S.; Karp, P.; et al. Plant Metabolic Network 15: A Resource of Genome-Wide Metabolism Databases for 126 Plants and Algae. *J. Integr. Plant Biol.* **2021**, *63*, 1888–1905. [CrossRef] [PubMed]
63. Tsugawa, H.; Ikeda, K.; Takahashi, M.; Satoh, A.; Mori, Y.; Uchino, H.; Okahashi, N.; Yamada, Y.; Tada, I.; Bonini, P.; et al. A Lipidome Atlas in MS-DIAL 4. *Nat. Biotechnol.* **2020**, *38*, 1159–1163. [CrossRef] [PubMed]
64. Blaženović, I.; Kind, T.; Ji, J.; Fiehn, O. Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics. *Metabolites* **2018**, *8*, 31. [CrossRef] [PubMed]
65. Pang, Z.; Chong, J.; Zhou, G.; de Lima Morais, D.A.; Chang, L.; Barrette, M.; Gauthier, C.; Jacques, P.É.; Li, S.; Xia, J. MetaboAnalyst 5.0: Narrowing the Gap between Raw Spectra and Functional Insights. *Nucleic Acids Res.* **2021**, *49*, W388–W396. [CrossRef] [PubMed]
66. Batista, B.L.; Nigar, M.; Mestrot, A.; Rocha, B.A.; Júnior, F.B.; Price, A.H.; Raab, A.; Feldmann, J. Identification and Quantification of Phytochelatins in Roots of Rice to Long-Term Exposure: Evidence of Individual Role on Arsenic Accumulation and Translocation. *J. Exp. Bot.* **2014**, *65*, 1467–1479. [CrossRef] [PubMed]
67. Zhang, D.; Zhao, L.; Wang, W.; Wang, Q.; Liu, J.; Wang, Y.; Liu, H.; Shang, B.; Duan, X.; Sun, H. Lipidomics Reveals the Changes in Non-Starch and Starch Lipids of Rice (*Oryza sativa* L.) during Storage. *J. Food Compos. Anal.* **2022**, *105*, 104205. [CrossRef]
68. Liu, L.; Waters, D.L.E.; Rose, T.J.; Bao, J.; King, G.J. Phospholipids in Rice: Significance in Grain Quality and Health Benefits: A Review. *Food Chem.* **2013**, *139*, 1133–1145. [CrossRef] [PubMed]
69. Rizov, L.; Doulis, A. Separation of Plant Membrane Lipids by Multiple Solid-Phase Extraction. *J. Chromatogr. A* **2001**, *922*, 347–354. [CrossRef]
70. Kobayashi, K. Role of Membrane Glycerolipids in Photosynthesis, Thylakoid Biogenesis and Chloroplast Development. *J. Plant Res.* **2016**, *129*, 565–580. [CrossRef] [PubMed]
71. Basnet, R.; Zhang, J.; Hussain, N.; Shu, Q. Characterization and Mutational Analysis of a Monogalactosyldiacylglycerol Synthase Gene OsMGD2 in Rice. *Front. Plant Sci.* **2019**, *10*, 992. [CrossRef] [PubMed]
72. Kadotani, N.; Akagi, A.; Takatsuji, H.; Miwa, T.; Igarashi, D. Exogenous Proteinogenic Amino Acids Induce Systemic Resistance in Rice. *BMC Plant Biol.* **2016**, *16*, 60. [CrossRef] [PubMed]
73. Guo, N.; Zhang, S.; Gu, M.; Xu, G. Function, Transport, and Regulation of Amino Acids: What Is Missing in Rice? *Crop J.* **2021**, *9*, 530–542. [CrossRef]
74. Tripathi, P.; Tripathi, R.D.; Singh, R.P.; Dwivedi, S.; Chakrabarty, D.; Trivedi, P.K.; Adhikari, B. Arsenite Tolerance in Rice (*Oryza sativa* L.) Involves Coordinated Role of Metabolic Pathways of Thiols and Amino Acids. *Environ. Sci. Pollut. Res.* **2013**, *20*, 884–896. [CrossRef]
75. de Juan, A.; Jaumot, J.; Tauler, R. Multivariate Curve Resolution (MCR). Solving the Mixture Analysis Problem. *Anal. Methods* **2014**, *6*, 4964–4976. [CrossRef]
76. Windig, W.; Stephenson, D.A. Self-Modeling Mixture Analysis of Second-Derivative Near-Infrared Spectral Data using the Simplisma Approach. *Anal. Chem.* **1992**, *64*, 2735–2742.

Supplementary Material A

Adverse effects of arsenic uptake in rice metabolome and lipidome revealed by untargeted liquid chromatography coupled to mass spectrometry (LC-MS) and regions of interest multivariate curve resolution

Miriam Pérez-Cova[1,2], Romà Tauler[1], Joaquim Jaumot[1]*

[1]Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, 08034 Barcelona, Spain

[2]Department of Chemical Engineering and Analytical Chemistry, University of Barcelona, Diagonal 647, Barcelona, E08028, Barcelona, Spain

* Correspondence: joaquim.jaumot@idaea.csic.es

1. Experimental conditions

Conditions of rice growth in chamber MLE-352H

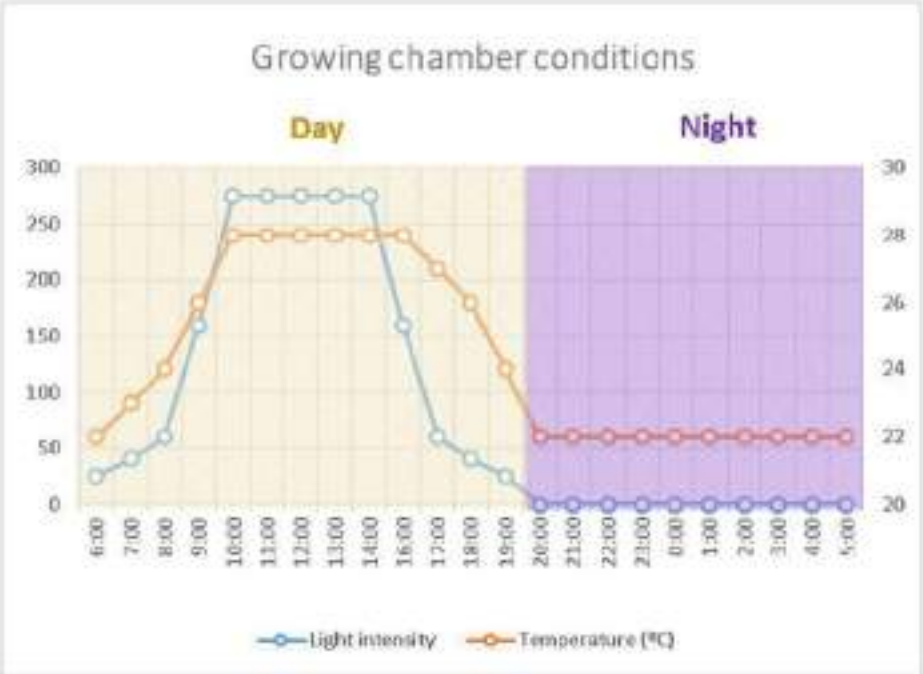


Figure S1. Experimental conditions employed for rice growing in the chamber MLE-352H: light intensity (blue) and temperature (orange). Day hours (yellow, left side of the graph), and night hours (purple, right side of the graph) are also distinguished.

Table S1. Summary of the As (V) concentration levels employed in this study.

Treatment applied	Concentration value
Watering Low (WL)	1 µM
Soil Low (SL)	15 µM
Soil High (SH)	155 µM
Watering High (WH)	1000 µM

2. Chemometric tools: the ROIMCR procedure

ROIMCR procedure is summarized in **Figure S2**. First, regions of interest, ROI, (purple section in **Figure S2**) is employed for compressing LC-MS datasets in the spectral dimension, while creating an augmented matrix which concatenates the different samples in a column-wise manner. Then, multivariate curve resolution alternating least squares, MCR-ALS, (blue section in **Figure S2**) is applied. The compressed LC-MS datasets are resolved, providing separate information about the spectra profiles (qualitative information for identification purposes) and elution profiles (quantitative information, from the resolved peak areas). Each component from the MCR-ALS ideally represents one potential compound, joining isotopic forms and adducts in the same component (i.e. componentization). Therefore, it can be associated with a compound (i.e. lipid or metabolite). Afterwards, multivariate analysis (e.g. statistical, exploratory or classification analysis) is performed on the area matrix obtained from MCR-ALS.

A more detail description of each chemometric method is included below.

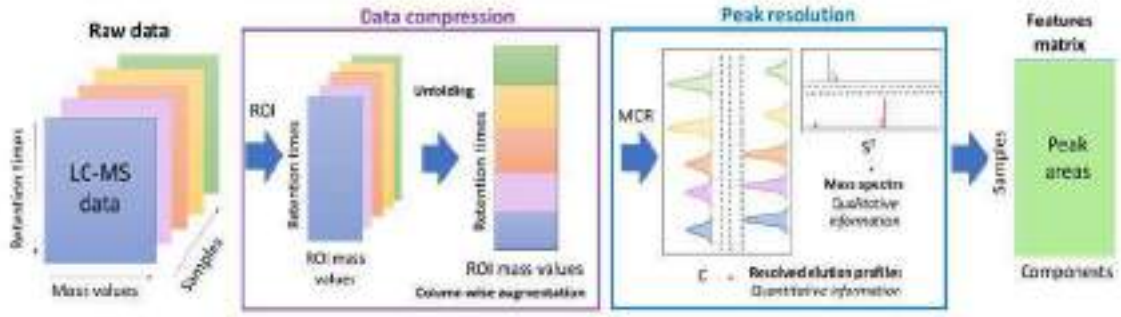


Figure S2. Scheme of the ROIMCR workflow. Step 1: spectral compression with regions of interest (purple). LC-MS datasets are reduced in the *m/z* direction by keeping only relevant *m/z* values. Samples are concatenated in a wise-column manner. Step 2: multivariate curve resolution alternating least square (blue). Resolution of the spectral and elution profiles. An area matrix is obtained from the integration of the resolved elution profiles.

Regions of interest (ROI) as spectral compression

ROI approach selects *m/z* values below an intensity threshold established *a priori* by the user according to the signal-to-noise ratio (SN threshold) for each dataset. ROI also takes into account a mass error tolerance, related to the mass accuracy of the mass spectrometer, and a minimum number of occurrences, required for defining a chromatographic peak. A factor can also be set to establish an intensity threshold low, but only considering the features whose intensities are a multiple of this factor (e.g. min max 2, means features kept have intensities at least twice the SN threshold). ROI *m/z* values are searched for each retention time, and the final value will be the mean (or the median) of all the values corresponding to the same chromatographic peak. If an *m/z* value is detected for some samples but others no, then non-present ROIs will be set to a low random intensity value at the noise level. With this strategy, the original *m/z* vector is reduced for all samples simultaneously; the new vector is composed of discrete *m/z* values. Hence, only relevant features remain at the end of the procedure, which are considered for further analysis. More information about the MSroi app and the ROI approach can be found elsewhere [1,2]. ROI parameters employed in the analysis of each dataset, according to the platform (lipidomics or metabolomics), the tissue (roots or aerial parts) or the electrospray ionization mode (positive and negative), are included in **Table S1**.

Table S2. *ROI parameters employed and MCR components obtained for each dataset.*

	Lipidomics				***Metabolomics***			
ROI parameters	**Roots ESI (+)**	**Roots ESI (-)**	**Aerial parts ESI (+)**	**Aerial parts ESI (-)**	**Roots ESI (+)**	**Roots ESI (-)**	**Aerial parts ESI (+)**	**Aerial parts ESI (-)**
SN threshold	1.00E+03	3.50E+02	1.00E+03	6.50E+02	2.00E+06	1.00E+07	2.00E+06	1.00E+07
Min max factor	2	2	2	2	2	2	2	2
Mass error (ppm)	30	30	30	30	10	10	10	10
Min occurences	100	100	100	100	100	100	100	100
Rois calculation	Median	Median	Median	Median	Median	Median	Median	Median
N° of rois obtained	217	297	221	187	347	137	345	212
*N° of **MCR** components*	150	100	110	65	150	80	150	100

Multivariate curve resolution alternating least squares (MCR-ALS) for the resolution of the spectral and elution profiles

The decomposition provided by MCR follows a bilinear model:

$$\mathbf{D} = \mathbf{CS}^{\mathbf{T}} + \mathbf{E} \qquad \textbf{Equation (1)}$$

Where **D** is the LC-MS data matrix with the different retention times in the rows and the *m/z* values in the columns; **C** is the matrix containing the resolved elution profiles for each of the MCR components (which can ideally be associated with a single chemical compound), $\mathbf{S^T}$ the matrix containing the spectral profiles for each component too, and **E** the matrix with the residuals not explained by the model. If multiple samples are analyzed simultaneously, then the bilinear model extends accordingly:

$$\mathbf{D_{aug}} = \begin{bmatrix} \mathbf{D_1} \\ \dots \\ \mathbf{D_L} \end{bmatrix} = \begin{bmatrix} \mathbf{C_1} \\ \dots \\ \mathbf{C_L} \end{bmatrix} \mathbf{S^T} + \begin{bmatrix} \mathbf{E_1} \\ \dots \\ \mathbf{E_L} \end{bmatrix} = \mathbf{C_{aug}} \mathbf{S^T} + \mathbf{E_{aug}} \qquad \textbf{Equation (2)}$$

Where $\mathbf{D_{aug}}$ is the augmented data matrix including all samples, each of them composed by a data matrix of the chromatographic run ($\mathbf{D_1}$, ...$\mathbf{D_L}$). Samples are concatenated vertically, in a column-wise manner. $\mathbf{C_{aug}}$ has the elution profiles of the components present in all analyzed samples, and $\mathbf{S^T}$ the spectral of these components. Again, $\mathbf{E_{aug}}$ is the residual matrix, containing the non-explained variances.

MCR-ALS is a specific version of the MCR method where the optimization step to resolve the component profiles employs an iterative alternating least squares algorithm (ALS). This approach has been described elsewhere [3]. In the first step of the MCR-ALS approach, the number of components is selected, which is initially estimated from the number of singular values, using the singular value decomposition (SVD) algorithm. In the next step, initial estimates (elution or spectral profiles) are obtained for the selected number of components. Then, the ALS iterative optimization begins. Here, optimization was started on initial estimates of pure spectra ($\mathbf{S^T}$), obtained from a purest spectra detection of pure variable detection approach [4]. When convergence criterion is achieved, the process finishes. Constraints are also frequently employed to reduce ambiguities associated with the bilinear model (i.e. it does not assure unique solutions) and provide chemical meaning to the mathematical solutions. In this case, non-negativity constraint was selected for both spectral and elution profiles. Equal height was also applied for spectral normalization. The number of components obtained for each dataset is included at the bottom part of **Table S1**.

3. Multivariate analysis

Statistical assessment and exploratory analysis of arsenic exposure

Table S3. *Statistical results from ASCA.*

ASCA results	Lipidomics				Metabolomics			
	Roots ESI (+)	Roots ESI (-)	Aerial parts ESI (+)	Aerial parts ESI (-)	Roots ESI (+)	Roots ESI (-)	Aerial parts ESI (+)	Aerial parts ESI (-)
C vs W vs S (treatment)	1	1	1	1	1	1	1	1
C vs WVL vs SL vs SM vs WH (concentration)	0.0001	0.0001	0.0001	0.003	0.0001	0.0405	0.0086	0.0253
Interaction	1	1	1	1	1	1	1	1
C vs WVL	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.2368	0.0062
C vs WH	0.0001	0.0002	0.0001	0.0001	0.0001	0.0001	0.0247	0.0036
C vs WVL vs WH	0.0001	0.0001	0.0008	0.0013	0.0001	0.0001	0.0045	0.0205
C vs SL	0.0001	0.0001	0.0001	0.0001	0.0014	0.3679	0.2867	0.0013
C vs SM	0.0001	0.0004	0.0016	0.0001	0.0001	0.0001	0.062	0.0346
C vs SL vs SM	0.0001	0.0001	0.0029	0.0003	0.0001	0.1478	0.0978	0.0094

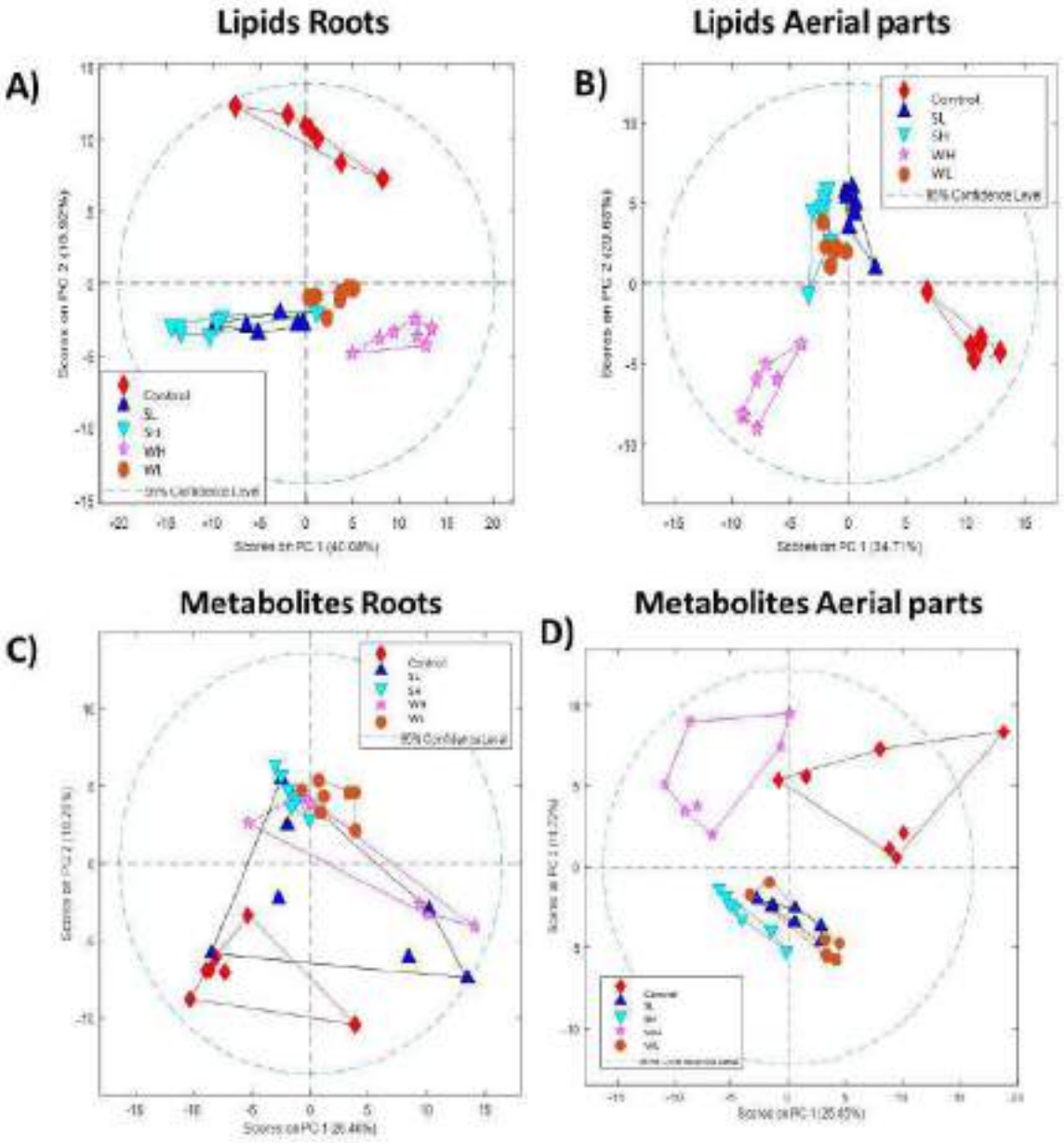


Figure S3. PCA score plots are shown for positive ionization mode obtained for lipidomic results for roots (A) and aerial parts (B), as well as metabolomic results for roots (C) and aerial parts (D). Both lipids and metabolites from root tissue are more affected by treatment rather than concentration level, whereas aerial parts have the opposite scenario.

Feature selection and annotation

Table S4. *PLS-DA results: n° of Variables Important in Projection > 1.0 and Matthew Correlation Coefficients.*

PLS-DA results		Lipidomics				Metabolomics			
		Roots ESI (+)	Roots ESI (-)	Aerial parts ESI (+)	Aerial parts ESI (-)	Roots ESI (+)	Roots ESI (-)	Aerial parts ESI (+)	Aerial parts ESI (-)
PLSDA Vips > 1.0	*C vs WH*	72	48	61	37	62	35	69	46
	C vs WVL	72	47	57	35	68	44	50	43
	C vs SM	74	47	58	37	65	37	59	53
	C vs SL	70	55	57	38	60	23	53	46
MCC	*C vs WH*	1.0	1.0	1.0	1.0	0.9	0.9	1.0	0.7
	C vs WVL	1.0	1.0	1.0	1.0	0.9	1.0	1.0	0.7
	C vs SM	1.0	1.0	1.0	1.0	0.7	1.0	1.0	0.9
	C vs SL	1.0	1.0	1.0	1.0	0.7	0.7	1.0	0.9

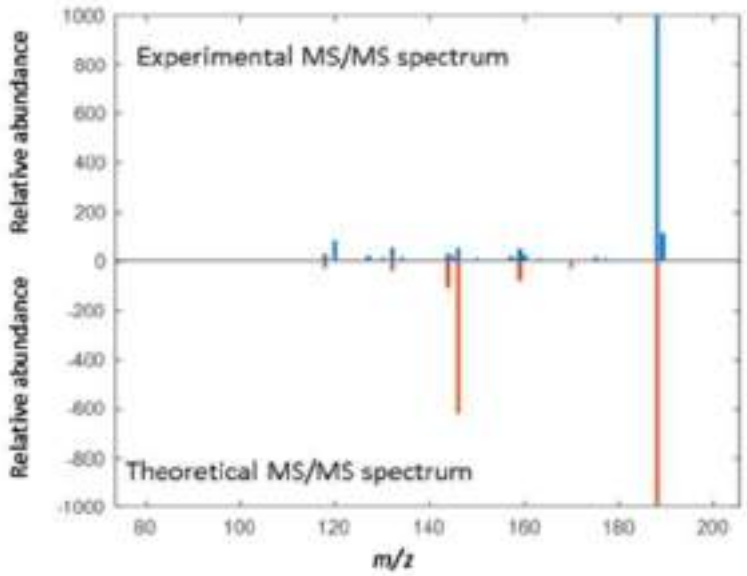


Figure S4. Experimental MS/MS spectrum of L-tryptophan compared to a theoretical MS/MS spectrum for the same compound in aerial tissues with ESI (+).

Table S5. MS-DIAL parameters used in metabolomic annotation.

Start up a project	HILIC-HRMS method ESI(+)	HILIC-HRMS method ESI(-)
Ionization type	Soft ionization	Soft ionization
Separation type	Chromatography (LC)	Chromatography (LC)
Method type	SWATH-MS or conventional All-ions method	SWATH-MS or conventional All-ions method
Data type (MS1)	Profile	Profile
Data type (MS/MS)	Profile	Profile
Ion mode	Positive ion mode	Negative ion mode
Target omics	Metabolomics	Metabolomics
	Data collection	
MS1 tolerance	0.01	0.01
MS2 tolerance	0.1	0.1
Retention time begin	0	0
Retention time end	20	20
Mass range begin	90	90
Mass range end	1000	1000
Maximum charged number	2	2
Consider Cl and Br elements	Unchecked	Unchecked
Number of threads	20	20
Execute retention time corrections	Unchecked	Unchecked
	Peak detection	
Minimum peak height	500	500
Mass slice width	0.1	0.1
Smoothing method	Linear weighted moving average	Linear weighted moving average
Smoothing level	3	3
Minimum peak width	5	5
Exclusion mass list (tolerance: 0.01Da)	Not used	Not used
	MS2Dec	
Sigma window value	0.5	0.5
MS2Dec amplitude cut off	100	100
Exclude after precursor	Checked	Checked
Keep isotope until	0.5	0.5
Keep the isotopic ion w/o MS2Dec	Unchecked	Unchecked
	Identification	
Retention time tolerance	0.5	0.5

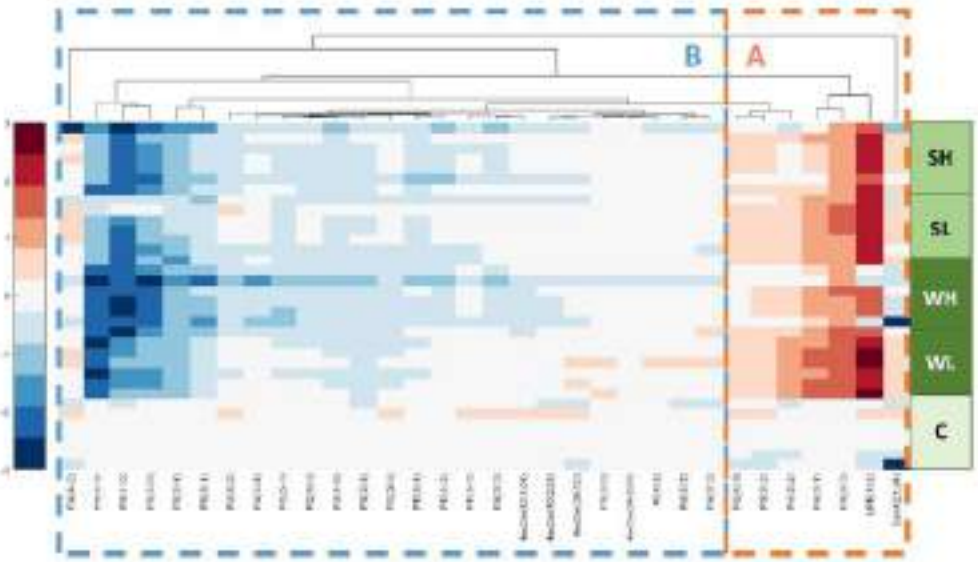
Accurate mass tolerance (MS1)	0.01	0.01
Accurate mass tolerance (MS2)	0.1	0.1
Identification score cut off	70	70
Use retention time for scoring	Unchecked	Unchecked
Use retention time for filtering	Unchecked	Unchecked
Postidentification	Not used	Not used
	Adduct	
Molecular species	$[M+H]^+$, $[M+NH_4]^+$, $[M+H-H_2O]^+$	$[M-H]^-$, $[M+Hac-H]^-$, $[M-H-H_2O]^-$
	Alignment	
Retention time tolerance	0.2	0.2
MS1 tolerance	0.02	0.02
Retention time factor	0	0
MS1 factor	1	1
Peak count filter	0	0
N% detected in at least one group	0	0
Remove feature based on blank information	Unchecked	Unchecked
Sample average / blank average	5	5
Keep "reference matched" metabolite features	Checked	Checked
Keep "suggested (w/o MS2)" metabolite features	Unchecked	Unchecked
Keep removable features and assign the tag	Checked	Checked
Gap filling by compulsion	Checked	Checked
	Isotope tracking	
	Not used	Not used

Table S6. *Experiment file used in MS method type section from start a project window in MS-DIAL*

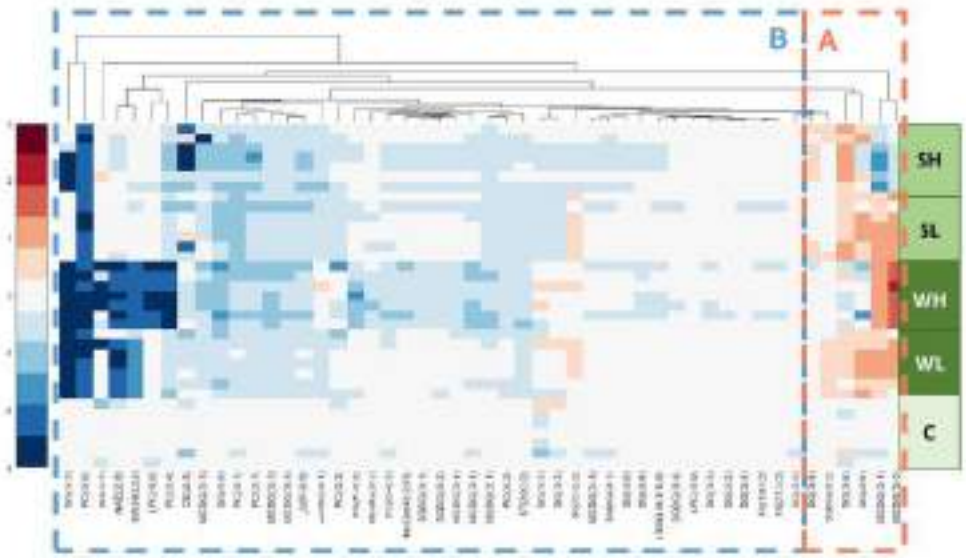
Experiment file			
Experiment	*MS Type*	*Min m/z*	*Max m/z*
0	SCAN	90	1000
1	MSMS	90	1000

Discussion of lipidomic results

Aerial negative mode



Roots positive mode



Roots negative mode

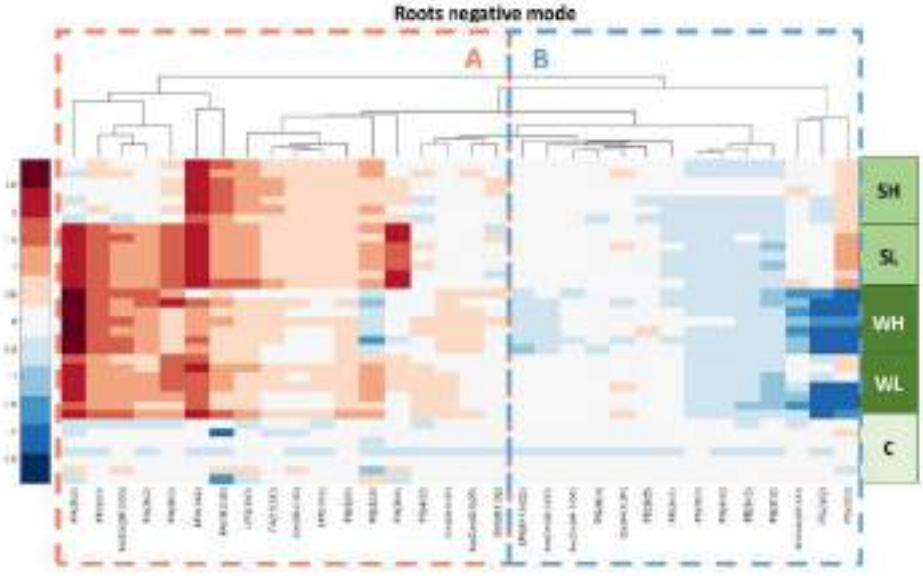


Figure S5. Hierarchical clustering heatmaps applied to the logarithm of the fold changes of the annotated lipids for aerial in positive mode, roots in positive and negative modes, respectively. A color intensity bar is included on the left side of the figure, indicating the relative abundance of the lipid regarding control samples (higher abundance in red, lower abundance in blue). Two clusters are differentiated in all cases, with an increasing abundance (A) and a diminishing abundance (B).

References

1. Pérez-Cova, M.; Bedia, C.; Stoll, D.R.; Tauler, R.; Jaumot, J. MSroi: A Pre-Processing Tool for Mass Spectrometry-Based Studies. *Chemometrics and Intelligent Laboratory Systems* **2021**, *215*, doi:10.1016/j.chemolab.2021.104333.
2. Gorrochategui, E.; Jaumot, J.; Tauler, R. ROIMCR: A Powerful Analysis Strategy for LC-MS Metabolomic Datasets. *BMC Bioinformatics* **2019**, *20*, 1–17, doi:10.1186/s12859-019-2848-8.
3. de Juan, A.; Jaumot, J.; Tauler, R. Multivariate Curve Resolution (MCR). Solving the Mixture Analysis Problem. *Analytical Methods* **2014**, *6*, 4964–4976, doi:10.1039/c4ay00571f.
4. Windig, W.; Stephenson, D.A. Self-Modeling Mixture Analysis of Second-Derivative Near-Infrared Spectral Data Using the Simplisma Approach. **1992**, *64*, 2735–2742.

Supplementary Material B

Adverse effects of arsenic uptake in rice metabolome and lipidome revealed by untargeted liquid chromatography coupled to mass spectrometry (LC-MS) and regions of interest multivariate curve resolution

Miriam Pérez-Cova[1,2], Romà Tauler[1], Joaquim Jaumot[1]*

[1]Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, 08034 Barcelona, Spain

[2]Department of Chemical Engineering and Analytical Chemistry, University of Barcelona, Diagonal 647, Barcelona, E08028, Barcelona, Spain

* Correspondence: joaquim.jaumot@idaea.csic.es

Table of Contents

Table SH. Metabolite identification of the most relevant features responsible for the changes induced by arsenic exposure on rice in aerial tissues, in negative ionization mode.

Table SA. Lipid identification of the most relevant features responsible for the changes induced by arsenic exposure on rice in root tissues, in positive ionization mode.									
Significant lipids (PLSDA, vips)				**Lipids ID**					
C-WH	**C-WVL**	**C-SM**	**C-SL**	**Sample type**	**MCR component**	**ROI ID**	**m/z**	**RT (min)**	**Lipid name**
x	x			Roots(+)	4	26	610.5402	7.77	DG(34:2)
		x	x	Roots(+)	16	36	634.5410	8.30	DG(36:4)
		x	x	Roots(+)	19	34	632.5306	7.37	DG(36:5)
		x		Roots(+)	22	40	638.5734	10.16	DG(36:2)
x		x		Roots(+)	30	174	934.6590	5.79	DGDG(34:2)
	x		x	Roots(+)	32	125	852.5716	4.87	DGDG(28:1)
	x	x		Roots(+)	34	185	958.6603	5.73	DGDG(36:4)
x	x	x	x	Roots(+)	41	85	756.5636	6.29	MGDG(33:3)
x	x	x	x	Roots(+)	44	75	718.5447	5.95	MGDG(30:1)
		x	x	Roots(+)	46	43	647.4626	12.05	DG(37:7)
		x		Roots(+)	50	15	510.3564	2.48	LPC(17:0)
x		x		Roots(+)	54	38	636.5582	9.26	DG(36:3)
x				Roots(+)	56	127	844.6961	9.23	MGDG(39:1)
x		x		Roots(+)	61	121	826.6831	9.79	HexCer(42:2;O3)
x	x	x	x	Roots(+)	63	92	760.5891	6.94	PC(34:1)
x				Roots(+)	66	14	496.3393	2.14	LPC(16:0)
x	x	x	x	Roots(+)	81	96	764.5518	5.08	PC(O-36:6) or PC(P-36:5)
x	x			Roots(+)	86	3	338.3385	3.84	NAE(20:0)
x		x	x	Roots(+)	87	176	936.6726	6.63	DGDG(34:1)
		x		Roots(+)	88	44	647.5161	4.62	SM(30:1;O2)
	x	x	x	Roots(+)	89	33	630.5157	6.44	DG(36:6)
	x	x		Roots(+)	93	89	742.5514	5.79	MGDG(32:3)
x	x	x	x	Roots(+)	96	80	740.5302	6.20	PC(33:4)
x	x		x	Roots(+)	99	77	720.5607	5.95	MGDG(30:0)
x	x	x	x	Roots(+)	100	101	778.5451	5.30	PC(36:6)
	x			Roots(+)	101	61	692.6376	16.17	CE(20:3)
x				Roots(+)	104	137	854.6949	11.43	PC(O-42:3)
x				Roots(+)	105	109	788.6353	6.47	MGDG(35:1)
x	x	x	x	Roots(+)	106	191	962.6876	7.03	DGDG(36:2)
x	x		x	Roots(+)	107	99	771.6293	7.13	SM(39:2;O2)
x	x		x	Roots(+)	108	120	816.6635	7.87	MGDG(37:1)
	x		x	Roots(+)	111	79	728.5301	5.39	PC(32:3)
x	x		x	Roots(+)	112	9	413.3763	5.52	ST(29:2;O)
x	x	x	x	Roots(+)	113	75	718.5447	6.63	PC(31:1)
x	x	x	x	Roots(+)	116	67	704.5332	6.29	MGDG(29:1)
	x	x	x	Roots(+)	118	46	663.4874	4.46	DG(38:6)
x	x		x	Roots(+)	122	18	573.4899	5.45	DG(O-32:2)
	x	x	x	Roots(+)	127	2	337.2675	8.61	FA(21:3;O)
x	x		x	Roots(+)	128	88	758.5759	7.28	PC(34:2)
x	x	x		Roots(+)	129	82	742.5830	5.95	PC(P-34:2)
	x		x	Roots(+)	131	60	691.5192	6.20	DG(40:6)
		x		Roots(+)	135	73	716.5305	5.79	MGDG(30:2)
x	x	x	x	Roots(+)	140	81	741.5357	6.20	DG(44:9)
x	x	x	x	Roots(+)	141	31	617.5154	7.96	DG(35:2)
		x		Roots(+)	142	30	615.4991	8.95	DG(34:2)
	x		x	Roots(+)	144	78	726.5131	4.87	MGDG(31:4)

x	x	x	x	Roots(+)	145	104	782.5789	6.78	MGDG(35:4)
			x	Roots(+)	146	35	633.5353	7.37	DG(35:0)
		x	x	Roots(+)	147	28	613.4841	7.96	DG(34:3)
	x	x	x	Roots(+)	149	1	313.2683	8.98	FA(19:1;O)

Table SA. Lipid identification of the most relevant features responsible for the changes induced by arsenic exposure on rice in root tissues, in positive ionization mode.

Lipids ID								
Molecular formula lipid	**Molecular formula adduct**	**Adduct**	**m/z theo adduct**	**Lipidmaps code**	**Lipid name**	**HMDB code**	**KEGG code**	**delta (ppm)**
C37H68O5	C37H72NO5	[M+NH4]+	610.541	LMGL02010021	DG(34:2)	NA	NA	0.49
C39H68O5	C39H72NO5	[M+NH4]+	634.541	LMGL02010063	DG(36:4)	HMDB0093295	NA	0.79
C39H66O5	C39H70NO5	[M+NH4]+	632.525	LMGL02010071	DG(36:5)	HMDB0007250	NA	9.17
C39H72O5	C39H76NO5	[M+NH4]+	638.572	LMGL02010049	DG(36:2)	HMDB0007218	C00165	2.51
C49H88O15	C49H92NO15	[M+NH4]+	934.646	LMGL05019F6T	DGDG(34:2)	NA	NA	13.80
C43H82NO15	C43H78O15	[M+NH4]+	852.568	LMGL05019FL2	DGDG(28:1)	NA	NA	4.34
C51H88O15	C51H92NO15	[M+NH4]+	958.646	LMGL05019GIT	DGDG(36:4)	NA	NA	14.81
C42H74O10	C42H78NO10	[M+NH4]+	756.562	LMGL05019AOY	MGDG(33:3)	NA	NA	2.11
C39H72O10	C39H76NO10	[M+NH4]+	718.546	LMGL05019AFB	MGDG(30:1)	NA	NA	2.37
C40H64O5	C40H64O5Na	[M+Na]+	647.465	LMGL02010473	DG(37:7)	NA	NA	3.09
C25H52NO7P	C25H53NO7P	[M+H]+	510.355	LMGP01050024	LPC(17:0)	HMDB0012108	NA	1.96
C39H70O5	C39H74NO5	[M+NH4]+	636.556	LMGL02010056	DG(36:3)	HMDB0007219	C00165	3.30
C48H90O10	C48H94NO10	[M+NH4]+	844.687	LMGL05019AI1	MGDG(39:1)	NA	NA	10.54
C48H91NO9	C48H92NO9	[M+H]+	826.677	LMSP05010110	HexCer(42:2;O3)	NA	NA	7.74
C42H82NO8P	C42H83NO8P	[M+H]+	760.585	LMGP01010005	PC(34:1)	HMDB0007972	C00157	5.26
C24H50NO7P	C24H51NO7P	[M+H]+	496.34	LMGP01050018	LPC(16:0)	HMDB0010382	C04230	1.01
C44H78NO7P	C44H78NO7P	[M+H]+	764.559	LMGP01030040	PC(O-36:6) or PC(P-36:5)	HMDB0011222	NA	9.29
C22H45NO2	C22H44NO	[M+H-H2O]+	338.342	LMFA08040038	NAE(20:0)	NA	NA	9.46
C49H90O15	C49H94NO15	[M+NH4]+	936.662	LMGL05019F4I	DGDG(34:1)	NA	NA	11.53
C35H71N2O6P	C35H71N2O6P	[M+H]+	647.512	LMSP03010002	SM(30:1;O2)	HMDB0012096	C00550	6.02
C39H64O5	C39H68NO5	[M+NH4]+	630.509	LMGL02010401	DG(36:6)	HMDB0007034	NA	10.31
C41H72O10	C41H76NO10	[M+NH4]+	742.546	LMGL05019AKC	MGDG(32:3)	NA	NA	6.72
C41H74NO8P	C41H75NO8P	[M+H]+	740.523	LMGP01011704	PC(33:4)	HMDB0008231	C00157	10.40
C39H74O10	C39H78NO10	[M+NH4]+	720.562	LMGL05019AFA	MGDG(30:0)	NA	NA	1.80
C44H76NO8P	C44H77NO8P	[M+H]+	778.538	LMGP01010512	PC(36:6)	HMDB0007892	C00157	8.99
C47H78O2	C47H82NO2	[M+NH4]+	692.634	LMST01020013	CE(20:3)	HMDB06736	C02530	5.20
C50H96NO7P	C50H97NO7P	[M+H]+	854.7	LMGP01020252	PC(O-42:3)	NA	NA	5.62
C44H82O10	C44H86NO10	[M+NH4]+	788.625	LMGL05019AZ2	MGDG(35:1)	NA	NA	13.57
C51H92O15	C51H96NO15	[M+NH4]+	962.677	LMGL05019GIL	DGDG(36:2)	NA	NA	10.60
C44H87N2O6P	C44H88N2O6P	[M+H]+	771.637	LMSP03010064	SM(39:2;O2)	NA	NA	10.50
C46H86O10	C46H90NO10	[M+NH4]+	816.656	LMGL05019A9T	MGDG(37:1)	NA	NA	9.31
C40H74NO8P	C40H75NO8P	[M+H]+	728.523	LMGP01010497	PC(32:3)	HMDB0007876	C00157	10.43
C29H48O	C29H49O	[M+H]+	413.378	LMST01010176	ST(29:2;O)	NA	NA	3.63
C39H76NO8P	C39H77NO8P	[M+H]+	718.538	LMGP01010535	PC(31:1)	HMDB0007936	C00157	9.15
C38H70O10	C38H74NO10	[M+NH4]+	704.531	LMGL05019AC0	MGDG(29:1)	NA	NA	3.55
C41H68O5	C41H68O5Na	[M+Na]+	663.496	LMGL02010162	DG(38:6)	HMDB0007121	NA	12.81
C35H66O4	C35H66O4Na	[M+Na]+	573.485	LMGL02040001	DG(O-32:2)	NA	C13864	8.02
C21H36O3	C21H37O3	[M+H]+	337.274	LMFA01070041	FA(21:3;O)	NA	NA	18.38
C42H80NO8P	C42H81NO8P	[M+H]+	758.569	LMGP01010585	PC(34:2)	NA	NA	8.57
C42H80NO7P	C42H80NO7P	[M+H]+	742.575	LMGP01030008	PC(P-34:2)	HMDB0011211	NA	11.45
C43H72O5	C43H72O5Na	[M+Na]+	691.527	LMGL02010216	DG(40:6)	HMDB0007179	NA	11.53
C39H70O10	C39H74NO10	[M+NH4]+	716.531	LMGL05019AFK	MGDG(30:2)	NA	NA	0.28
C47H74O5	C47H74O5Na	[M+Na]+	741.543	LMGL02010300	DG(44:9)	NA	NA	9.57
C38H70O5	C37H70O5Na	[M+Na]+	617.512	LMGL02010373	DG(35:2)	NA	NA	6.32
C37H68O5	C37H68O5Na	[M+Na]+	615.496	LMGL02010349	DG(34:2)	NA	NA	5.17
C40H68O10	C40H72NO10	[M+NH4]+	726.515	LMGL05019AH3	MGDG(31:4)	NA	NA	2.75

C44H76O10	C44H80NO10	[M+NH4]+	782.578	LMGL05019A0J	MGDG(35:4)	NA	NA	1.53
C38H74O5	C38H74O5Na	[M+Na]+	633.543	LMGL02010371	DG(35:0)	HMDB0093295	NA	11.84
C37H66O5	C37H66NaO5	[M+Na]+	613.48	LMGL02010350	DG(34:3)	NA	NA	6.36
C19H36O3	C19H37O3	[M+H]+	313.274	LMFA01060128	FA(19:1;O)	NA	NA	17.20

Table SA. Lipid identification of the most relevant features responsible for the changes induced by arsenic exposure on rice in root tissues, in positive ionization mode.

	Log 10(Fold changes)				General trend of fold changes			
Lipid name	**WH**	**WVL**	**SM**	**SL**	**WH**	**WVL**	**SM**	**SL**
DG(34:2)	-3.54	-2.15	-0.17	-3.54	DOWN	DOWN	DOWN	DOWN
DG(36:4)	-0.16	-0.06	-0.09	-0.05	DOWN	DOWN	DOWN	DOWN
DG(36:5)	-0.17	0.29	0.21	0.08	DOWN	UP	UP	UP
DG(36:2)	0.11	-0.31	-0.42	-0.01	UP	DOWN	DOWN	DOWN
DGDG(34:2)	-0.60	-0.25	-0.22	-0.15	DOWN	DOWN	DOWN	DOWN
DGDG(28:1)	-0.19	-0.33	-0.16	0.02	DOWN	DOWN	DOWN	UP
DGDG(36:4)	-0.20	-0.13	-0.22	0.00	DOWN	DOWN	DOWN	UP
MGDG(33:3)	-0.94	-0.57	-0.68	-0.56	DOWN	DOWN	DOWN	DOWN
MGDG(30:1)	1.01	-1.11	0.82	0.76	UP	DOWN	UP	UP
DG(37:7)	-2.45	0.04	0.07	-0.72	DOWN	UP	UP	DOWN
LPC(17:0)	-0.18	-0.11	-0.10	0.01	DOWN	DOWN	DOWN	UP
DG(36:3)	0.10	-0.31	-0.37	-0.01	UP	DOWN	DOWN	DOWN
MGDG(39:1)	-0.61	-0.25	-0.20	-0.18	DOWN	DOWN	DOWN	DOWN
HexCer(42:2;O3)	-0.56	-0.24	-0.19	-0.09	DOWN	DOWN	DOWN	DOWN
PC(34:1)	-0.86	-0.94	-0.99	-0.47	DOWN	DOWN	DOWN	DOWN
LPC(16:0)	-2.40	-0.14	-0.04	-0.09	DOWN	DOWN	DOWN	DOWN
PC(O-36:6) or PC(P-36:5)	-0.64	-0.78	-0.73	-0.42	DOWN	DOWN	DOWN	DOWN
NAE(20:0)	-2.46	-0.28	-0.23	-2.09	DOWN	DOWN	DOWN	DOWN
DGDG(34:1)	-0.53	-0.24	-0.12	-0.06	DOWN	DOWN	DOWN	DOWN
SM(30:1;O2)	-0.29	-0.20	-0.18	-0.10	DOWN	DOWN	DOWN	DOWN
DG(36:6)	-0.19	0.95	0.77	0.46	DOWN	UP	UP	UP
MGDG(32:3)	-0.90	-1.06	-0.30	-0.55	DOWN	DOWN	DOWN	DOWN
PC(33:4)	-2.53	-0.86	-0.61	-0.64	DOWN	DOWN	DOWN	DOWN
MGDG(30:0)	1.80	-0.42	1.02	0.80	UP	DOWN	UP	UP
PC(36:6)	-2.55	-2.29	-2.40	-2.38	DOWN	DOWN	DOWN	DOWN
CE(20:3)	-0.28	-1.66	-0.49	-0.37	DOWN	DOWN	DOWN	DOWN
PC(O-42:3)	-0.59	-0.32	-0.20	-0.14	DOWN	DOWN	DOWN	DOWN
MGDG(35:1)	-0.82	-0.32	-0.23	-0.23	DOWN	DOWN	DOWN	DOWN
DGDG(36:2)	0.06	0.18	0.21	0.35	UP	UP	UP	UP
SM(39:2;O2)	-2.13	-0.18	-0.07	-1.59	DOWN	DOWN	DOWN	DOWN
MGDG(37:1)	-0.69	-0.45	-0.43	-0.28	DOWN	DOWN	DOWN	DOWN
PC(32:3)	-0.49	-0.28	-0.50	-0.40	DOWN	DOWN	DOWN	DOWN
ST(29:2;O)	-0.56	-0.24	-0.57	-0.41	DOWN	DOWN	DOWN	DOWN
PC(31:1)	-0.81	-1.02	-0.79	-0.46	DOWN	DOWN	DOWN	DOWN
MGDG(29:1)	0.13	-0.68	-0.63	-0.57	UP	DOWN	DOWN	DOWN
DG(38:6)	-0.32	-0.10	0.06	0.09	DOWN	DOWN	UP	UP
DG(O-32:2)	0.12	-0.28	0.27	0.25	UP	DOWN	UP	UP
FA(21:3;O)	-0.16	0.01	-0.04	-0.02	DOWN	UP	DOWN	DOWN
PC(34:2)	-0.57	-0.44	-0.51	-0.31	DOWN	DOWN	DOWN	DOWN
PC(P-34:2)	-1.08	-0.19	-0.26	-0.20	DOWN	DOWN	DOWN	DOWN
DG(40:6)	-0.49	0.11	0.61	0.90	DOWN	UP	UP	UP
MGDG(30:2)	-0.56	-0.09	-0.25	-0.07	DOWN	DOWN	DOWN	DOWN
DG(44:9)	-1.21	-0.82	-0.92	-0.80	DOWN	DOWN	DOWN	DOWN
DG(35:2)	0.01	0.10	0.02	0.02	UP	UP	UP	UP
DG(34:2)	-0.17	-0.07	-0.10	-0.03	DOWN	DOWN	DOWN	DOWN
MGDG(31:4)	-0.13	-0.33	-0.07	0.04	DOWN	DOWN	DOWN	UP

MGDG(35:4)	-0.73	-0.59	-0.75	-0.48	DOWN	DOWN	DOWN	DOWN
DG(35:0)	-0.19	-0.28	0.00	0.06	DOWN	DOWN	UP	UP
DG(34:3)	-0.27	-0.06	-0.03	-0.04	DOWN	DOWN	DOWN	DOWN
FA(19:1;O)	-0.21	-0.02	-0.05	0.00	DOWN	DOWN	DOWN	UP

Table SB. Lipid identification of the most relevant features responsible for the changes induced by arsenic exposure on rice in root tissues, in negative ionization mode.

Lipids ID								
Molecular formula lipid	**Molecular formula adduct**	**Adduct**	**m/z theo adduct**	**Lipidmaps code**	**Lipid name**	**HMDB code**	**KEGG code**	**delta (ppm)**
C40H75O10P	C40H74O10P	[M-H]-	745.5025	LMGP04010066	PG(34:2)	NA	NA	3.15
C48H91NO9	C49H92NO11	[M+HCOO]-	870.6676	LMSP05010056	HexCer(36:1;O)	NA	NA	1.82
C38H71O8P	C38H70O8P	[M-H]-	685.4814	LMGP10010088	PA(35:2)	NA	NA	3.23
C42H79NO9	C43H80NO11	[M+HCOO]-	786.5737	LMSP0501AA67	HexCer(36:2;O3)	NA	NA	4.23
C41H73O8P	C41H72O8P	[M-H]-	723.497	LMGP10010970	PA(38:4)	HMDB0114846	NA	2.18
C18H32O3	C18H31O3	[M-H]-	295.2279	LMFA01060232	FA(18:2;O)	NA	NA	8.87
C21H39O7P	C21H38O7P	[M-H]-	433.2361	LMGP10050044	LPA(18:2)	HMDB07852	C00416	6.24
C39H72NO8P	C39H71NO8P	[M-H]-	712.4923	LMGP02010417	PE(34:3)	HMDB0008837	C00350	2.30
C45H82NO10P	C45H81NO10P	[M-H]-	826.5604	LMGP03010275	PS(39:3)	NA	NA	2.04
C50H97NO10	C51H98NO12	[M+HCOO]-	916.7095	LMSP05010047	HexCer(44:1;O4)	NA	NA	1.07
C22H45O9P	C22H44O9P	[M-H]-	483.2729	LMGP04050008	LPG(16:0)	NA	NA	5.49
C18H32O4	C18H31O4	[M-H]-	311.2228	LMFA02000248	FA(18:2;O2)	NA	NA	6.74
C38H70NO8P	C38H69NO8P	[M-H]-	698.4766	LMGP02010398	PE(33:3)	NA	NA	10.26
C46H85NO9	C46H84NO9	[M-H]-	794.6152	LMSP05010169	HexCer(40:3;O3)	NA	NA	6.02
C40H81NO4	C41H82NO6	[M+HCOO]-	684.6148	LMSP02020032	Cer(40:0;O3)	NA	NA	4.54
C41H77O8P	C41H76O8P	[M-H]-	727.5283	LMGP10010198	PA(38:2)	HMDB0114845	NA	3.83
C39H68NO10P	C39H67NO10P	[M-H]-	740.4508	LMGP03010086	PS(33:4)	NA	NA	5.53
C38H73NO4	C39H74NO6	[M+HCOO]-	652.5522	LMSP02010080	Cer(38:2;O3)	NA	NA	4.23
C41H75O8P	C41H74O8P	[M-H]-	725.5127	LMGP10010222	PA(38:3)	HMDB0114866	NA	5.81
C41H72NO8P	C41H71NO8P	[M-H]-	736.4923	LMGP02010450	PE(36:5)	HMDB0008877	C00350	2.06
C39H70NO10P	C39H69NO10P	[M-H]-	742.4665	LMGP03010085	PS(33:3)	NA	NA	7.39
C44H87NO5	C45H88NO7	[M+HCOO]-	754.6566	LMSP02010179	Cer(44:1;O4)	NA	NA	5.32
C25H52NO9P	C25H51NO9P	[M-H]-	540.3307	LMGP03060017	LPS(O-19:0;O)	NA	NA	5.78
C39H76NO8P	C39H75NO8P	[M-H]-	716.5236	LMGP02010378	PE(34:1)	NA	NA	8.83
C23H44NO7P	C23H43NO7P	[M-H]-	476.2783	LMGP02050011	LPE(18:2)	HMDB0011507	NA	5.02
C35H71N2O6P	C36H72N2O8P	[M+HCOO]-	691.5032	LMSP03010002	SM(30:1;O2)	HMDB0012096	C00550	3.48
C47H81O8P	C47H80O8P	[M-H]-	803.5596	LMGP10010043	PA(44:6)	HMDB0115266	NA	3.36
C43H80NO10P	C43H79NO10P	[M-H]-	800.5447	LMGP03010158	PS(37:2)	HMDB0116749	NA	4.62
C48H78NO10P	C48H77NO10P	[M-H]-	858.5291	LMGP03010588	PS(42:8)	HMDB0116765	NA	8.41
C44H83NO9	C45H84NO11	[M+HCOO]-	814.605	LMSP0501AA69	HexCer(38:2;O3)	NA	NA	2.95
C46H83O10P	C46H82O10P	[M-H]-	825.5651	LMGP04010883	PG(40:4)	HMDB0010611	NA	0.12
C38H71O10P	C38H70O10P	[M-H]-	717.4712	LMGP04010060	PG(32:2)	NA	NA	3.35
	Log 10(Fold changes)				**General trend of fold changes**			
Lipid name	**WH**	**WVL**	**SM**	**SL**	**WH**	**WVL**	**SM**	**SL**
PG(34:2)	-2.01	-0.29	-0.18	-1.36	DOWN	DOWN	DOWN	DOWN
HexCer(36:1;O)	-1.37	0.17	0.24	-0.54	DOWN	UP	UP	DOWN
PA(35:2)	2.66	-0.37	2.08	1.94	UP	DOWN	UP	UP
HexCer(36:2;O3)	-0.33	-0.01	-0.07	-0.13	DOWN	DOWN	DOWN	DOWN
PA(38:4)	-0.09	0.31	1.94	0.23	DOWN	UP	UP	UP
FA(18:2;O)	0.37	0.75	0.69	0.33	UP	UP	UP	UP
LPA(18:2)	1.13	2.06	2.12	1.77	UP	UP	UP	UP

PE(34:3)	-0.40	-0.31	-0.55	-0.61	DOWN	DOWN	DOWN	DOWN
PS(39:3)	-0.35	-0.35	-0.57	-0.39	DOWN	DOWN	DOWN	DOWN
HexCer(44:1;O4)	-0.19	-0.04	0.06	0.03	DOWN	DOWN	UP	UP
LPG(16:0)	0.45	1.02	1.15	0.78	UP	UP	UP	UP
FA(18:2;O2)	0.80	1.43	1.31	0.92	UP	UP	UP	UP
PE(33:3)	1.71	0.17	1.47	1.25	UP	UP	UP	UP
HexCer(40:3;O3)	0.38	0.11	0.14	0.06	UP	UP	UP	UP
Cer(40:0;O3)	0.50	-0.05	0.03	0.47	UP	DOWN	UP	UP
PA(38:2)	1.15	0.05	1.05	0.88	UP	UP	UP	UP
PS(33:4)	0.33	0.55	0.70	0.59	UP	UP	UP	UP
Cer(38:2;O3)	0.18	0.58	0.63	0.29	UP	UP	UP	UP
PA(38:3)	0.82	0.30	1.61	1.37	UP	UP	UP	UP
PE(36:5)	0.10	-0.13	0.01	0.06	UP	DOWN	UP	UP
PS(33:3)	-0.82	0.70	0.88	0.91	DOWN	UP	UP	UP
Cer(44:1;O4)	-0.15	0.05	0.20	0.14	DOWN	UP	UP	UP
LPS(O-19:0;O)	-0.65	-0.01	0.04	0.09	DOWN	DOWN	UP	UP
PE(34:1)	-0.28	-0.33	-0.28	-0.07	DOWN	DOWN	DOWN	DOWN
LPE(18:2)	0.09	0.58	0.51	0.27	UP	UP	UP	UP
SM(30:1;O2)	0.15	0.15	-0.10	-0.05	UP	UP	DOWN	DOWN
PA(44:6)	-0.47	-0.31	-0.53	-0.40	DOWN	DOWN	DOWN	DOWN
PS(37:2)	-0.68	-0.41	-0.71	-0.74	DOWN	DOWN	DOWN	DOWN
PS(42:8)	0.12	-0.09	0.19	0.43	UP	DOWN	UP	UP
HexCer(38:2;O3)	1.38	0.01	1.05	1.00	UP	UP	UP	UP
PG(40:4)	-0.15	-0.06	0.13	0.07	DOWN	DOWN	UP	UP
PG(32:2)	-2.05	0.66	0.80	-0.92	DOWN	UP	UP	DOWN

Table SC. Lipid identification of the most relevant features responsible for the changes induced by arsenic exposure on rice in aerial tissues, in positive ionization mode.

Significant lipids (PLSDA, vips)				**Lipids ID**					
C-WH	**C-WVL**	**C-SM**	**C-SL**	**Sample type**	**MCR component**	**ROI ID**	**m/z**	**RT (min)**	**Lipid name**
x	x	x	x	Aerials(+)	9	105	760.5199	4.307	PC(O-34:5)
x	x	x	x	Aerials(+)	10	55	630.5172	6.319	DG(36:6)
x	x	x	x	Aerials(+)	12	45	608.5277	7.773	DG(34:3)
x	x	x		Aerials(+)	15	131	792.5728	5.144	PG(36:2)
		x	x	Aerials(+)	19	197	960.6777	6.196	DGDG(36:3)
	x	x	x	Aerials(+)	20	59	634.5431	8.176	DG(36:4)
x	x	x	x	Aerials(+)	25	115	770.6211	7.094	PC(P-36:2) or PC(O-36:3)
x				Aerials(+)	27	1	338.3377	3.812	NAE(20:0)
x	x	x	x	Aerials(+)	28	118	780.5660	5.947	PC(36:5)
x	x	x	x	Aerials(+)	31	57	632.5293	7.217	DG(36:5)
x	x	x	x	Aerials(+)	34	120	782.5779	7.278	PC(36:4)
	x	x	x	Aerials(+)	38	20	520.3398	1.986	LPC(18:2)
x		x		Aerials(+)	44	157	858.5433	4.120	SQDG(36:5)
x	x	x	x	Aerials(+)	45	14	496.3394	2.140	LPC(16:0)
	x	x	x	Aerials(+)	46	18	518.3242	1.706	LPC(18:3)
x	x	x	x	Aerials(+)	52	101	758.5745	7.278	PC(34:2)
	x			Aerials(+)	55	125	784.6628	10.778	HexCer(40:1;O2)
x	x	x		Aerials(+)	58	79	716.5325	6.691	MGDG(30:2)
x	x	x	x	Aerials(+)	59	151	850.5566	4.400	DGDG(28:2)
x	x	x	x	Aerials(+)	60	50	612.5564	9.630	DG(34:1)
	x	x	x	Aerials(+)	65	116	778.5505	5.452	PG(35:2)
x	x	x	x	Aerials(+)	70	3	374.3228	2.975	MG(18:1)
	x		x	Aerials(+)	84	77	714.5172	5.947	MGDG(30:3)
x		x	x	Aerials(+)	86	23	532.3484	1.614	LPC(19:3)
x	x		x	Aerials(+)	88	153	852.5735	4.741	DGDG(28:1)
x	x		x	Aerials(+)	94	76	704.5354	5.640	MGDG(29:1)

x	x			Aerials(+)	106	73	688.5009	5.359	PC(29:2)
x		x	x	Aerials(+)	110	109	764.5526	4.864	PC(O-36:6) or PC(P-36:5)
x	x	x	x	Aerials(+)	101	93	744.5638	6.350	MGDG(32:2)

Table SC. Lipid identification of the most relevant features responsible for the changes induced by arsenic exposure on rice in aerial tissues, in positive ionization mode.

Lipids ID								
Molecular formula lipid	**Molecular formula adduct**	**Adduct**	**m/z theo adduct**	**Lipidmaps code**	**Lipid name**	**HMDB code**	**KEGG code**	**delta (ppm)**
C42H76NO7P	C42H76NO7PNa	[M+Na]+	760.5252	LMGP01030033	PC(O-34:5)	HMDB0011214	NA	6.93
C39H64O5	C39H68NO5	[M+NH4]+	630.5092	LMGL02010401	DG(36:6)	HMDB0007034	NA	12.61
C37H66O5	C37H70NO5	[M+NH4]+	608.5248	LMGL02010350	DG(34:3)	NA	NA	4.78
C42H79O10P	C42H83NO10P	[M+NH4]+	792.5749	LMGP04010107	PG(36:2)	NA	NA	2.63
C51H90O15	C51H94NO15	[M+NH4]+	960.6618	LMGL05019GIR	DGDG(36:3)	NA	NA	16.52
C39H68O5	C39H72NO5	[M+NH4]+	634.5405	LMGL02010063	DG(36:4)	HMDB0093295	NA	4.03
C44H84NO7P	C44H85NO7P	[M+H]+	770.6058	LMGP01030038	PC(P-36:2) or PC(O-36:3)	HMDB0011217	NA	19.85
C22H45NO2	C22H44NO	[M+H-H2O]+	338.3417	LMFA08040038	NAE(20:0)	NA	NA	11.71
C44H78NO8P	C44H79NO8P	[M+H]+	780.5538	LMGP01010633	PC(36:5)	HMDB0007984	C00157	15.68
C39H66O5	C39H70NO5	[M+NH4]+	632.5248	LMGL02010071	DG(36:5)	HMDB0007250	NA	7.06
C44H80NO8P	C44H81NO8P	[M+H]+	782.5694	LMGP01010629	PC(36:4)	NA	NA	10.80
C26H50NO7P	C26H51NO7P	[M+H]+	520.3398	LMGP01050034	LPC(18:2)	NA	NA	0.03
C45H76O12S	C45H80NO12S	[M+NH4]+	858.5396	LMGL05019K7S	SQDG(36:5)	NA	NA	4.28
C24H50NO7P	C24H51NO7P	[M+H]+	496.3398	LMGP01050018	LPC(16:0)	HMDB0010382	C04230	0.83
C26H48NO7P	C26H49NO7P	[M+H]+	518.3241	LMGP01050038	LPC(18:3)	HMDB0010388	C04230	0.23
C42H80NO8P	C42H81NO8P	[M+H]+	758.5694	LMGP01010585	PC(34:2)	NA	NA	6.67
C46H89NO8	C46H90NO8	[M+H]+	784.6661	LMSP0501AC04	HexCer(40:1;O2)	NA	NA	4.23
C39H70O10	C39H74NO10	[M+NH4]+	716.5307	LMGL05019AFK	MGDG(30:2)	NA	NA	2.56
C43H76O15	C43H80NO15	[M+NH4]+	850.5522	LMGL05019FMB	DGDG(28:2)	NA	NA	5.15
C37H70O5	C37H74NO5	[M+NH4]+	612.5561	LMGL02010004	DG(34:1)	NA	NA	0.48
C41H77O10P	C41H81NO10P	[M+NH4]+	778.5593	LMGP04010090	PG(35:2)	NA	NA	11.33
C21H40O4	C21H44NO4	[M+NH4]+	374.3265	LMGL01010024	MG(18:1)	HMDB0011537	NA	9.76
C39H68O10	C39H72NO10	[M+NH4]+	714.5151	LMGL05019ABG	MGDG(30:3)	NA	NA	2.96
C27H50NO7P	C27H51NO7P	[M+H]+	532.3398	LMGP01050003	LPC(19:3)	NA	NA	16.13
C43H82NO15	C43H78O15	[M+NH4]+	852.5679	LMGL05019FL2	DGDG(28:1)	NA	NA	6.55
C38H70O10	C38H74NO10	[M+NH4]+	704.5307	LMGL05019AC0	MGDG(29:1)	NA	NA	6.67
C37H70NO8P	C37H71NO8P	[M+H]+	688.4912	LMGP01011322	PC(29:2)	NA	NA	14.16
C44H78NO7P	C44H78NO7P	[M+H]+	764.5589	LMGP01030040	PC(O-36:6) or PC(P-36:5)	HMDB0011222	NA	8.22
C41H74O10	C41H78NO10	[M+NH4]+	744.562	LMGL05019AL8	MGDG(32:2)	NA	NA	2.39

	Log 10(Fold changes)				General trend of fold changes			
Lipid name	**WH**	**WVL**	**SM**	**SL**	**WH**	**WVL**	**SM**	**SL**
PC(O-34:5)	-0.90	-0.19	-0.14	-0.17	DOWN	DOWN	DOWN	DOWN
DG(36:6)	1.32	1.93	1.78	1.55	UP	UP	UP	UP
DG(34:3)	0.29	0.53	0.37	0.34	UP	UP	UP	UP
PG(36:2)	-2.15	-1.73	-0.70	-1.47	DOWN	DOWN	DOWN	DOWN
DGDG(36:3)	-0.14	0.23	0.17	0.15	DOWN	UP	UP	UP
DG(36:4)	-0.09	0.42	0.29	0.37	DOWN	UP	UP	UP
PC(P-36:2) or PC(O-36:3)	-0.11	0.25	0.19	0.24	DOWN	UP	UP	UP
NAE(20:0)	0.13	0.14	0.05	-0.02	UP	UP	UP	DOWN
PC(36:5)	-0.63	-0.29	-0.30	-0.28	DOWN	DOWN	DOWN	DOWN
DG(36:5)	0.14	0.59	0.39	0.34	UP	UP	UP	UP
PC(36:4)	-0.76	-0.40	-0.31	-0.17	DOWN	DOWN	DOWN	DOWN
LPC(18:2)	-0.52	1.24	1.25	1.30	DOWN	UP	UP	UP
SQDG(36:5)	-0.77	-0.11	-0.05	-0.05	DOWN	DOWN	DOWN	DOWN
LPC(16:0)	0.72	1.58	1.42	1.37	UP	UP	UP	UP
LPC(18:3)	1.68	3.52	3.48	3.42	UP	UP	UP	UP
PC(34:2)	-2.24	-1.99	-1.48	-1.64	DOWN	DOWN	DOWN	DOWN
HexCer(40:1;O2)	-0.24	0.30	-0.03	0.17	DOWN	UP	DOWN	UP
MGDG(30:2)	-0.15	-0.20	-0.07	0.21	DOWN	DOWN	DOWN	UP
DGDG(28:2)	0.15	0.29	0.23	0.36	UP	UP	UP	UP
DG(34:1)	0.05	0.51	0.36	0.51	UP	UP	UP	UP
PG(35:2)	-0.39	-0.16	-0.26	-0.33	DOWN	DOWN	DOWN	DOWN
MG(18:1)	-1.00	-0.20	-0.19	-0.34	DOWN	DOWN	DOWN	DOWN
MGDG(30:3)	-0.13	-0.09	-0.19	0.22	DOWN	DOWN	DOWN	UP
LPC(19:3)	-1.85	0.41	0.39	0.27	DOWN	UP	UP	UP
DGDG(28:1)	0.15	0.22	0.27	0.46	UP	UP	UP	UP
MGDG(29:1)	0.13	0.10	0.24	0.64	UP	UP	UP	UP
PC(29:2)	0.04	0.06	0.10	0.41	UP	UP	UP	UP
PC(O-36:6) or PC(P-36:5)	-0.78	-0.38	-0.23	0.01	DOWN	DOWN	DOWN	UP
MGDG(32:2)	-0.37	-0.29	-0.15	0.14	DOWN	DOWN	DOWN	UP

Table SD. Lipid identification of the most relevant features responsible for the changes induced by arsenic exposure on rice in aerial tissues, in negative ionization mode.

Significant lipids (PLSDA, vips)				**Lipids ID**					
C-WH	**C-WVL**	**C-SM**	**C-SL**	**Sample Type**	**MCR component**	**ROI ID**	**m/z**	**RT (min)**	**Lipid name**
x	x	x	x	Aerials(-)	9	92	860.6432	7.85	PS(41:0)
x	x	x	x	Aerials(-)	10	65	814.6031	7.14	HexCer(38:2;O3)
	x		x	Aerials(-)	12	55	797.5387	5.99	PG(38:4)
	x			Aerials(-)	16	103	888.6792	9.18	HexCer(42:1;O4)
x	x	x	x	Aerials(-)	17	68	822.5324	5.65	PS(39:5)
	x	x		Aerials(-)	18	87	842.6367	8.47	HexCer(40:2;O3)
x		x		Aerials(-)	20	16	699.4935	5.96	PA(36:2)
	x	x	x	Aerials(-)	23	4	476.2774	1.81	LPE(18:2)
x	x	x	x	Aerials(-)	24	72	826.5603	6.46	PS(39:3)
	x	x		Aerials(-)	26	69	823.5528	5.65	PG(40:5)
x	x	x	x	Aerials(-)	27	15	697.4775	5.22	PA(36:3)
x	x	x		Aerials(-)	31	97	870.6695	9.80	HexCer(36:1;O)
x			x	Aerials(-)	32	71	825.5699	7.29	PG(40:4)
		x	x	Aerials(-)	34	27	714.5062	6.71	PE(34:2)
x	x	x	x	Aerials(-)	39	74	828.5782	7.48	PS(39:2)
	x	x	x	Aerials(-)	40	62	811.6252	10.79	PA(44:2)
x				Aerials(-)	43	21	707.4651	4.73	PA(37:5)
x	x	x	x	Aerials(-)	44	60	804.5748	8.28	PS(37:0)
x	x	x	x	Aerials(-)	46	78	832.6017	6.46	PS(39:0)
x				Aerials(-)	47	23	709.4771	5.22	PA(37:4)
	x	x	x	Aerials(-)	48	42	738.5049	6.21	PE(36:4)
x	x	x	x	Aerials(-)	51	70	824.5445	5.81	PS(39:4)
x	x	x	x	Aerials(-)	52	93	861.6506	7.85	PG(42:0)

x	x			Aerials(-)	53	10	683.4625	4.97	PA(35:3)
x			x	Aerials(-)	54	24	712.4907	5.96	PE(34:3)
x	x	x	x	Aerials(-)	56	12	685.4774	5.65	PA(35:2)
x	x	x	x	Aerials(-)	57	76	830.5916	8.62	PS(39:1)
x			x	Aerials(-)	58	47	747.5145	5.47	PG(34:1)
x	x		x	Aerials(-)	60	125	961.6047	5.81	PI(40:3)
x	x	x	x	Aerials(-)	61	18	701.5089	6.80	PA(36:1)
x	x	x	x	Aerials(-)	64	59	803.5629	6.86	PA(44:6)

Table SD. Lipid identification of the most relevant features responsible for the changes induced by arsenic exposure on rice in aerial tissues, in negative ionization mode.

Lipids ID								
Molecular formula lipid	**Molecular formula adduct**	**Adduct**	**m/z theo adduct**	**Lipidmaps code**	**Lipid name**	**HMDB code**	**KEGG code**	**delta (ppm)**
C47H92NO10P	C47H91NO10P	[M-H]-	860.6386	LMGP03010524	PS(41:0)	NA	NA	5.40
C44H83NO9	C45H84NO11	[M+HCOO]-	814.605	LMSP0501AA69	HexCer(38:2;O3)	NA	NA	2.27
C44H79O10P	C44H78O10P	[M-H]-	797.5338	LMGP04010968	PG(38:4)	HMDB0010581	NA	6.14
C48H93NO10	C49H94NO12	[M+HCOO]-	888.6782	LMSP05010058	HexCer(42:1;O4)	NA	NA	1.16
C45H78NO10P	C45H77NO10P	[M-H]-	822.5291	LMGP03010276	PS(39:5)	NA	NA	4.02
C46H87NO9	C47H88NO11	[M+HCOO]-	842.6363	LMSP0501AA81	HexCer(40:2;O3)	NA	NA	0.42
C39H73O8P	C39H72O8P	[M-H]-	699.497	LMGP10010105	PA(36:2)	HMDB0116699	NA	4.96
C23H44NO7P	C23H43NO7P	[M-H]-	476.2783	LMGP02050011	LPE(18:2)	HMDB0011507	NA	1.94
C45H82NO10P	C45H81NO10P	[M-H]-	826.5604	LMGP03010275	PS(39:3)	NA	NA	0.12
C46H81O10P	C46H80O10P	[M-H]-	823.5495	LMGP04010339	PG(40:5)	HMDB0010641	NA	4.00
C39H71O8P	C39H70O8P	[M-H]-	697.4814	LMGP10010134	PA(36:3)	HMDB0114805	NA	5.62
C48H91NO9	C49H92NO11	[M+HCOO]-	870.6676	LMSP05010056	HexCer(36:1;O)	NA	NA	2.18
C46H83O10P	C46H82O10P	[M-H]-	825.5651	LMGP04010883	PG(40:4)	HMDB0010611	NA	5.87
C39H74NO8P	C39H73NO8P	[M-H]-	714.5079	LMGP02010379	PE(34:2)	NA	NA	2.34
C45H84NO10P	C45H83NO10P	[M-H]-	828.576	LMGP03010246	PS(39:2)	NA	NA	2.60
C47H89O8P	C47H88O8P	[M-H]-	811.6222	LMGP10010722	PA(44:2)	HMDB0115262	NA	3.75
C40H69O8P	C40H68O8P	[M-H]-	707.4657	LMGP10010188	PA(37:5)	NA	NA	0.87
C43H84NO10P	C43H83NO10P	[M-H]-	804.576	LMGP03010156	PS(37:0)	HMDB0112334	NA	1.46
C45H88NO10P	C45H87NO10P	[M-H]-	832.6073	LMGP03010244	PS(39:0)	NA	NA	6.67
C40H71O8P	C40H70O8P	[M-H]-	709.4814	LMGP10010157	PA(37:4)	HMDB0114827	NA	6.04
C41H74NO8P	C41H73NO8P	[M-H]-	738.5079	LMGP02010421	PE(36:4)	HMDB0008844	C00350	4.09
C45H80NO10P	C45H79NO10P	[M-H]-	824.5447	LMGP03010247	PS(39:4)	NA	NA	0.25
C48H95O10P	C48H94O10P	[M-H]-	861.659	LMGP04010947	PG(42:0)	NA	NA	9.74
C38H69O8P	C38H68O8P	[M-H]-	683.4657	LMGP10010151	PA(35:3)	HMDB0115509	NA	4.65
C39H72NO8P	C39H71NO8P	[M-H]-	712.4923	LMGP02010417	PE(34:3)	HMDB0008837	C00350	2.21
C38H71O8P	C38H70O8P	[M-H]-	685.4814	LMGP10010088	PA(35:2)	NA	NA	5.81
C45H86NO10P	C45H85NO10P	[M-H]-	830.5917	LMGP03010245	PS(39:1)	NA	NA	0.18
C40H77O10P	C40H74O10P	[M-H]-	747.5182	LMGP04010066	PG(34:1)	NA	NA	4.91
C49H89O13P	C50H90O15P	[M+HCOO]-	961.6023	LMGP06010306	PI(40:3)	NA	NA	2.50
C39H75O8P	C39H74O8P	[M-H]-	701.5127	LMGP10010104	PA(36:1)	NA	NA	5.40
C47H81O8P	C47H80O8P	[M-H]-	803.5596	LMGP10010043	PA(44:6)	HMDB0115266	NA	4.14

Table SD. Lipid identification of the most relevant features responsible for the changes induced by arsenic exposure on rice in aerial tissues, in negative ionization mode.

	Log 10(Fold changes)				General trend of fold changes			
Lipid name	**WH**	**WVL**	**SM**	**SL**	**WH**	**WVL**	**SM**	**SL**
PS(41:0)	-2.39	-2.20	-1.47	-1.51	DOWN	DOWN	DOWN	DOWN
HexCer(38:2;O3)	-0.12	0.07	0.07	0.15	DOWN	UP	UP	UP
PG(38:4)	-0.49	-0.27	-0.20	-0.29	DOWN	DOWN	DOWN	DOWN
HexCer(42:1;O4)	-0.45	-0.25	-0.12	0.09	DOWN	DOWN	DOWN	UP
PS(39:5)	-0.55	-0.48	-0.38	-0.43	DOWN	DOWN	DOWN	DOWN
HexCer(40:2;O3)	-0.31	-0.10	-0.03	0.11	DOWN	DOWN	DOWN	UP
PA(36:2)	-0.71	-0.43	-0.01	-0.14	DOWN	DOWN	DOWN	DOWN
LPE(18:2)	0.65	1.98	2.18	2.46	UP	UP	UP	UP
PS(39:3)	-2.25	-1.34	-0.96	-2.46	DOWN	DOWN	DOWN	DOWN
PG(40:5)	-0.42	-0.41	-0.15	0.11	DOWN	DOWN	DOWN	UP
PA(36:3)	1.28	1.09	1.35	1.57	UP	UP	UP	UP
HexCer(36:1;O)	-0.23	-0.08	0.02	0.22	DOWN	DOWN	UP	UP
PG(40:4)	-0.68	-0.39	-0.25	-0.19	DOWN	DOWN	DOWN	DOWN
PE(34:2)	-0.42	-0.52	-0.31	-0.17	DOWN	DOWN	DOWN	DOWN

PS(39:2)	0.29	0.36	0.48	0.56	UP	UP	UP	UP
PA(44:2)	-0.17	-2.03	0.14	0.19	DOWN	DOWN	UP	UP
PA(37:5)	0.08	-0.13	0.03	0.08	UP	DOWN	UP	UP
PS(37:0)	-2.04	-1.68	-0.98	-1.17	DOWN	DOWN	DOWN	DOWN
PS(39:0)	-0.06	0.10	0.12	0.20	DOWN	UP	UP	UP
PA(37:4)	-0.72	-0.36	-0.11	-0.11	DOWN	DOWN	DOWN	DOWN
PE(36:4)	-0.50	-0.54	-0.48	-0.24	DOWN	DOWN	DOWN	DOWN
PS(39:4)	-1.19	-1.27	-0.84	-1.20	DOWN	DOWN	DOWN	DOWN
PG(42:0)	0.08	0.32	0.34	0.44	UP	UP	UP	UP
PA(35:3)	0.12	-0.16	-0.08	0.17	UP	DOWN	DOWN	UP
PE(34:3)	-0.20	-0.32	-0.23	-0.09	DOWN	DOWN	DOWN	DOWN
PA(35:2)	0.40	0.10	0.49	1.01	UP	UP	UP	UP
PS(39:1)	-1.23	-0.85	-0.55	-0.45	DOWN	DOWN	DOWN	DOWN
PG(34:1)	-0.85	-0.41	-0.36	-0.20	DOWN	DOWN	DOWN	DOWN
PI(40:3)	-0.07	-0.07	0.01	0.11	DOWN	DOWN	UP	UP
PA(36:1)	0.74	0.55	1.17	1.44	UP	UP	UP	UP
PA(44:6)	-0.63	-0.50	-0.37	-0.27	DOWN	DOWN	DOWN	DOWN

Table SE. Metabolite identification of the most relevant features responsible for the changes induced by arsenic exposure on rice in root tissues, in positive ionization mode.

Significant metabolites (PLSDA, vips)				**Metabolite ID**							
C-WH	**C-SM**	**C-SL**	**C-WVL**	**MCR components**	**ROI ID**	**m/z**	**RT(min)**	**MS/MS fragments, ordered by intensity**			
x	x	x	x	67	4	102.0563	13.47	102.06			
x	x	x	x	70	5	104.0721	13.62	104.07	103.01		
x	x	x	x	22	9	106.0513	13.63	106.04			
x	x		x	14	14	118.0877	12.37	118.04	95.05		
x	x	x	x	35	16	120.0670	13.45	120.06	102.03	< 90	
x	x	x	x	39	25	130.0876	12.69	94.05	110.07	112.05	114.06
x	x	x	x	57	33	134.0463	12.51	< 90			
x	x		x	33	72	166.0879	8.82	120.08	103.06	91.06	102.06
	x	x	x	6	83	177.1040	14.43	115.04	159.07	130.04	105.03
x	x		x	41	90	182.0831	10.57	91.06	95.05	107.74	119.05
x	x	x	x	21	162	268.1068	7.27	134.06	268.16		

Table SE. Metabolite identification of the most relevant features responsible for the changes induced by arsenic exposure on rice in root tissues, in positive ionization mode.

Metabolite ID								
Metabolite name	**Adduct**	**Plantcyc (*Oryza L. sativa*)**	**Chemical Formula**	**Monoisotopic-Molecular-Weight**	**m/z theo adduct**	**delta (ppm)**	**HMDB**	**Kegg**
1-Aminocyclopropane-1-carboxylic acid	[M+H]+	yes	C4H7NO2	101.0477	102.0550	13.16	HMDB36458	C01234
4-aminobutyric acid	[M+H]+	yes	C4H9NO2	103.0633	104.0706	14.18	HMDB00112	C00334
Serine	[M+H]+	yes	C3H7NO3	105.0426	106.0499	13.61	HMDB0000187	C00065
Betaine	[M+H]+	yes	C5H11NO2	117.0790	118.0863	11.93	HMDB00043	C00719
Threonine	[M+H]+	yes	C4H9NO3	119.0582	120.0655	12.09	HMDB00167	C00188
N4-Acetylaminobutanal	[M+H]+	yes	C6H11NO2	129.0790	130.0863	11.00	HMDB0004226	C05936
Aspartic acid	[M+H]+	yes	C4H6NO4	133.0375	134.0448	10.98	HMDB00191	C00049
Phenylalanine	[M+H]+	yes	C9H11NO2	165.0790	166.0863	9.67	HMDB00159	C00079
Serotonin	[M+H]+	yes	C10H13N2O	176.0950	177.1022	9.99	HMDB00259	C00780
Tyrosine	[M+H]+	yes	C9H11NO3	181.0739	182.0812	10.43	HMDB00158	C00082
Adenosine		yes	C10H13N5O4	267.0968	268.1040	10.00	HMDB0000050	C00212
	Log 10(Fold changes)				**General trend of fold changes**			
Metabolite name	**WH**	**WVL**	**SM**	**SL**	**WH**	**WVL**	**SM**	**SL**

1-Aminocyclopropane-1-carboxylic acid	0.08	0.59	0.86	0.90	UP	UP	UP	UP
4-aminobutyric acid	-3.23	-3.63	-1.27	-3.25	DOWN	DOWN	DOWN	DOWN
Serine	-2.81	-3.83	-2.09	-2.99	DOWN	DOWN	DOWN	DOWN
Betaine	0.91	0.35	0.61	0.98	UP	UP	UP	UP
Threonine	-2.26	-2.86	-0.87	-2.46	DOWN	DOWN	DOWN	DOWN
N4-Acetylaminobutanal	1.46	0.62	0.87	1.48	UP	UP	UP	UP
Aspartic acid	0.62	0.33	0.40	0.86	UP	UP	UP	UP
Phenylalanine	0.37	-0.02	0.13	0.43	UP	DOWN	UP	UP
Serotonin	0.70	0.17	0.43	0.83	UP	UP	UP	UP
Tyrosine	0.70	0.31	0.37	0.97	UP	UP	UP	UP
Adenosine	0.57	0.34	0.31	0.77	UP	UP	UP	UP

Table SF. Metabolite identification of the most relevant features responsible for the changes induced by arsenic exposure on rice in root tissues, in negative ionization mode.

Significant metabolites (PLSDA, vips)				**Metabolite ID**							
C-WH	**C-SM**	**C-SL**	**C-WVL**	**MCR components**	**ROI ID**	**m/z**	**RT(min)**	**MS/MS fragments, ordered by intensity**			
x	x	x	x	11	3	104.0329	13.65	104.02	< 90 Da		
			x	35	5	114.0536	12.59	114.02	< 90 Da		
x	x		x	14	6	116.0692	11.80	116.07			
x	x		x	13	7	117.0168	5.19	117.02	99.01		
x	x		x	9	11	130.0848	10.15	130.05	112.98	115.04	
			x	22	14	132.0277	12.56	132.06	114.93		
x				30	18	135.0274	7.58	134.96	91.04	117.02	
x	x		x	20	19	137.0218	2.98	137.09	93.00	94.04	
x				6	23	145.0594	13.54	145.06	127.05	128.03	109.04
x	x	x	x	19	28	150.0205	6.94	123.90	125.90		
x	x		x	18	33	157.0339	7.53	157.05	96.96	113.99	140.01
x	x	x		61	39	173.0430	8.29	173.10	93.02	111.01	137.02
x	x		x	27	44	179.0536	6.86	179.03	161.04	125.02	178.02
	x	x		48	46	181.0691	12.07	181.07	113.02	101.07	97.03
x	x		x	58	55	203.0800	8.25	203.11	159.09	142.07	116.05
		x	x	34	68	233.1270	13.77	131.08			
x	x		x	39	74	255.2308	3.01	219.98			
		x		59	77	267.0697	8.09	135.03	92.06		

Table SF. Metabolite identification of the most relevant features responsible for the changes induced by arsenic exposure on rice in root tissues, in negative ionization mode.

Metabolite ID								
Metabolite name	**Adduct**	**Plantcyc (*Oryza L. sativa*)**	**Chemical Formula**	**Monoisotopic-Molecular-Weight**	**m/z theo adduct**	**delta (ppm)**	**HMDB**	**Kegg**
Serine	[M-H]-	yes	C3H7NO3	105.0426	104.0353	23.71	HMDB0000187	C00065
Proline	[M-H]-	yes	C5H9NO2	115.0633	114.0561	21.81	HMDB00162	C00148
Norvaline	[M-H]-	yes	C5H11NO2	117.0790	116.0717	21.34	HMDB0013716	C01826
Succinic acid	[M-H]-	yes	C4H4O4	118.0266	117.0193	21.33	HMDB00254	C00042
Isoleucine	[M-H]-	yes	C6H13NO2	131.0946	130.0874	19.28	HMDB00172	C00407
Aspartate	[M-H]-	yes	C4H6NO4	133.0375	132.0302	19.49	HMDB00191	C00049
Threonic acid	[M-H]-	yes	C4H7O5	136.0372	135.0299	18.63	HMDB00943	C01620
4-hydroxybenzoate	[M-H]-	yes	C7H5O3	138.0317	137.0244	19.14	HMDB00500	C00156
Glutamine	[M-H]-	yes	C5H10N2O3	146.0691	145.0619	16.95	HMDB00641	C00064

2,3-Dihydrodipicolinate	[M-H20-H]-	yes	C7H7NO4	169.0375	150.02	9.21	HMDB0012247	C03340
allantoin	[M-H]-	yes	C4H6N4O3	158.0440	157.0367	17.87	HMDB0000462	C01551
Shikimic acid	[M-H]-	yes	C7H9O5	174.0528	173.0455	14.96	HMDB03070	C00493
myo-Inositol	[M-H]-	yes	C6H12O6	180.0634	179.0561	13.80	HMDB00211	C00137
Sorbitol	[M-H]-	yes	C6H14O6	182.0790	181.0718	14.52	HMDB00247	C00794
Tryptophan	[M-H]-	yes	C11H12N2O2	204.0899	203.0826	12.90	HMDB00929	C00078
Arginine	[M+Hac-H]-	yes	C6H14N4O2	174.1117	233.13	6.00	HMDB0000517	C00062
Palmitic acid	[M-H]-	yes	C16H32O2	256.2402	255.2330	9.00	HMDB0000220	C00249
Inosine	[M-H]-	yes	C10H12N4O5	268.0808	267.0735	14.29	HMDB00195	C00294
	Log 10(Fold changes)				**General trend of fold changes**			
Metabolite name	**WH**	**WVL**	**SM**	**SL**	**WH**	**WVL**	**SM**	**SL**
Serine		0.59	-0.07	0.26	UP	DOWN	UP	UP
Proline		0.85	0.09	0.62	UP	UP	UP	UP
Norvaline		0.72	0.19	0.61	UP	UP	UP	UP
Succinic acid		0.67	0.05	0.48	UP	UP	UP	UP
Isoleucine		0.92	0.17	0.60	UP	UP	UP	UP
Aspartate		-0.90	-2.15	0.50	DOWN	DOWN	UP	DOWN
Threonic acid		-1.82	-1.92	0.29	DOWN	DOWN	UP	DOWN
4-hydroxybenzoate		0.69	0.33	0.67	UP	UP	UP	UP
Glutamine		0.53	-0.46	0.64	UP	DOWN	UP	UP
2,3-Dihydrodipicolinate		-2.08	-2.22	0.47	DOWN	DOWN	UP	DOWN
allantoin		0.27	-0.45	0.55	UP	DOWN	UP	DOWN
Shikimic acid		-1.84	-2.09	-1.96	DOWN	DOWN	DOWN	DOWN
myo-Inositol		0.50	-0.69	0.81	UP	DOWN	UP	UP
Sorbitol		-2.10	-2.22	0.46	DOWN	DOWN	UP	DOWN
Tryptophan		-0.69	-2.70	-0.11	DOWN	DOWN	DOWN	DOWN
Arginine		-0.87	-1.43	0.49	DOWN	DOWN	UP	DOWN
Palmitic acid		-2.38	-2.95	-1.17	DOWN	DOWN	DOWN	DOWN
Inosine		-0.14	-0.45	0.25	DOWN	DOWN	UP	DOWN

Table SG. Metabolite identification of the most relevant features responsible for the changes induced by arsenic exposure on rice in aerial tissues, in positive ionization mode.

Significant metabolites (PLSDA, vips)				**Metabolite ID**							
C-WH	**C-SM**	**C-SL**	**C-WVL**	**MCR components**	**ROI ID**	**m/z**	**RT(min)**	**MS/MS fragments, ordered by intensity**			
x	x	x	x	54	6	104.0719	13.61	104.07	< 90		
x	x	x	x	24	10	106.0511	13.65	106.05	105.07		
	x		x	18	12	116.0718	11.48	115.03	116.06	< 90	
	x	x	x	34	18	120.0667	13.48	120.02	101.97	119.02	
		x	x	12	28	130.0511	13.54	130.05	< 90		
x	x	x	x	10	34	133.0619	13.87	104.12	105.07	115.05	133.07
x	x	x		4	70	160.0769	16.30	116.07	118.09	133.06	142.04
x	x			9	78	166.0877	8.89	103.07	95.07	120.08	91.06
x	x			102	88	177.1035	14.17	115.07	130.07	159.07	143.07
x	x			51	110	205.0989	8.26	143.11	115.09	91.06	130.07

Table SG. Metabolite identification of the most relevant features responsible for the changes induced by arsenic exposure on rice in aerial tissues, in positive ionization mode.

Metabolite ID								
Metabolite name	**Adduct**	**Plantcyc (*Oryza L. sativa*)**	**Chemical Formula**	**Monoisotopic-MW**	**m/z theo**	**delta (ppm)**	**HMDB**	**Kegg**
Dimethylglycine	[M+H]+	yes	C4H9NO2	103.0633	104.07	12.05	HMDB00092	C01026
Serine	[M+H]+	yes	C3H7NO3	105.0426	106.05	11.60	HMDB00187	C00065
Proline	[M+H]+	yes	C5H9NO2	115.0633	116.07	9.93	HMDB00162	C00148
Beta-Homoserine	[M+H]+	yes	C4H9NO3	119.0582	120.07	9.82	HMDB00719	C00263
Pyroglutamic acid	[M+H]+	yes	C5H6NO3	129.0426	130.05	9.22	HMDB00267	C01879
Cinnamaldehyde	[M+H]+	yes	C9H8O	132.0575	133.06	21.58	HMDB03441	C00903
Indoleacetaldehyde	[M+H]+	yes	C10H9NO	159.0684	160.08	8.00	HMDB0001190	C00637
Phenylalanine	[M+H]+	yes	C9H11NO2	165.0790	166.09	8.68	HMDB00159	C00079
Serotonin	[M+H]+	yes	C10H13N2O	176.0950	177.10	7.37	HMDB00259	C00780
Tryptophan	[M+H]+	yes	C11H12N2O2	204.0899	205.10	8.45	HMDB00929	C00078
	Log 10(Fold changes)				**General trend of fold changes**			
Metabolite name	**WH**	**WVL**	**SM**	**SL**	**WH**	**WVL**	**SM**	**SL**
Dimethylglycine	1.87	1.80	-1.62	-0.90	UP	UP	DOWN	DOWN
Serine	0.15	0.27	-0.53	-0.89	UP	UP	DOWN	DOWN
Proline	1.46	1.73	1.33	1.20	UP	UP	UP	UP
Beta-Homoserine	-0.23	-0.04	-0.22	-0.46	DOWN	DOWN	DOWN	DOWN
Pyroglutamic acid	-0.43	0.36	-0.25	-0.37	DOWN	UP	DOWN	DOWN
Cinnamaldehyde	1.12	0.26	-0.06	-0.34	UP	UP	DOWN	DOWN
Indoleacetaldehyde	0.31	0.42	0.11	-0.22	UP	UP	UP	DOWN
Phenylalanine	1.64	0.94	-0.59	-0.77	UP	UP	DOWN	DOWN
Serotonin	0.93	0.67	0.10	-0.44	UP	UP	UP	DOWN
Tryptophan	3.12	1.57	0.12	-0.08	UP	UP	UP	DOWN

Table SH. Metabolite identification of the most relevant features responsible for the changes induced by arsenic exposure on rice in aerial tissues, in negative ionization mode.

Significant metabolites (PLSDA, vips)				Metabolite ID								
C-WH	**C-SM**	**C-SL**	**C-WVL**	**MCR component**	**ROI ID**	**m/z**	**RT(min)**	**MS/MS fragments**				
x	x	x	x	22	6	104.0331	13.62	104.02	<90			
x	x	x	x	69	12	113.0334	13.83	113.02	112.02			
x	x			27	13	114.0538	12.56	114.02	< 90 Da			
x	x		x	14	17	117.0172	5.31	126.97	99.01			
x	x	x	x	53	20	121.0272	4.20	118.98	94.98	92.03	120.04	93.04
x	x	x		10	28	130.0850	10.07	130.09				
	x	x	x	4	30	131.0440	13.83	131.08	114.03	95.01	113.03	111.05
	x	x	x	15	32	132.0280	12.36	114.98	114.02	< 90		
	x			97	38	134.0450	7.02	134.06	106.98	133.06		
	x	x		33	39	135.0277	7.34	135.03	91.06	117.05		
	x	x		32	44	145.0121	5.87	144.99	101.02			
x		x	x	1	46	145.0598	13.66	145.11	127.00	109.04	128.05	
			x	3	55	153.0172	2.76	152.97	109.03			
x	x	x	x	24	59	157.0343	7.52	156.94	113.99	97.06		
	x			25	65	164.0695	8.80	164.07	147.04	91.05	103.05	
x	x	x		9	73	173.0432	8.34	173.04	137.02	127.05	111.04	93.03
	x			37	88	203.0806	8.20	203.00	116.05	142.07	159.09	186.06
x	x	x	x	69	116	285.0799	10.43	119.01	93.02	120.04		

Table SH. Metabolite identification of the most relevant features responsible for the changes induced by arsenic exposure on rice in aerial tissues, in negative ionization mode.

Metabolite ID								
Metabolite name	**Adduct**	**Plantcyc (*Oryza L. sativa*)**	**Chemical Formula**	**Monoisotopic-Molecular-Weight**	**m/z theo adduct**	**delta(ppm)**	**HMDB**	**Kegg**
Serine	[M-H]-	yes	C3H7NO3	105.0426	104.04	21.31	HMDB00187	C00065
5,6-dihydrouracil	[M-H]-	yes	C4H6N2O2	114.0429	113.04	19.61	HMDB00076	C00429
Proline	[M-H]-	yes	C5H9NO2	115.0633	114.06	19.75	HMDB00162	C00148
Succinic acid	[M-H]-	yes	C4H4O4	118.0266	117.02	18.02	HMDB00254	C00042
Benzoic acid	[M-H]-	yes	C7H5O2	122.0368	121.03	19.20	HMDB01870	C00180
Leucine	[M-H]-	yes	C6H13NO2	131.0946	130.09	18.08	HMDB00687	C00123
Asparagine	[M-H]-	yes	C4H8N2O3	132.0535	131.05	16.91	HMDB00168	C00152
L-Aspartic acid	[M-H]-	yes	C4H6NO4	133.0375	132.03	16.89	HMDB0019	C00049
Adenine	[M-H]-	yes	C5H5N5	135.0545	134.05	16.50	HMDB00034	C00147
Threonic acid	[M-H]-	yes	C4H7O5	136.0372	135.03	16.13	HMDB00943	C01620
2-Oxoglutarate	[M-H]-	yes	C5H4O5	146.0215	145.01	14.58	HMDB00208	C00026
Glutamine	[M-H]-	yes	C5H10N2O3	146.0691	145.06	14.41	HMDB00641	C00064
2,5-dihydroxybenzoic acid	[M-H]-	yes	C7H5O4	154.0266	153.02	13.76	HMDB00152	C00628
Allantoin	[M-H]-	yes	C4H6N4O3	158.0440	157.04	15.37	HMDB00462	C01551
Phenylalanine	[M-H]-	yes	C9H11NO2	165.0790	164.07	13.42	HMDB00159	C00079
Shikimic Acid	[M-H]-	yes	C7H9O5	174.0528	173.05	13.56	HMDB03070	C00493
Tryptophan	[M-H]-	yes	C11H12N2O2	204.0899	203.08	9.97	HMDB00929	C00078
Sakuranetin	[M-H]-	yes	C16H14O5	286.0841	285.08	10.85	HMDB30090	C09833
	Log 10(Fold changes)				**General trend of fold changes**			
Metabolite name	**WH**	**WVL**	**SM**	**SL**	**WH**	**WVL**	**SM**	**SL**
Serine	-2.69	-1.96	0.43	-3.00	DOWN	DOWN	UP	DOWN
5,6-dihydrouracil	0.34	0.16	0.27	0.26	UP	UP	UP	UP
Proline	0.67	1.87	1.63	1.73	UP	UP	UP	UP
Succinic acid	-0.10	-0.06	0.30	0.07	DOWN	DOWN	UP	UP
Benzoic acid	1.57	0.84	0.41	0.15	UP	UP	UP	UP
Leucine	0.04	-0.19	-0.31	-0.34	UP	DOWN	DOWN	DOWN
Asparagine	-0.92	-0.22	-0.23	-0.43	DOWN	DOWN	DOWN	DOWN
L-Aspartic acid	1.18	0.64	0.49	0.42	UP	UP	UP	UP
Adenine	1.58	1.30	1.11	1.28	UP	UP	UP	UP
Threonic acid	-0.20	0.88	-0.75	-1.32	DOWN	UP	DOWN	DOWN
2-Oxoglutarate	1.86	1.19	-0.70	-0.66	UP	UP	DOWN	DOWN
Glutamine	0.04	-0.58	-0.54	0.13	UP	DOWN	DOWN	UP
2,5-dihydroxybenzoic acid	0.52	0.00	0.06	0.07	UP	UP	UP	UP
Allantoin	0.95	1.63	0.07	0.28	UP	UP	UP	UP
Phenylalanine	2.00	1.47	0.37	0.21	UP	UP	UP	UP
Shikimic Acid	2.42	2.04	1.27	1.68	UP	UP	UP	UP
Tryptophan	2.42	2.04	1.27	1.68	UP	UP	UP	UP
Sakuranetin	0.87	2.19	0.39	0.17	UP	UP	UP	UP

VIII. SCIENTIFIC PUBLICATION VIII

Title: Environmental metabolomic and sphingolipids assessment of three pharmaceutical compounds, amoxicillin, carbamazepine, and trazodone in human hepatic cells

Authors: Miriam Pérez-Cova, Carmen Bedia, Antonio Checa, Isabel Meister, Romà Tauler, Craig E Wheelock, Joaquim Jaumot

To be submitted – June 2022

Title:

Metabolomics and sphingolipidomics study of human hepatoma cells exposed to environmental concentrations of pharmaceutical compounds

Authors:

Miriam Pérez-Cova[a,b], Carmen Bedia[a], Antonio Checa[c], Isabel Meister[c], Romà Tauler[a], Craig E Wheelock[c] , Joaquim Jaumot[a]

[a]Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, E08034 Barcelona, Spain

[b]Department of Chemical Engineering and Analytical Chemistry, University of Barcelona, Diagonal 647, Barcelona, E08028, Barcelona, Spain

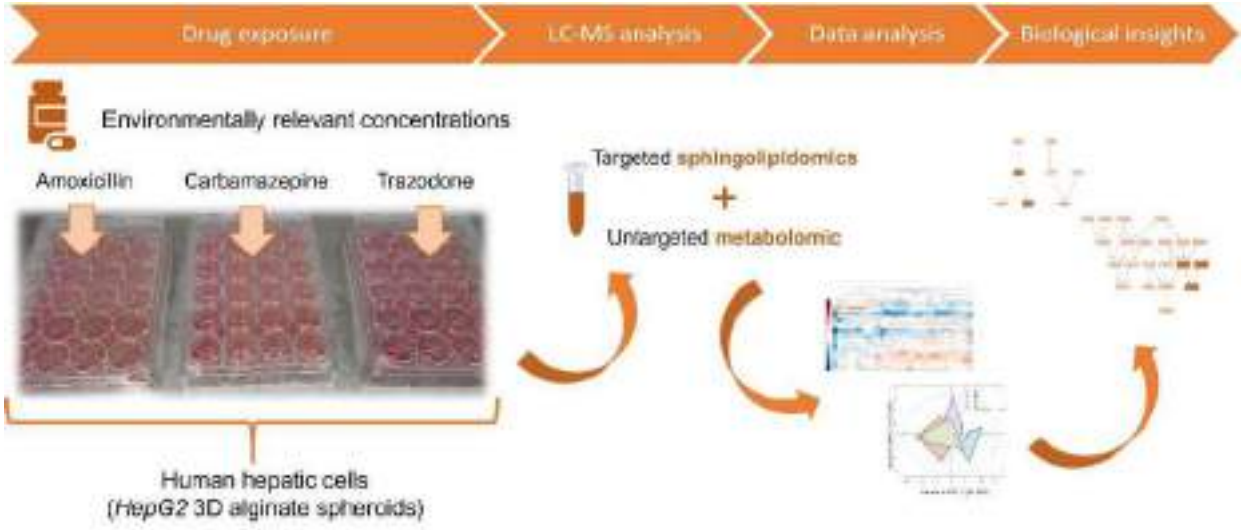
[c]Division of Physiological Chemistry II, Department of Medical Biochemistry and Biophysics, Karolinska Institute, 17165, Solna, Sweden

* Correspondence: joaquim.jaumot@idaea.csic.es

0. Abstract

Pharmaceutical compounds have arisen as one of the main emerging contaminants (ECs) because their high use and release into the environment has considerably increased worldwide. The goal of this study was to assess the effects caused by three widely consumed hepatoxic pharmaceutical compounds: an antibiotic (amoxicillin), an antiepileptic (carbamazepine), and an antidepressant (trazodone), at environmentally relevant concentrations. A combination of an untargeted metabolomic and a targeted sphingolipid analyses were selected to unravel the metabolic alterations in human hepatic cells exposed to these ECs at three concentrations for 24 hours. HepG2 hepatoma cells were encapsulated in sodium alginate spheroids to improve the physiological relevance of this in-vitro approach. Univariate and multivariate statistical methods were employed for discriminating the most affected metabolites and sphingolipids for each drug exposure. Therefore, this study allowed identifying the main metabolic pathways altered by the drug exposure, including glycerophospholipid metabolism, sphingolipid metabolism, alanine aspartate and glutamate metabolism, taurine and hypotaurine metabolism, and lysine degradation.

Graphical abstract



Keywords: HepG2 cells, pharmaceuticals, metabolomics, sphingolipids, amoxicillin, carbamazepine, trazodone.

1. Introduction

Pharmaceuticals are one important group of worldwide emerging contaminants (ECs) due to their increased use and amplified release into the environment in the last decades [1–3]. Although there have been remarkable contributions in novel strategies for their removal from wastewaters [4–6],

there is still a long way before the spread of these compounds and their transformation products is finally contained. In the meantime, these compounds are found in rivers and other water bodies, as well as in sewage sludge or sediments [7]. Furthermore, as humans can end up by being affected by pharmaceutical exposure (through raw and drinking water, for instance) chemical risk assessments are needed [8]. These emerging contaminants can produce addiction, bioaccumulation and antibiotic resistance among other severe consequences [9]. Hence, pharmaceuticals pose a threat to both the aquatic ecosystems and human populations [10]. A considerable effort has been made in developing analytical methods for detecting these compounds in wastewater [11,12].

Besides from detection and quantification of pharmaceutical compounds in the environment and wastewater treatment plants, it is also crucial to study the effects they may have on aquatic organisms and human health. Metabolomic analyses aim to decipher the biological role of metabolites present in cells, tissues and biofluids, ranging from small polar molecules up to lipids; it also looks into the regulation into the metabolic pathways to which they are associated. In particular, sphingolipids play a crucial role in cell metabolism. These lipids are involved in cell signaling and recognition, as well as many biological processes such as growth regulation, cell migration, adhesion, apoptosis, senescence and inflammatory responses [13]. Metabolomics can be applied to assess the effects of environmental stressors, such as ECs, in a more immediate manner than other omics. The reason is that changes in the metabolome are quicker, proving a snapshot of what is happening in the cell. Metabolomics also allows to identify the metabolic pathways affected by the contaminants and discover potential biomarkers of the exposure [14]. Hence, it is a very useful tool for unraveling toxicity and mechanism of action of ECs. Indeed, metabolomics have already been employed in ecotoxicological evaluations of pharmaceutical compounds at environmentally relevant concentrations, for instance, in mussels [15,16], crustaceans [17,18] and fish [19–21].

Following the European legislation (Directive 2010/63/EU) that has promoted the use of non-animal models, *in-vitro* models are a good alternative to replace traditional experimental testing. More specifically, cell lines derived from liver and kidney are commonly selected for toxicity assessments. In this case, the immortal and nontumorigenic human hepatocellular carcinoma (HepG2) cell line has been chosen. This cell line has been used in previous environmental studies for evaluating the effects of other pollutants, such as polycyclic aromatic hydrocarbons (PAHs) [22], flame retardants and dust extracts [23], natural pyrethrins [24], or microplastics [25]. With the aim of improving the physiological relevance of *in-vitro* approaches, 3D liver spheroids have arisen as a more robust liver model [26–28].

Three widely consumed hepatoxic pharmaceutical compounds were selected for their study: the antibiotic amoxicillin (AMOX), the antiepileptic carbamazepine (CBZ), and the antidepressant

trazodone (TRA). The ranges of concentration in which these compounds have been found in Catalonian wastewater treatment plants and senior residence wastewaters are, respectively: 0.68-2.56 and 0-0.5 μg L^{-1} for AMOX, 0.14-0.74 and 0-5.4 μg L^{-1} for CBZ, 0.035-0.46 and 0-314 μg L^{-1} for TRA [29,30]. All three pharmaceutical compounds can cause drug-induced liver injury [31,32].

In this work, we employed a combination of LC-MS platforms for targeted sphingolipid analysis and untargeted metabolomics to assess the effects of AMOX, CBZ, and TRA on HepG2 cell line using 3D spheroids.

2. Materials and methods

2.1 Chemicals and reagents

AMOX, CBZ, and TRA (purity > 98%) were purchased from Merck (Darmstadt, Germany), and their stock solutions were prepared at 100 mM in DMSO.

Dulbecco's Modified Eagle's Medium (DMEM) with Ultraglutamine and fetal bovine serum were supplied by Lonza (Basel, Switzerland). Phosphate-buffered saline (PBS), trypsin, dimethyl sulfoxide (DMSO), resazurin sodium salt, sodium alginate, sodium citrate, calcium chloride ($CaCl_2$), and sodium chloride (NaCl) were purchased from Merck (Darmstadt, Germany). Pierce ™ BCA Protein Assay kit was purchased from Thermo Fisher (Waltham, Massachusetts, USA). CellTiter-Blue® cell viability assay was provided by Promega (Madrid, Spain). For the 3D spheroids protocol, sodium alginate, sodium chloride, and calcium chloride solutions were set at 2.4%, 9% and 101 mM respectively. Lysis solution is composed of sodium citrate at 55 mM and sodium chloride at 150 mM.

LC-MS grade solvents water, acetonitrile, isopropanol, were purchased from Merck (Darmstadt, Germany); LC-MS methanol and 25% NH_4OH solution were acquired from Honeywell Fluka (Seelze, Germany) and ammonium acetate 1M from Fujifilm Wako (Osaka, Japan).

The sphingolipid internal standard mix, prepared in MeOH, contained Cer(d18:1/16:0)-d_7, Cer(d18:1/18:0)-d_7, Cer(d18:1/24:1)-d_7, Cer(d18:1/24:0)-d_7, GlcCer(d18:1/18:0)-d_5, Sphingosine-d_7, Sphinganine-d_7, Sphingosine-1-phosphate-d_7, C15 Ceramide-1-Phosphate-d_7 and SM(d18:1/18:1(9Z))-d_9, all from Avanti Lipids (distributed by Merck KGaA, Darmstadt, Germany), and LacCer(d18:1/16:0)-d_3 and GlcCer(d18:1/16:0)-d_3 from Matreya (distributed by Larodan ,Solna, Sweden). The following lipid abbreviation are used from now on: SM: Sphingomyelin; Cer: Ceramide; DhCer: Dihydroceramide; GlcCer: Glucosylceramide; LacCer:Lactosylceramide; S1P: Sphingosine-1-Phosphate; Spa1P: Sphinganine-1-Phosphate.

Metabolite internal standards piperazine-N,N′-bis-2-ethanesulfonic acid (PIPES), N-Cyclohexyl-2-aminoethanesulfonic acid (CHES) and 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) were purchased from Dojindo (Kumamoto, Japan). A stock solution of 10 mM for the three standards was prepared in water, and it was further diluted in a acetonitrile:water (8:2) prior to its use in the metabolomic analysis.

2.2 Cell culture and viability assay

Human hepatoma cell line (HepG2) from the American Type Culture Collection (ATCC, HB-8065) was cultured in DMEM with Ultraglutamine 1 supplemented with 10% of fetal bovine serum (DMEM10). Cells incubation was performed at 37 °C in a humidified incubator set at 5% CO_2 and passaged every 3-4 days.

The acute toxicity of each pharmaceutical compound in 2D cultures was assessed. Cells were rinsed with PBS, trypsinized, counted and seeded into 96-well plates at a density of 1 x 10^4 cells per well. The next day, cells were exposed to pharmaceutical compounds in a concentration range from 1 to 500 μM. The maximum percentage of DMSO used as vehicle was 0.5% (v/v). Four biological replicates per compound and per concentration level were tested. At 24 h of exposure, the CellTiter-Blue® cell viability assay was performed. The fluorescence at 560/590 nm (excitation/emission wavelengths) of the wells was measured after 4 h of incubation in a microplate well reader (Agilent Cary Eclipse Fluorescence Spectrometer (Agilent Technologies Inc., Santa Clara, CA, USA)). This process was also carried out at 48 h of exposure.

2.3 Exposure conditions and 3D cultures

To ensure working under sublethal doses, the highest exposure concentrations for the 3D cultures with no effects on cell viability were chosen for each drug. These concentration levels were set at 30, 15 and 7.5 μM for CBZ and AMOX, whereas the levels for TRA were 4, 2, and 1 μM.

Final exposures were performed on 3D cultures. In this case, sodium alginate was used for forming the gel in the spheroids [33]. The protocol for generating the 3D cultures started trypzining the 2D cultures and counting the number of cells. The suspension of cells in DMEM10 and a 2.4% solution of sodium alginate were mixed in 1:1 ratio to obtain a spherification mixture at 7 million cells mL^{-1}. The 24 well-plates (one drug) were filled with a 102 mM solution of $CaCl_2$. Alginate cell suspension was then spilled through a 26-gauge needle into the $CaCl_2$ solution, in a ratio of 10 drops per well. After 10 min, the $CaCl_2$ was removed and spheroids were washed with a 0.9% NaCl solution. Again, 10 min later the NaCl was removed and substituted by DMEM10 solution containing the different drugs at the concentration levels previously specified. 6 replicates per dose, including control samples, were analyzed. The spheroids were incubated for 24 h at 37 °C. Then, the drug solutions were removed, spheroids were washed with PBS (1x), and 1 mL of lysis solution was added per well. Next steps were performed under cold conditions (ice

bath). The dissolved spheres were centrifuged (5 min at 1507 gat 4 ºC), pellets were washed with PBS and centrifuged again. Washed cell pellets were frozen at -80 ºC until extraction.

2.4 Extraction protocol

Before the extraction, samples were placed on a tray with ice. Then, 250 µL of LC-MS grade methanol and 10 µL of the sphingolipid internal standard mix were added to each sample. Cells were vortexed, sonicated for 15 minutes and centrifuged at 12000 *g* for 15 minutes at 6 ºC. Extraction solvent and material blanks following the same steps without sample were performed simultaneously. The supernatant of each sample was divided into two vials: 60 µL for sphingolipid analysis (SL), 60 µL for metabolomic analysis (MT). Additionally, a 60 µL aliquot from each extract was pooled to be used as Quality Control (QC pool). Two types of QCs of injection were aliquoted from the QC pool, according to the LC-MS analysis platform (SL or MT). SL samples and their QCs were analyzed directly. A dilution 1:1 was performed by adding a mixture of internal standards for the MT analysis to the metabolomic samples and QCs. Sample extracts were stored at -20 ºC until LC-HRMS analysis.

2.5 LC-MS/MS method for sphingolipid analysis

Targeted sphingolipid analysis was performed on a LC-MS/MS platform [34]. A detailed description of the analytical method employed can be found in **Supplementary Material A Section 1**.

2.6 LC-HRMS method for untargeted metabolomic analysis

The LC-HRMS analysis were performed on an Agilent 1290 Affinity II HPLC system coupled to an Agilent 6550 iFunnel QTOF mass spectrometer equipped with a dual AJS electrospray ionization source used in positive and negative modes. The method employed,proposed by Meister et al. [35] with minor modifications, can be found in **Supplementary Material A Section 1**.

2.7 Data analysis

2.7.1 Data conversion and preprocessing

Targeted sphingolipid analysis

LC-MS/MS raw files were preprocessed with Masslynx and Targetlynx v4.1 (Waters Corporation). A table with the absolute concentration or the ratios of the areas of the compounds and the internal standards were reported

Untargeted metabolomic analysis

LC-HRMS raw files were converted to (.mzML) using ProteoWizard for a first quality check on MZmine 2.53 [36]. Then, the (.mzML) files were converted to “Analysis Base File” (.abf) format with Reifycs Abf Converter. Preprocessing and integration was performed in MS-DIAL 4.20 [37],

using the parameters set by Meister et al. [35], which are also included in **Supplementary Material A Section 2 (Tables S1-S3)**. Annotation was performed based on in-house retention time and MS/MS spectral libraries [38,39].

A table containing the peak areas was exported from MS-DIAL. A batch correction based on the QCs was performed using the MATLAB algorithm proposed by Broadhurst et al. [40].

For the sake of clarity, from now on, the different sets will be referred as SL in the case of sphingolipid analysis, whereas MT pos and MT neg will be designed for metabolite analysis in positive and negative ionization modes, respectively.

2.7.2 Statistical assessment and multivariate analysis

The post-processing analysis was performed on the concentration or ratio matrices in the case of SL platform, or on the obtained areas after QC batch correction for MT platform.

Statistical assessment was carried out using SPSS 27.0.1.0 (©Copyright IBM Corporation). On one side, ANOVA tests were employed for the simultaneous evaluation of all doses (including control samples) per each pharmaceutical compound. On the other side, t-tests were used for studying the significant features between the highest concentration of exposure for each drug and control samples.

Multivariate analysis with the normalized areas were performed in MATLAB environment (Release 2020b, The Mathworks Inc, Natick, MA, US) and PLS Toolbox 8.9.1 (Eigenvector Research Inc, Wenatchee, WA, US). A first exploratory analysis was conducted through principal component analysis (PCA) (Joliffe and Morgan, 1992). PCA analysis was performed on a matrix containing both the concentration or ratio results from the SL platform and the areas values from the MT platform (from both ionization modes). In this case, the biological replicate from each sample class (regarding the administered dose for each drug) should cluster together. Also, a differentiation according to the dose can also be expected.

Partial least square discriminant analysis (PLS-DA) [43] is a supervised multivariate classification method that discriminates between groups of samples. PLS-DA was employed to assess the significant variables when comparing control samples (C) and the highest concentration level (H and M doses). The variables important in projections (VIPs) indicate the significance of the variables, i.e., VIP value > 1 is equivalent to a p-value < 0.05. The models were built using *leave-one-out* as internal cross validation method. PLS-DA analysis was performed on a matrix containing only the variables with a p-value lower than 0.05 in the univariate statistical assessment, including the variables from both platforms (SL and MT, in the two ionization modes). The quality of the binary classifications was assessed with the Matthews Correlation Coefficient (MCC), whose values are comprehended from -1 to 1 [44,45]. A perfect model would be represented with a 1, whereas a wrong prediction model would be assigned to a -1 of MCC,

and therefore, should be discarded. Acceptable prediction models can be considered for MCC values higher than 0.7.

Autoscaling pre-processing was applied in PCA and PLS-DA analysis.

3. Results and discussion

3.1 Preliminary range-finding test for drug exposure

Cell viability results are included in **Figure 1**, expressed as the % of cell viability relative to control cells treated with vehicle. The range of concentrations tested was set from 500 to 1 µM for each drug (AMOX, CBZ, TRA) at 24 and 48 hours of exposure in 2D cultures.

For AMOX and CBZ, the same cell viability as in the controls was observed up to 30 µM at 24 hours of exposure, whereas the viability was reduced to approximately 70% at the maximum concentration tested. TRA was more toxic to cells and doses larger than 30 µM induced a more pronounced cytoxicity, leaving only 20% of alive cells at 500 µM. Exposures longer than 48 h produced higher toxicity in all drugs. Therefore, this study was limited to only 24 h for the three drugs. Hence, the final concentration levels were set at 30, 15 and 7.5 µM for AMOX and CBZ, and 4, 2, and 1 µM for TRA at 24h, henceforth referred as High (H), Medium (M), Low (L), plus control samples (C).

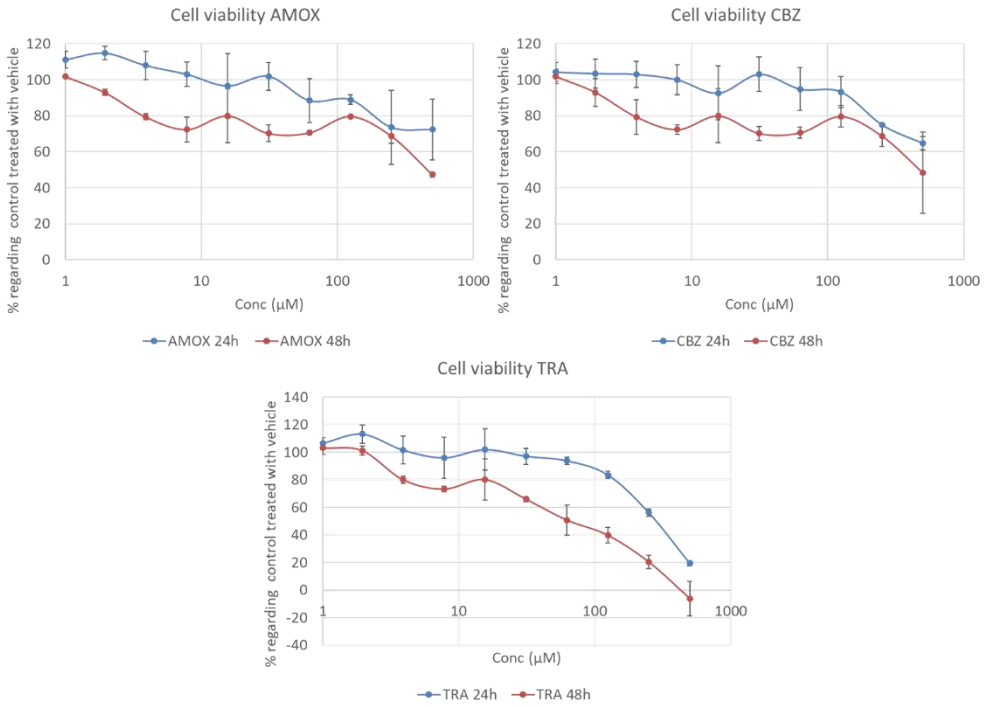


Figure 1. Cell viability assessed using the resazurin assay (expressed as the % respect control cells treated with vehicle) for the three drugs tested: amoxicillin (AMOX), carbamazepine (CBZ) and trazodone (TRA) from 500 µM to 1 µM, and for 24 and 48 h of exposure. The final concentration levels for this study were set at 30, 15 and 7.5 µM for AMOX and CBZ, and 4, 2, and 1 µM for TRA at 24h.

3.2 Exploratory analysis of drug exposures

We used PCA to explore the behavior of the different dosages of drug exposures in an unsupervised manner. PCA analysis was performed on a unique matrix composed by the concentrations or ratios from SL platform and the areas from the MT platform, after an autoscale normalization to compensate the differences in scale between the two platform outputs. The PCA scores plot obtained for each drug are shown in **Figure 2**: **A)** AMOX, **B)** CBZ and **C)** TRA. In all cases, the explained variance regarding the first two principal components (PC1 and PC2, respectively) was larger than 55%. There was no clear cluster separation according to the concentration dosage for AMOX, which seemed to indicate that there was no evident effect in the metabolome or sphingolipidome caused by the selected doses in the hepatic cells at this time of exposure. On the contrary, a trend Control-Low-Medium-High of clustered samples was observed for CBZ, as expected. Besides, for this drug, PC2 separated the two lowest doses (and Control samples) from the highest dose. In the case of TRA, although Control and High dose samples clustered very close (even partially overlapping), there was a strong differentiation between Control and Medium dose samples.

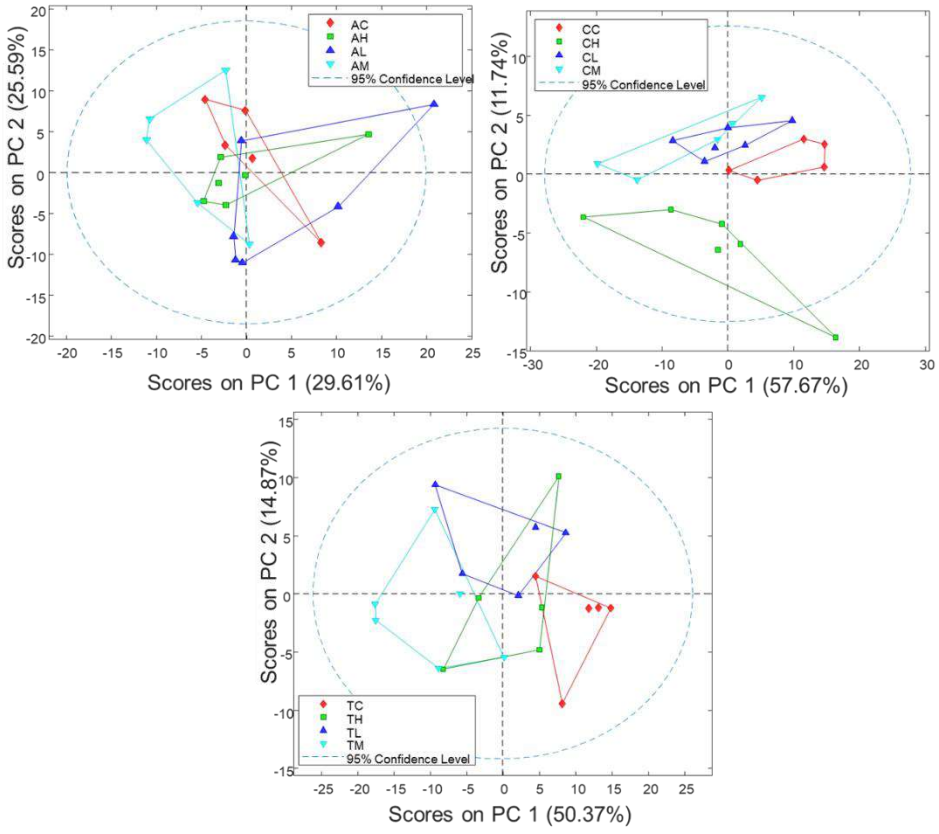


Figure 2. PCA scores plot of the three doses of exposure for each drug: A) amoxicillin, B) carbamazepine, and C) trazodone. No differentiation is observed regarding the dosage for AMOX, but there is a trend from Control to High dose samples for CBZ, and a clear separation of Control and Medium doses for TRA.

3.3 Selection of potential markers of drug exposures

ANOVAs analyses for multiple-level comparisons and t-test analyses for two-level comparisons were performed on the matrices with the concentration or ratio quantification results from the SL platform or the areas from the MT platform in both ionization modes. **Tables SA-SC from Supplementary Material** contain information on the annotated lipids and metabolites (including compound ID, chemical formula, other relevant information such as HMDB or Lipid Maps codes, *m/z* and retention times), as well as the univariate statistical results. A reduced matrix including only the variables from SL and MT analysis with a *p*-value lower than 0.05 were selected for a further assessment with PLS-DA. PLS-DA models Control *versus* High and Control *versus* Medium were built for the three drugs, selected as the most interesting doses comparisons regarding the PCA scores plots. The models obtained for AMOX were discarded, as the MCC values were lower than 0.7. Consequently, this drug was not considered for the rest of the analysis from now on, as no reliable markers of its exposure were obtained. The best model for CBZ was the C vs H combination (MCC=1.0), whereas the best model for TRA was the C vs M (MCC=1.0). A summary of the MCC for all the models tested can be found in **Table S4 from Supplementary Material A**. These results are in agreement with the previously obtained with PCA, where the separation control and highest dose was very clear for CBZ, whereas in the case of TRA, the control and medium dose samples separation was more evident. Hence, the variables (i.e., sphingolipids or metabolites) that presented a VIP value higher than 1.0 with the best PLS-DA models and had a fold-change value lower than 0.9 or higher than 1.1 were selected as potential markers of the exposure of CBZ and TRA, which are included in **Table 1.** The total number of significant compounds for each drug were 18 for CBZ and 28 for TRA. In addition, six compounds that were detected as significantly altered due to the exposure of both drugs: Carnitine, N-acetyl-aspartic acid, Norvaline betaine, Glycerophosphocholine, 2-Aminoadipic acid and Taurine. A graphical representation of fold-change values obtained for the significant features is displayed in **Figure 3**, for each drug exposure: **3.A)** CBZ, **3.B)** TRA. All compounds presented a lower abundance regarding control samples, except taurocholic acid for TRA exposure. The effects of TRA on the sphingolipidome were more severe (i.e., 10 sphingolipids altered *versus* only 3 for CBZ exposure, mainly ceramides and lactosylceramides in both cases).

Table 1. Compounds with VIP > 1.0 in PLS-DA analysis for CBZ and TRA exposures. The common significant compounds from both exposures are marked in bold and italics.

CBZ expsoure C vs H PLS-DA model			TRA expsoure C vs M PLS-DA model		
Platform	**Compound name**	**VIP**	**Platform**	**Compound name**	**VIP**
MT_pos	***Carnitine***	1.46	MT_neg	***Glycerophosphocholine***	1.48
SL	Cer(d18:0/16:0)	1.43	MT_pos	Ethanolamine	1.27
MT_pos	Iminodiacetic acid	1.42	MT_neg	Glycerophosphorylethanolamine	1.22
MT_neg	***2-Aminoadipic acid***	1.35	SL	Glucosyl sphingosine	1.18
MT_neg	Benzoic acid	1.33	MT_neg	4-Hydroxyphenyllactic acid	1.18
MT_pos	***Norvaline betaine***	1.32	MT_pos	Pyridoxine	1.17
MT_neg	Aspartic acid	1.30	MT_pos	Taurocholic acid	1.15
MT_neg	***Taurine***	1.28	MT_neg	N-Acetylthreonine	1.15
MT_pos	Pro-Gly	1.27	MT_neg	Fructoseglycine	1.15
MT_pos	***Glycerophosphocholine***	1.25	SL	LacCer(d18:1/18:1)	1.14
MT_pos	CDP-Choline	1.12	MT_pos	Valeryl-carnitine	1.12
SL	CerP(d18:1/24:0)	1.09	SL	Cer(d18:1/24:1)	1.11
MT_pos	Creatinine	1.08	MT_neg	N-Acetylaspartylglutamic acid	1.09
MT_neg	Pyroglutamic acid	1.07	SL	Cer(d16:1/24:1)	1.07
MT_neg	***N-Acetylaspartic acid***	1.05	SL	Cer(d18:1/22:0)	1.07
SL	LacCer(d18:1/17:0)	1.04	MT_neg	***N-Acetylaspartic acid***	1.06
MT_pos	Phosphorylcholine	1.04	SL	LacCer(d18:1/18:0)	1.06
MT_pos	Butyryl-carnitine	1.02	MT_neg	Glucose 6-sulfate	1.06
			MT_pos	***Carnitine***	1.04
			MT_pos	***Norvaline betaine***	1.04
			MT_neg	Creatine	1.04
			MT_neg	N-Acetylglutamine	1.03
			SL	Cer(d18:1/26:1)	1.03
			MT_neg	***Taurine***	1.03
			MT_pos	***2-Aminoadipic acid***	1.02
			SL	Cer(d18:0/18:0)	1.02
			SL	GlcCer(d18:1/18:0)	1.02
			SL	LacCer(d18:1/20:0)	1.00

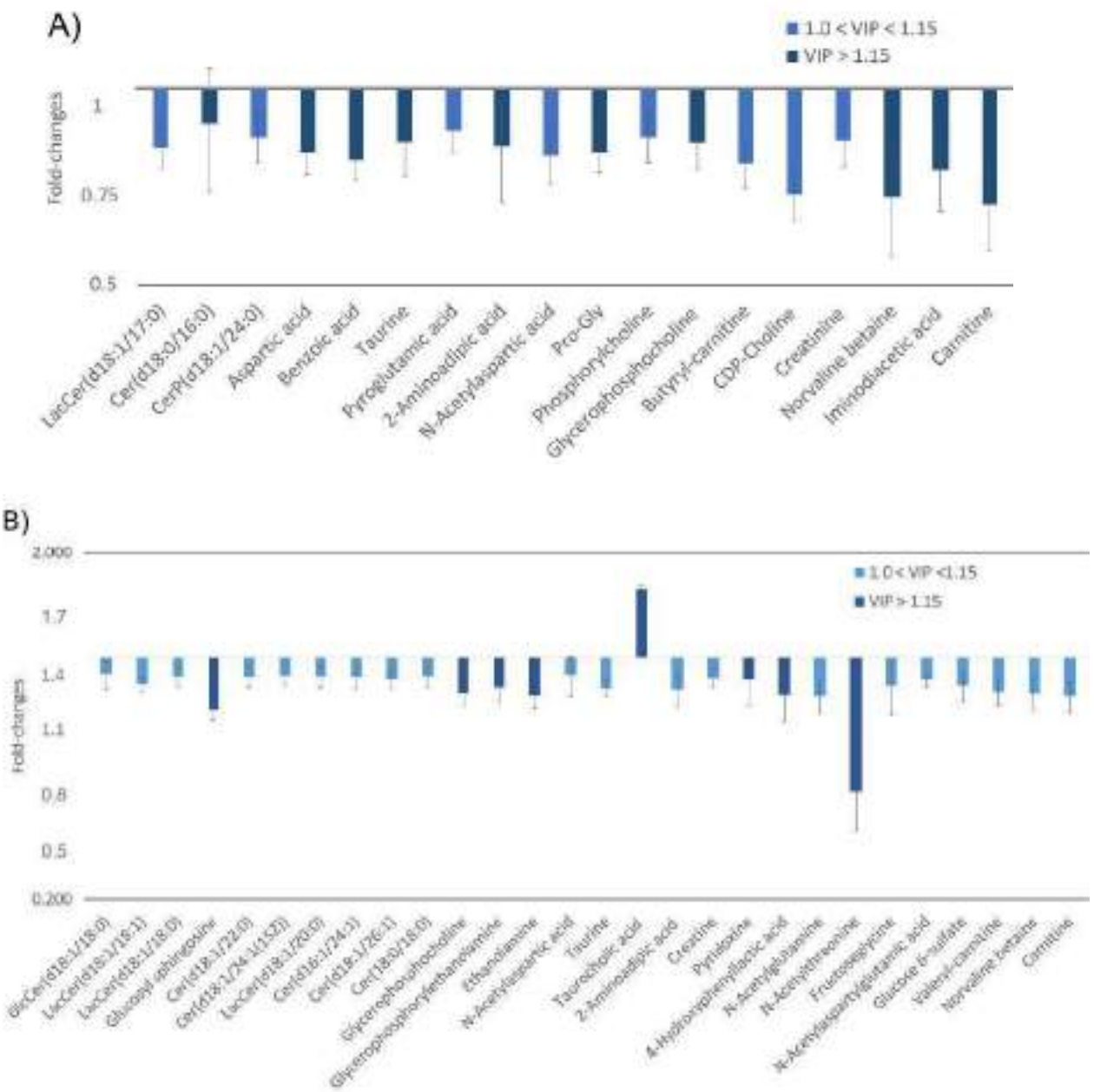


Figure 3. Graphical display of fold-changes values for the significant compounds associates with A) CBZ C vs H ezposure and B) TRA C vs M exposure, incluidng VIP values in a colored scale (higher VIPs present a darker tone of blue). The X axis is expressed in a logaritmic scale.

3.4 Discussion of the sphingolipids and metabolites altered by the drug exposures

A pathway analysis was then carried out using MetaboAnalyst [46], to evaluate the principal metabolic pathway affected by these drug exposures. The metabolic pathway analyses from Metaboanalyst for CBZ and TRA are included in **Tables S5 and S6 from Supplementary Material A**, respectively. **Figure 4** shows the pathway analysis, for each drug exposure. CBZ

exposure in this study affected mainly to the following pathways (see **Figure 4.A**): glycerophospholipid metabolism (**1**), sphingolipid metabolism (**2**), alanine aspartate and glutamate metabolism (**3**), taurine and hypotaurine metabolism (**4**), and lysine degradation (**5**). TRA exposure altered the same metabolic pathways than CBZ, but also some others (see **Figure 4.B**), such as vitamin B6 metabolism (**6**), ether lipid metabolism (**7**), primary bile acid biosynthesis (**8**), and glycine serine and threonine metabolism (**9**).

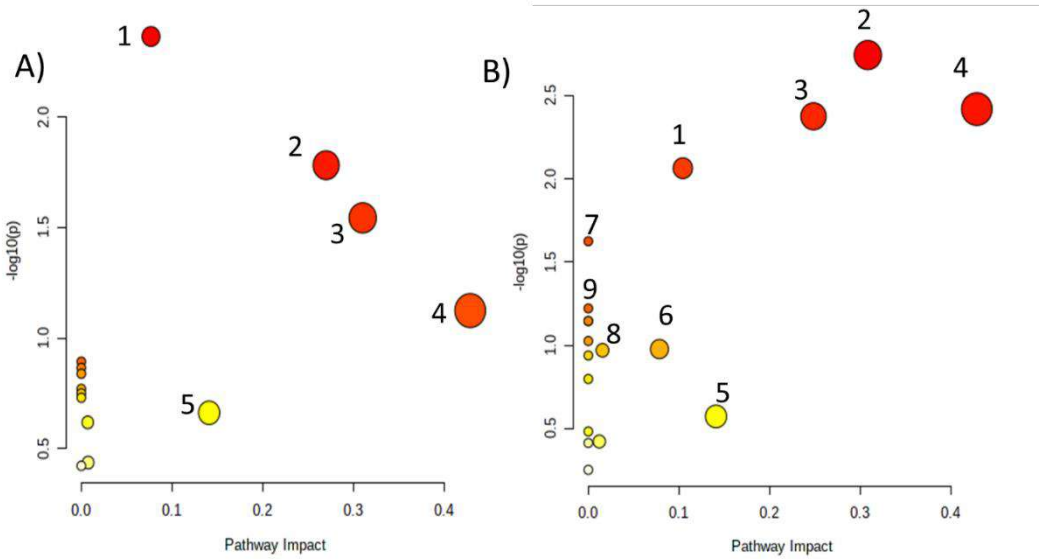


Figure 3. Pathway analysis obtained with Metaboanalyst of the significant compounds related to the exposure to: **A)** CBZ and **B)** TRA. The numeric code corresponds to the following metabolic pathways: **1)** Glycerophospholipid metabolism, **2)** Sphingolipid metabolism, **3)** Alanine, aspartate and glutamate metabolism, **4)** Taurine and hypotaurine metabolism, **5)** Lysine degradation, **6)** Vitamin B6 metabolism, **7)** Ether lipid metabolism, **8)** Primary bile acid biosynthesis, **9)** Glycine serine and threonine metabolism.

3.5 Biological insights of the drug exposures

Trazadone and carbamazepine exposures resulted in small but significant fold changes in sphingolipid and metabolite concentrations.

Regarding the effects of trazodone on sphingolipids, the levels of ceramides of long fatty acyl chain (22:0, 24:1 and 26:1) were notably reduced (0.8 fold). Ceramides can be produced by cell membrane sphingomyelin hydrolysis through the action of acid sphingomyelinase activity, and can also be *de novo* synthesized through different biosynthetic steps from L-serine and palmitoyl-CoA [47]. In relationship with the first option, previous reports have demonstrated that antidepressant drugs reduce acid sphingomyelinase activity and the release of ceramides from sphingomyelin in the hippocampus [48,49]. These studies demonstrated that this acid sphingomyelinase/ceramide system mediates the effects of antidepressants on neuronal

proliferation, maturation, and survival. The possibility that trazodone also affects the biosynthesis *de novo* is supported by the reduced levels of dihydroceramide 18:0 (0.8 fold), a direct precursor of ceramide 18:0. Lower levels of ceramide could explain the decreased levels of their glucosylated products glucosylceramide and lactosylceramide species (0.8 fold). These are important components of cell membrane and participate as signalling molecules in cellular processes such apoptosis. Also, they are the precursors of more complex glycosphingolipids, known as globosides and gangliosides, which have essential roles in cell membrane in the mediation of cell-cell interactions and regulation of membrane proteins activity [50]. The decrease of important chemical structures that belong to glycerophospholipid metabolism (0.7-0.8 fold), such as ethanolamine, glycerophosphorylethanolamine and glycerophosphocholine suggests changes in the levels of this lipid subfamily, of which most of their members have essential functions in cell membrane. This, together with the changes in ceramide glucosylation indicate that the structure and functions of plasmatic membrane could be affected by trazodone exposure.

In the metabolite analysis, one of the main changes observed under trazodone is the increase in the conjugation of cholic acid with taurine, as reflected by the rise of taurocholic acid (1.5 fold) and the decrease of taurine (0.8 fold). The biosynthesis of bile acid conjugates is an exclusive function of hepatocytes, which is catalyzed by the enzyme bile acid coenzyme A: amino acid *N*-acyltransferase (BAAT) [51]. Bile acids are required for the absorption of digested lipids and fat-soluble vitamins in the intestine, and can also affect the proportions of the different species of intestinal bacteria. The conjugation of bile acids with taurine or glycine allows bile acids in the intestine to form micelles, which are necessary for the absorption of lipids [52]. In this case, higher levels of taurocholic acid suggest an increase in BAAT activity as a result of trazadone exposure, which in the context of human chronic exposure to trazadone could have an effect on the normal absorption of lipids in the intestine and on the composition of gut microbiome.

L-carnitine and valeryl-carnitine levels were also reduced under trazodone exposure (0.7 fold). L- carnitine plays a key role in lipid metabolism since its function is the transport of long-chain fatty acids across the inner mitochondrial membrane for β- oxidation and generation of ATP energy. Carnitine is a non-essential amino acid that is mainly produced in the liver and kidneys. Therefore, a reduction of carnitine levels in the hepatocytes due to trazodone exposure could have an effect on the levels of this amino acid in the whole organism, and have consequences on the energy production rates through lipid beta-oxidation [53].

Another interesting observation is the decrease of N-acetylation of amino acids, such as N-acetylglutamine, N-acetylaspartic and N-acetylthreonine (0.7, 0.8, and 0.4 fold, respectively). N-acetylation of free amino acids can occur by the action of N-acetyltransferases, but these species can be also generated via the proteolytic degradation of N-acetylated proteins by specific hydrolases. The N-terminal acetylation of proteins by N-acetyl transferases has important

biological functions in cells, such as targeting proteins for degradation, proper folding of proteins, protein-protein interactions, or targeting proteins to membranes. Our results suggest the possibility that the activity of N-acetyl transferases under trazadone exposure might be altered, which could have an impact on the function of some proteins [54].

Concerning the exposure to carbamazepine, some similarities to trazadone treatment have been found. Hence, the glycerophospholipid metabolism resulted altered, as shown by the reduced levels of CDP-choline, glycerophosphocholine, and phosphocholine (0.7, 0.8 and 0.8 fold, respectively). Regarding the sphingolipid family, only DhCer 16:0 and LacCer 17:0 species were found decreased (0.8 fold), but no changes were detected on ceramide levels, as observed with trazadone.

Similar to trazadone, the levels of L-carnitine and another short chain acyl carnitine (in this case butyl-carnitine) were found reduced (0.6 and 0.7 fold, respectively), which may indicate that the exposure to this anti-epileptic drug could also cause alterations in beta-oxidation and energy production. This agreed with previous clinical reports in which a decrease of serum carnitine levels in epileptic children was found under carbamazepine therapy [55,56]. The decrease of taurine levels was also detected (0.8 fold), although in this case it was not accompanied with a rise of taurocholic acid levels, suggesting a potential effect of carbamazepine on its enzymatic biosynthesis from cysteine. Although in a very different context, the decrease of taurine levels had also been observed in rat hippocampus under carbamazepine exposure [57].

The decrease of aspartic acid and N-acetylaspartate (NAA) observed (0.8 fold) might have consequences on the cellular energy production. In the brain, NAA is considered an important energy metabolite for lipid synthesis. However, its role in peripheral tissues is not well known. In a recent work, NAA has been described as an important energy metabolite for the regulation of whole-body energy homeostasis [58]. In this study, carried out in brown adipocytes, the genetic disruption of NAA pathway resulted in reduced cytosolic acetyl-CoA levels and lipid synthesis. Also, NAA reduced the glucose incorporation into acyl glycerol species. If these mechanisms were similar in hepatocytes, this could be related to the alteration of the glycerophospholipid metabolism observed under carbamazepine exposure.

Globally, the fold changes in metabolites and lipids observed in the present study are very slight. This most likely due to the low doses of drugs applied, which tried to mimic realistic environmental exposures. However, these differences are statistically significant, and small changes may be able to destabilize cell homeostasis. The common features observed under both exposures affect lipid composition (changes in sphingolipids and glycerophospholipids precursors) and energy production through beta-oxidation (decrease of carnitine levels), which can affect cellular functions and energy production not only in hepatic cells, but also in other

organs. The fact that trazodone affects sphingomyelinase activity, which is considered a mechanism of antidepressant drugs, and that carbamazepine alters the levels of N-acetylaspartate, an important energy metabolite in brain, indicates that, at the concentrations applied, the changes previously described could have important consequences in neural cells.

4. Conclusions

Cell viability of HepG2 cells was assessed for the exposure of the three drugs (amoxicillin, carbamazepine and trazodone) ranging from 500 µM to 1µM at 24 and 48 hours. The same cell viability as in the controls was observed at 24 hours of exposure up to 30 µM for amoxicillin and carbamazepine, and to 4 µM for trazodone. No clear differentiation between the doses tested was observed in the metabolomic and sphingolipid analyses performed on amoxicillin. On the contrary, alterations in the following metabolic pathways were observed for carbamazepine and trazodone: glycerophospholipid metabolism, sphingolipid metabolism, alanine aspartate and glutamate metabolism, taurine and hypotaurine metabolism, and lysine degradation. Although administrated at lower doses, trazodone seems to affect other specific pathways as well (i.e., vitamin B6 metabolism, ether lipid metabolism, primary bile acid biosynthesis, and glycine serine and threonine metabolism). In addition, trazodone exposure produced significant changes in ceramides and neutral glycosphingolipids.

Regarding the specific effects both drugs provoked on the HepG2 cells, trazadone seems to have a negative impact on the structure and functions of plasmatic membrane, on the normal absorption of lipids in the intestine and on the composition of gut microbiome, and on energy production rates through lipid beta-oxidation. On the other hand, carbamazepine exposure seems also linked to alterations in beta-oxidation and cellular energy production. All in all, in spite of the low doses employed in this study, small but significant changes were observed at molecular level. Both drugs exposure may be related to alterations in cellular functions and energy production not only in hepatic cells, but also in neural cells, or even in other organs.

Acknowledgments

The research leading to these results has received funding from the Spanish Ministry of Science and Innovation (MCI, Grants CTQ2017-82598-P). The authors also want to grant support from the Catalan Agency for Management of University and Research Grants (AGAUR, Grant 2017SGR753) and the Spanish MCI (Severo Ochoa Project CEX2018-000794-S). MPC acknowledges a predoctoral FPU 16/02640 scholarship from the Spanish Ministry of Education and Vocational Training (MEFP), and Post-graduate department from CSIC for the funding of the research stay in Karolinska Institute, *via* the award to best outreach video in the YoInvestigoYosoyCsic contest, 2019 edition.

Conflict of interests: The authors declare no conflict of interest.

References

[1] H. Ramírez-Malule, D.H. Quiñones-Murillo, D. Manotas-Duque, Emerging contaminants as global environmental hazards. A bibliometric analysis, Emerging Contaminants. 6 (2020) 179–193. https://doi.org/10.1016/j.emcon.2020.05.001.

[2] K. Świacka, J. Maculewicz, D. Kowalska, M. Caban, K. Smolarz, J. Świeżak, Presence of pharmaceuticals and their metabolites in wild-living aquatic organisms – Current state of knowledge, Journal of Hazardous Materials. 424 (2022). https://doi.org/10.1016/j.jhazmat.2021.127350.

[3] P. Chaturvedi, P. Shukla, B.S. Giri, P. Chowdhary, R. Chandra, P. Gupta, A. Pandey, Prevalence and hazardous impact of pharmaceutical and personal care products and antibiotics in environment: A review on emerging contaminants, Environmental Research. 194 (2021) 110664. https://doi.org/10.1016/j.envres.2020.110664.

[4] R.B. González-González, A. Sharma, R. Parra-Saldívar, R.A. Ramirez-Mendoza, M. Bilal, H.M.N. Iqbal, Decontamination of emerging pharmaceutical pollutants using carbon-dots as robust materials, Journal of Hazardous Materials. 423 (2022). https://doi.org/10.1016/j.jhazmat.2021.127145.

[5] R. Ricky, S. Shanthakumar, Phycoremediation integrated approach for the removal of pharmaceuticals and personal care products from wastewater – A review, Journal of Environmental Management. 302 (2022) 113998. https://doi.org/10.1016/j.jenvman.2021.113998.

[6] N. Koch, N.F. Islam, S. Sonowal, R. Prasad, H. Sarma, Environmental antibiotics and resistance genes as emerging contaminants: Methods of detection and bioremediation, Current Research in Microbial Sciences. 2 (2021) 100027. https://doi.org/10.1016/j.crmicr.2021.100027.

[7] J.P. Fernandes, C.M.R. Almeida, M.A. Salgado, M.F. Carvalho, A.P. Mucha, Pharmaceutical compounds in aquatic environments— occurrence, fate and bioremediation prospective, Toxics. 9 (2021) 1–26. https://doi.org/10.3390/toxics9100257.

[8] S.F. de Aquino, E.M.F. Brandt, S.E.C. Bottrel, F.B.R. Gomes, S. de Q. Silva, Occurrence of pharmaceuticals and endocrine disrupting compounds in brazilian water and the risks they may represent to human health, International Journal of Environmental Research and Public Health. 18 (2021) 1–27. https://doi.org/10.3390/ijerph182211765.

[9] A. Shraim, A. Diab, A. Alsuhaimi, E. Niazy, M. Metwally, M. Amad, S. Sioud, A. Dawoud, Analysis of some pharmaceuticals in municipal wastewater of Almadinah Almunawarah, Arabian Journal of Chemistry. 10 (2017) S719–S729. https://doi.org/10.1016/j.arabjc.2012.11.014.

[10] M. Carere, A. Antoccia, A. Buschini, G. Frenzilli, F. Marcon, C. Andreoli, G. Gorbi, A. Suppa, S. Montalbano, V. Prota, F. de Battistis, P. Guidi, M. Bernardeschi, M. Palumbo, V. Scarcelli, M. Colasanti, V. D'Ezio, T. Persichini, M. Scalici, A. Sgura, F. Spani, I. Udroiu, M. Valenzuela, I. Lacchetti, K. di Domenico, W. Cristiano, V. Marra, A.M. Ingelido, N. Iacovella, E. de Felip, R. Massei, L. Mancini, An integrated approach for chemical water quality assessment of an urban river stretch through Effect-Based Methods and emerging pollutants analysis with a focus on genotoxicity, Journal of Environmental Management. 300 (2021) 113549. https://doi.org/10.1016/j.jenvman.2021.113549.

[11] C.A. Marasco Júnior, D.M. Sartore, R.S. Lamarca, B.F. da Silva, Á.J. Santos-Neto, P.C.F. de Lima Gomes, On-line solid-phase extraction of pharmaceutical compounds from wastewater treatment plant samples using restricted access media in column-switching liquid chromatography-tandem mass spectrometry, Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences. 1180 (2021). https://doi.org/10.1016/j.jchromb.2021.122896.

[12] A.B. Martínez-Piernas, P. Plaza-Bolaños, A. Gilabert, A. Agüera, Application of a fast and sensitive method for the determination of contaminants of emerging concern in wastewater using a quick, easy, cheap, effective, rugged and safe-based extraction and liquid chromatography coupled to mass spectrometry, Journal of Chromatography A. 1653 (2021). https://doi.org/10.1016/j.chroma.2021.462396.

[13] Y.A. Hannun, L.M. Obeid, Sphingolipids and their metabolism in physiology and disease, Nat Rev Mol Cell Biol. 19 (2018) 175. https://doi.org/10.1038/NRM.2017.107.

[14] L.J. Zhang, L. Qian, L.Y. Ding, L. Wang, M.H. Wong, H.C. Tao, Ecological and toxicological assessments of anthropogenic contaminants based on environmental metabolomics, Environmental Science and Ecotechnology. 5 (2021) 100081. https://doi.org/10.1016/j.ese.2021.100081.

[15] T. Dumas, F. Courant, C. Almunia, J. Boccard, D. Rosain, G. Duporté, J. Armengaud, H. Fenet, E. Gomez, An integrated metabolomics and proteogenomics approach reveals molecular alterations following carbamazepine exposure in the male mussel Mytilus galloprovincialis, Chemosphere. 286 (2022). https://doi.org/10.1016/j.chemosphere.2021.131793.

[16] B. Bonnefille, E. Gomez, M. Alali, D. Rosain, H. Fenet, F. Courant, Metabolomics assessment of the effects of diclofenac exposure on Mytilus galloprovincialis: Potential effects on osmoregulation and reproduction, Science of the Total Environment. 613–614 (2018) 611–618. https://doi.org/10.1016/j.scitotenv.2017.09.146.

[17] I. Fuertes, B. Piña, C. Barata, Changes in lipid profiles in Daphnia magna individuals exposed to low environmental levels of neuroactive pharmaceuticals, Science of the Total Environment. 733 (2020) 139029. https://doi.org/10.1016/j.scitotenv.2020.139029.

[18] M.N. Sheikholeslami, C. Gómez-Canela, L.P. Barron, C. Barata, M. Vosough, R. Tauler, Untargeted metabolomics changes on Gammarus pulex induced by propranolol, triclosan, and nimesulide pharmaceutical drugs, Chemosphere. 260 (2020). https://doi.org/10.1016/j.chemosphere.2020.127479.

[19] H. Ziarrusta, A. Ribbenstedt, L. Mijangos, S. Picart-Armada, A. Perera-Lluna, A. Prieto, U. Izagirre, J.P. Benskin, M. Olivares, O. Zuloaga, N. Etxebarria, Amitriptyline at an Environmentally Relevant Concentration Alters the Profile of Metabolites Beyond Monoamines in Gilt-Head Bream, Environmental Toxicology and Chemistry. 38 (2019) 965–977. https://doi.org/10.1002/ETC.4381.

[20] E.J. Ussery, K.M. Nielsen, D. Simmons, Z. Pandelides, C. Mansfield, D. Holdway, An 'omics approach to investigate the growth effects of environmentally relevant concentrations of guanylurea exposure on Japanese medaka (Oryzias latipes), Aquatic Toxicology. 232 (2021) 105761. https://doi.org/10.1016/J.AQUATOX.2021.105761.

[21] Y. Song, T. Chai, Z. Yin, X. Zhang, W. Zhang, Y. Qian, J. Qiu, Stereoselective effects of ibuprofen in adult zebrafish (Danio rerio) using UPLC-TOF/MS-based metabolomics, Environmental Pollution. 241 (2018) 730–739. https://doi.org/10.1016/j.envpol.2018.06.009.

[22] Å. Mattsson, S. Lundstedt, U. Stenius, Exposure of HepG2 cells to low levels of PAH-containing extracts from contaminated soils results in unpredictable genotoxic stress responses, Environ Mol Mutagen. 50 (2009) 337–348. https://doi.org/10.1002/EM.20486.

[23] J. Zhang, M. Abou-Elwafa Abdallah, T.D. Williams, S. Harrad, J.K. Chipman, M.R. Viant, Gene expression and metabolic responses of HepG2/C3A cells exposed to flame retardants and dust extracts at concentrations relevant to indoor environmental exposures, Chemosphere. 144 (2016) 1996–2003. https://doi.org/10.1016/J.CHEMOSPHERE.2015.10.014.

[24] J. Lu, Y. Yang, L. Zhu, M. Li, W. Xu, C. Zhang, J. Cheng, L. Tao, Z. Li, Y. Zhang, Exposure to environmental concentrations of natural pyrethrins induces hepatotoxicity: Assessment in HepG2 cell lines and zebrafish models, Chemosphere. 288 (2022) 132565. https://doi.org/10.1016/J.CHEMOSPHERE.2021.132565.

[25] A. Menéndez-Pedriza, J. Jaumot, C. Bedia, Lipidomic analysis of single and combined effects of polyethylene microplastics and polychlorinated biphenyls on human hepatoma cells, Journal of Hazardous Materials. 421 (2022) 0–3. https://doi.org/10.1016/j.jhazmat.2021.126777.

[26] F. Li, L. Cao, S. Parikh, R. Zuo, Three-Dimensional Spheroids With Primary Human Liver Cells and Differential Roles of Kupffer Cells in Drug-Induced Liver Injury, Journal of Pharmaceutical Sciences. 109 (2020) 1912–1923. https://doi.org/10.1016/j.xphs.2020.02.021.

[27] N.M. Tran, M. Dufresne, G. Duverlie, S. Castelain, C. Défarge, P. Paullier, C. Legallais, An appropriate selection of a 3d alginate culture model for hepatic Huh-7 cell line encapsulation intended for viral studies, Tissue Engineering - Part A. 19 (2013) 103–113. https://doi.org/10.1089/TEN.TEA.2012.0139/ASSET/IMAGES/LARGE/FIGURE7.JPEG.

[28] N. Chaicharoenaudomrung, P. Kunhorm, P. Noisa, Three-dimensional cell culture systems as an in vitro platform for cancer and stem cell modeling, World J Stem Cells. 11 (2019) 1065–1083. https://doi.org/10.4252/wjsc.v11.i12.1065.

[29] C. Gómez-Canela, T. Sala-Comorera, V. Pueyo, C. Barata, S. Lacorte, Analysis of 44 pharmaceuticals consumed by elderly using liquid chromatography coupled to tandem mass spectrometry, Journal of Pharmaceutical and Biomedical Analysis. 168 (2019) 55–63. https://doi.org/10.1016/j.jpba.2019.02.016.

[30] C. Gómez-Canela, S. Edo, N. Rodríguez, G. Gotor, S. Lacorte, Comprehensive Characterization of 76 Pharmaceuticals and Metabolites in Wastewater by LC-MS/MS, Chemosensors 2021, Vol. 9, Page 273. 9 (2021) 273. https://doi.org/10.3390/CHEMOSENSORS9100273.

[31] C.S. Voican, E. Corruble, S. Naveau, G. Perlemuter, Antidepressant-induced liver injury: A review for clinicians, American Journal of Psychiatry. 171 (2014) 404–415. https://doi.org/10.1176/appi.ajp.2013.13050709.

[32] S. Higuchi, A. Yano, S. Takai, K. Tsuneyama, T. Fukami, M. Nakajima, T. Yokoi, Metabolic activation and inflammation reactions involved in carbamazepine-induced liver injury, Toxicological Sciences. 130 (2012) 4–16. https://doi.org/10.1093/toxsci/kfs222.

[33] T. Andersen, P. Auk-Emblem, M. Dornish, 3D Cell Culture in Alginate Hydrogels, Microarrays 2015, Vol. 4, Pages 133-161. 4 (2015) 133–161. https://doi.org/10.3390/MICROARRAYS4020133.

[34] N. Akawi, A. Checa, A.S. Antonopoulos, I. Akoumianakis, E. Daskalaki, C.P. Kotanidis, H. Kondo, K. Lee, D. Yesilyurt, I. Badi, M. Polkinghorne, N. Akbar, J. Lundgren, S. Chuaiphichai, R. Choudhury, S. Neubauer, K.M. Channon, S.S. Torekov, C.E. Wheelock, C. Antoniades, Fat-Secreted Ceramides Regulate Vascular Redox State and Influence Outcomes in Patients With Cardiovascular Disease, J Am Coll Cardiol. 77 (2021) 2494–2513. https://doi.org/10.1016/j.jacc.2021.03.314.

[35] I. Meister, P. Zhang, A. Sinha, C.M. Sköld, Å.M. Wheelock, T. Izumi, R. Chaleckis, C.E. Wheelock, High-Precision Automated Workflow for Urinary Untargeted Metabolomic Epidemiology, Cite This: Anal. Chem. 93 (2021) 5258. https://doi.org/10.1021/acs.analchem.1c00203.

[36] T. Pluskal, S. Castillo, A. Villar-Briones, M. Orešič, MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data, BMC Bioinformatics. 11 (2010). https://doi.org/10.1186/1471-2105-11-395.

[37] H. Tsugawa, T. Cajka, T. Kind, Y. Ma, B. Higgins, K. Ikeda, M. Kanazawa, J. Vandergheynst, O. Fiehn, M. Arita, MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis, Nature Methods 2015 12:6. 12 (2015) 523–526. https://doi.org/10.1038/nmeth.3393.

[38] I. Tada, H. Tsugawa, I. Meister, P. Zhang, R. Shu, R. Katsumi, C.E. Wheelock, M. Arita, R. Chaleckis, Creating a Reliable Mass Spectral-Retention Time Library for All Ion Fragmentation-Based Metabolomics, Metabolites. 9 (2019). https://doi.org/10.3390/METABO9110251.

[39] S. Naz, H. Gallart-Ayala, S.N. Reinke, C. Mathon, R. Blankley, R. Chaleckis, C.E. Wheelock, Development of a Liquid Chromatography-High Resolution Mass Spectrometry Metabolomics Method with High Specificity for Metabolite Identification Using All Ion Fragmentation Acquisition, Analytical Chemistry. 89 (2017) 7933–7942. https://doi.org/10.1021/acs.analchem.7b00925.

[40] D. Broadhurst, R. Goodacre, S.N. Reinke, J. Kuligowski, I.D. Wilson, M.R. Lewis, W.B. Dunn, Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies, Metabolomics. 14 (2018) 1–17. https://doi.org/10.1007/s11306-018-1367-3.

[41] T. Cserháti, Data evaluation in chromatography by principal component analysis, Biomedical Chromatography. 24 (2010) 20–28. https://doi.org/10.1002/bmc.1294.

[42] I.T. Joliffe, B. Morgan, Principal component analysis and exploratory factor analysis, Statistical Methods in Medical Research. 1 (1992) 69–95. https://doi.org/10.1177/096228029200100105.

[43] M. Barker, W. Rayens, Partial least squares for discrimination, Journal of Chemometrics. 17 (2003) 166–173. https://doi.org/10.1002/cem.785.

[44] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genomics. 21 (2020) 1–13. https://doi.org/10.1186/s12864-019-6413-7.

[45] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, BBA - Protein Structure. 405 (1975) 442–451. https://doi.org/10.1016/0005-2795(75)90109-9.

[46] J. Xia, I. v. Sinelnikov, B. Han, D.S. Wishart, MetaboAnalyst 3.0-making metabolomics more meaningful, Nucleic Acids Research. 43 (2015) W251–W257. https://doi.org/10.1093/nar/gkv380.

[47] Y.A. Hannun, L.M. Obeid, The Ceramide-centric Universe of Lipid-mediated Cell Regulation: Stress Encounters of the Lipid Kind *, Journal of Biological Chemistry. 277 (2002) 25847–25850. https://doi.org/10.1074/JBC.R200008200.

[48] E. Gulbins, M. Palmada, M. Reichel, A. Lüth, C. Böhmer, D. Amato, C.P. Müller, C.H. Tischbirek, T.W. Groemer, G. Tabatabai, K.A. Becker, P. Tripal, S. Staedtler, T.F. Ackermann, J. van Brederode, C. Alzheimer, M. Weller, U.E. Lang, B. Kleuser, H. Grassmé, J. Kornhuber, Acid sphingomyelinase–ceramide system mediates effects of antidepressant drugs, Nature Medicine 2013 19:7. 19 (2013) 934–938. https://doi.org/10.1038/nm.3214.

[49] H. Grassmé, P.L. Jernigan, R.S. Hoehn, B. Wilker, M. Soddemann, M.J. Edwards, C.P. Müller, J. Kornhuber, E. Gulbins, Inhibition of Acid Sphingomyelinase by Antidepressants Counteracts Stress-Induced Activation of P38-Kinase in Major Depression, Neurosignals. 23 (2015) 84–92. https://doi.org/10.1159/000442606.

[50] C.-H. Kim, Glycosphingolipids (GSLs), Glycosphingolipids Signaling. (2020) 1–8. https://doi.org/10.1007/978-981-15-5807-8_1.

[51] A. Pellicoro, F.A.J. van den Heuvel, M. Geuken, H. Moshage, P.L.M. Jansen, K.N. Faber, Human and rat bile acid–CoA:amino acid N-acyltransferase are liver-specific peroxisomal enzymes: Implications for intracellular bile salt transport, Hepatology. 45 (2007) 340–348. https://doi.org/10.1002/HEP.21528.

[52] A.F. Hofmann, L.R. Hagey, Key discoveries in bile acid chemistry and biology and their clinical applications: history of the last eight decades, J Lipid Res. 55 (2014) 1553–1595. https://doi.org/10.1194/JLR.R049437.

[53] M.A. Virmani, M. Cirulli, The Role of l-Carnitine in Mitochondria, Prevention of Metabolic Inflexibility and Disease Initiation, International Journal of Molecular Sciences 2022, Vol. 23, Page 2717. 23 (2022) 2717. https://doi.org/10.3390/IJMS23052717.

[54] R. Ree, S. Varland, T. Arnesen, Spotlight on protein N-terminal acetylation, Experimental & Molecular Medicine 2018 50:7. 50 (2018) 1–13. https://doi.org/10.1038/s12276-018-0116-z.

[55] G. Hug, C.A. McGraw, S.R. Bates, E.A. Landrigan, Reduction of serum carnitine concentrations during anticonvulsant therapy with phenobarbital, valproic acid, phenytoin, and carbamazepine in children, The Journal of Pediatrics. 119 (1991) 799–802. https://doi.org/10.1016/S0022-3476(05)80306-3.

[56] M. Castro-Gago, J. Eirís-Puñal, M.I. Novo-Rodríguez, J. Couceiro, F. Camiña, S. Rodríguez-Segade, Serum carnitine levels in epileptic children before and during treatment with valproic acid, carbamazepine, and phenobarbital, Journal of Child Neurology. 13 (1998) 546–549. https://doi.org/10.1177/088307389801301104.

[57] S. Ahmad, L.J. Fowler, P.S. Whitton, Lamotrigine, carbamazepine and phenytoin differentially alter extracellular levels of 5-hydroxytryptamine, dopamine and amino acids, Epilepsy Research. 63 (2005) 141–149. https://doi.org/10.1016/J.EPLEPSYRES.2005.02.002.

[58] D.C. Hofer, G. Zirkovits, H.J. Pelzmann, K. Huber, A.R. Pessentheiner, W. Xia, K. Uno, T. Miyazaki, K. Kon, H. Tsuneki, T. Pendl, W. al Zoughbi, C.T. Madreiter-Sokolowski, G. Trausinger, M. Abdellatif, G. Schoiswohl, R. Schreiber, T. Eisenberg, C. Magnes, S. Sedej, M. Eckhardt, M. Sasahara, T. Sasaoka, A. Nitta, G. Hoefler, W.F. Graier, D. Kratky, J. Auwerx, J.G. Bogner-Strauss, N-acetylaspartate availability is essential for juvenile survival on fat-free diet and determines metabolic health, The FASEB Journal. 33 (2019) 13808–13824. https://doi.org/10.1096/FJ.201801323R.

Supplementary Material A

Metabolomics and sphingolipidomics study of human hepatoma cells exposed to environmental concentrations of pharmaceutical compounds

Miriam Pérez-Cova[a,b], Carmen Bedia[a], Antonio Checa[c], Isabel Meister[c], Romà Tauler[a], Craig E Wheelock[c] , Joaquim Jaumot[a]

[a]Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, E08034 Barcelona, Spain

[b]Department of Chemical Engineering and Analytical Chemistry, University of Barcelona, Diagonal 647, Barcelona, E08028, Barcelona, Spain

[c]Division of Physiological Chemistry II, Department of Medical Biochemistry and Biophysics, Karolinska Institute, 17165, Solna, Sweden

* Correspondence: joaquim.jaumot@idaea.csic.es

1. LC-MS methodology

1.1 LC-MS/MS method for sphingolipid analysis

Chromatographic separation was carried out on an ACQUITY UPLC System with a sample manager cooled to 8°C (both from Waters Corporation, Milford, MA, USA). Sphingolipids were separated on a Zorbax Rapid Resolution RRHD C18 Column (80Å, 1.8 µm, 2.1 mm X 100 mm) using a guard column (5 × 2 mm, 1.8 µm particle size) (both from Agilent Technologies). Mobiles phases A and B consisted of 5mM ammonium formate (Sigma-Aldrich) / 0.2% formic acid (Optima, Fisher-Scientific) in water and in methanol (VWR), respectively. Separation was carried out at a 450 µl min^{-1} flowrate and column temperature was held at 40°C. The following chromatographic gradient was used: 0 min, 75% B; time range 0 → 1 min, 75% B (constant); time range 1 → 5 min, 85 → 100% B (linear increase); time range 5 to 15.2 min, 100% B (isocratic range); time range 15.2 → 15.3 min, 100 → 75% B (linear decrease); time range 15.3 → 16 min, 75% B (isocratic column conditioning). Samples were analyzed on a Waters Xevo® TQ-S system equipped with an Electrospray Ion Source (ESI) and ScanWave™ collision cell technology operating in the positive mode. A class specific single reaction monitoring (SRM) transition for each sphingolipid and internal standard was used. The method does not distinguish glycosylated species (GlcCer) from galactosylated species (GalCer), and Glc sphingolipids are therefore potentially a mixture of the two species.

1.2 LC-HRMS method for untargeted metabolomic analysis

Chromatographic separation was carried out on pre column SeQuant® ZIC®-pHILIC Guard Kit 20 x 2.1 mm, and a column SeQuant® ZIC®-pHILIC 5 µm polymer 100 x 2.1 mm were used. Mobile phases composition were: A) 5 mM ammonium acetate in water with 0.04% NH4OH (pH 9.3), B) 100% Acetonitrile. Gradient conditions expressed as time in minutes (%A, flow mL/min) : initial (12%, 0.28), 8.50 (40%, 0.28), 9.30 (95%, 0.28), 9.90 (95%, 0.20), 10.60 (95%, 0.17), 13.00 (95%, 0.20), 15.00 (95%, 0.2), 17.00 (12%, 0.2), 19.00 (12%, 0.5) and held until 28.50 min. Injection volume was set at 1 or 3 µL in positive or negative mode respectively, and the column oven at 35 ºC.

MS acquisition was performed in data independent analysis (DIA) mode. A mass range of 40-1200 m/z and an acquisition rate of 6 spectra/s were selected. MS/MS was acquired in all ion fragmentation (AIF) mode at three collision energies (0, 10, 30 eV).

2. MS-DIAL settings

Table S1. *MS-DIAL parameters used in metabolomic annotation.*

Start up a project	HILIC-HRMS method ESI(+)	HILIC-HRMS method ESI(-)
Ionization type	Soft ionization	Soft ionization
Separation type	Chromatography (LC)	Chromatography (LC)
Method type	All-ions with multiple CEs	All-ions with multiple CEs
Data type (MS1)	Centroid	Centroid
Data type (MS/MS)	Centroid	Centroid
Ion mode	Positive ion mode	Negative ion mode
Target omics	Metabolomics	Metabolomics
	Data collection	
MS1 tolerance	0.01	0.01
MS2 tolerance	0.01	0.01
Retention time begin	0.5	0.5
Retention time end	15	14
Mass range begin	40	40
Mass range end	1200	1200
Maximum charged number	2	2
Consider Cl and Br elements	Unchecked	Unchecked
Number of threads	20	20
Execute retention time corrections	Unchecked	Unchecked
	Peak detection	
Minimum peak height	800	800
Mass slice width	0.1	0.1
Smoothing method	Linear weighted moving average	Linear weighted moving average
Smoothing level	3	3
Minimum peak width	8	8
Exclusion mass list (tolerance: 0.01Da)	121.051 and 922.0098	112.9856, 1033.988
	MS2Dec	
Sigma window value	0.5	0.5
MS2Dec amplitude cut off	800	800
Exclude after precursor	Checked	Checked
Keep isotope until	0.5	0.5
Keep the isotopic ion w/o MS2Dec	Unchecked	Unchecked
	Identification	
Retention time tolerance	2	1
Accurate mass tolerance (MS1)	0.01	0.01

Accurate mass tolerance (MS2)	0.01	0.01
Identification score cut off	70	70
Use retention time for scoring	Unchecked	Unchecked
Use retention time for filtering	Unchecked	Unchecked
Postidentification	Not used	Not used
	Adduct	
Molecular species	[M+H]+, [M+NH4]+, [M+Na]+, [M+K]+, [M+H-H2O]+, [2M+H]+	[M-H]-, [M+Na-2H]-, [M-H2O-H]+, [2M-H]-
	Alignment	
Retention time tolerance	0.2	0.2
MS1 tolerance	0.02	0.02
Retention time factor	0.5	0.5
MS1 factor	0.5	0.5
Peak count filter	5	5
N% detected in at least one group	5	5
Remove feature based on blank information	Checked	Checked
Sample average / blank average	5	5
Keep "reference matched" metabolite features	Unchecked	Unchecked
Keep "suggested (w/o MS2)" metabolite features	Unchecked	Unchecked
Keep removable features and assign the tag	Checked	Checked
Gap filling by compulsion	Checked	Checked
	Isotope tracking	
	Not used	Not used

Table S2. *Correlation deconvolution parameters.*

	Correlation deconvolution	
MS2 tolerance	0.01	0.01
MS2 minimal peak intensity	800	800
Min. number of detected samples	4	4
Exclude highly correlated spots	0.9	0.9
Min. correlation coefficient (MS2)	0.7	0.7
Margins	0.1 (target and co-eluted precursors)	0.1 (target and co-eluted precursors)
Min. detected rate	0.1	0.1
Min. MS2 relative intensity	1%	1%
Remove peaks larger than precursor	Checked	Checked

Table S3. *Experiment file used in MS method type section from start a project window in MS-DIAL.*

Experiment file					
ID	*MS Type*	*Start m/z*	*End m/z*	*Collision energy (eV)*	*Deconvolution target (0: No; 1:Yes)*
0	MSMS	40	1200	10	1
1	MSMS	40	1200	30	1
2	SCAN	40	1200	0	1

3. Selection of potential markers of drug exposures

Table S4. *Matthews correlation coefficient of the PLS-DA, for the comparisons Control vs Medium dose and Control vs High.*

MCC values (PLS-DA)		**AMOX**	**CBZ**	**TRA**
Doses comparisons	C vs M	0.311	0.816	1.000
	C vs H	0.408	1.000	0.816

4. Discussion of the sphingolipids and metabolites altered by the drug exposures

Table S5. *Metaboanalyst pathway analysis of CBZ exposure.*

CBZ pathway analysis	Total	Expected	Hits	Raw p	-log(p)	Holm adjust	FDR	Impact
Glycerophospholipid metabolism	36	0.348	3	0.004	2.364	0.363	0.363	0.077
Sphingolipid metabolism	21	0.203	2	0.017	1.782	1.000	0.694	0.270
Alanine aspartate and glutamate metabolism	28	0.271	2	0.029	1.544	1.000	0.800	0.310
Taurine and hypotaurine metabolism	8	0.077	1	0.075	1.125	1.000	1.000	0.429
Arginine biosynthesis	14	0.135	1	0.128	0.894	1.000	1.000	0.000
Nicotinate and nicotinamide metabolism	15	0.145	1	0.136	0.865	1.000	1.000	0.000
Histidine metabolism	16	0.155	1	0.145	0.839	1.000	1.000	0.000
Pantothenate and CoA biosynthesis	19	0.184	1	0.170	0.771	1.000	1.000	0.000
Ether lipid metabolism	20	0.194	1	0.178	0.750	1.000	1.000	0.000
beta-Alanine metabolism	21	0.203	1	0.186	0.731	1.000	1.000	0.000
Lysine degradation	25	0.242	1	0.217	0.663	1.000	1.000	0.141
Glutathione metabolism	28	0.271	1	0.240	0.619	1.000	1.000	0.007
Primary bile acid biosynthesis	46	0.445	1	0.365	0.438	1.000	1.000	0.008
Aminoacyl-tRNA biosynthesis	48	0.465	1	0.378	0.423	1.000	1.000	0.000

Table S6. *Metaboanalyst pathway analysis of TRA exposure.*

CBZ pathway analysis	Total	Expected	Hits	Raw p	-log(p)	Holm adjust	FDR	Impact
Sphingolipid metabolism	21	0.257	3	0.002	2.743	0.152	0.118	0.308
Taurine and hypotaurine metabolism	8	0.098	2	0.004	2.418	0.317	0.118	0.429
Alanine aspartate and glutamate metabolism	28	0.343	3	0.004	2.375	0.346	0.118	0.2484
Glycerophospholipid metabolism	36	0.441	3	0.009	2.064	0.700	0.181	0.104
Ether lipid metabolism	20	0.245	2	0.024	1.625	1.000	0.399	0.000
Glycine serine and threonine metabolism	33	0.405	2	0.060	1.222	1.000	0.750	0.000
D-Glutamine and D-glutamate metabolism	6	0.074	1	0.071	1.146	1.000	0.750	0.000
Nitrogen metabolism	6	0.074	1	0.071	1.146	1.000	0.750	0.000
Valine leucine and isoleucine biosynthesis	8	0.098	1	0.094	1.026	1.000	0.805	0.000
Vitamin B6 metabolism	9	0.110	1	0.105	0.977	1.000	0.805	0.078
Primary bile acid biosynthesis	46	0.564	2	0.107	0.970	1.000	0.805	0.016
Aminoacyl-tRNA biosynthesis	48	0.588	2	0.115	0.939	1.000	0.805	0.000
Arginine biosynthesis	14	0.172	1	0.159	0.798	1.000	1.000	0.000
Lysine degradation	25	0.306	1	0.267	0.573	1.000	1.000	0.141
Glyoxylate and dicarboxylate metabolism	32	0.392	1	0.329	0.483	1.000	1.000	0.000
Arginine and proline metabolism	38	0.466	1	0.378	0.423	1.000	1.000	0.012
Pyrimidine metabolism	39	0.478	1	0.386	0.414	1.000	1.000	0.000
Purine metabolism	65	0.797	1	0.559	0.253	1.000	1.000	0.000

5.3 Discussion

This section focuses on the metabolomic and lipidomic workflows employed in **scientific publications VII and VIII,** with a special emphasis on the aspects learnt during the 2021 research stay. Firstly, a workflow for a quick evaluation of data quality for large untargeted metabolomic studies (e.g., clinical cohorts) is discussed. Secondly, two data analysis workflows are compared for their use in untargeted metabolomics datasets: one based on the ROIMCR strategy and the other based on the analysis provided by the MS-DIAL software. Finally, the targeted *versus* untargeted analysis problematic is discussed.

5.3.1 Quick check of data quality for large untargeted metabolomics

When performing the instrumental analysis of large untargeted metabolomics studies, such as clinical cohorts where multiple batches are analyzed non-stop during several months, it is crucial to ensure intra- and inter- batch reproducibility of the results. **Figure 5.2** summarizes the basic steps to consider for evaluating data quality, designed by the research group from KI [13]. Although considerably shorter than a clinical cohort, this data quality evaluation protocol was also employed to assess the quality of the untargeted metabolomic study from **scientific publication VIII.** In this case, the samples from each drug exposure (i.e., amoxicillin, carbamazepine, and trazodone) were considered as an entire batch, in each ionization mode (i.e., positive and negative). Therefore, six batches were separately evaluated. Besides, a replicate for each condition (drug and dose) was re-injected at the end of the sequence, to further verify the method's reproducibility.

The first step in this evaluation workflow could be checking the pressure profiles (see **Figure 5.3**). The registered pressure along the whole analysis sequence should be stable and reproducible among samples. If the pressure of an individual sample increases suddenly, it could indicate an augment in column pressure. This issue is often encountered in HILIC separations, where the high percentage of the aqueous phase needed for an optimal gradient usually forces the user to work at high pressures, closer to the maximum pressure that the column can hold. This is the case when high-performance liquid chromatography (HPLC) columns are used instead of ultra-high-performance liquid chromatography (UHPLC). Column pressure profile was stable during the metabolomic study from **scientific publications VIII.**

The superposition of base peak chromatograms (BPCs) or total ion chromatograms (TICs) is useful to quickly determine whether all samples have been injected, and if they present similar intensities. In the case that overall profiles of BPCs (or TICs) of all samples do not at least partially overlap, retention times may have shifted among samples, and a proper alignment and/or peak modelling could be necessary during data processing. If only an individual sample presents a different chromatographic profile and/or intensity from its replicates (or from other samples expected to have a similar profile), the user could decide to re-inject it. None of these scenarios was encountered for the dataset of **scientific publications VIII.**

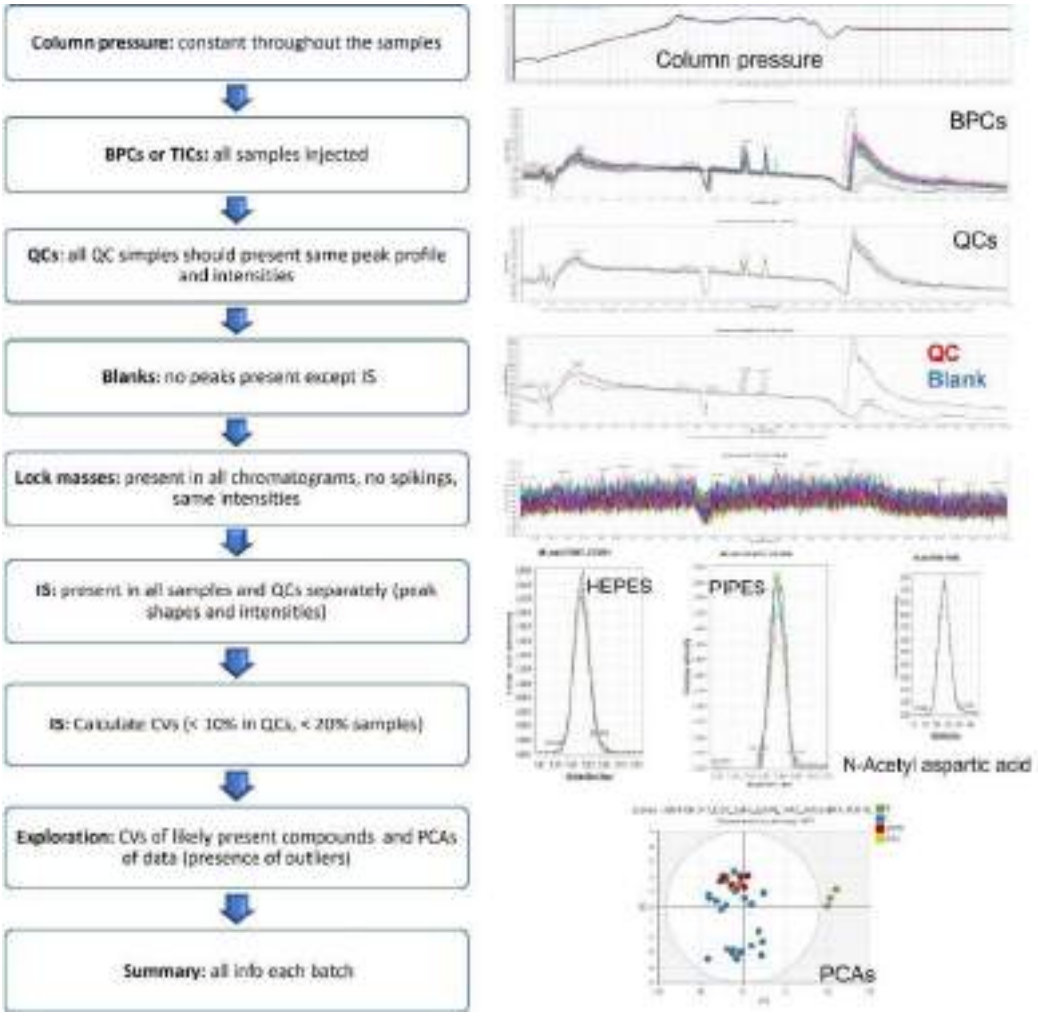


Figure 5.2. Workflow for a quick data analysis check of large untargeted metabolomic studies. BPC: base peak chromatograms; TIC: total ion chromatogram; QC: Quality control; IS: internal standard; CV: coefficient of variance; PCA: principal component analysis.

Quality controls (QCs) are representative mixtures of aliquots composed of a pool of all types of samples, which are run repeatedly along the analysis sequence

(e.g., every 8-10 samples). Hence, they can be used for intra- and inter- batch correction when required. QCs should have identical elution profiles and intensities, as the same homogeneous solution is injected repeatedly. Otherwise, QCs batch correction can be applied to compensate, for instance, instrumental drifts [15,16]. In the case of **scientific publications VIII,** a QC batch correction [17] was applied to the different batches, for each ionization mode separately for inter-batch corrections.

Blanks are often analyzed at the beginning and end of each batch. They are useful to detect carryover effect (i.e., if compounds from a sample are not properly eluted and they appear in the following chromatogram) and to check for IS reproducibility (if spiked with the IS solution). It is recommended to pre-condition the column in the sample matrix after running a blank, which can be performed by running a certain number of QCs at the beginning of the sequence. Although the composition of these initial QCs is the same as the other QCs, they are usually referred to as conditioning QCs. In **scientific publications VIII**, blanks were only measured at the beginning and at the end of each ionization mode analysis instead of between each drug sample set, due to the short length of the analysis sequence. In addition, fifteen conditioning QCs were added before running the samples in each ionization mode.

Lock masses are calibrant solutions used by the instrument to calibrate the mass values throughout the analysis. Whenever the lock masses are detected, it is possible to re-calibrate the datasets after the analysis is complete. Spikes in the lock masses can be caused by ionization suppression regions in the chromatogram, and special attention is required to the *m/z* values of the compounds present in these regions, as they present larger mass errors compared to other regions. No spiking regions were found for the cell samples from **scientific publications VIII** (see **Figure 5.2**).

The internal standards (IS) added to the samples just before the injection are useful to calculate the reproducibility of the analysis and correct for instrumental drifts during the QC batch correction already mentioned. Coefficient of variances (CVs) of these compounds are often considered acceptable under 10% in QCs and under 20% in the rest of samples, in the cases of large studies. If fewer samples are analyzed, then the reproducibility should be higher. For instance, PIPES and HEPES, two compounds used as IS, presented CVs lower than 5% for negative ionization batches after QC batch correction. A visual inspection of the peak shapes and intensities is also useful for a qualitative overview of the reproducibility among QCs

and all samples from the batch. In the case that the user has some a priori knowledge of compounds expected to be present in the samples, a quick screening of their CVs values and an exploration of their elution profiles is also desirable. For instance, N-acetyl aspartic acid and taurine were used to check the reproducibility of peak shapes and intensities across cells samples, in negative ionization mode, as shown in **Figure 5.2**.

Lastly, a PCA scores plot can help identify outliers and detect some trends in the data (e.g., if the QCs cluster together). A summary of all the previous information is also handy, and a table with all the CVs is also recommended for a global inspection of multiple batches. Although a report with such a level of detail as the one proposed here is unnecessary in practice, it becomes useful because information on every batch is collected and data quality is ensured, especially in large cohorts. All in all, the minimum steps that require a thorough inspection are blanks and QCs profiles, CVs of IS (in samples, blanks and QCs) and confirmation that the lock mass is detected in all chromatograms.

5.3.2 Data analysis workflows for untargeted metabolomics

Two data analysis workflows have been employed in the publications included in this Chapter, which are summarized in **Figure 5.3.** Workflow A (**scientific publication VII**) in the Figure represents a ROIMCR based pipeline. In contrast, workflow B represents an MS-DIAL based pipeline (**scientific publications VIII**). This section aims to highlight the main advantages and limitations of each workflow.

The use of the ROI approach in combination with MCR-ALS has been widely discussed throughout this PhD Thesis (see **Chapter 3** for more details). Although MS-DIAL software [18,19] has also been used in different PhD Thesis publications for compound annotation, only in **scientific publications VIII** is also employed for data pre-processing. Briefly, the MS-DIAL pipeline incorporates a between samples alignment step for the chromatographic peaks, and the possibility to filter the signals present in a blank (e.g., extraction blank). Besides, MS-DIAL also performs peak integration, and, more interestingly, compound annotation based on retention time (RT) and MS/MS libraries.

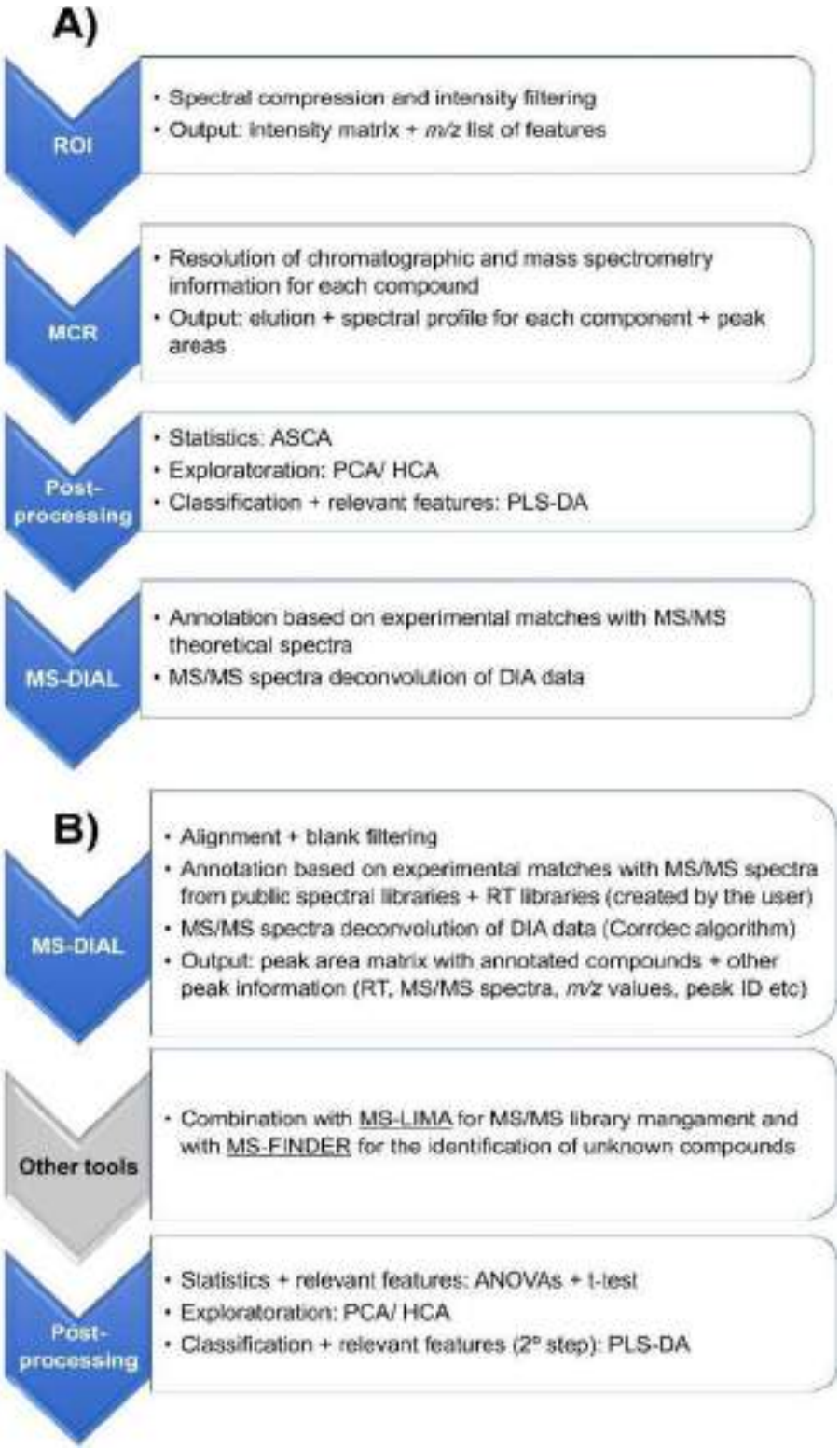


Figure 5.3. Comparison of data analysis workflows employed in **scientific publications VII** (workflow A) **and VIII** (workflow B). ROI: regions of interest; MCR: multivariate curve resolution; ASCA: ANOVA simultaneous component analysis; PCA: principal component analysis; HCA: hierarchical clustering; DIA: data independent acquisition; RT: retention times; ANOVAs: ANalysis Of Variance; PLS-DA: partial least squares discriminant analysis.

An important benefit of using ROI compared to MS-DIAL on raw data is the reduction in data dimensionality. When high-resolution mass spectrometry (HRMS) is employed after chromatography, the resulting datasets are usually very big. In addition, all precursor ions are fragmented in all ion fragmentation (AIF) acquisition mode (used in both metabolomic studies from this Chapter). In **scientific publications VII and VIII**, AIF provided MS1 (*m/z* spectrum) and MS2 (MS/MS patterns from the *m/z* values from MS1) levels for each retention time, which increments data size. For instance, the entire metabolomic raw dataset size of both studies (cells and rice) is approximately 40 GB each (including the two ionization modes). Therefore, importing and processing these datasets into MS-DIAL can last several hours, whereas the pre-processing with ROI can be much shorter (if the appropriate settings are selected, see discussion about ROI parameters in **Chapters 3 and 4**).

The use of MCR-ALS provides an elution profile and a mass spectrum (MS1) for each component. Ideally, each chemical compound present in the mixture would be represented by one component, and the different adducts and isotopic forms of the same compound would be joined into the same component, i.e., what is known by componentization. On the contrary, MS-DIAL can furnish several hits for the same compound, according to different adducts forms, or even the same adduct at different retention times if MS/MS hits are found. Moreover, sometimes the correct adducts are not well assigned automatically (e.g., the software may assign a $[M+Na]^+$ adduct when the *m/z* value obtained corresponds to a $[M+H]^+$ form of the same compound), and a further check should be carried out with complementary information from databases, such as the Human Metabolome DataBase (HMDB) [20,21] or Lipid MAPS [22,23]). Besides, the compounds found by MS-DIAL (e.g., lipid annotation based on MS1 level only) should be checked carefully in the absence of MS/MS information. For instance, in the untargeted rice lipidomic analysis from **scientific publications VII** (where no information at MS/MS level was acquired), the tentative annotation provided by Lipid MAPS [22,23] and an in-house library was preferred to the provided by MS-DIAL.

However, one of the main limitations of the ROIMCR workflow is the annotation step, especially for untargeted analysis. Until now, MCR-ALS only provided compound resolution and quantification at the MS1 level, and the annotation has been confirmed based on the MS2 information through other software. For instance,

in a preliminary compound annotation of the rice dataset, a list of relevant components was obtained with partial least squares discriminant analysis (PLS-DA) (after evaluating the differences between arsenic treatments supplied to rice) and selected for annotation. Then, the MS2 spectra for each retention time (associated with the relevant components) were acquired from the raw data through vendor software. These MS/MS spectra were compared to spectral libraries from HMDB [20,21], Global Natural Product Social Molecular Networking (GNPS) [24], and Massbank [25,26]. However, a low number of hits were obtained and, therefore, only a few metabolites were tentatively annotated. The main reason was a lack of prior deconvolution of the AIF data, which means that the used MS2 spectra were a mix of MS/MS fragmentation patterns of all the detected *m/z* values from the MS1 level at that retention time. This deconvolution step, especially useful in data independent acquisition (DIA) such as AIF, is included in MS-DIAL, which makes it appealing for compound annotation. In addition, a new algorithm, known as Corrdec [27], has been recently proposed for an alternative deconvolution of DIA data at the MS/MS level, and has been incorporated into the MS-DIAL software. Thus, this deconvolution step becomes necessary for compound annotation (especially for DIA data). Current work is being pursued for using MCR-ALS for the resolution of the MS2 level too. The aim is to obtain for each compound an elution profile, an MS1 mass spectra profile, and a resolved MS2 mass spectra profile containing only the fragments from that specific MS1 profile (instead of a mix of all the MS1 profiles from all the compounds fragmented at the same retention time).

Apart from the deconvolution step, another main benefit of MS-DIAL is the possibility of incorporating compound libraries when processing the data, in order to automatize compound annotation. On the one hand, the user can create RT libraries in the same chromatographic conditions that in the new samples by injecting standard solutions of as many analytes as possible. On the other hand, MS/MS libraries are also available. MS-DIAL already includes a rich MS/MS library of lipids [19] that is constantly being updated with new lipids. For metabolites, MS/MS compilations from the main public spectral libraries are available on the MS-DIAL website for each ionization mode [28]. Alternatively, MS/MS libraries created by the user could also be uploaded. This functionality of MS-DIAL is especially useful for pseudo-targeted and untargeted approaches in routine metabolomic platforms (e.g., when metabolomic analyses from sample extraction to data analysis steps are offered as a service). The users could create their own libraries using the

chromatographic and mass spectrometric methods of their routine analysis and continue to complement them by discovering unknown compounds from the new analyses, cyclically. The KI research group has implemented this strategy in collaboration with Gunma University in Japan [13].

Although not employed in **scientific publications VIII**, there are other two valuable tools related to MS-DIAL (marked in the grey section of **Figure 5.3**). On the one side, MS-LIMA [29] is a software that allows the management of the existing MS/MS libraries, and could also be used to create new libraries, based on MS/MS information acquired by the user in different modes and collision energies. On the other side, MS-FINDER [30] is a tool for discovering unknown compounds, especially useful in untargeted analysis (i.e., provides structural information and MS/MS fragmentation patterns of potential candidates, plus other information such as InChIKeys of these potential compounds).

All in all, a proper comparison of the capabilities of ROIMCR *versus* MS-DIAL is still lacking. The main advantages of MS-DIAL processing are related to compound annotation, which is straightforward when MS/MS information from the samples is acquired (for both, data dependent and data independent acquisitions). However, the ROIMCR provides more complete and transparent processing that includes compression and resolution of the data, although including the resolution at the MS2 level and an automatic annotation step from the resolved spectra are needed. Future work will be pursued in this direction.

5.3.3 The targeted *versus* untargeted analysis problematic

In recent years, the tendency has been to move towards fully untargeted analyses, for the discovery of unknown compounds to expand current metabolome knowledge and provide a holistic overview of the effects of a certain exposure (e.g., exposure to an environmental stressor, or differences between healthy and ill patients/tissues), as shown in **Figure 5.4**. Both analytical and data analysis methods have been developed to detect and annotate as many compounds as possible from many different families. Consequently, the amount of information that can be achieved per analyzed sample has considerably increased, but also data size. In the era of big data, automatic workflows (e.g., based on executing multiple commands *via* scripts in which the user merely supervises each step to ensure the quality of the results) are required, for the sake of time.

The main question here is: are we detecting new compounds with these approaches, or are we mainly detecting the same compounds over and over again?

Sometimes, aiming for a bigger picture, the actual limitations of a one-for-all solution can be ignored. As stated at the beginning of this Chapter, the variety of physicochemical properties of the analytes makes it difficult to implement a unique analytical workflow. Moreover, these generical conditions can detriment the analytical measurements compared to targeted approaches where the experimental conditions for each family or group of compounds are optimized. In addition, metabolites present in the samples in low abundances may be lost with untargeted analysis due to the unspecific conditions and the ion suppression caused by other predominant compounds. On the contrary to polymerase chain reaction (PCR), which copies and amplifies the signal of deoxyribonucleic acid (DNA) fragments, there is no current technology able to increase these signals of the less abundant metabolites without a specific pre-treatment of the sample.

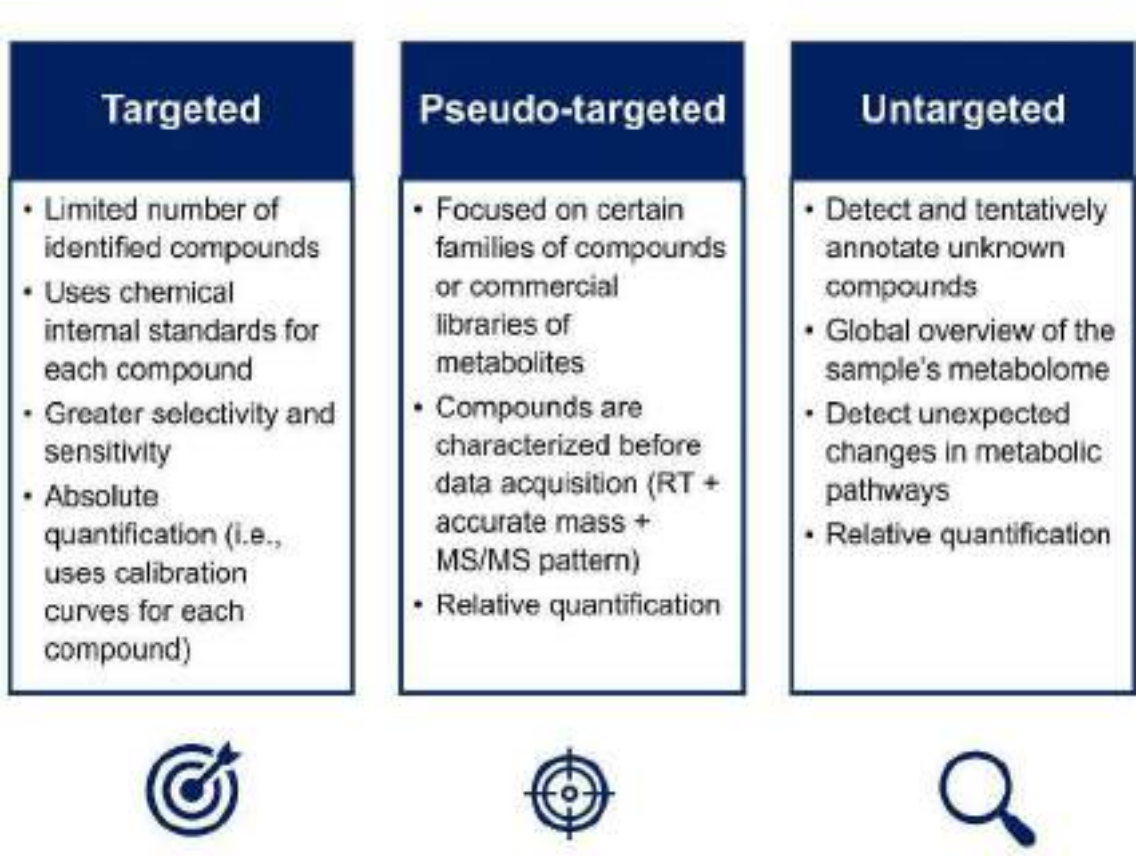


Figure 5.4. Comparison of targeted, pseudo-targeted and untargeted approaches for metabolomic analysis.

Strategies that aim to solve the problem of ion suppression may involve data dependent acquisition (DDA). In these cases, samples can be reinjected cyclically, establishing different intensities thresholds. For instance, the first fifty more intense compounds would be detected and fragmented. Then, in the second injection, the next fifty more intense are detected, and so on. An iterative auto MS/MS, a form of DDA, was employed in **scientific publication V** for the fragmentation of QC samples. These approaches are more effective when a list of interesting compounds is available for fragmentation (otherwise, low abundant compounds may be lost), but to know which are these interesting compounds, a preliminary untargeted approach is needed. In contrast, DIA acquisitions fragment every compound eluting at that retention time (even the less abundant), but a deconvolution step is crucial in order to associate the fragments to the compound, as already discussed.

Another aspect to keep in mind is that every potential discovery found with untargeted approaches should be further validated in a targeted manner. The differences between untargeted and targeted analysis are summarized in **Figure 5.4.** In the untargeted analysis, the aim is a global overview of the samples' metabolome, whereas the targeted provides identification and quantification of a reduced number of compounds (e.g., lower than a hundred). Regarding untargeted discovery validation, if the finding is related to specific metabolic pathways altered by a certain stressor, the solution would be to study these pathways individually, delimiting the number of compounds detected with the same method and optimizing the analytical conditions for their quantification. In the case of detecting new metabolites, then the confirmation needs to be *via* structural elucidation based, for instance, on NMR analysis. However, for achieving a complete validation of the annotated compounds, a standard of each new metabolite is needed, which requires the synPhD Thesis of these compounds if they are not commercially available (in most cases). Hence, these discoveries are not as fast as we may want them to be, or as if the effort would be concentrated on a smaller picture (e.g., first, focusing on one family of relevant compounds on a certain matrix, then another etc., instead of looking for a needle in a haystack). In both scenarios, the targeted second step is more expensive due to the need for standard solutions for each analyte and certified reference materials. In the end, the effort may seem doubled, as the first untargeted analysis provides only preliminary results.

An intermediate solution may be using pseudo-targeted approaches (as used in **scientific publication VIII** for the metabolomic analysis). In this scenario (see **Figure 5.4**.), the aim is to screen as many compounds as possible with a generic method, but with the guarantee of previously characterizing standard of these compounds with the same method (e.g., retention time information, MS/MS patterns in different acquisition modes and collision energies). Metabolite libraries of hundreds of compounds are already commercially available. Although it would seem expensive, this strategy is worthy of routine analyses. Instead of injecting the calibration curves of the standard solutions in each analysis, these standards are only injected when optimizing and validating the method, but their characterization will still be available afterwards. However, an absolute quantification or correction based on internal standards for each compound (e.g., correct extraction losses or ion suppression effects) is not possible with this approach. Another disadvantage of this approach is that the slightest change in the analytical workflow (e.g., using a brand new chromatographic column although in theory identical to the previous one) may need re-adjustments in the whole characterization process. Therefore, there is little room for improvements once the method has been validated. Hence, the use of this approach has an expiration date dependent on the current technology (e.g., new stationary phases or packing technologies for chromatographic columns, new mass spectrometers with novel capabilities).

After these considerations, the way out of this dead-end may be in the use of multidimensional approaches discussed in the previous Chapter, or the simultaneous combination of multiple platforms, as pointed out in the publications from this Chapter. Nevertheless, there is no doubt that the application of these novel approaches will need automatic, straightforward, accessible, flexible, and reliable data analysis workflows from pre-processing to compound annotation. Otherwise, the analytical developments will not achieve their full potential.

5.4 Conclusions

This section includes the specific conclusions drawn throughout this Chapter about the different metabolomics data analysis workflows employed.

- A quick check of the quality of the data right after the acquisition is essential to ensure the reliability of the results and enable the re-injection of certain samples if needed (otherwise, the information from these samples may be lost due to instrumental errors). This quality check is especially crucial in the case of large metabolomic studies (e.g., clinical cohorts).
- ROIMCR is a powerful approach for the analysis of untargeted metabolomic datasets, from the pre-processing (i.e., compression) to the resolution of compounds in complex mixtures (i.e., elution and spectral profiles). However, it currently lacks the resolution of information at the MS2 level and an automatic annotation step.
- On the contrary, MS-DIAL is a user-friendly tool for metabolomic data processing, including deconvolution of MS/MS spectra and compound annotation. However, no compression is applied, which slows down the whole analysis time. Besides, it is only recommended when MS/MS information from the samples is available, and a double-check of the assigned adducts is desirable. In addition, special attention is required when several hits with the MS/MS libraries at different retention times are found.
- Regarding the targeted *versus* untargeted dilemma, the decision needs to be made on a case-by-case basis. Still, all untargeted analyses need to be further validated using targeted methods afterwards. Current trends are moving towards pseudo-targeted approaches, an in-between solution, where a large scope of metabolites and pathways are targeted, and relative quantification and annotation based on RT, MS and MS/MS information are provided.

References

[1] J. Villaret-Cazadamont, N. Poupin, A. Tournadre, A. Batut, L. Gales, D. Zalko, N.J. Cabaton, F. Bellvert, J. Bertrand-Michel, An optimized dual extraction method for the simultaneous and accurate analysis of polar metabolites and lipids carried out on single biological samples, Metabolites. 10 (2020) 1–19. https://doi.org/10.3390/metabo10090338.

[2] M.A. Salem, J. Jüppner, K. Bajdzienko, P. Giavalisco, Protocol: A fast, comprehensive and reproducible one-step extraction method for the rapid preparation of polar and semi-polar metabolites, lipids, proteins, starch and cell wall polymers from a single sample, Plant Methods. 12 (2016) 1–15. https://doi.org/10.1186/s13007-016-0146-2.

[3] J. Kang, L. David, Y. Li, J. Cang, S. Chen, Three-in-One Simultaneous Extraction of Proteins, Metabolites and Lipids for Multi-Omics, Frontiers in Genetics. 12 (2021) 313. https://doi.org/10.3389/FGENE.2021.635971/BIBTEX.

[4] M. Schwaiger, H. Schoeny, Y. el Abiead, G. Hermann, E. Rampler, G. Koellensperger, Merging metabolomics and lipidomics into one analytical run, Analyst. 144 (2019) 220–229. https://doi.org/10.1039/c8an01219a.

[5] R. Tuli, D. Chakrabarty, P.K. Trivedi, R.D. Tripathi, Recent advances in arsenic accumulation and metabolism in rice, Molecular Breeding. 26 (2010) 307–323. https://doi.org/10.1007/s11032-010-9412-6.

[6] FDA, Arsenic in Rice and Rice Products Risk Assessment Report, Center for Food Safety and Applied Nutrition of the Food and Drug Administration. 1 (2016) 1–284.

[7] C.A. Marasco Júnior, D.M. Sartore, R.S. Lamarca, B.F. da Silva, Á.J. Santos-Neto, P.C.F. de Lima Gomes, On-line solid-phase extraction of pharmaceutical compounds from wastewater treatment plant samples using restricted access media in column-switching liquid chromatography-tandem mass spectrometry, Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences. 1180 (2021). https://doi.org/10.1016/j.jchromb.2021.122896.

[8] M. Navarro-Reig, J. Jaumot, B. Piña, E. Moyano, M.T. Galceran, R. Tauler, Metabolomic analysis of the effects of cadmium and copper treatment in: Oryza sativa L. using untargeted liquid chromatography coupled to high resolution mass spectrometry and all-ion fragmentation, Metallomics. 9 (2017) 660–675. https://doi.org/10.1039/c6mt00279j.

[9] M. Navarro-Reig, R. Tauler, G. Iriondo-Frias, J. Jaumot, Untargeted lipidomic evaluation of hydric and heat stresses on rice growth, Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences. 1104 (2019) 148–156. https://doi.org/10.1016/j.jchromb.2018.11.018.

[10] E. Gorrochategui, J. Jaumot, R. Tauler, ROIMCR: A powerful analysis strategy for LC-MS metabolomic datasets, BMC Bioinformatics. 20 (2019) 1–17. https://doi.org/10.1186/s12859-019-2848-8.

[11] E. Ortiz-Villanueva, L. Navarro-Martín, J. Jaumot, F. Benavente, V. Sanz-Nebot, B. Piña, R. Tauler, Metabolic disruption of zebrafish (Danio rerio) embryos by bisphenol A. An integrated metabolomic and transcriptomic approach, Environmental Pollution. 231 (2017) 22–36. https://doi.org/10.1016/J.ENVPOL.2017.07.095.

[12] F. Puig-Castellví, C. Bedia, I. Alfonso, B. Piña, R. Tauler, Deciphering the Underlying Metabolomic and Lipidomic Patterns Linked to Thermal Acclimation in Saccharomyces cerevisiae, Journal of Proteome Research. 17 (2018) 2034–2044. https://doi.org/10.1021/acs.jproteome.7b00921.

[13] I. Meister, P. Zhang, A. Sinha, C.M. Sköld, Å.M. Wheelock, T. Izumi, R. Chaleckis, C.E. Wheelock, High-Precision Automated Workflow for Urinary Untargeted Metabolomic Epidemiology, Cite This: Anal. Chem. 93 (2021) 5258. https://doi.org/10.1021/acs.analchem.1c00203.

[14] N. Akawi, A. Checa, A.S. Antonopoulos, I. Akoumianakis, E. Daskalaki, C.P. Kotanidis, H. Kondo, K. Lee, D. Yesilyurt, I. Badi, M. Polkinghorne, N. Akbar, J. Lundgren, S. Chuaiphichai, R. Choudhury, S. Neubauer, K.M. Channon, S.S. Torekov, C.E. Wheelock, C. Antoniades, Fat-Secreted Ceramides Regulate Vascular Redox State and Influence Outcomes in Patients With Cardiovascular Disease, J Am Coll Cardiol. 77 (2021) 2494–2513. https://doi.org/10.1016/j.jacc.2021.03.314.

[15] R. Wehrens, J.A. Hageman, F. van Eeuwijk, R. Kooke, P.J. Flood, E. Wijnker, J.J.B. Keurentjes, A. Lommen, H.D.L.M. van Eekelen, R.D. Hall, R. Mumm, R.C.H. de Vos, Improved batch correction in untargeted MS-based metabolomics, Metabolomics. 12 (2016). https://doi.org/10.1007/s11306-016-1015-8.

[16] J. Kuligowski, Á. Sánchez-Illana, D. Sanjuán-Herráez, M. Vento, G. Quintás, Intra-batch effect correction in liquid chromatography-mass spectrometry using quality control samples and support vector regression (QC-SVRC), Analyst. 140 (2015) 7810–7817. https://doi.org/10.1039/c5an01638j.

[17] D. Broadhurst, R. Goodacre, S.N. Reinke, J. Kuligowski, I.D. Wilson, M.R. Lewis, W.B. Dunn, Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies, Metabolomics. 14 (2018) 1–17. https://doi.org/10.1007/s11306-018-1367-3.

[18] and M.A. Hiroshi Tsugawa, Tomas Cajka, Tobias Kind, Yan Ma, Brendan Higgins, Kazutaka Ikeda, Mitsuhiro Kanazawa, Jean VanderGheynst, Oliver Fiehn, MS-DIAL: Data Independent MS/MS Deconvolution for Comprehensive, Nat Methods. 12 (2015) 523–526. https://doi.org/10.1038/nmeth.3393.MS-DIAL.

[19] H. Tsugawa, K. Ikeda, M. Takahashi, A. Satoh, Y. Mori, H. Uchino, N. Okahashi, Y. Yamada, I. Tada, P. Bonini, Y. Higashi, Y. Okazaki, Z. Zhou, Z.J. Zhu, J. Koelmel, T. Cajka, O. Fiehn, K. Saito, M. Arita, M. Arita, A lipidome atlas in MS-DIAL 4, Nature Biotechnology. 38 (2020) 1159–1163. https://doi.org/10.1038/s41587-020-0531-2.

[20] D.S. Wishart, T. Jewison, A.C. Guo, M. Wilson, C. Knox, Y. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong, S. Bouatra, I. Sinelnikov, D. Arndt, J. Xia, P. Liu, F. Yallou, T. Bjorndahl, R. Perez-Pineiro, R. Eisner, F. Allen, V. Neveu, R. Greiner, A. Scalbert, HMDB 3.0-The Human Metabolome Database in 2013, Nucleic Acids Research. 41 (2013) 801–807. https://doi.org/10.1093/nar/gks1065.

[21] D.S. Wishart, A.C. Guo, E. Oler, F. Wang, A. Anjum, H. Peters, R. Dizon, Z. Sayeeda, S. Tian, B.L. Lee, M. Berjanskii, R. Mah, M. Yamamoto, J. Jovel, C. Torres-Calzada, M. Hiebert-Giesbrecht, V.W. Lui, D. Varshavi, D. Varshavi, D. Allen, D. Arndt, N. Khetarpal, A. Sivakumaran, K. Harford, S. Sanford, K. Yee, X. Cao, Z. Budinski, J. Liigand, L. Zhang, J. Zheng, R. Mandal, N. Karu, M. Dambrova, H.B. Schiöth, R. Greiner, V. Gautam, HMDB 5.0: the Human Metabolome Database for 2022, Nucleic Acids Res. 50 (2022) D622–D631. https://doi.org/10.1093/NAR/GKAB1062.

[22] E. Fahy, S. Subramaniam, R.C. Murphy, M. Nishijima, C.R.H. Raetz, T. Shimizu, F. Spener, G. van Meer, M.J.O. Wakelam, E.A. Dennis, Update of the LIPID MAPS comprehensive classification system for lipids, Journal of Lipid Research. 50 (2009) 9–14. https://doi.org/10.1194/jlr.R800095-JLR200.

[23] E. Fahy, M. Sud, D. Cotter, S. Subramaniam, LIPID MAPS online tools for lipid research, Nucleic Acids Research. 35 (2007). https://doi.org/10.1093/nar/gkm324.

[24] M. Wang, J.J. Carver, V. v. Phelan, L.M. Sanchez, N. Garg, Y. Peng, D.D. Nguyen, J. Watrous, C.A. Kapono, T. Luzzatto-Knaan, C. Porto, A. Bouslimani, A. v. Melnik, M.J. Meehan, W.T. Liu, M. Crüsemann, P.D. Boudreau, E. Esquenazi, M. Sandoval-Calderón, R.D. Kersten, L.A. Pace, R.A. Quinn, K.R. Duncan, C.C. Hsu, D.J. Floros, R.G. Gavilan, K. Kleigrewe, T. Northen, R.J. Dutton, D. Parrot, E.E. Carlson, B. Aigle, C.F. Michelsen, L. Jelsbak, C. Sohlenkamp, P. Pevzner, A. Edlund, J. McLean, J. Piel, B.T. Murphy, L. Gerwick, C.C. Liaw, Y.L. Yang, H.U. Humpf, M. Maansson, R.A. Keyzers, A.C. Sims, A.R. Johnson, A.M. Sidebottom, B.E. Sedio, A. Klitgaard, C.B. Larson, C.A.P. Boya, D. Torres-Mendoza, D.J. Gonzalez, D.B. Silva, L.M. Marques, D.P. Demarque, E. Pociute, E.C. O'Neill, E. Briand, E.J.N. Helfrich, E.A. Granatosky, E. Glukhov, F. Ryffel, H. Houson, H. Mohimani, J.J. Kharbush, Y. Zeng, J.A. Vorholt, K.L. Kurita, P. Charusanti, K.L. McPhail, K.F. Nielsen, L. Vuong, M. Elfeki, M.F. Traxler, N. Engene, N. Koyama, O.B. Vining, R. Baric, R.R. Silva, S.J. Mascuch, S. Tomasi, S. Jenkins, V. Macherla, T. Hoffman, V. Agarwal, P.G. Williams, J. Dai, R. Neupane, J. Gurr, A.M.C. Rodríguez, A. Lamsa, C. Zhang, K. Dorrestein, B.M. Duggan, J. Almaliti, P.M. Allard, P. Phapale, L.F. Nothias, T. Alexandrov, M. Litaudon, J.L. Wolfender, J.E. Kyle, T.O. Metz, T. Peryea, D.T. Nguyen, D. VanLeer, P. Shinn, A. Jadhav, R. Müller, K.M. Waters, W. Shi, X. Liu, L. Zhang, R. Knight, P.R. Jensen, B. Palsson, K. Pogliano, R.G. Linington, M. Gutiérrez, N.P. Lopes, W.H. Gerwick, B.S. Moore, P.C. Dorrestein, N. Bandeira, Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking, Nature Biotechnology. 34 (2016) 828–837. https://doi.org/10.1038/nbt.3597.

[25] T. Schulze, R. Meier, N. Alygizakis, E. Schymanski, E. Bach, D.H. LI, lauperbe, raalizadeh, S. Tanaka, M. Witting, MassBank/MassBank-data: Release version 2021.12, (2021). https://doi.org/10.5281/ZENODO.5775684.

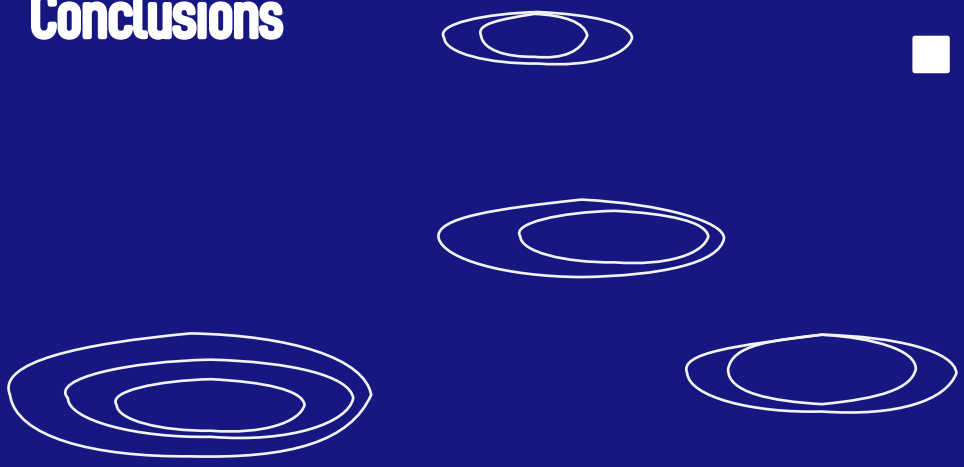
[26] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M.Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, T. Nishioka, MassBank: A public repository for sharing mass spectral data for life sciences, Journal of Mass Spectrometry. 45 (2010) 703–714. https://doi.org/10.1002/jms.1777.

[27] I. Tada, R. Chaleckis, H. Tsugawa, I. Meister, P. Zhang, N. Lazarinis, B. Dahlén, C.E. Wheelock, M. Arita, Correlation-Based Deconvolution (CorrDec) to Generate High-Quality MS2 Spectra from Data-Independent Acquisition in Multisample Studies, Analytical Chemistry. 92 (2020) 11310–11317. https://doi.org/10.1021/acs.analchem.0c01980.

[28] MS-DIAL web page. http://prime.psc.riken.jp/compms/msdial/main.html, (n.d.).
[29] GitHub - tipputa/MS-LIMA-Standard: Mass spectral library manager, (n.d.). https://github.com/tipputa/MS-LIMA-Standard (accessed April 20, 2022).
[30] CompMS | MS-FINDER, (n.d.). http://prime.psc.riken.jp/compms/msfinder/main.html (accessed April 20, 2022).

Chapter .six

Conclusions



The **analytical methodologies** and **data analysis strategies** developed in this PhD Thesis, based on the use of one- and two-dimensional liquid chromatography coupled to mass spectrometry (LC-MS and LC×LC-MS, respectively), have been successfully optimized and validated for **metabolomic** (and lipidomic) **studies**. Furthermore, analytical and chemometric pipelines have been proposed to assess the effects) that **environmental stressors** produce in the metabolome and lipidome of model biosystems. In conclusion, the new approaches methodologies (NAMs) followed in this PhD Thesis have facilitated the extraction of the most relevant analytical and biological information related to the exposure of the investigated emerging contaminants (ECs).

Conclusions related to the analytical and data analysis strategies employed in this PhD Thesis are described in detail below:

Analytical conclusions

- Two-dimensional liquid chromatography coupled to high-resolution mass spectrometry (LC×LC-HRMS) has been successfully applied for untargeted lipidomics analyses. An RP×HILIC configuration has been preferred because it provided a higher chromatographic resolution power for lipid analysis. The use of the active solvent modulation (ASM) strategy led to an increase in sensitivity and solvent compatibility, as well as the decrease of the total run time. However, more research is still required to optimize the conditions allowing a better retention of the less polar lipids in the second column dimension (e.g., ceramides and triacylglycerides).
- There is a need for reducing the analysis time of metabolites using LC×LC-MS while enhancing solvent compatibility between the two dimensions. The ASM strategy is highly recommended. However, more research is also required to assess the efficacy of ASM, in particular, for small and polar molecules.
- Individual metabolomic and lipidomic workflows are conditioned by the wide variety of physicochemical properties of the analytes. Even if a common extraction step is used to improve the efficiency and analysis times of untargeted studies, LC-MS analyses should still be performed on different

MS platforms or with multidimensional separations to achieve a broader coverage. The combined information about metabolic and lipidic pathways affected by emerging contaminants provides a holistic coverage and assessment of their mode of action at molecular level.

Chemometric conclusions

- The assessment of the multiway structure and behavior of LC×LC MS and UV datasets showed deviations from the ideal full trilinear factor decomposition model. Hence, trilinear model-based methods such as parallel factor analysis (PARAFAC) are not recommended in general for LC ×LC data. In contrast, acceptable results can always be obtained using a bilinear model factor decomposition, for instance using the multivariate curve resolution alternating least squares (MCR-ALS) method.
- Data fusion strategies combining the information from multiple detectors provided a more complete resolution of the sample constituents and an easier identification of the compounds associated with each MCR-ALS resolved component.
- Multivariate analysis of variance (ANOVA) – based methods gave acceptable results for the statistical assessment of statistically significant effects produced by the tested factors (environmental stressors) in designed experiments and allowed variable selection. However, further work is required to establish the best method to fulfil both goals. The workflow proposed in this PhD Thesis combines ANOVA simultaneous component analysis (ASCA) and partial least squares discriminant analysis (PLS-DA) for optimal statistical analysis and variable selection, respectively.
- The ROI approach for filtering and spectral compression provided a sufficient dimensionality reduction of LC×LC-HRMS datasets. This ROI strategy demonstrated its usefulness for targeted, pseudo-targeted and untargeted approaches, also allowing the proper recovery of absolute and relative quantification information.
- Acceptable quantification results can be obtained in the analysis of LC×LC-MS datasets with ROIMCR based strategies. The use of the peak area correlation constraint for the known calibration samples (applied during the alternating least squares optimization) is recommended for strongly

coeluted analytes because, in these cases (e.g., biological matrices with multiple overlapping compounds, some of them isobaric), more accurate absolute quantifications of the analytes are obtained.

- Different data analysis workflows were tested for untargeted metabolomic studies. ROIMCR is a very powerful approach for analyzing untargeted metabolomic datasets, but further work is required to extract information at the MS2 level and incorporate an automatic annotation step. Meanwhile, the ROIMCR approach can be complemented with other approaches such as MS-DIAL, to improve the deconvolution of MS/MS spectra and the proper compound annotation. In conclusion, results obtained with the proposed metabolomic workflows allowed the assessment of the effects produced by the environmental stressors investigated in this Thesis.

Chapter seven

Annexes

Annex 1. Experimental conditions employed for the acquisition of the chromatograms shown in Figure 4.3.

The experimental conditions in which the chromatograms of **Figure 4.3** were acquired are described below. Visualization of the chromatograms shown in **Figure 4.3** were obtained with MassHunter Workstation Software, Qualitative Analysis Navigator, version B.08.00 (Agilent TechnologiesSanta Clara, CA, USA).

A) and B). HILIC 1DLC separation of phospholipid extracts. The HILIC column was prepared in-house by slurry packing unmodified bare Zorbax silica (3.5 μm, 80 Å pore size) (Agilent Technologies, Santa Clara, CA, USA) into a small column (50 mm x 2.1 mm i.d.). Mobile phases employed were: A) Ammonium formate 30 mM, B) AcN. Injection volume: 2 μL. Flow rate: 0.4 mL min^{-1}. Temperature: 40 °C. Gradient (percentage of B): 0 (98), 1 (80), 5 (60), 10 (60), 10.1 (98), 15 (98) min. **RP 1DLC** separation of phospholipid extracts. HPH C18 column (50 x 2.1mm i.d., 2.7 μm) (Agilent Technologies, Santa Clara, CA, USA). Mobile phase composition: A) 30 mM Ammonium formate (pH 4.5), B) AcN/IPA (1:2). Injection volume: 2 μL. Flow rate: 0.03 mL min^{-1}. Temperature: 50 °C. Gradient (percentage of B): 0 (70), 8 (90), 9 (100), 10 (100), 10.1 (70), 13 (70) min.

C) The HILIC column was prepared in-house by slurry packing unmodified bare Zorbax silica (3.5 μm, 80 Å pore size) (Agilent Technologies, Santa Clara, CA, USA) into a small column (20 mm x 2.1 mm i.d.). Mobile phases employed were: A) Ammonium formate 30 mM, B) AcN. ASM step was set to 0.22min, ASM factor at 5, and flow rate at 2 mL min^{-1}, with a split ratio of 1:2 (one part to MS, two parts to waste). Gradient was established as follows (percentages of B):0 min (98), 0.22 (98), 0.85 (80), 0.97 (60), 0.98 (98). In this example, four EICs of lipids are shown: 16:0 D31-18:1 PE, 16:0 D31-18:1 PG, 16:0 D31-18:1 PC, and 1,2,3-17:0 TG, purchased from Avanti Polar Lipids (Merck KGaA, Darmstadt, Germany).

All chromatograms from parts A, B and C were acquired with a 6545XT AdvanceBio LC/Q-TOF (Agilent Technologies, Santa Clara, CA, USA) mass spectrometer with an Agilent JetStream (AJS) electrospray ionization source in positive mode. Full scan spectra were acquired from 150 to 1500 Da, with an acquisition frequency of 125 ms per spectrum and a resolution of over 20000 FWHM in the *m/z* lipid range.

Annex 2. Experimental conditions employed for the acquisition of the chromatogram shown in Figure 4.4.

The RP × HILIC-HRMS chromatogram shown in **Figure 4.4** was included by Pérez-Cova et al. in the publication about the MSroi GUI as pre-processing step in all types of MS data [1]. The experimental conditions in which this chromatogram was acquired are summarized below. Visualization in **Figure 4.4** was performed with GC Image™ LC × LC version software (GC Image, LLC, Lincoln, USA).

In this example, a mixture of nine lipids (17:0 MG, 17:0 Lyso PA, 17:1 Lyso PE, 17:1 Lyso PG, 17:1 Lyso PS, 17:0 Lyso PC, 1,3-17:0 D5 DG, 17:0 cholesteryl ester, 16:0 D31-18:1 PE, 16:0 D31-18:1 PG, 16:0 D31-18:1 PC, 16:0 D31-18:1 PS and 1,2,3-17:0 TG) was analyzed to evaluate the distribution of these compounds into the two-dimensional chromatographic space. LC × LC analysis was carried out using a combination of chromatographic modes configurated in an RP × HILIC setting. For the first-dimension separation, an Agilent Poroshell HPH C18 (150 mm x 2.1 mm i.d., 1.9 μm) (Agilent Technologies, Santa Clara, CA, USA) using a flow rate of 40 $\mu L \cdot min^{-1}$ and a temperature of 50 ºC. The two mobile phases were: A) 30 mM ammonium formate (pH 4.5); B) ACN/IPA 1:2 (v/v). Elution gradient was as follows (percentages of A): 0 min (30), 60 min (10), 75 min (0), 120 min (0), 121 min (30), 150 min (30) (re-equilibration). The total chromatographic analysis time was 150 min. In the second chromatographic dimension, a HILIC column was prepared in-house by slurry packing unmodified bare Zorbax silica (3.5 μ, 80 Å pore size) (Agilent Technologies, Santa Clara, CA, USA) into a small column (20 mm x 2.1 mm i.d.). The column was used at 40˚C, and a flow rate of 2 mL min^{-1}, with a split ratio of 1:2 (one part to MS, two parts to waste), by passive flow splitting using a simple T-piece before detection. Mobile phases consisted of A: 30 mM ammonium formate (pH 4.5); and phase B: ACN, with the following gradient (percentages of A): 0 min (0), 0.22 min (0), 0.8 min (35), 1.00 min (0), where the initial 0.22 min corresponds to active solvent modulation (ASM) step, and modulation time was 1 min. The ASM factor was set to 5. The mass spectrometer was a 6545XT AdvanceBio LC/Q-TOF (Agilent Technologies, Santa Clara, CA, USA) with an Agilent JetStream (AJS) electrospray ionization source in positive mode. Full scan spectra were acquired from 100 to 1500 Da, with an acquisition frequency of 125 ms per spectrum and a resolution of over 20000 FWHM in the *m/z* lipid range.

Annex 3. Experimental conditions employed for the acquisition of the chromatograms shown in Figure 4.5.

The experimental conditions in which the chromatograms of **Figure 4.5** were acquired, are described below. Visualization of the chromatograms shown in **Figure 4.5** were obtained with **A)** GC Image™ LC×LC version software (GC Image, LLC, Lincoln, USA); **B**) MassHunter Workstation Software, Qualitative Analysis Navigator, version B.08.00 (Agilent Technologies, Santa Clara, CA, USA).

A) HILIC (^{1}D) conditions: HILIC column was prepared in-house by slurry packing unmodified bare Zorbax silica (3.5 μm, 80 Å pore size) (Agilent Technologies, Santa Clara, CA, USA) into a small column (50 mm x 2.1 mm i.d.). Mobile phases composition: A) Ammonium formate 30 mM, B) AcN. ASM step: 0.22min, Temperature (column oven): 40 °C. Injection Injection volume: 4 μL. Gradient (percentage of B): 0(98), 60(89), 140(70), 170(40), 171(98), 200 (98) min.

RP (^{2}D) conditions: Carbon prototype column (30 x 2.1 mm i.d., 2.7 μm). Mobile phases composition: A) 30 mM Ammonium formate (pH 4.5), B) ACN. Temperature (column oven): 50 °C. Gradient (percentage of B): 0 (0), 0.44 (0), 0.87 (20), 1.07 (100), 1.22 (100), 1.3 (0), 1.5 (0) min. ASM step: 0.44 min. ASM factor 5. Modulation time: 1.5 min. Flow rate:1 mL min^{-1} with a split ratio of 1:1 (one part to MS, one part to waste).

B) Mobile phases composition: A) 30 mM Ammonium formate (pH 4.5), B) ACN. Temperature (column oven): 50 °C. ASM factor at 5. Injection volume: 2 μL (in 100% AcN).

Column 1 (left side): Carbon prototype column (30 x 2.1 mm, 2.7 μm). Gradient (percentage of B): 0 (0), 0.44 (0), 0.87 (20), 1.07 (100), 1.22 (100), 1.3 (0), 1.5 (0) min. ASM step: 0.44 min. Modulation time: 1.5 min. Injection volume: 2 μL (in 100% AcN). Flow rate:1 mL min^{-1}, with a split ratio of 1:1 (one part to MS, one part to waste).

Column 2 (right side): Zorbax SB-C18 (2.1x30mm, 3.5μm) (Agilent Technologies, Santa Clara, CA, USA). Gradient (percentage of B): 0 (2), 0.29 (20), 0.69 (100), 0.79 (100), 0.89 (2), 1 (2) min. ASM step: 0.29 min. Modulation time: 1 min. Flow rate:1.5 mL min^{-1} with a split ratio of 1:1 (one part to MS, one part to waste).

All chromatograms from parts A and B were acquired with a 6545XT AdvanceBio LC/Q-TOF (Agilent Technologies, Santa Clara, CA, USA) mass spectrometer with an Agilent JetStream (AJS) electrospray ionization source in positive mode. Full scan spectra were acquired from 150 to 1500 Da, with an acquisition frequency of 125 ms per spectrum.

References

[1] M. Pérez-Cova, C. Bedia, D.R. Stoll, R. Tauler, J. Jaumot, MSroi: A pre-processing tool for mass spectrometry-based studies, Chemometrics and Intelligent Laboratory Systems. 215 (2021). https://doi.org/10.1016/j.chemolab.2021.104333.