



UNIVERSITAT_{DE}
BARCELONA

Development and application of Nuclear Magnetic Resonance spectroscopy and chemometric methods for the analysis of the metabolome of *Saccharomyces cerevisiae* under different growing conditions

Francesc Puig Castellví



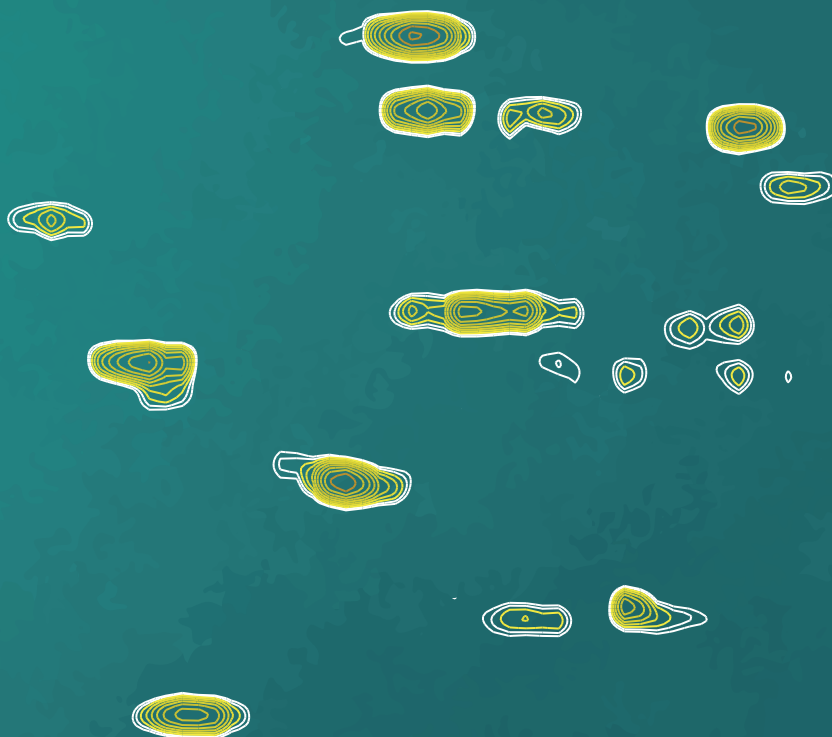
Aquesta tesi doctoral està subjecta a la llicència **Reconeixement- Compartiqual 3.0. Espanya de Creative Commons.**

Esta tesis doctoral está sujeta a la licencia **Reconocimiento - Compartiqual 3.0. España de Creative Commons.**

This doctoral thesis is licensed under the **Creative Commons Attribution-ShareAlike 3.0. Spain License.**

Development and application of Nuclear Magnetic Resonance spectroscopy and chemometric methods for the analysis of the metabolome of *Saccharomyces cerevisiae* under different growing conditions

Francesc Puig Castellví



UNIVERSITAT DE
BARCELONA



CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS



UNIVERSITAT_{DE}
BARCELONA

**Development and application of Nuclear Magnetic Resonance
spectroscopy and chemometric methods for the analysis of the
metabolome of *Saccharomyces cerevisiae* under different
growing conditions**

Francesc Puig Castellví



UNIVERSITAT DE
BARCELONA



CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

**DEVELOPMENT AND APPLICATION OF NUCLEAR
MAGNETIC RESONANCE SPECTROSCOPY AND
CHEMOMETRIC METHODS FOR THE ANALYSIS OF
THE METABOLOME OF *Saccharomyces cerevisiae* UNDER
DIFFERENT GROWING CONDITIONS**

Francesc Puig Castellví

Doctoral programme: “Química Analítica i Medi Ambient (HDK15)”

**DEVELOPMENT AND APPLICATION OF NUCLEAR
MAGNETIC RESONANCE SPECTROSCOPY AND
CHEMOMETRIC METHODS FOR THE ANALYSIS OF
THE METABOLOME OF *Saccharomyces cerevisiae* UNDER
DIFFERENT GROWING CONDITIONS**

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

Francesc Puig Castellví

Supervisors:

Prof. Romà Tauler Ferré

Department of Environmental Chemistry

Institute of Environmental Assessment and Water Research (IDAEA)

Spanish National Research Council (CSIC)

Dr. Ignacio Alfonso Rodríguez

Department of Biological Chemistry and Molecular Modelling

Institute of Advanced Chemistry of Catalonia (IQAC)

Spanish National Research Council (CSIC)

Tutor:

Prof. Anna Maria De Juan Capdevila

Department of Chemical Engineering and Analytical Chemistry (UB)

Barcelona, May of 2018

Prof. Romà Tauler Ferré, Professor of the Department of Environmental Chemistry of the Institute of Environmental Assessment and Water Research, and **Dr. Ignacio Alfonso Rodríguez**, Research Scientist from the Department of Biological Chemistry and Molecular Modelling of the Institute of Advanced Chemistry of Catalonia,

STATE THAT:

the current PhD report entitled “*Development and application of Nuclear Magnetic Resonance spectroscopy and chemometric methods for the analysis of the metabolome of *Saccharomyces cerevisiae* under different growing conditions*” has been elaborated under our supervision by Mr. **Francesc Puig Castellví** 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, May of 2018

Prof. Romà Tauler Ferré

Dr. Ignacio Alfonso Rodríguez

*“Errors using inadequate
data are much less than
those using no data at all”*

Charles Babbage, Mathematician

Agraïments

Primer de tot, vull agrair als meus dos directors de Tesi, Romà i Nacho, per haver-me donat l'oportunitat de realitzar aquesta tesi, per ser els meus guies durant aquests anys, i per tot el suport rebut durant aquesta etapa. Gràcies per tot el que m'heu ensenyat, per confiar amb mi, per la vostra paciència, i per estar allí sempre que ho he necessitat. De la mateixa manera, també vull agrair a la Dra. Anna de Juan Capdevila per haver acceptat ser la meva tutora durant aquesta Tesi; i a en Benjamí, li dono també les gràcies per haver estat el “tercer director”.

A tota la família de la quarta planta, us vull agrair els bons moments compartits. Us trobaré a faltar a tots. A la Carma (i les nostres converses), al Cristian (i els nostres “piques”), al meu company de riu l'Stefan (i el seu ampli coneixement), al Joaquim (i les seves visites al laboratori a les 16h20), al Marc (i les seves “coses”), i a la Míriam (i la seva energia). Als últims dos, gaudiu de l'experiència de la Tesi, però tampoc no us encanteu, que quatre anys passen molt ràpid. També vull agrair a la Laia i a l'Andrés pels bons moments viscuts. Gràcies també als qui ja va vau volar: Alejandro, Yahya i Igor. Menció especial a la Mireia, companya de racó i de tardes, que em va ajudar amb els inicis al grup (i amb el llevat). No paris mai de riure. Per últim, no em podia deixar a les “nenes” del grup (Elena, Elba, Núria, Meritxell i Eva) i al Víctor, gràcies per ser com sou. Vam començar alhora, i ens hem tingut uns als altres, en tot moment, però sobretot als descansos, als cafès de les 14, als cafès de les 12, les escapades a la fleca, les cerveses, els sopars, els berenars, les converses sobre inquietuds, plans, confidències i desesperacions, el viatge a Lisboa, el body step i el body pump. D'aquesta aventura m'emporto uns quants tresors, i vosaltres sou un d'ells. Podria dir molt més, però si començo no acabo. Moltes gràcies.

Del CSIC, també, però amb el despatx situat una mica més amunt, vull donar les gràcies a en Benjamí, una altra vegada, per totes les converses enriquidores que hem tingut. A la Marta, per estar sempre disposada a ajudar i a escoltar-me. I a en Rubén, per no saber dir que no a cap pla d'última hora.

Per acabar amb la gent del CSIC, vull donar les gràcies a la Yolanda per tota la seva dedicació, pels seus consells, i per ser capaç de resoldre'm tots els dubtes que m'han sorgit durant aquesta Tesi.

Agraeixo també a la gent de fora del CSIC ha jugat un paper important: les companyes de pis (Mireia, Sheila i Maria), els companys de pintxo-pàdel (Albert, Leire, Jenny, Guille), els amics de la URV (Sandra, Monty, Mireia, Pepe, Alex, Ana, Alba) i els de Riba-Roja (Abraham, Aïda, Albert, Andreu, Anna, Blai, Carlos, Cristina, Eva, Javi, Leire, Lidia, Mariol, Pilar, Xavi i Yolanda).

I would also like to acknowledge Dr. Jesús Angulo for the opportunity he gave me to learn more about NMR during my research stay in the University of East Anglia, and also to acknowledge Dr. Chris Hamilton for the enriching projects they let me to pursue in his laboratory. During this stay, I also met other unforgettable people that make the cold days much cozier. Juanca, the Spanish Post-doc; the always optimistic Valeria, Susanita, and Serena. To Serena, I am very grateful for all the support she gave me during the stay, at both personal and professional levels. In this list, I have to include also my two flatmates, Laura and Asli (and her Turkish coffees). I wish you the best in your PhD. Despite being in a different country, all of you made me feel like home. Thank you.

Finalment, no puc deixar d'agrair als més importants: a la meva família. A vosaltres, per tot el recolzament rebut durant aquests anys. Gràcies per haver estat sempre al meu costat.

INDEX

Resum	IV
Abstract	VI
Abbreviations	VIII
Notation	XII
CHAPTER 1: Objectives and Thesis structure	1
1 SCOPE AND OBJECTIVES	3
2 THESIS STRUCTURE	5
3 LIST OF SCIENTIFIC PUBLICATIONS	7
CHAPTER 2: Introduction	9
1 HISTORY OF NUCLEAR MAGNETIC RESONANCE	11
2 NMR METABOLOMICS	13
3 WORKFLOW OF NMR METABOLOMICS STUDIES	14
3.1 SAMPLE PREPARATION	16
3.2 SPECTRA ACQUISITION	17
3.3 PREPROCESSING	20
3.4 NMR RESONANCES ASSIGNMENT	31
3.5 INTEGRATION	37
3.6 IMPORTING AND EXPORTING NMR DATA	45
4 CHEMOMETRICS	50
4.1 THE DEFINITION OF CHEMOMETRICS	50
4.2 DATA STRUCTURES	51
4.3 BILINEARITY	54
4.4 CHEMOMETRIC DATA ANALYSIS METHODS	57
5 ENVIRONMENTAL METABOLOMICS	78
5.1 <i>Saccharomyces cerevisiae</i> (YEAST)	78

CHAPTER 3: Current data analysis strategies for the investigation of ¹H NMR datasets. The *Saccharomyces cerevisiae* case-study **83**

1	INTRODUCTION	86
1.1	YEAST STRESS (TEMPERATURE AND STARVATION)	86
1.2	ANALYTICAL STRATEGIES USED	90
2	SCIENTIFIC RESEARCH	92
2.1	SCIENTIFIC ARTICLE I	92
2.2	SCIENTIFIC ARTICLE II	118
2.3	SCIENTIFIC ARTICLE III	152
3	DISCUSSION OF THE RESULTS	180
3.1	NMR IS A POWERFUL ANALYTICAL TECHNIQUE TO IDENTIFY BIOMARKERS	180
3.2	MCR-ALS HIGHLIGHTS THE AFFECTED METABOLIC PATHWAYS UNDER DIFFERENT STRESSES	185
3.3	BIOLOGICAL INTERPRETATION OF THE TEMPERATURE ADAPTATION	187
3.4	BIOLOGICAL INTERPRETATION OF THE STARVATION STRESS	190
4	CONCLUSIONS	193

CHAPTER 4: Development and application of data analysis strategies for the investigation of ¹H NMR metabolomics datasets **195**

1	INTRODUCTION	198
1.1	NUCLEAR RELAXATION AND RESONANCE WIDTH	198
1.2	NMR DATA	200
1.3	RESOLUTION OF NMR DATA BY CHEMOMETRICS: PREVIOUS WORK	203
1.4	PROPOSED CHEMOMETRIC STRATEGIES	205
2	SCIENTIFIC RESEARCH	208
2.1	SCIENTIFIC ARTICLE IV	208
2.2	SCIENTIFIC ARTICLE V	252
2.3	SCIENTIFIC ARTICLE VI	276

3	DISCUSSION OF THE RESULTS	301
3.1	MCR-ALS AND NMR DATA	301
3.2	MCR-ALS AS A BIOMARKER DETECTION TOOL	303
3.3	MCR-ALS AS A RESONANCES INTEGRATION TOOL	307
3.4	NOISE INFLUENCES THE RANK OF 2D NMR DATA	309
3.5	VOI APPROACH IS A ROBUST 2D INTEGRATION METHOD	312
3.6	^1H NMR AND ^1H - ^{13}C HSQC NMR METABOLOMICS: DOES THE DIMENSIONALITY MATTER?	314
4	CONCLUSIONS	317
	CHAPTER 5: Conclusions	319
	CHAPTER 6: References	325

Resum

L'espectroscòpia de ressonància magnètica nuclear (RMN) és capaç de generar mitjançant una mesura simple i directa una gran quantitat d'informació química. Tanmateix, aquesta informació no sempre és fàcil d'interpretar. De fet, la complexitat de l'anàlisi espectral és proporcional al nombre de compostos presents en la mostra analitzada, ja que les ressonàncies dels diferents compostos es troben superposades. Una de les situacions més extremes la podem trobar en el cas dels espectres de RMN de mostres obtingudes en estudis de metabolòmica, en les que es poden arribar a detectar al voltant d'una cinquantena de compostos en una sola mesura.

En l'estudi dels processos químics relacionats amb els metabòlits (metabolòmica), els espectres de RMN més utilitzats són els espectres monodimensionals de protó (1D ^1H), ja que són relativament ràpids d'adquirir i la sensibilitat del protó és la més alta. És també corrent utilitzar en estudis de metabolòmica els espectres de RMN bidimensionals ^1H - ^{13}C heteronuclears de coherència quàntica única (2D ^1H - ^{13}C HSQC), els quals permeten obtenir una millor caracterització estructural dels metabòlits detectats.

En aquesta Tesi, s'han desenvolupat diferents estratègies d'anàlisi d'espectres de RMN de ^1H i de ^1H - ^{13}C HSQC de mostres de metabolòmica. Els espectres de RMN van ser adquirits d'extractes de llevat *Saccharomyces cerevisiae* que prèviament havia estat exposat a diferents pertorbacions mediambientals. L'objectiu d'aquests estudis ha estat millorar la comprensió dels diferents processos metabòlics que regulen l'aclimatació de les cèl·lules de llevat a diferents condicions de creixement.

A partir d'aquests estudis de metabolòmica realitzats, es van dissenyar noves estratègies i protocols d'anàlisi de dades de RMN que inclouen la seva importació, el seu preprocessament, l'assignació de les ressonàncies i la seva integració. A més, es van aplicar diferents mètodes quimiomètrics que van permetre identificar els biomarcadors de l'estat metabòlic de les cèl·lules del llevat i extreure els principals perfils metabòlics que descriuen els canvis en el seu metabolisme. Es van proposar a més, dues estratègies quimiomètriques per a l'anàlisi no dirigida d'espectres de RMN de ^1H i de ^1H - ^{13}C HSQC, respectivament.

En el cas dels estudis d'espectres de RMN de ^1H , l'aplicació del mètode de resolució multivariant de corbes per mínims quadrats alternats (MCR-ALS) va permetre resoldre satisfactòriament les concentracions i els espectres individuals dels diferents metabòlits.

D'altra banda, la investigació de l'estructura de les dades dels espectres de RMN de ^1H - ^{13}C HSQC va revelar que la majoria dels valors espectrals són descriptius del soroll, cosa que dificulta la seva anàlisi. En aquest context, s'ha desenvolupat una nova estratègia per filtrar

les variables descriptives del soroll, anomenada selecció de les variables d'interès (Variables of Interest, VOI). Després d'aplicar aquest procediment, es va observar que l'anàlisi dels espectres ^1H - ^{13}C HSQC filtrats produeix resultats similars als obtinguts amb els espectres corresponents de ^1H . Degut a l'existència de la segona dimensió en els espectres de ^1H - ^{13}C HSQC, les ressonàncies estan menys solapades i es poden integrar sense fer servir estratègies basades en la seva deconvolució. Degut a tot això i al fet que els espectres de ^1H - ^{13}C HSQC contenen més informació química que els de ^1H , l'anàlisi dels espectres de ^1H - ^{13}C HSQC filtrats amb aquest procediment permet una caracterització del sistema metabolòmic més acurada i amb temps d'anàlisis més curts, en comparació a l'anàlisi dels espectres de ^1H corresponents.

Abstract

Nuclear Magnetic Resonance (NMR) spectroscopy is able to produce by a single direct measurement a very high amount of chemical information. However, this information is not always easy to interpret. In fact, the complexity of the NMR spectral data analysis is proportional to the number of compounds present simultaneously in the analyzed sample, as resonances from different compounds overlap. One of the most extreme situations can be found for NMR spectra of samples from metabolomics studies, from which approximately fifty compounds can be detected in a single measurement.

In the study of the chemical processes involving metabolites (metabolomics), the most commonly used NMR spectra are the one-dimensional proton (1D ^1H) NMR spectra, since they are relatively fast to acquire and proton sensitivity is the highest. The ^1H - ^{13}C Heteronuclear Single Quantum Coherence (HSQC) NMR spectra are also frequently used in metabolomics for an improved structural characterization of the detected metabolites.

In this Thesis, we have developed different data analysis strategies of ^1H NMR and ^1H - ^{13}C HSQC NMR metabolomics datasets. The investigated NMR spectra were acquired from extracts of *Saccharomyces cerevisiae* cells previously exposed to different environmental perturbations. The aim of these studies was to better understand the different metabolic processes that regulate the yeast metabolism acclimation to different growing conditions.

From the study of these NMR metabolomics experiments, we designed new strategies and protocols for the analysis of these datasets that include the steps of data import, data pre-treatment, resonance assignment and metabolite quantification. Moreover, different chemometric methods were applied for the identification of the possible biomarkers that define the metabolic states of yeast cells and to extract the main metabolic profiles that describe the observed changes in the metabolome. Furthermore, two chemometric strategies were proposed for the untargeted analysis of ^1H NMR and ^1H - ^{13}C HSQC NMR, respectively.

For the study of ^1H NMR spectra of metabolomics samples, the application of the Multivariate Curve Resolution–Alternating Least Squares (MCR-ALS) chemometric method allowed the satisfactory resolution of the individual ^1H NMR spectra and concentrations of the different metabolites.

On the other hand, the investigation of metabolomics datasets by ^1H - ^{13}C HSQC NMR revealed that most of the data values in these NMR spectra are only descriptive of noise, hampering their chemometric data analysis. In this context, a new strategy to filter the variables relative to noise, named ‘Variables of Interest’ (or VOI) is proposed. After the application of this procedure, we observed that the analysis of the noise-filtered ^1H - ^{13}C HSQC

NMR spectra produced similar results to the corresponding analysis of ^1H NMR spectra. Due to the existence of the second dimension in the ^1H - ^{13}C HSQC NMR spectra, resonances are less overlapped and they could be integrated without using deconvolution approaches. For all these reasons, and linked to the fact that more chemical information is contained in the ^1H - ^{13}C HSQC NMR spectra than in the ^1H NMR spectra, the analysis of noise-filtered ^1H - ^{13}C HSQC NMR spectra allow a more accurate characterization of the metabolomic system, in a reduced amount of time in comparison to the analysis of the corresponding ^1H NMR spectra.

Abbreviations

^1H	Proton
$^1\text{H NMR}$	Proton Nuclear Magnetic Resonance
^{13}C	Carbon-13
^{15}N	Nitrogen-15
^{19}F	Fluor-19
2D NMR	two-dimensional Nuclear Magnetic Resonance
^{31}P	Phosphorus-31
ALS	Alternating Least Squares
AMP	Adenosine 5'-monophosphate
AMS	Absolute Minimal Sampling
ANOVA	ANalysis Of Variance
ASCA	ANOVA-Simultaneous Component Analysis
B_0	Magnetic field
BATMAN	Bayesian AuTomatic Metaoblite Analyser for NMR spectra
B_i	Induced magnetic field
BML-NMR	Birmingham Metabolite Library of Nuclear Magnetic Resonance spectra
BMRB	Biological Magnetic Resonance Data Bank
B_{nucleus}	Magnetic field applied to the nucleus
c	concentration
C	Matrix of concentrations
Cer	Ceramide
CH_3Cl	Chloroform
CH_3OD	Methanol with exchangeable protons exchanged with deuterium
CLS	Classical Least Squares
COSY	Correlation Spectroscopy
COW	Correlation Optimized <i>Warping</i>
CPMG	Carr-Purcell-Meiboom-Gill
CV	Cross Validation
CW	continuous-wave
D	Matrix of data
DG	Diacylglycerol
DM	Drop-out medium
DOSY	Diffusion-ordered spectroscopy
DTC	Decision Tree of Correlations
DTC-MCR-ALS	Decision Tree of Correlations combined with MCR-ALS

DSS	4,4-Dimethyl-4-silapentane-1-sulfonic acid
<i>e</i>	residual
E	Matrix of the residual information
ECMDB	<i>Escherichia coli</i> Metabolome Database
ERETIC	Electronic REference To access In vivo Concentrations
ESI	Electrospray ionization
f_1	Direct dimension in a multi-dimensional NMR spectrum
f_2	First indirect dimension in a multi-dimensional NMR spectrum
f_3	Second indirect dimension in a multi-dimensional NMR spectrum
FA	Fatty acyls
FID	Free Induction Decay
FT	Fourier Transform
FT-NMR	Fourier Transformed Nuclear Magnetic Resonance
GC	Gas Chromatography
GC-MS	Gas Chromatography coupled to Mass Spectrometry
GP	Glycerophospholipids
HATS	Hierarchical Alignment of Two-dimensional Spectra
HDO	Water with protons partially exchanged with deuterium
HMBC	Heteronuclear Multiple Bond Correlation
HMDB	Human Metabolome Database
HPLC	High Performance Liquid Chromatography
HPLC-MS	High Performance Liquid Chromatography coupled to Mass Spectrometry
HSQC	Heteronuclear Single Quantum Coherence
Hz	hertz
<i>icoshift</i>	Interval Correlation Optimised Shifting
IN	Integral Normalization
<i>J</i>	Indirect <i>spin-spin</i> coupling
KEGG	Kyoto Encyclopedia of Genes and Genomes
$l(\delta)$	Lorentzian function of the resonance centered at δ
LC	Liquid Chromatography
LOF	Lack-Of-Fit
LV	Latent Variable
<i>m</i>	metabolites
<i>m/z</i>	mass-to-charge ratio
MANOVA	Multivariate ANOVA
MCR	Multivariate Curve Resolution

MCR-ALS	Multivariate Curve Resolution by Alternating Least Squares
MHz	Megahertz
MQMCD	Madison-Qingdao Metabolomics Consortium Database
MRI	Magnetic Resonance Imaging
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
NMR	Nuclear Magnetic Resonance
NMC	Number of misclassification
NOESY	Nuclear Overhauser Effect Spectroscopy
OD600	Optical density at 600 nm
OSC	Orthogonal Signal Correction
OSC-PLS-DA	Orthogonal Signal Correction-Partial Least Squares-Discriminant Analysis
P	Loadings matrix
PC	Principal Component
PCA	Principal Component Analysis
PCR	Principal Component Regression
PGSE	Pulsed Field Gradient Echo
PLS	Partial Least Squares
PLS-DA	Partial Least Squares-Discriminant Analysis
<i>ppm</i>	Parts per million
PQN	Probabilistic Quotient Normalization
PRESAT	presaturation
PRESS	Predicted Residual Sum of Squares
PRIME	Platform for RIKEN Metabolomics
qNMR	Quantitative Nuclear Magnetic Resonance
R ²	Percentage of explained variance
RD	Relaxation delay
RF	Radio frequency
RMSEC	Root Mean Square Error of Calibration
RMSECV	Root Mean Square Error of Cross Validation
ROI	Regions of Interest
SCA	Simultaneous Component Analysis
SIMPLISMA	SIMPLe-to-use Iterative Self-Modeling Analysis
S	Matrix of Spectra
SM	Submatrix
SNR	Signal-to-noise ratio
SNV	Standard Normal Variate

SVD	Singular Value Decomposition
T	Scores matrix
T_1	<i>spin-lattice</i> relaxation
T_2	<i>spin-spin</i> relaxation
t_{acq}	Acquisition time
t_{mu}	spin-spin splitting transition of a <i>u</i> resonance from a <i>m</i> metabolite.
TG	Triacylglycerol
TIC	Total Ion Current
TOCSY	Total Correlation Spectroscopy
TOF	Time of Flight
TMS	Tetramethylsilane
TSP	trimethylsilylpropanoic acid
<i>u</i>	resonances
UHPLC	Ultra High Performance Liquid Chromatography
UHPLC-MS	Ultra High Performance Liquid Chromatography coupled to Mass Spectrometry
UHPLC-MS-ROI	ROIs for a sample or set of samples acquired in a UHPLC-MS instrument
VIP	Variable Important on Projection
VOI	Variables of Interest
w	weight vector
WET	Water Excitation Technique
YMDB	Yeast Metabolome Database
YNB	Yeast Nitrogen Base
YPD	Yeast Peptone Dextrose
YSC	Yeast Synthetic Complete
γ	Gyromagnetic ratio
δ	Chemical shift
Φ_0	zero-order phase correction
Φ_1	first-order phase correction
ν	fundamental resonance frequency
$\nu_{1/2}$	width at half height
σ	shielding
ω	frequency

Notation

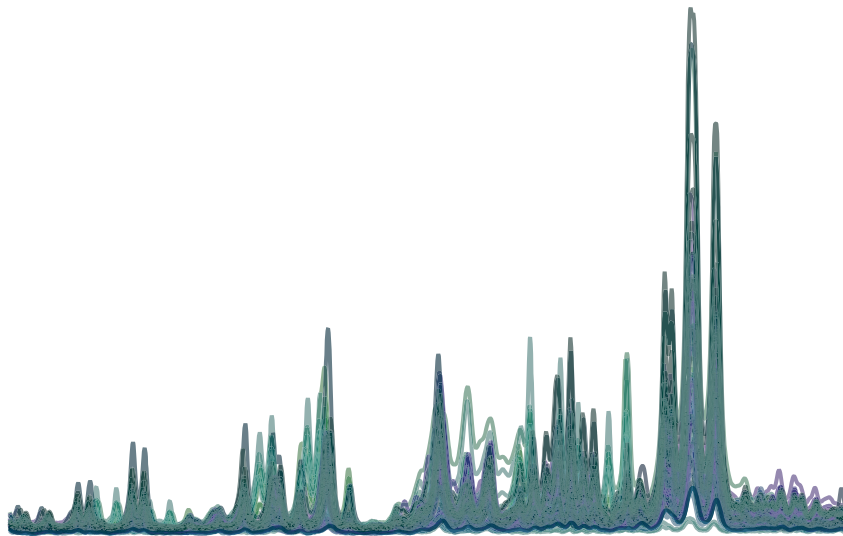
In this section the mathematical and gene notations used in this thesis are presented. These conventions are the ones commonly accepted by the scientific community.

Regarding the mathematical notation, italic lowercase letters (*e.g.*, x) indicate scalars, bold lowercase letters (*e.g.*, \mathbf{x}) indicate vectors, and bold uppercase letters (*e.g.*, \mathbf{X}) indicate matrices. The mean of a variable is indicated with an overline (*e.g.*, \bar{x}). The transposition of a vector or matrix is symbolized by a superscripted “T” (*e.g.*, \mathbf{X}^T). The inverse of a matrix is symbolized by a superscript “-1” (*e.g.*, \mathbf{X}^{-1}). A matrix enclosed by two verticals bars on each side (*e.g.*, $\|\mathbf{X}\|$) represents the square root of the sum of squares of the given matrix.

Regarding the gene notation, genes are indicated with italic uppercase letters (*e.g.*, *URA3*). The loss-of-function of the gene is indicated in italic lowercase letters, followed by the “ Δ ” symbol (*e.g.*, *ura3 Δ*). The protein resulting the transcription and translation of the gene is indicated in roman letters with an initial capital letter, followed by the suffix “p” (*e.g.*, Ura3p).

Chapter 1

Objectives and Thesis structure



1 SCOPE AND OBJECTIVES

Nowadays, Nuclear Magnetic Resonance (NMR) spectroscopy has become one of the most preferred instrumental techniques in different scientific fields, since it allows obtaining qualitative and quantitative chemical information directly from measurements. For this reason, NMR spectroscopy (and more specifically, 1D ^1H NMR) is a commonly chosen technique for the analysis of complex samples, such as metabolomics samples.

Despite being so informative, NMR spectra are sometimes difficult to interpret because resonances from several dozens of compounds are overlapped, hampering the identification process of the sample constituents. With two-dimensional NMR, the spectral overlapping can be reduced, but metabolomics studies usually do not take advantage of these NMR data because they present a lower sensitivity, they need longer acquisition times, and their processing is more demanding than for 1D ^1H NMR data.

Moreover, with (NMR) metabolomics, information from several metabolites are obtained simultaneously for each sample, and understanding this information from a biological point of view can be challenging, not only because of the large amount of data information available, but also because the measured metabolite perturbations are the direct consequence of different biological processes going on simultaneously in the studied organism.

Considering these situations, the two general goals of this Thesis are:

- To develop new data analysis strategies based on chemometric methods to investigate NMR metabolomics datasets and extract biochemical knowledge from them.
- To study the effects of environmental perturbations on the metabolome of yeast as a representative biological organism by using the proposed chemometrics strategies.

These two general goals can be divided into specific (analytical and biological) goals, presented below:

Analytical goals

- To design metabolomics experiments that provide relevant information from the metabolic state of the investigated representative organisms.
- To develop a protocol to prepare NMR metabolomics samples from representative organisms.
- To define a data analysis workflow to study these NMR samples. This workflow will include the data import of NMR spectra acquired in Bruker or Varian NMR

instruments to MATLAB[®], the alignment and normalization of the NMR data, the resonance assignment and integration, the investigation of the dataset with chemometric methods, such as Principal Component Analysis (PCA) and Partial Least Squares – Discriminant Analysis (PLS-DA), and the export of the processed data to the NMR suites TopSpin[®] (for Bruker NMR data) or MestReNova (for Bruker and Varian NMR data).

- To investigate the differences in the data analysis of 1D (specifically, ¹H) NMR and 2D (specifically, ¹H-¹³C HSQC) NMR metabolomics datasets.
- To propose new chemometric-based methods to improve the analysis of 1D and 2D NMR metabolomics datasets.
- To establish a data-analysis strategy to fuse data from NMR spectroscopy with data obtained from other analytical platforms with the aim of generating a more comprehensive characterization of the metabolic perturbations in the studied organisms.

Biological objectives

- To evaluate and interpret the metabolic response of *Saccharomyces cerevisiae* cells (strain BY4741) to temperature acclimation.
- To evaluate and interpret the metabolic response of *Saccharomyces cerevisiae* cells (strain BY4741) cultured in four different drop-out media (lacking L-leucine, L-methionine, L-histidine or uracil in the used media) over time.
- To evaluate and interpret the metabolic response of *Saccharomyces cerevisiae* cells (strain S288C) cultured in two different growth conditions (minimal and rich media) over time.

2 THESIS STRUCTURE

This Thesis is structured in six chapters that are described below:

In the first chapter, the aim of this Thesis, its structure and the list of scientific publications derived from this work are presented.

In the second chapter, a background introduction of NMR spectroscopy, metabolomics, and NMR metabolomics is given, and the workflow of a typical NMR metabolomics study is detailed. The chemometric methods used in NMR metabolomics are reviewed. A focus in environmental metabolomics is given, and the usefulness of *Saccharomyces cerevisiae* as a representative model organism in environmental metabolomics studies is explained.

In the third chapter of the Thesis, the metabolic response of yeast when exposed to two different environmental stresses, temperature acclimation and nutrient starvation, is investigated. At the beginning of this chapter, the concepts concerning cell cycle regulation are introduced, and precedent published work regarding these two stresses is presented. In addition, different strategies used to assign resonances from the NMR metabolomics datasets are shown and discussed. Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS) is presented as a compelling method for untangling the set of underlying biological processes that occur simultaneously in the metabolism of the studied organisms. This strategy was applied to two datasets. The first dataset consists in a time-series experiment of yeast cells cultured in five different media (one standard and four starvation conditions). The second dataset contains the relative concentrations of the metabolites (including lipids) from yeast cells at four different temperatures. Results from these two MCR-ALS analyses are presented and compared with the results from the precedent literature. Finally, the specific conclusions that can be drawn from the results discussed in this chapter are given.

In the fourth chapter of the Thesis, concepts regarding NMR relaxation are introduced. The quality of the NMR spectra is investigated from a data analysis point of view, covering aspects of spectral resolution, resonance width, signal-to-noise ratio and data multidimensionality. Precedent work regarding the resolution of complex NMR spectra by chemometric methods is presented. The particularities of NMR metabolomics data, specifically related to signal overlapping and to inter-sample metabolic variance, are explained. The application of the MCR-ALS method using spectral window constraints is proposed as a feasible approach to resolve ^1H NMR metabolomics datasets into the set of pure concentrations and ^1H NMR spectra of the pure metabolites. The suitability of this method is investigated with two simulated and one real ^1H NMR datasets of yeast metabolic extracts. In addition, it is demonstrated that the vast amount of noise in 2D NMR spectra hinders their data analysis. In this respect, a noise-filtering approach for 2D (and 3D) NMR

spectra is proposed. The differences between 1D NMR and 2D NMR metabolomics are mentioned and compared. Finally, specific conclusions drawn from the set of results discussed in this chapter are given.

In the fifth chapter of this Thesis, the main conclusions resulting from the present work are presented.

Finally, in the sixth and final chapter of this Thesis, the references of the publications mentioned in the previous chapters are given.

3 LIST OF SCIENTIFIC PUBLICATIONS

The work performed in this Thesis resulted in six scientific publications, which are listed in chronological order below:

A quantitative ^1H NMR approach for evaluating the metabolic response of *Saccharomyces cerevisiae* to mild heat stress.

Authors: Puig-Castellví F., Alfonso I., Piña B., Tauler R.

Citation reference: *Metabolomics* (2015). 11:1612–1625.

DOI: 10.1007/s11306-015-0812-9

^1H NMR metabolomic study of auxotrophic starvation in yeast using Multivariate Curve Resolution-Alternating Least Squares for Pathway Analysis.

Authors: Puig-Castellví F., Alfonso I., Piña B., Tauler R.

Citation reference: *Scientific Reports* (2016), 6:30982.

DOI: 10.1038/srep30982

Untargeted Assignment and Automatic Integration of ^1H NMR metabolomic datasets using a Multivariate Curve Resolution Approach.

Authors: Puig-Castellví F., Alfonso I., Tauler R.

Citation reference: *Analytica Chimica Acta* (2017), 964:55-66.

DOI: 10.1016/j.aca.2017.02.010

Compression of multidimensional NMR spectra allows a faster and more accurate analysis of complex samples.

Authors: Puig-Castellví F., Pérez Y., Piña B, Alfonso I., Tauler R.

Citation reference: *Chemical communications* (2018), 54:3090-3093.

DOI: 10.1039/C7CC09891J

Deciphering the underlying metabolomic and lipidomic patterns linked to thermal acclimation in *Saccharomyces cerevisiae*.

Authors: Puig-Castellví F., Bedia C., Alfonso I., Piña B., Tauler R.

Citation reference: *Journal of Proteome Research* (2018)

DOI: 10.1021/acs.jproteome.7b00921

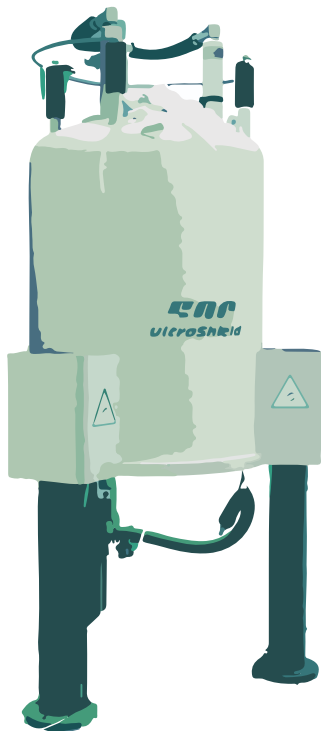
Comparative analysis of ^1H NMR and ^1H - ^{13}C HSQC NMR metabolomics to understand the effects of medium composition in yeast growth.

Authors: Puig-Castellví F., Pérez, Y., Piña B., Tauler R., Alfonso I.

Citation reference: *Submitted*

Chapter 2

Introduction



1 HISTORY OF NUCLEAR MAGNETIC RESONANCE

Nuclear Magnetic Resonance (NMR) is nowadays a powerful analytical tool for organic synthesis [1], monitoring kinetics [2], structural biology [3], diagnostics [4] and, of course, metabolomics [5], among others. However, NMR origins emerged from the physics realm, grounded in electricity, magnetism, classical mechanics and quantum mechanics.

NMR was discovered prior confirmation of the nuclear spin existence, made by Stern and Gerlach in 1933 [6]. Six years later, Rabi demonstrated the principle of NMR: the nuclear moment can be measured if the particles are subjected to a homogeneous magnetic field and irradiated by a radiofrequency electromagnetic energy [7]. These two important discoveries were recognized in 1943 and in 1944, respectively, with the Nobel Prize in Physics [8].

Despite the outstanding revelation made by Rabi, his approach had a limited application, since only nuclei from small molecules could be analyzed. In the following decade, Purcell and Block groups took a different approach.

In this approach, they applied a radiofrequency magnetization energy to bulk materials, and the nuclear magnetization could be rotated away from its equilibrium parallel to the applied magnetic field, and then precess about the magnetic field at a well-defined frequency (ω). This experiment was tested by Bloch with water [9], while Purcell did the same with paraffin [10]. The importance of this discovery was recognized again in 1952 with the Nobel Prize in Physics [11].

It was first assumed that the relationship between the resonance frequency, ω , and the magnetic field applied to the nucleus, B_{nucleus} , was constant. This constant was named the gyromagnetic ratio, γ (**eq. 2.1**).

$$\omega = \gamma B_{\text{nucleus}} \quad \text{eq. 2.1}$$

However, important deviations of the predicted values in further experiments measuring ^{19}F and ^{31}P led to postulate that the magnetic properties of the electrons surrounding the nucleus provide a shielding, σ , of the applied magnetic field, and that this shielding depends on the density and electron configuration (**eq. 2.2**) [11].

$$\omega = \gamma B_0(1 - \sigma) \quad \text{eq. 2.2}$$

This shift in the resonance frequency was called chemical shift, δ , and was expressed in parts per million (ppm) relative to the fundamental resonance frequency, ν (**eq. 2.3**).

$$\omega = \nu(1 + \delta \times 10^{-6}) \quad \text{eq. 2.3}$$

The first NMR instruments operate in a continuous-wave (CW) mode, in which the spectrum is recorded by slowly changing the irradiation frequency. Then, when the frequency passed

through a resonant frequency for a particular nucleus in the sample, the oscilloscope recorded a peak in the spectrum. After improvements in the homogeneity and stability of magnetic fields, the chemical shifts for ^1H was demonstrated in 1951, with the acquisition of the ^1H NMR spectra of ethanol [12]. Further improvements in resolution revealed that the detected resonances are, in some cases, formed by sets of resonance lines, which derived from the foundations of the concept of indirect *spin-spin* coupling.

In 1957, Lowe and Norberg postulated that the nuclei precession after a 90° pulse, measured as a Free Induction Decay (FID) spectrum, can be transformed into a spectrum of resonances line-shapes after application of the Fourier Transform (FT) [13] and, in 1966, Richard Ernst finally made it into practice [14]. The advantages of this method to CW NMR were important: the entire spectrum could be recorded in a single scan in 2-3 s rather than the circa 5 min for the frequency sweep in the CW mode. Richard Ernst was awarded with the Nobel Prize in Chemistry in 1991 for developing FT-NMR spectroscopy [15].

In the mid-1960s, the superconducting magnets appeared, pushing forward the NMR field. The first 500 MHz NMR spectrometer was introduced around 1978, whereas the first commercial 600 MHz instrument appeared in 1987. The introduction of superconducting magnets allowed to distinguish between resonances that would be coincident at the lower magnetic field, enabling the analysis of complex chemical compounds, such as proteins and oligonucleotides. In the next decade, two-dimensional (2D) NMR spectroscopy appeared [16], which allowed an even better characterization of the measured compounds. Pushed by these latest advances, many research groups jumped on the analysis of biological samples [17,18], which could be considered as the first steps of NMR metabolomics.

NMR is still in ongoing development, reflected by the fact that two more Nobel Prizes were conceded in the XXIst century because of advances in the NMR field: a Nobel Prize in Chemistry was awarded in 2002 to Kurt Wüthrich for his NMR methods to study biological macromolecules [19], and another in Physiology or Medicine in 2003 to Lauterbur and Mansfield for developing Magnetic Resonance Imaging (MRI) [20].

2 NMR METABOLOMICS

In 1999, Jeremy Nicholson defined Metabolomics as [21]:

“The quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification”.

In this word, the use of the suffix ‘-omics’ served to give a complementary insight to the other ‘Omics’ sciences, such as Genomics (the study of genomes) and Proteomics (the study of proteomes). Metabolomics, in contrast to the other ‘Omics’ sciences, provides a closer understanding of the cellular functions of the living systems since it does not ignore the dynamics of the metabolism.

High-resolution ^1H NMR spectroscopy is particularly appropriate for the investigation of the metabolic states of an organism since a wide range of metabolites can be quantified simultaneously with minimal or no sample preparation. Other techniques such as Mass Spectrometry (MS) may also be useful, but differential ionization efficiencies may affect detectability and quantitation in some circumstances.

NMR metabolomics has proven to be a powerful tool to study the developmental stages of organisms [22,23], to reveal the metabolomic characteristics associated to a specific genotype [24,25], and to evaluate metabolic changes due to environmental or external factors [26,27]. Due to its versatility, NMR metabolomics has become a rapidly growing area of research, playing an important role in the studies of complex mixtures of small biological molecules, their metabolic networks, and their interactions with biomacromolecules [28]. Furthermore, in the field of environmental sciences, NMR metabolomics has resulted in a notable breakthrough for the assessment of environmental stressors to a variety of organisms, since it enabled the identification of perturbations in the metabolic networks of these organisms, transforming our fundamental understanding of the impacts of these environmental stressors [29].

3 WORKFLOW OF NMR METABOLOMICS STUDIES

The generation of biological information from the NMR spectra is not immediate. The biological results will depend first on the data analyses (*chemometrics*), which can be performed directly from the raw NMR spectral data or from the already ‘curated’ concentrations dataset of the detected metabolites. Nevertheless, in any of these two situations, the generation of any of these two datasets is neither easy nor straightforward.

Extracting conclusions from spectral data implies a shorter and faster workflow than the equivalent for concentrations datasets, since only the detected biomarkers will be assigned and the resonances integration step is then avoided. Having said this, it has been proven that the analysis of ‘curated’ concentrations data can provide more information than the analysis of the spectral data [30].

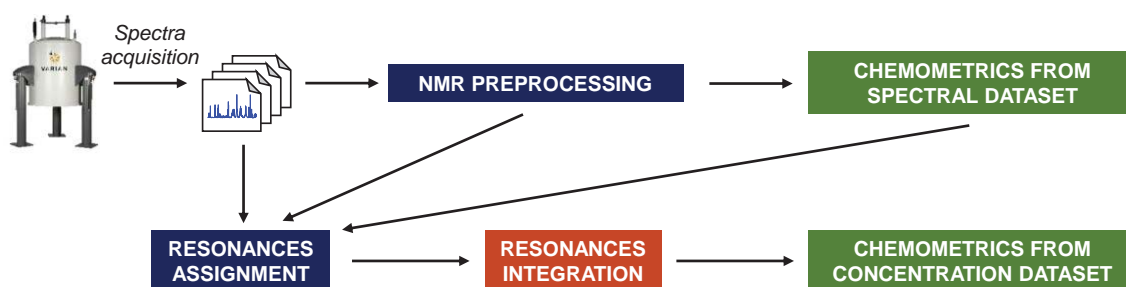


Figure 2.1. Workflow of an NMR metabolomics experiment.

In an NMR metabolomics study, metabolites are first extracted and dissolved in a deuterated solvent (or in a solvent containing a fraction of deuterated solvent), and then an FID is acquired from these prepared samples in the NMR spectrometer.

In order to obtain a proper spectral dataset, first the set of acquired FID NMR signals need to be Fourier transformed to generate the corresponding set of NMR spectra. Then, a set of NMR-specific preprocessing methods are used on these spectra to obtain better interpretable spectra. The NMR-specific preprocessing methods commonly applied are the reference to an NMR standard, apodization, phasing, and baseline correction, among others. Other more general preprocessing methods can also be applied on these spectral datasets, such as binning, peak alignment, normalization, and scaling.

The most important step in an NMR metabolomics study is the resonance assignment. This step may require the acquisition of complementary experiments (2D NMR homonuclear or heteronuclear) to provide a robust assignment. The resonance assignment can be extensively executed for all resonances in the (preprocessed or not) NMR spectra, or only focus on the resonances that were highlighted in the chemometric analysis of the spectral data.

These assigned resonances can be integrated, and these integrals can be directly interpreted as relative metabolite concentrations. There exist several programs that perform resonances integration (see **Figure 2.2**), and not all of them require an exact spectral alignment. Finally, the table with the curated resonance integrals can be investigated with univariate or multivariate data analysis methods, and conclusions from the metabolic fingerprint of every group of samples can be deduced.

NMR metabolomics is a multidisciplinary research field where NMR spectroscopy, biochemistry, and multivariate data analysis fields have converged. Nowadays, there is no single software that allows for the complete workflow analysis of the NMR spectra (see **Figure 2.2**).

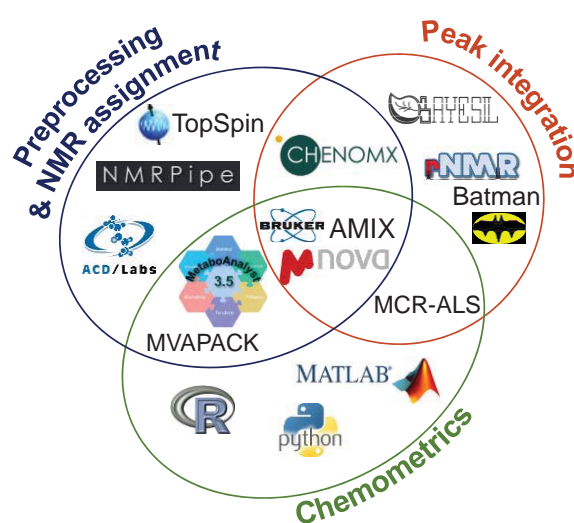


Figure 2.2. Computational tools used in NMR metabolomics. For specific details of the depicted software tools, see [31-48].

The most complete programs for the analysis of NMR spectra at present are MestreNova (MestreLab, Spain) and AMIX (Bruker Inc., US), both capable to preprocess spectra and integrate resonances, although their data analysis modules include only simple statistical data analysis methods.

In this Thesis, the NMR spectra have been preprocessed using both MestreNova software and scripts written in MATLAB[®] (Mathworks Inc., US) computer environment. Proton resonances have been integrated with the Bayesian AuTomed Metabolite Analyser for NMR spectra (BATMAN) [35,39] R-Package or using the Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) [49] chemometrics toolbox under the MATLAB[®] environment. Cross-peak resonances from ¹H-¹³C HSQC NMR spectra have been integrated with in-house scripts written in MATLAB[®]. Finally, most of the data analysis and processing

steps have been performed using different MATLAB[®] toolboxes (PLS Toolbox[™] and Bioinformatics Toolbox[™]) and R-packages (cluster [50], gplots [51], igraph [52]).

In the following sections, a more detailed explanation of the different steps of this workflow is provided.

3.1 SAMPLE PREPARATION

In metabolomics studies, biological changes are detected after characterization of a representative fraction of the metabolome.

To obtain an accurate measurement of this metabolome, it is important to use a robust sample preparation method. Different extraction methods have been proposed depending on the type of sample or metabolite class, and even several articles on the literature have been focused on designing an optimal sample preparation method. Examples of these optimization studies can be found for full organisms (bacteria [53,54], yeasts [55,56], nematodes [57], plants [58], plant seeds [59], worms [60]), tissues (brain [61,62], vein [63], muscle [61]), fluids (urine [64,65], serum [66]), and feces [67].

The sample preparation method can be divided into three different steps: the collection step, the extraction step, and the NMR sample preparation step.

Each organism or tissue has its own particularities and, thus, sample preparation protocol is organism-specific. For instance, in the collection step for microbial cell cultures, cells are collected by filtering or by centrifugation, washed with phosphate buffer saline to remove traces of medium, flash frozen in liquid N₂ and kept at -80°C [55]; while for solid samples such as tissues, they can be directly flash frozen in liquid N₂ and stored at -80°C [61].

It is important to design the protocol of the collection step in such a way that it does not introduce more variance on the studied system. For instance, in plants, all samples are usually collected at the same harvesting time (the same growth stage, and at the same period of the day) to reduce biological variability [58]. Moreover, in order to collect consistent samples, the enzymatic activity of the samples must be arrested. This can be performed with flash freezing, although freezing destroys the cells, and some biochemical conversions may occur after thawing. A more suitable technique is freeze-drying, which prevents sample degradation because the absence of water reduces enzymatic activity [68]. Freeze-dried samples yield flatter baselines because some unwanted macromolecules do not re-dissolve efficiently [61]. However, this step may lead to the irreversible adsorption of metabolites on cell walls and membranes [69]. To avoid these *ex vivo* degradations, denaturing solvents (*e.g.*,

organic solvents or solvents at 70-80°C [55]), microwaves [70], cold or acid treatments [71] can be used during the extraction.

To extract metabolites, cells need to be disrupted. For hard tissues, like plant stems and leaves, they are ground in a liquid N₂-cooled mortar and pestle, or homogenized with an electric tissue homogenizer [59]. For yeast cells, cell wall disruption can be achieved using a freeze and thaw strategy [55], glass beads [72], or sonication [56].

Due to the different physicochemical properties of the metabolites, there is no ideal method to simultaneously extract all classes of metabolites with high efficiency: polar organic solvents are typically mixed with water to extract hydrophilic metabolites, while chloroform can be used to extract hydrophobic metabolites [61]. An aqueous buffer extraction is sufficient to obtain a polar metabolite profile, but a more rigorous extraction involving a mixture of polar and nonpolar solvents is required to extract both polar and non-polar metabolites [73]. Evaluation of different extraction solvents with different biological samples has been performed in [74].

To increase sample stability, the obtained extracts are usually dried and stored at -80°C until the NMR analysis.

In the last step, the preparation of the NMR sample, extracts are dissolved in a deuterated solvent (deuterated phosphate buffer [30], deuterated methanol in deuterated phosphate buffer [27], deuterated chloroform [75]) that contains a NMR standard, such as DSS (4,4-Dimethyl-4-silapentane-1-sulfonic acid) or TSP (trimethylsilylpropanoic acid). Deuterated solvents are used to avoid the dominance of the solvent resonance over the spectrum and, in addition, the deuterium signal is employed to 'lock' the magnetic field strength and keep it from changing over time.

With the aim of reducing sample handling and increasing repeatability, single-step extraction methods have been proposed [76,77]. In single-step sample preparation for NMR metabolomics studies, the extraction solvents are deuterated, avoiding the need of freeze-drying the extract to remove the presence of non-deuterated solvents [76]. Then, the NMR standard used must be added after removal of proteins, because they tend to bind to proteins, leading to large variations in the quantification of metabolite concentrations [78].

3.2 SPECTRA ACQUISITION

After the sample is placed into the NMR spectrometer, previously to the spectra acquisition, the 'lock' and the 'shimming' are performed in order to generate good quality spectra. With the 'lock', the deuterium frequency is 'locked' and the magnetic field cannot drift even for

long-term acquisitions. On the other hand, with ‘shimming’, the best possible magnetic field homogeneity is achieved on all the sample.

Typically, in metabolomics studies, the conventional pulse sequence is used to acquire one-dimensional (1D) ^1H NMR spectra. This simple pulse sequence consists of a relaxation delay (RD), followed by a radiofrequency 90° pulse, and the acquisition time. With this experiment, the radio frequency (RF) pulse emitted by the coil of the probe surrounding the NMR spectrometer excites the proton nuclei, and after the RF pulse, the relaxation from this excited state is measured as a decaying sine wave during the acquisition delay time that represents this free induction decay. This pulse sequence is repeated a number of times (or scans) and added together in order to magnify the intensity of the measurements and to improve the signal-to-noise ratio.

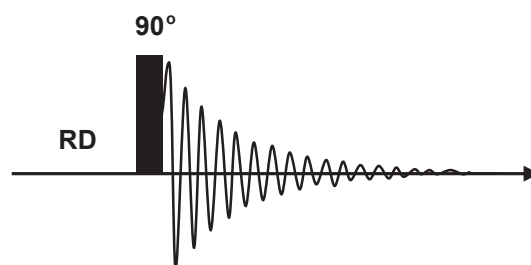


Figure 2.3. 1D ^1H NMR pulse sequence.

Other 1D ^1H NMR pulse sequences have been used in NMR metabolomics studies. For example, for plasma samples, the current standard procedure is to use the 1D Carr-Purcell-Meiboom-Gill (CPMG) pulse sequence, which allows the removal of broad protein signals. For urine samples, which contain vast amounts of non-deuterated water, pulse sequences that include solvent signal suppression schemes are preferred (*e.g.*, 1D- ^1H NOESY, WET, PRESAT) [79]. With this approach, dynamic range problems and baseline distortions in the spectrum region close to the water peak are minimized.

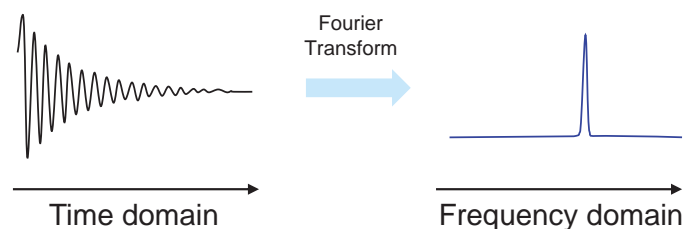
Two-dimensional (2D) NMR experiments have been also carried out to improve the characterization of the detected metabolites [80,81], and only a few number of articles in the literature evaluate the metabolic changes in the investigated biological organisms from these data [82-88]. A table summarizing the most used 2D NMR experiments in metabolomics, their purpose, and some references of interest is presented below.

Table 2.1. Most common 2D NMR experiments in metabolomics.

	x-axis	y-axis	Correlations or interactions established	References
^1H - ^1H COSY	$\delta(^1\text{H})$	$\delta(^1\text{H})$	Two non-equivalent ^1H nuclei coupled over two or three bonds.	[82,83]
^1H - ^1H TOCSY	$\delta(^1\text{H})$	$\delta(^1\text{H})$	Two non-equivalent ^1H nuclei connected within the same spin coupling system.	[89,90]
2D J -resolved	$\delta(^1\text{H})$	J_{HH}	Coupling constant (J_{HH}) and chemical shift (δ) information in two different axes.	[84,85,91]
^1H - ^{13}C HSQC	$\delta(^1\text{H})$	$\delta(^{13}\text{C})$	A ^1H and a ^{13}C nuclei connected through a C-H bond.	[86-88]
^1H - ^{13}C HMBC	$\delta(^1\text{H})$	$\delta(^{13}\text{C})$	A ^1H and a ^{13}C nuclei connected through two, three or four bonds.	[92,93]

2D NMR experiments can be classified as correlation experiments or resolved experiments. In correlation experiments, detected signals reflect the magnetization transfer between two nuclei. Correlation experiments are divided in homonuclear experiments (the two nuclei involved are of the same type), or heteronuclear (different type). On the other hand, in resolved experiments, the second dimension is used to plot a second variable relative to the same nuclei measured in the first dimension. Therefore, ^1H - ^1H COSY, ^1H - ^1H TOCSY are homonuclear correlation 2D NMR experiments; 2D J -resolved is a 2D NMR resolved experiment, and ^1H - ^{13}C HSQC and ^1H - ^{13}C HMBC are heteronuclear correlation 2D NMR experiments.

Regardless of the pulse sequence used, the detected data is always an FID function, which depends on time. By application of the FT, the time domain function is converted into a frequency domain function. In this new domain, every decay for every measured magnetically inequivalent nucleus can be spotted as a resonance.

**Figure 2.4.** Fourier Transform (FT).

One of the greatest challenges in NMR metabolomics lies in consistency and reproducibility on the acquired spectral data. To obtain consistent and reliable results, identical NMR tubes must be used as well as identical instrumental parameters and identical data processing steps for all the samples [64].

3.3 PREPROCESSING

The major goal of NMR metabolomics is to evaluate the abundance of metabolites from NMR data and to interpret variations in these abundances as alterations in metabolism of the investigated organisms. In order to achieve this goal, the acquired data need to be accurate. The quality of the data does not only depend on the experimental design, on the instrumentation and on the pulse sequences used, as stated in the previous section, but it can also be modified and improved by the applied data pretreatments.

There are several different NMR data preprocessing methods used to improve sensitivity, spectral resolution and peak shapes, to remove artifacts, and to align shifted resonances. These preprocessing tools can be either applied in the time-domain (in the FID spectrum) or in the frequency-domain (in the NMR spectrum). In addition, data can be also normalized or scaled with more general preprocessing methods. The aim of these second preprocessing methods is to obtain even more significant and representative results from the applied chemometric analyses or resonances integration strategies.

3.3.1 NMR Preprocessing methods applied before Fourier Transform

- 1) Zero-filling: the size of the FID can be artificially increased by adding zeros at the end of the measured data-points. However, since all nuclei complete their decay before finishing the acquisition of the FID, the added zero values would correspond to the expected measured values if the acquisition time would have been longer, and therefore, the added points have no effect on the peak positions, intensities, or linewidths of the spectrum. On the other hand, the increase of the number of data-points in the FID results in an increase of the digital resolution (fewer hertz per data point) in the spectrum after application of FT (**Fig. 2.5**). Normally, the number of added zeros is the same as the number of real points in the original FID and, as a result of this zero-filling operation, the spectral resolution is doubled.

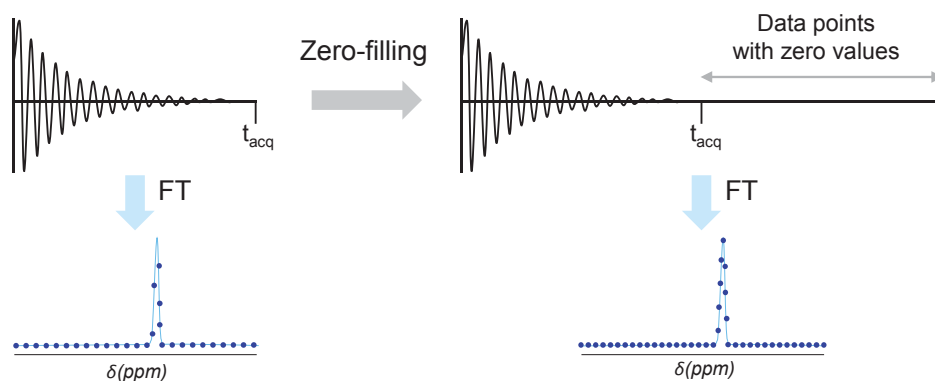


Figure 2.5. Zero-filling.

- 2) Apodization: the FID is multiplied by a weighting function, allowing to emphasize some parts of the spectrum at the expense of the others. After a certain acquired time, all nuclei have been completely decayed and all measured data points are representative of noise. The presence of noise in these last data-points in the FID only contributes to noise in the spectrum, therefore reducing the signal-to-noise ratio of the measured resonances. Exchanging these noise values by zero would introduce a sharp discontinuity in the FID, which could introduce artifacts into the spectrum. Instead, the FID can be multiplied by a negative exponential function that emphasizes the early data in the FID and deemphasizes the latter. This is a much smoother strategy than the exchange with zeros that increases the SNR, although the final resonances are broader and smaller in their absolute peak height (**Fig. 2.6**). It is also possible to use a weighting to produce the opposite effect: deemphasize the beginning of the FID and amplify the later part. After FT, resonances are sharper, resulting on an improvement of the resonance resolution.

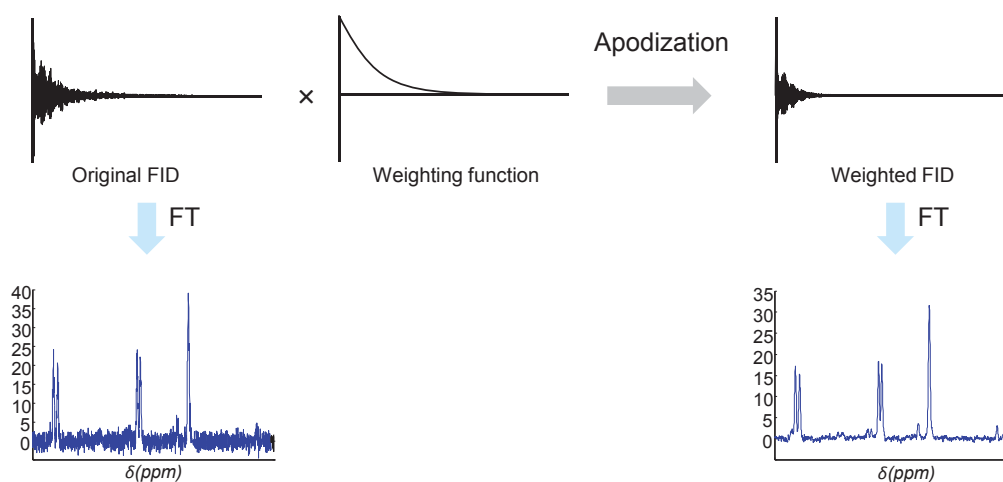


Figure 2.6. Apodization with an exponential negative weighting function.

3.3.2 NMR Preprocessing methods applied after Fourier Transform

- 1) Phasing: after FT, some peaks may not have the expected peak-shape (*e.g.*, some may be half up-half down or in dispersive mode) due to problems in phase. These phase problems usually come from a misadjustment of the phase detector, from delays between the initial RF pulse and the start of data acquisition, and from the electronic filtering of the NMR signal [94]. With spectral phasing, the shape of the resonances can be corrected to be in the absorptive mode.

This difference between absorptive and dispersive peaks is caused because for each data point in the raw FID contains one real value and one imaginary value, and therefore for each frequency point after FT. In the ideal situation, the absorptive spectrum would be obtained by representing the real values. However, due to these phase problems, in the real practice, a linear combination between the real and the imaginary spectra needs to be calculated to recover the absorptive spectrum. This linear combination can be expressed as in **equation 2.4**.

$$\text{Absorptive spectrum} = \text{real spectrum} \times \cos(\theta) + \text{imaginary spectrum} \times \sin(\theta)$$

eq. 2.4

In this equation, the angle θ represents the rotation angle between the two mutually perpendicular vectors of the real and imaginary spectra, and it is usually referred as the phase rotation angle θ . In addition, this phase rotation angle θ required to obtain the absorptive mode linearly depends on the chemical shift, as defined in **equation 2.5**.

$$\theta(\delta) = \Phi_1 \delta + \Phi_0$$

eq. 2.5

where the intercept Φ_0 is called the zero-order phase correction, and the slope Φ_1 is called the first-order phase correction. In consequence, a dispersive spectrum can be phase-corrected after establishing the Φ_0 and Φ_1 parameters. In the real practice, these two parameters can be estimated by slightly changing their values progressively and observing how the shapes of the resonances improve or not [95].

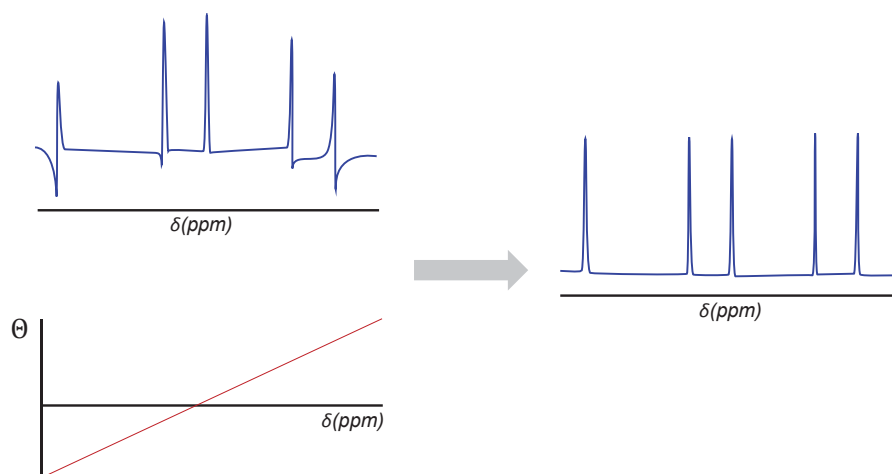


Figure 2.7. Phasing.

- 2) Reference to an NMR standard: if the sample contains an NMR standard (*e.g.*, DSS, TMS, or TSP), the methyl groups of these standards can be set to 0.00 ppm in the proton and carbon ppm scale. With this referencing, the chemical shift of different compounds in different spectra can be compared.

Alternatively, the solvent peak can be used as a chemical-shift reference. However, this can only be used in dilute solutions, where there is only one solvent, and with solvents with known resonance chemical shifts [96]. Another existing option, for Bruker NMR spectrometers, is to use as a reference the signal of electronic origin known as ERETIC [97].

- 3) Baseline correction: baseline distortions are mainly due to the corruption of the first few data points in the FID [98], but it can also be caused by instabilities, or by the presence of macromolecules that decay much faster than the other low molecular weight metabolites. To correct this, a smooth function (red line in **Figure 2.8**) that represents the offset between the original spectrum and the ideal spectrum is built, and this function is subsequently subtracted from the original spectrum. If the baseline correction is properly performed, the resulting corrected spectrum will have their noise values centered to zero (left figure in **Figure 2.8**).

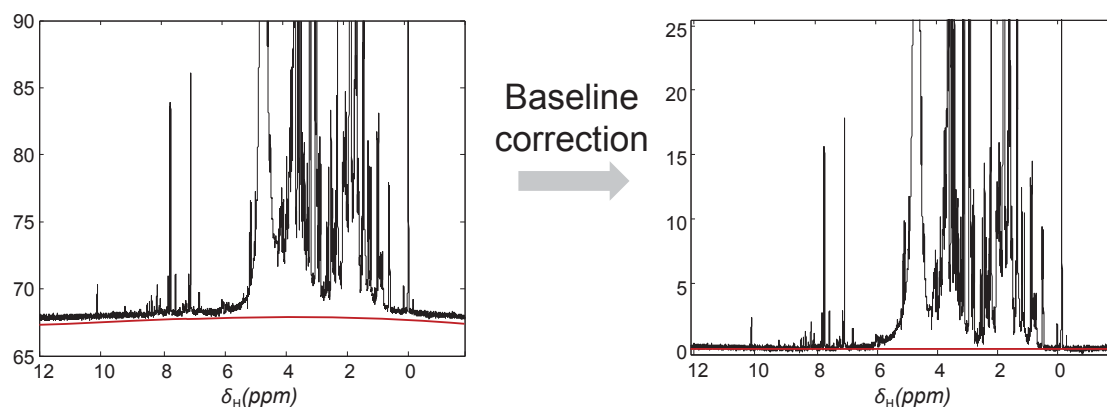


Figure 2.8. Baseline correction.

Different methods for baseline correction exist, that differ in the smoothing function used for subtracting. Examples of these methods are Bernstein polynomial [99], Whittaker smoother [100] and the splines method [101], among others.

- 4) Resonance alignment: even though standardized protocols are used in NMR metabolomics studies, spectral misalignments can still occur due to little pH changes and intermolecular interactions among the metabolites from these biologically complex samples. Most robust NMR integration tools can deal with these chemical shiftings, but chemometric analyses cannot be applied on the raw spectral data without leading to aberrant results because changes in peak position disrupt the usually assumed bilinear type of model (see the bilinear model in [section 4.3](#) of this chapter). Several algorithms have been proposed to correct these misalignments, such as *icoshift* [102], FOCUS [103], HATS [90] and COW [104], among others [105].

From all resonance alignment algorithms, *icoshift* is one of the most preferred options used in ^1H NMR metabolomics data. One of the advantages of *icoshift* in comparison to others is that NMR regions where the algorithm will be applied can be well defined (blue dashed boxes in [Figure 2.9](#)), and peaks cannot shift out of their corresponding NMR regions. With this strategy, the undesired convergence of resonances from different compounds to the same chemical shift is avoided.

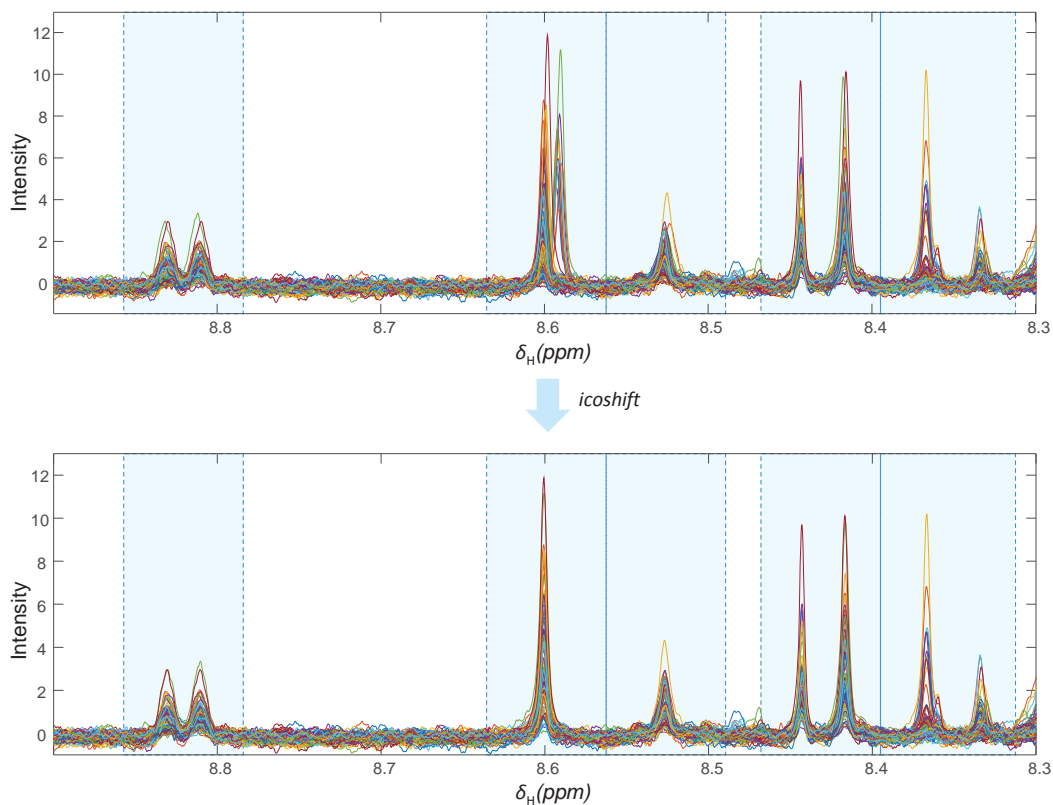


Figure 2.9. *icoshift* algorithm.

- 5) Bucketing (also known as binning): NMR spectra are segmented into a desired number of buckets (bins) and all intensity values inside each bucket are summed. For historical reasons, ^1H NMR spectra are usually bucketed with equidistant buckets of 0.04 ppm [106].

Application of bucketing produces a much smaller dataset, but since the resolution is simultaneously reduced, it leads to a loss of information and it may generate undesired artifacts. For instance, different resonances can be included into the same buckets, in which the smaller resonances will be obscured by the bigger ones; or a single resonance may be split into two or more consecutive buckets [107].

In order to avoid that bucketing becomes a source of errors, intelligent or adaptive bucketing algorithms have been proposed [106,107]. These bucketing methods capture single metabolite resonances into single buckets, adapting the width to each bucketed resonance. Bucket widths are calculated after searching for local minima between resonances [107]. Because of its apparent simplicity and rapidness, several NMR metabolomics studies [108,109] obtain biological interpretations from analyzing these buckets, without going further into proper resonances integration tools. Even though this approach results faster, it should not be assumed that each bucket is a totally correct resonance integral since spectral noise is captured into the bucket along with the considered peak.

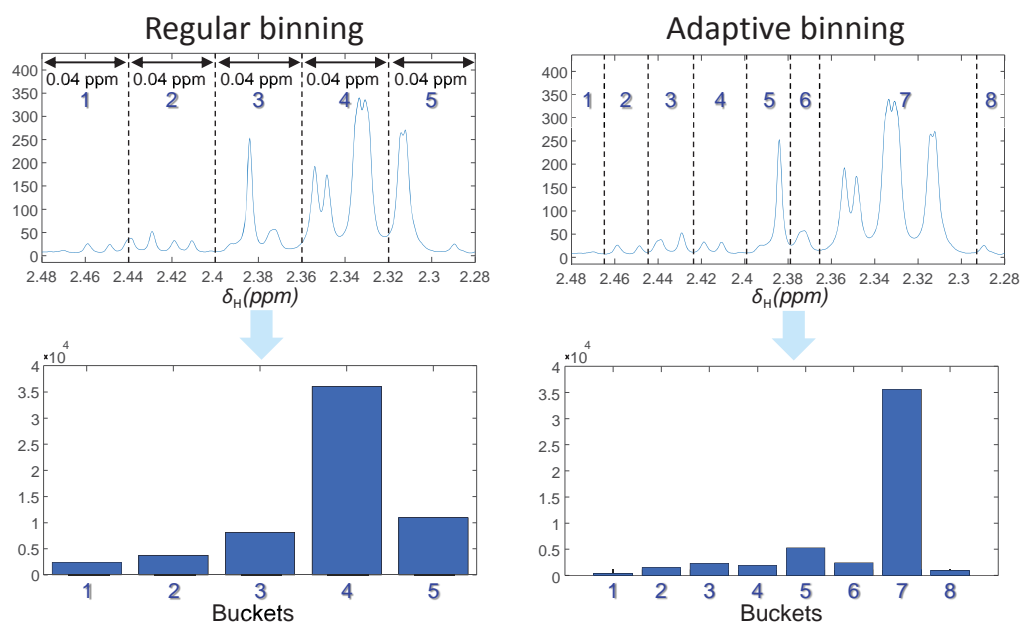


Figure 2.10. Regular bucketing (left) and adaptive bucketing (right).

- 6) Removal of undesired regions: before chemometric analysis, NMR spectral data are usually refined by removing all those NMR regions that do not contain resonances from metabolites [110,111]. Therefore, the outer regions, which only contain noise, are discarded, as well as the regions with the solvent resonances or any other interfering metabolite (**Fig. 2.11**).

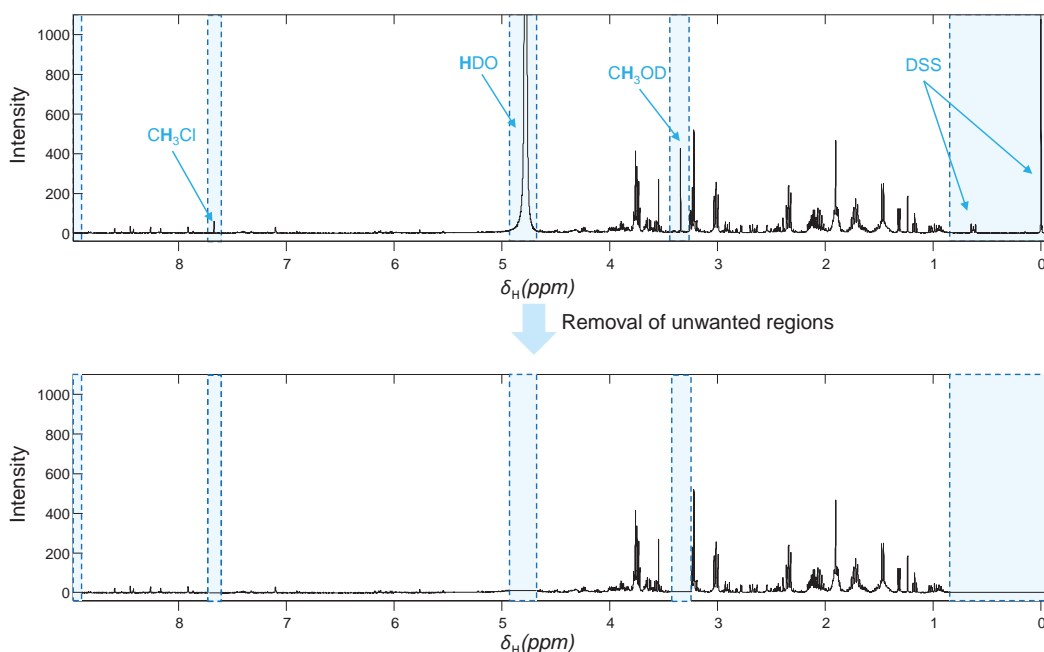


Figure 2.11. Removal of undesired regions.

3.3.3 Normalization methods for NMR spectral datasets

The aim of normalization is to remove variations attributed to global differences in sample concentrations due to dilution or size effects. Thus, after normalization, the relative intensity of the overall spectrum becomes similar for all samples.

Normalization methods are crucial in the analysis of samples in which the metabolite amount cannot be easily controlled. These situations are typical for NMR metabolomics experiments involving urine samples (there is an uncontrollable dilution variation) and in time-course experiments of growing organisms, where the biomass increases over time. The most used normalization methods are the following:

- 1) Integral or Total Sum normalization (IN or TSN, respectively): each NMR spectrum is divided by the total sum of intensities of the spectrum.
- 2) Constant integral normalization: each NMR spectrum is divided by the concentration of a metabolite that is intrinsically related to the sample dilution factor. For instance, ^1H NMR spectra of urine samples are usually normalized by the concentration of creatinine [112].
- 3) Standard Normal Variate (SNV): each NMR spectrum is corrected by its mean and divided by its standard deviation as depicted in **equation 2.6**.

$$NMR_{new} = \frac{NMR_{old} - \overline{NMR_{old}}}{\sigma(NMR_{old})} \quad \text{eq. 2.6}$$

In **equation 2.6**, NMR_{old} is the vectorized form of the NMR to be normalized, NMR_{new} is the normalized spectrum, and σ is the standard deviation of the intensity variables of NMR_{old} . This normalization is suitable for sets of samples in which the relative concentration of the distinct metabolites is expected to be similar [113].

- 4) Probabilistic quotient normalization (PQN): in PQN [114], a target NMR spectrum (green spectrum in **Figure 2.12**) is compared to a reference NMR spectrum (blue spectrum in **Figure 2.12**). In this comparison, the quotient between the target spectrum and the reference spectrum is calculated for every spectral data point, and the median quotient is established thereafter. This median quotient points out the median difference in intensity between this target spectrum and the reference spectrum. Finally, the target spectrum is divided by the median quotient in order to compensate for this median difference. This process is repeated for all the NMR spectra of the dataset (**Fig. 2.12**).

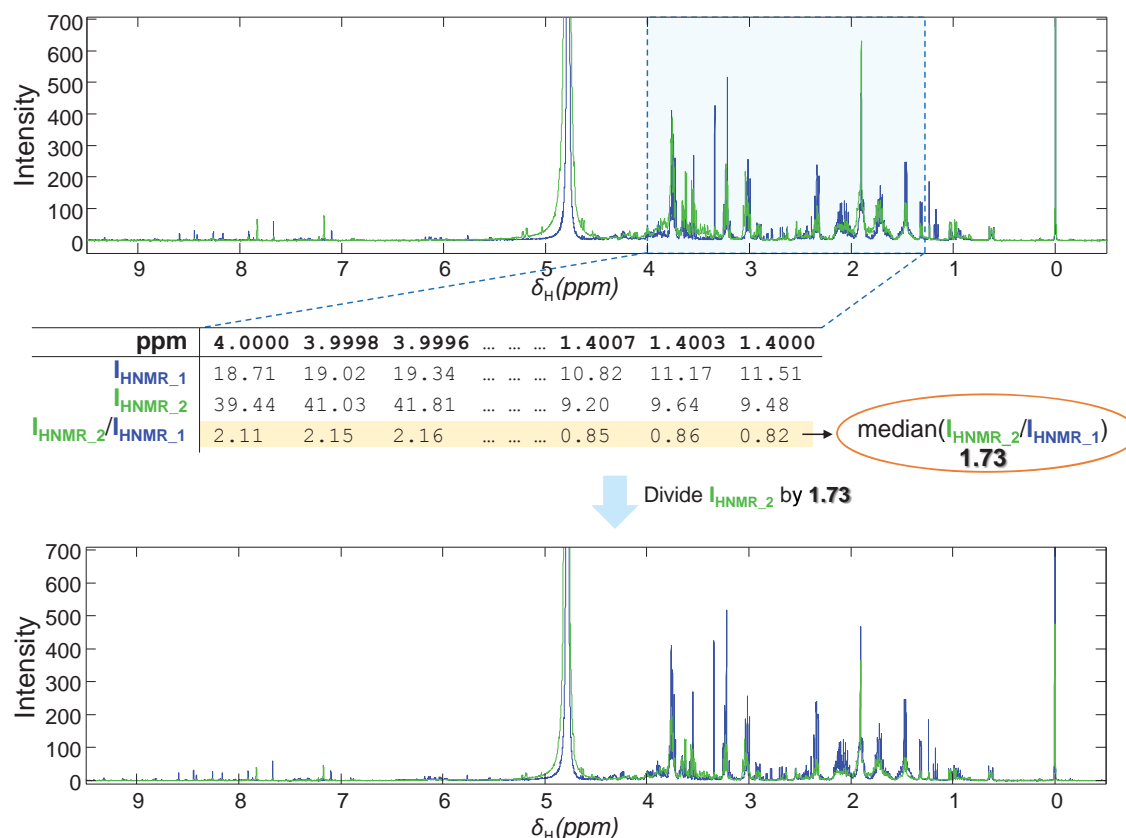


Figure 2.12. Probabilistic Quotient Normalization.

The reasoning behind PQN is that biologically relevant concentration changes influence only few parts of the spectrum (or intensity data-points), while global intensity changes in the spectrum are mostly caused by dilution effects. Thus, since all intensity data points are used in the calculation of the median quotient, this median quotient can be considered the most probable dilution factor between the two compared samples.

This method is more efficient than IN (or TSN) [114]. IN produces unreliable results when unexpected resonances (*e.g.*, secondary solvents, chemical interferences) appear since the area of the normalized spectra become then underestimated and the relation among different common resonances of different samples is artificially changed. On the other hand, the influence of these unexpected resonances is minimized with PQN normalization because these peaks are usually sharp and therefore they do not have an important weight in the calculation of the median quotient.

The calculation of the median quotient can be distorted if a significant amount of intensity signals only relative to noise is considered in its calculation. For this reason, for the calculation of this quotient, the selection of a portion of the spectrum containing minimal noise contribution is preferred [115]. This concept is illustrated in **Figure 2.12** with the blue box in the upper spectrum.

In addition, when applying PQN normalization, an adequate reference spectrum is required. Commonly, this reference spectrum is the median spectrum of the analyzed dataset [116] or the median spectrum of the control group dataset [117].

For time-course experiments, reference spectrum can be built from the samples collected at the initial time-point. However, since observed metabolic changes are progressive over time, in this Thesis we have applied a progressive PQN normalization in the analysis of datasets containing time-course data [118,119]. In this strategy, we have normalized every spectrum with the median spectrum obtained from the group of samples collected in the previous time-point exposed to the same treatment. For samples collected at the initial time point (Time0 in **Figure 2.13**), the reference spectrum was calculated using all the samples collected this time-point (**Fig. 2.13**).

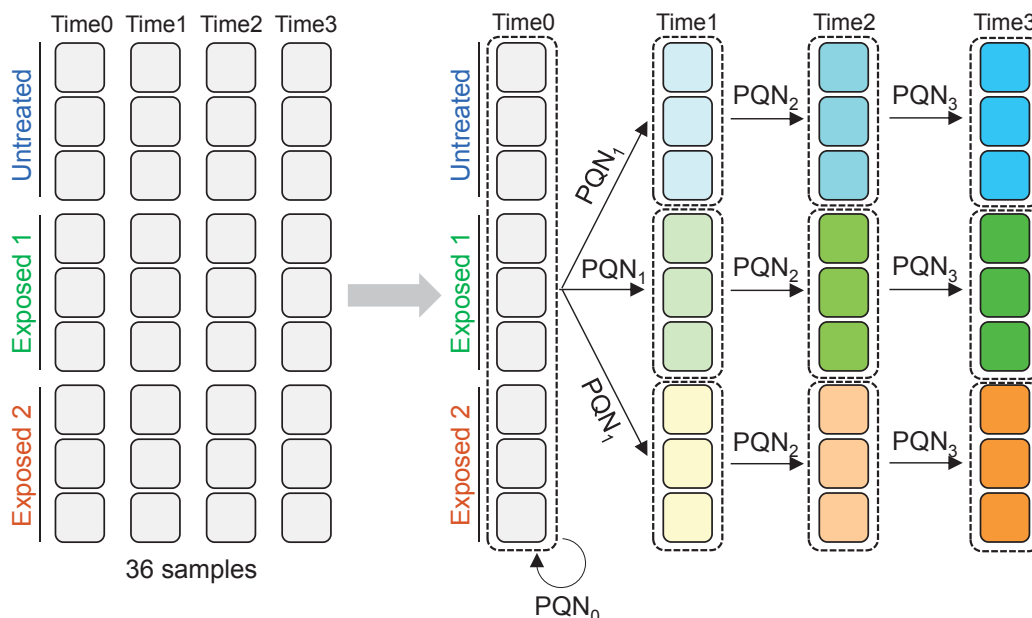


Figure 2.13. Progressive PQN normalization strategy for time-course experiments. Subscripts denote the order of application of the PQN normalization, and the dashed rectangles group the spectra used to calculate a reference spectrum for every PQN step. Blue, green and orange color-schemes denote that the observed response is exposition-dependent and that it is magnified over time.

3.3.4 Mean-centering

This pretreatment centers the raw measurements to zero. Therefore, with this pretreatment, the offset difference between high and low abundant metabolites is removed, and the relevant variation between the samples for analysis is the only information left for the analysis. In the literature, mean-centering is widely used for the pretreatment of NMR spectral data [120,121].

The formula used to mean-center the data is the following:

$$x_{ij_new} = x_{ij_old} - \bar{x}_j \quad \text{eq. 2.7}$$

where x_{ij_old} represents every element of the $\mathbf{X}(i, j)$ matrix before mean-centering, x_{ij_new} is the corresponding mean-centered x_{ij} element, and x_j is the j^{th} column vector of the \mathbf{X} matrix.

3.3.5 Scaling methods

Scaling methods are data pretreatments used to standardize all measured variables contained in a dataset. To compensate for the differences in magnitude among the different variables, the scaling methods correct the data based on their variation. In most of the scaling methods, the measurements are divided by a factor (the scaling factor), which is different for each variable [122]. As a result of this division, the untreated measurements are converted to relative measurements of the used scaling factor. The most used scaling methods in NMR metabolomics are auto-scaling and Pareto scaling, among others.

- 1) Scaling: the data are divided by the standard deviation of the x_j column vector. After this pretreatment is applied, variables will have a standard deviation of one.

$$x_{ij_new} = \frac{x_{ij_old}}{std(x_j)} \quad \text{eq. 2.8}$$

- 2) Auto-scaling: with this pretreatment, the data are mean-centered and divided by the standard deviation of the x_j column vector. Therefore, after auto-scaling, all variables have a standard deviation of one and they are centered to zero.

$$x_{ij_new} = \frac{x_{ij_old} - \bar{x}_j}{std(x_j)} \quad \text{eq. 2.9}$$

Auto-scaling and scaling are not recommended for raw NMR spectral data, in which some of the variables are only representative of noise, and these noisy variables will become as important in the dataset as the variables representative of meaningful signals after application of any of these two scaling methods. By contrast, auto-scaling or scaling are widely used in the pretreatment of ‘curated’ resonance integrals data, such as in concentrations datasets. After application of this pretreatment, significant relative variations of the relative measurements will be more easily detected, since all variables will be found at the same scale.

- 3) Pareto scaling: in Pareto scaling, every measurement is converted into a relative dimensionless measurement as in auto-scaling. The formula used in this pretreatment is similar to the one in auto-scaling, but the square root of the standard deviation is used as the scaling factor instead (eq. 2.10).

$$x_{ij_new} = \frac{x_{ij_old} - \bar{x}_j}{\sqrt{std(x_j)}} \quad \text{eq. 2.10}$$

Due to the square root in the formula, the variance for all the variables will not be the same. Because of this, variables associated with a higher variance (such as noise) will be down-weighted, whereas those associated with a lower variance will be inflated.

This pretreatment is commonly used for NMR spectral data [123,124].

- 4) Min-max scaling: this pretreatment sets the minimal and maximal values for each variable at 0 and 1, respectively. The transformation formula used is presented in **equation 2.11**.

$$x_{ij_new} = \frac{x_{ij_old} - \min(x_j)}{\max(x_j) - \min(x_j)} \quad \text{eq. 2.11}$$

In this Thesis, this transformation method was used for scaling metabolite concentrations because the non-negativity of the data is maintained after the application of this scaling method [125].

3.4 NMR RESONANCES ASSIGNMENT

The NMR spectrum of any compound is unique, and therefore, distinguishable from other NMR signatures from other compounds in an NMR spectrum of a complex mixture.

Even though its uniqueness, the assignment of a set of resonances to a given compound is not trivial. A surprising amount of information can be extracted from the NMR spectrum, which needs to be examined altogether.

The set of spectroscopic parameters that unequivocally identify a compound are the measured chemical shifts, the *spin-spin* splitting pattern or multiplicity, the coupling constants, and the resonance integrals.

3.4.1 Chemical shift

When the sample is placed under the influence of a magnetic field, B_0 , the electronic cloud surrounding a nucleus begin to circulate, creating an induced current that generates a magnetic field opposed to the B_0 field, B_i , reducing the effective magnetic field felt by the nucleus. This reduction (or deshielding) is measured in the order of parts per million (ppm), and it varies from nuclei to nuclei, depending on the electron density in the neighboring environment. If two or more chemically equivalent atoms have exactly the same electronic neighboring environment, they are magnetically equivalent and they are represented by the same resonance in the NMR spectrum.

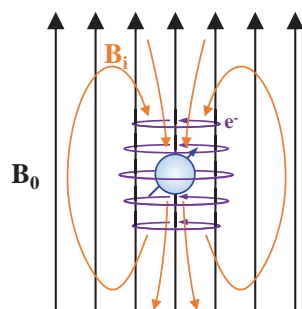


Figure 2.14. Deshielding. Black arrows represent the applied magnetic field (B_0), purple arrows represent the electronic cloud around the nucleus, and orange arrows represent the magnetic field (B_i) induced by the electronic current.

Neighboring electronegative groups weakens the electronic shield of the measured nuclei, which in addition reduces the induced magnetic field, and therefore increasing the exposition to the B_0 field. This is translated on a higher resonance frequency for the measured nucleus. Since all atoms of the studied molecule contribute to the electronic cloud, and the electronic cloud is molecule-specific, the chemical shifts are also molecule-specific. Having said this, it is possible to predict chemical shifts, with acceptable results, by considering the structure and the different functional groups of the molecule [126,127]. In fact, the major differences in measured frequencies account for the different functional groups, meaning that protons and carbons from certain functional groups appear always in the same frequency range [128]. For instance, proton resonances from methyl (CH_3) groups appear always around $\delta_{\text{H}} = 0.8\text{-}1.5$ ppm, whereas proton resonances from methyl ether (OCH_3) groups appear around $\delta_{\text{H}} = 3\text{-}4.5$ ppm.

The chemical shift can also be explained and predicted using a Quantum model. For simplicity, we will only consider the Quantum model for nuclear spins of $S = \pm 1/2$, since it is the common spin of nuclei measured in metabolomics (e.g., ^1H , ^{13}C). In the absence of an external magnetic field (B_0), nuclei spin in a random orientation. However, when an external magnetic field is applied, they spin at only two different orientations (α and β). Nuclei with α orientation spin parallel to B_0 , whereas nuclei with β orientation spin antiparallel to B_0 . Because of this alignment to B_0 , α protons have a lower energy than β protons.

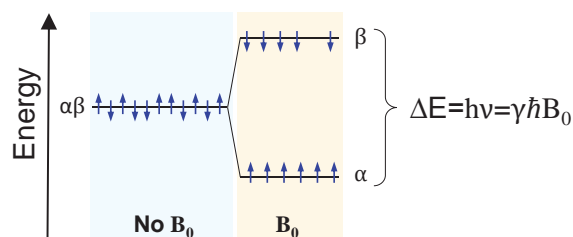


Figure 2.15. Quantum model of the B_0 effect on nuclear spins.

The energy separation between α and β spin states or transitions is proportional to the strength of B_0 , and, in the absence of B_0 , all protons have the same energy. However, in the influence of B_0 and in the equilibrium, slightly more than half of the population will be in the α transition, while slightly less than half population will be in the higher energy state. In this situation, a nucleus in the lower state can absorb a photon of electromagnetic energy and be promoted to the higher energy state, and this energy corresponds to the resonance frequency. Since the effective magnetic field felt by each (non-equivalent) nucleus is different due to the different chemical environments, the energy separation between the α and β spin states of each nucleus will be also different, resulting in resonances detected at different chemical shifts.

3.4.2 Coupling constant

Indirect *spin-spin* coupling (indirect dipole-dipole interaction or just *J*-coupling) refers to the magnetic interaction between individual nuclear spins transmitted by the bonding electrons through which the nuclear spins are indirectly connected.

This coupling depends on the hybridization of the atoms involved in the coupling, the bond angles, the dihedral angles, the C-C bond length and the effect derived from substituent atoms, such as electronegativity and neighboring π -bonds. Lists with detailed examples of *J*-couplings can be found elsewhere [129,130].

In the NMR spectrum, this coupling is observed as a splitting of the resonance in a structured pattern with a defined spacing between splits that coincides with the *J*-coupling measured in Hertz units. Because of the latter, the splitting pattern in lower magnetic fields is broader than in higher magnetic fields, where each of the split resonances are thinner and sharper.

3.4.3 Spin-spin splitting pattern

If the resonances from all the nuclei that are coupled to a given resonance are distant in the spectrum from this resonance, the coupling is weak (first order *spin-spin* splitting pattern) and the splitting pattern relatively simple. On the contrary, if resonances are coupled to nearby peaks, more complex patterns are observed. Since the order of the splitting pattern depends on the proton distance measured in hertz, a signal observed of non-first-order splitting patterns may become of first-order in a higher magnetic field.

Splitting patterns depends on the chemical structure. This is here illustrated with the example of L-lactic acid ((2S)-2-hydroxypropanoic acid in IUPAC nomenclature).

For this compound, two proton resonances are expected (one for the CH_3 group and another for the CH group). Since the three protons in the CH_3 group are magnetically equivalent, they will resonate at the same frequency. On the other hand, the proton resonance from the

hydroxyl group may not be detected if the equilibrium rate of the proton with an exchangeable deuterium from the medium (*e.g.*, CD₃OD) is fast [131].

Then, for this compound, the two detected resonances are coupled, since the nuclear spins of the protons from the two groups interact (they belong to the same spin system). Since the two resonances are very distant to each other, the splitting patterns that define these couplings are of first-order. In these couplings, each of the protons from the CH₃ group (which can be at either α or β) interact with the proton in the CH group (also at α or β), and vice versa.

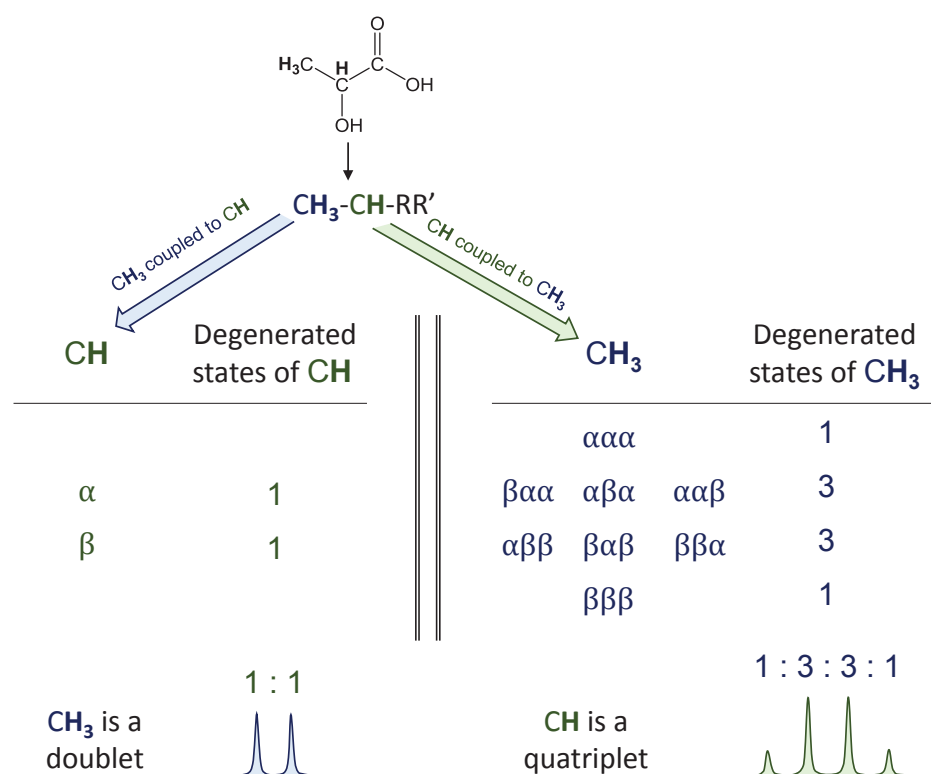


Figure 2.16. Proton nuclear spin states combinations for the inequivalent protons in L-lactic acid.

Accounting for all the possible spin combinations, the proton from the CH group, which can be found at two different spin states, couples to the 8 spin states combinations possible in the CH₃ group. Since these 8 spin states combinations are found in four energetic states because of the spin degeneration, the resonance from the CH is detected as a quadruplet (four peaks) resonance with an intensity ratio that coincides with the degeneration pattern (1:3:3:1). Using the same reasoning, the protons from the CH₃ group are detected as a doublet (two peaks) resonance with the same intensity.

The splitting pattern for L-lactic acid is one of the simplest because only two sets of equivalent nuclei are coupled. When the spin system is composed of more than two different equivalent nuclei, each set will pair to the rest of nuclei and will show a coupling in the spectrum if this

interaction is sufficiently strong. Thus, the resonance will be divided into smaller peaks according to the possible couplings.

For instance, if a proton H_a is coupled to a proton H_b and to a proton H_c at the same time, and all of them are chemically inequivalent, then the resonance from H_a is doubled twice, once due to the H_a-H_b coupling and another one due to the H_a-H_c coupling, resulting in four observed signals. The described spin system is also known as a double-doublet (**Fig. 2.17**).

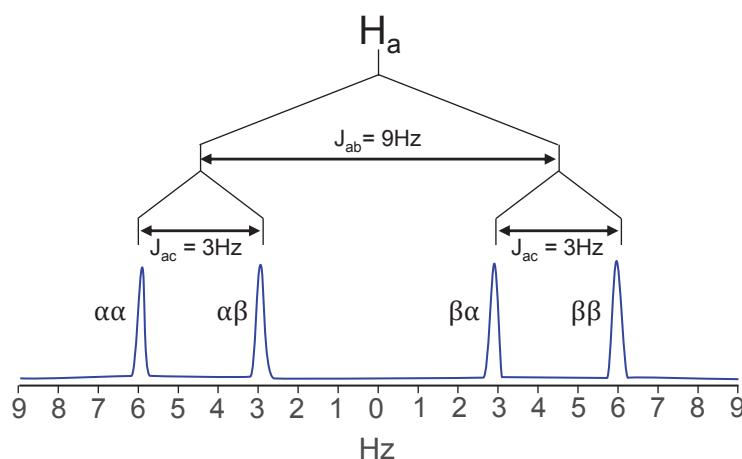


Figure 2.17. Double-doublet splitting pattern.

Lastly, a *spin-spin* splitting pattern can be much complex (non-first-order splitting pattern) if the difference in hertz between the frequency energy of the involved nuclei is smaller than the J -coupling in hertz. Precisely, it has been established that, for two nuclei A and B, the frequency energy must be five times or less than the J -coupling (**eq. 2.12**) [132].

$$\nu_A - \nu_B \leq 5 J_{AB}$$

eq. 2.12

Hence, since the *spin-spin* splitting pattern type depend on the chemical difference expressed in hertz, it means that the type of the splitting pattern depends on the strength of the magnetic field.

In a non-first order splitting pattern, all transitions are not equally favored, and therefore all peaks inside a multiplet are not equally intense. This difference in intensity depends on the proximity to the other resonances from the same spin system, being the most intense peaks those located closer to the resonance they couple with.

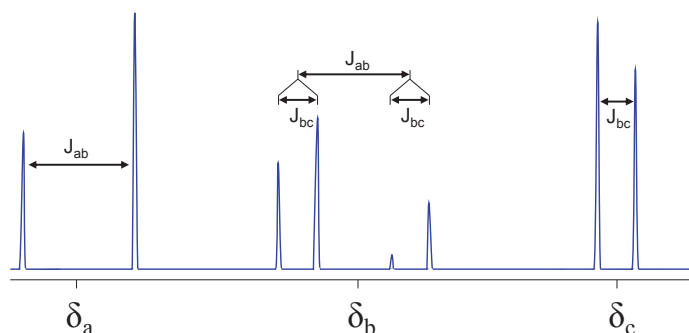


Figure 2.18. Non-first-order *spin-spin* splitting pattern for a H_a - H_b - H_c system.

3.4.4 Resonances integral

Measured nuclei can achieve full relaxation after each successive scan if enough relaxation delay time is left. In this situation, resonance integrals are quantitative, meaning that the areas of the different resonances are comparable. For example, the integral relative to the resonance associated to three magnetically equivalent protons (CH_3) will be three times larger than the area of a resonance relative to one proton (CH).

The concept of quantitative NMR (qNMR) is more profoundly covered in [section 3.5](#) of this chapter.

3.4.5 NMR databases

NMR assignment is a laborious task, but for metabolomics analysis, this step is even more challenging [133,134] because metabolomics samples are complex. Typically, in an NMR spectrum of a metabolomics origin, hundreds of resonances can be detected from tens to hundreds of compounds [135-137], but only a fraction of these resonances are finally assigned.

To cope with this severe problematic, various NMR databases have been created in the latest years with the aim of facilitating this analysis. These proposed databases store NMR spectra of common metabolites, and they all include a tool to search for candidate metabolites based on an input list of spectroscopic parameters.

In order to be this tool reliable, both data in these repositories and in our analyzed 1H NMR experiments should be from samples prepared similarly. Commonly, reference NMR spectra stored in these NMR databases are acquired from metabolites resuspended in phosphate buffered deuterated water at neutral or physiological pH and at 298 K. For organic compounds, deuterated chloroform is usually used as the solvent.

The most used public NMR Metabolomics databases are the following, including among all thousands of NMR spectra:

- Metabolomics databases from the Wishart Lab at the University of Alberta. The group created 18 metabolomics databases, such as the Human Metabolome Database (HMDB) [138,139], the DrugBank database [140], the *E. coli* Metabolome Database (ECMDB) [141], the Yeast Metabolome Database (YMDB) [142] and the Urine Metabolome Database [143].
- The Madison-Qingdao Metabolomics Consortium Database (MQMCD) [144].
- The Platform for RIKEN Metabolomics (PRIMe) [145].
- The Birmingham Metabolite Library (BML-NMR) [146].
- The Biological Magnetic Resonance Data Bank (BMRB) [147].

On the other hand, commercial software such as Chenomx (Chenomx Inc., Canada) and AMIX (Bruker Inc., US) also include NMR spectral libraries.

The most complete NMR spectral database at the moment is the Nuclear Magnetic Resonance Shift Data Base (NMRShiftDB) [148] that covers a wide range of chemical compounds (>40,000) [149], although the sample preparation is not always as consistent as for the NMR metabolomics databases, since it was created as a repository of NMR data of natural products.

Despite all these available databases, sometimes it is still not easy to confirm the NMR assignment of a resonance or set of resonances in a metabolomics sample. In these particular cases, acquiring complementary 2D NMR experiments on the same sample or spiking the sample with the candidate compound are recommended.

3.5 INTEGRATION

In NMR metabolomics, the ultimate purpose of acquiring and processing the NMR spectra is to establish relationships of biological interest between the studied samples and their corresponding metabolic profiles. This goal can be achieved by applying data analysis strategies on the set of NMR spectra. However, due to the fact that most of the detected resonances are overlapped, the obtained results may be misleading. For this reason, it is preferable to apply the data analysis strategies on the set of resonance integrals (obtained by separation, deconvolution or resolution approaches) of the detected metabolites relative to the same set of samples rather than directly in the set of raw NMR spectra.

The existing integration approaches are discussed below.

3.5.1 Resonances integration of non-overlapped signals

Traditionally, resonance integrals have been calculated by summing the intensity values comprised within the range of each resonance. Graphically, it is usually represented as integral lines (red lines in **Figure 2.19**) of the cumulative sum of intensities, and these integral values are proportional to their height.

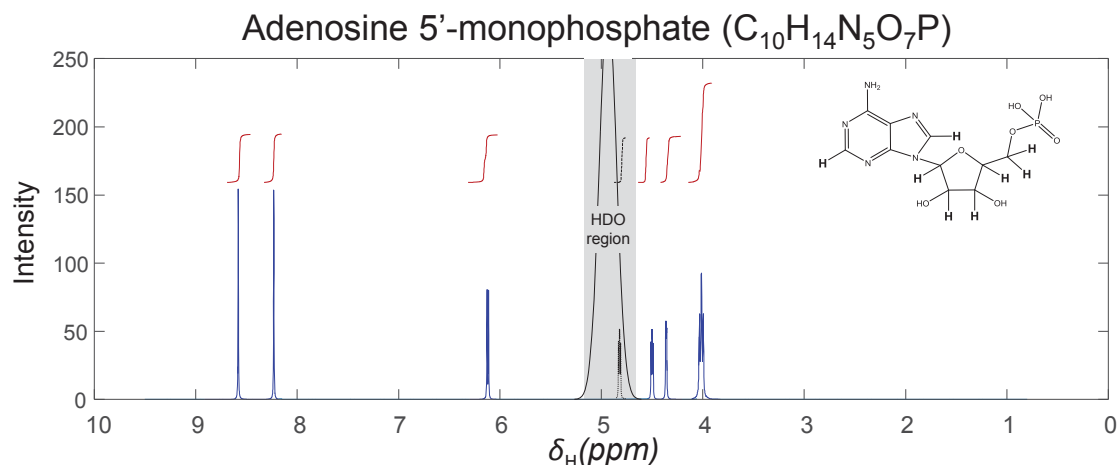


Figure 2.19. ¹H NMR spectrum of adenosine 5'-monophosphate (AMP).

For a simple ¹H NMR spectrum, the calculated integrals are proportional to the proton concentration of the measured compounds. Thus, for an individual compound, all resonances should be of the same height, except for magnetically equivalent protons, in which the area will be proportional to the number of magnetically equivalent protons.

In reality, NMR analyses are not that simple, as there are several situations to consider.

First, when metabolites are dissolved in totally deuterated solvents containing exchangeable protons, resonances from exchangeable protons (*e.g.*, protons from hydroxyl and amine groups) may not be detected. In **Figure 2.19**, AMP has 14 protons, but 6 are not detected in the ¹H NMR spectrum because they were exchanged with deuterium. Moreover, a non-exchangeable proton from AMP ($\delta = 4.77$ ppm) is also not detected in the spectrum because it is masked by the HDO signal.

Second, in ¹H NMR spectra from metabolomics samples, where proton resonances are in the order of the hundreds, most of the proton resonances appear overlapped. Because of this, to calculate the integrals for each of the resonances in the overlapped region, they need first to be separated using computer-derived approaches (either using spectral deconvolution methods or spectral resolution methods).

3.5.2 Spectral deconvolution

A ^1H NMR spectrum can be considered as a linear combination of proton resonances from a defined group of metabolites in presence of some residual noise. When the spectroscopic parameters (δ , multiplicity, J) that define the resonances from these metabolites are known, either because they have been determined using a set of complementary pulse sequences or because they can be consulted in spectroscopic databases, it is possible to decompose (or deconvolute) the real NMR spectrum into a set of line functions. Then, the corresponding integrals can be calculated by estimating the area under the curve of these line functions.

In NMR spectroscopy, a resonance is the product of applying the FT on the decaying sine wave function of the magnetization, and the resulting transformed function follows the Lorentzian model of [equation 2.13](#) [150].

$$l(\delta_{\text{mu}}) = \frac{2}{\pi} \frac{v_{1/2(\text{mu})}}{4(\delta_{\text{mu}} - \delta_0)^2 + v_{1/2(\text{mu})}^2} \quad \text{eq. 2.13}$$

where the $l(\delta_{\text{mu}})$ function is centered at δ_{mu} with a full width at half height $v_{1/2(\text{mu})}$.

Any existent resonance splitting pattern can be described as the sum of line-shape functions that represent each of the *spin-spin* splitting transitions (*e.g.*, a doublet can be defined as the sum of two line-shape functions of equal intensity) [151].

In reality, each *spin-spin* splitting transition is detected at a distribution of frequencies that can be approximated by a Gaussian function, so each line shape is, in reality, a convolution of Lorentzian peaks with a Gaussian distribution that jointly follows the Lorentzian-Gaussian (or Voigt) line-shape [152]. Having said that, differences in deconvolution between the Lorentzian and Voigt model are not very significant, and most existing deconvolution tools have assumed the Lorentzian model since it is simpler. Also for simplicity, only the peak fitting assuming the Lorentzian model is shortly described below.

Basically, a ^1H NMR of a metabolites mixture, with m metabolites at c_m concentration, where each individual spectrum contains u resonances with t_{mu} *spin-spin* splitting transitions, can be decomposed into two parts. These two parts are (i) the parametrized part, that includes the Lorentzian functions; and (ii) the non-parametrized part, e , or the observed differences in the spectra not seen in the predicted theoretical model such as noise [35]. The non-parametrized part is assumed to follow a σ Normal distribution with the intensity values centered to zero. This model is depicted in [equation 2.14](#).

$$y = \sum_{m=1}^M c_m \sum_{u=1}^U t_{\text{mu}} l(\delta_{\text{mu}}) + e, e \sim N(0, \sigma) \quad \text{eq. 2.14}$$

Since the chemical shifts of the resonances, δ_{mu} , as well as their t_{mu} *spin-spin* splitting transitions are known, the only parameters left to deduce are the $v_{1/2(\text{mu})}$. The values for these

half heights $v_{1/2(\text{mu})}$ can be determined by either manually fitting all the ^1H NMR templates of the individual metabolites to the experimental spectrum, or it can be automatically performed by computational means. For the latter, the process is driven iteratively accounting for all the resonances at the same time by reducing the ε contribution after every iteration, and by restraining the set of Lorentzian functions to functions with the positive domain. If a proper set of ^1H NMR spectral templates is used, upon reaching the optimal solution, the sum of fitted Lorentzian functions must resemble the original ^1H NMR spectrum (**Fig. 2.20**).

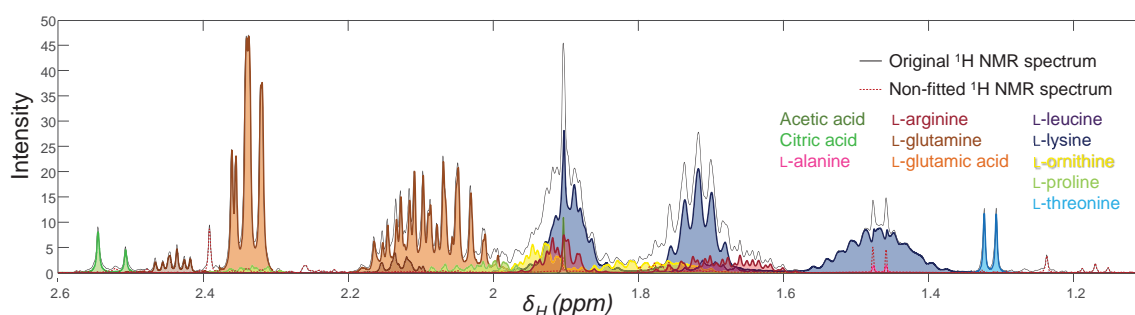


Figure 2.20. Deconvolution of a spectral region of a ^1H NMR spectrum.

Several tools exist for deconvolution of ^1H NMR spectra, such as BATMAN R-package, Bayesil, Chenomx, AMIX (Bruker Inc., US), and MestreNova (MestreLab, Spain), and opting for one or another method depends on several criteria. For instance:

- BATMAN and Bayesil are free, whereas the other ones are commercially available.
- In Chenomx, deconvolution can be done either manually or computationally, in the other methods, deconvolution is only performed computationally.
- MestreNova is yet not optimized for deconvoluting ^1H NMR spectra in a high-throughput mode, but the model used can be either the Lorentzian or the Voigt model.
- Bayesil is restricted to the analysis of cerebrospinal fluid since the query dataset only contains ^1H NMR spectra from metabolites found in this fluid.
- AMIX works only with Bruker data.

In this Thesis, we have decided to perform ^1H NMR integration with the BATMAN R-package for the following reasons:

- It is flexible: it does not rely on a fixed spectral database, and adding new query resonances is simple.
- It is free.

- A wide variety of output objects (apart from the concentration estimates), such as the individual fitted ^1H NMR spectra or the remaining residuals can be consulted by the user.
- There is no need to export the outputs to a statistical software since it is executed in R environment, a free software environment widely used for statistical computing since it already includes many statistical tools.

3.5.3 Spectral resolution by chemometric methods

In the previous section, it is stated that spectral deconvolution can be used only if an NMR spectral model of target metabolites are available. However, sometimes overlapping (or even just the spin systems of isolated resonances) are too complex, causing that determining the spectroscopic parameters that define the analyzed resonances becomes challenging. In these particular cases, it may still be possible by resolving the pure resonances independently and calculating their integral by chemometric means, by employing the Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS) chemometric method. This approach has been further explained in [section 4.4](#) of this chapter.

3.5.4 Resonances integration in 2D NMR datasets

In NMR metabolomics studies, quantitative analysis of 2D NMR data is not frequent, since most of these studies rely solely on ^1H NMR spectra for obtaining quantitation estimates, while 2D NMR data is only used to improve NMR resonances assignment. This prevalence is not trivial: ^1H NMR spectra are acquired faster, with higher sensitivity, they have a direct correspondence to concentration estimates and they are easier to analyze due to the absence of the second dimension. Despite this, some metabolomics studies use 2D NMR spectra because the signal resolution is greater than in the equivalent ^1H NMR dataset counterpart.

The few published 2D NMR metabolomics studies used two main analytical strategies to investigate the 2D NMR datasets. In most of them, the resonances integration step is avoided, and the data is analyzed by exploratory chemometric methods prior bucketing the 2D NMR spectra [85,153]. The second used strategy is based on a first careful resonance assignment, followed by a Regions of Interest (or ROI) analysis [34,88]. In the ROI analysis, assigned resonances are individually enclosed in ROI segments. Then, the integral can be obtained by summing all the intensity values contained in each ROI segment [87,88,154-157], or by deconvoluting every resonance using a Voigt function (parametrized for the two-dimensional space) with advanced deconvolution approaches [158].

In ^1H - ^{13}C HSQC NMR spectra, signals are very sparse and spectral overlapping is not as problematic as in ^1H NMR datasets. For this reason, in this Thesis, we have employed the

ROI strategy without peak deconvolution for the analysis of ^1H - ^{13}C HSQC NMR spectra. Nevertheless, in order to produce more reliable integral estimates, the surrounding noise of these resonances was previously filtered by using the Variables of Interest (or VOI) strategy designed and proposed in this Thesis [159].

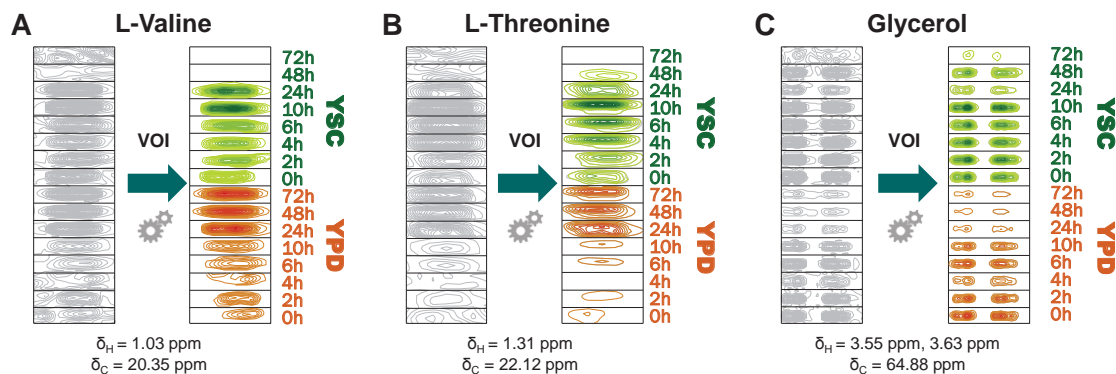


Figure 2.21. Noise filtering with VOI strategy. **A-C)** ROI segments relative to **A)** L-valine, **B)** L-threonine and **C)** glycerol for 16 ^1H - ^{13}C HSQC NMR spectra before (left) and after (right) noise filtering with VOI strategy.

A detailed explanation of the VOI strategy fundamentals is described in Puig-Castellví *et al.* (2018) [159] and in **Chapter 4**.

On the other hand, soft-modeling (chemometric) approaches can be also used for obtaining the resonance integrals from 2D NMR spectra, although very few number of studies can be found in the literature [160-162]. Moreover, these results cannot be directly extrapolated to the metabolomics field, since the analyzed samples were rather simple (less than 7 metabolites per sample). In this Thesis, the applicability of chemometrics for integration purposes has been discussed more in detail in **Chapter 4**.

3.5.5 Generating absolute concentrations from resonance integrals

Since ^1H NMR is inherently quantitative, absolute concentrations can be directly retrieved for all the metabolites contained in the measured samples. In other words, there is no need of using calibration standards for every metabolite. To achieve this, a metabolite of known concentration (*e.g.*, an NMR standard) has to be measured simultaneously with the sample. This metabolite can be introduced into the sample, or into a coaxial NMR tube that will be inserted into the NMR sample. With this second approach, the interaction of the standard with the metabolic mixture is avoided.

Thus, after establishing the link between the measured intensity (*i.e.*, the proton integral) and the concentration for the NMR standard, the absolute concentration for the rest of the metabolites present in the sample can be calculated as depicted below (**eq. 2.15**).

$$[\text{metabolite}] = \frac{[\text{NMR_standard}]}{I_{\text{NMR_standard}}} I_{\text{metabolite}} \quad \text{eq. 2.15}$$

It is important to state, however, that this method is only directly applicable to 1D ^1H NMR spectra, and only if certain requirements are met. For 1D ^{13}C NMR spectra, a typical approach is to use the inverse-gated proton decoupling and account for the natural abundance of the carbon-13 (1.1%). For 2D NMR spectra, 2D cross-peaks intensities depend on metabolite relaxation times and acquisition relaxation parameters, since RD relaxation delay is minimized in 2D experiments to reduce acquisition time for each t_1 increment. Moreover, pulses excitation profiles and diversity in the actual values of J -couplings also produce differences in the intensity of the observed signals. For example, for ^1H - ^{13}C heteronuclear experiments, the evolution time used for the experiments is normally optimized for an average $^1J_{\text{CH}}$ of 145 Hz, which is a compromise between the real values that can be observed for the different C-H groupings within molecule and between different molecules. For all these reasons, ratios between the metabolite concentrations and the detected intensities are usually not conserved for all the 2D cross-peak resonances, and calibration curves for every metabolite are needed in these cases [163]. Examples of quantitative 2D NMR spectra are the J -resolved experiments [164], and the zero-quantum experiments [165,166], among others.

Integrals obtained from ^1H NMR spectra are only reliable in quantitative terms when adequate acquisition parameters and processing methods are used. The exhaustive list of recommendations to obtain quantitatively accurate integral measurements can be found elsewhere [167]. Some examples of these mentioned recommendations are the following:

- The relaxation delay must be at least 5 times longer than the T_1 of the most slowly relaxing nuclei to ensure that at least 99 % of the nuclei have reached the equilibrium magnetization. For ^1H NMR, it implies that the relaxation delay should be equal or longer than 5 seconds [64].
- The recommended pulse angle to use is 90° . If a different angle is used, the ratio among resonances will still be maintained, but since measured intensities will be smaller, the accuracy of the integral for the less intense signals will be compromised.
- Inverse-gated decoupling should be applied to eliminate ^{13}C satellite on ^1H NMR spectra. In complex metabolomics mixtures, ^{13}C satellites of highly concentrated metabolites may be found in the same intensity range than the proton resonances of the lowest concentrated metabolites. So, if both these proton resonances and ^{13}C satellite resonances appear in the same spectral region, resonances assignment can be troublesome and the estimation of the resonances integrals may result even more challenging due to the additional presence of these ^{13}C satellite resonances.

- A number of 64k spectral data-points is recommended. With a bigger number of data-points defining a resonance, the resonance will be better resolved, increasing the accuracy of the resulting integral.

In addition, in samples containing a high concentration of non-deuterated water, the intense water signal dominates the spectrum, causing that the receiver gain needs to be decreased to not overflow the digitizer, which worsens the spectral sensitivity. To avoid this, solvent pre-saturation pulse sequences that suppress the water signal are recommended. For instance, for the acquisition of 1D ^1H NMR spectra, the 1D ^1H NOESY [168] pulse sequence (*noesygppr1d* in Bruker NMR spectrometers) has become the preferred option nowadays. With the solvent pre-saturation, the water signal will be suppressed and the spectral sensitivity will not be affected by water, but as a side effect, resonances from exchangeable protons will be also affected by the pre-saturation. Therefore, in these cases, quantitative measurements should not be performed from these integrals.

Finally, absolute quantitation can only be pursued from NMR spectra normalized to a resonance of a reference compound with known concentration. For NMR data normalized with methods based on other criteria (*e.g.*, accounting for sample dilution effects, such as TSN and PQN normalization methods), the normalized data are not absolute quantitative. In the latter situation, the concentration of all the metabolites present in the sample will be modified, including the standard, and **equation 2.15** cannot be applied.

Moreover, scaled NMR spectra are not absolute quantitative either. By application of scaling methods, the factor or ratio that links the measured integral with the real concentration will be different for every resonance. Hence, for every resonance, real concentrations can only be estimated by using their specific factor, and therefore, the factor found for the NMR standard cannot be extrapolated to the other metabolites.

However, in NMR metabolomics, (PQN-)normalized and scaled spectral data can be very useful despite this loss of absolute quantitative information because relative changes in the concentrations (fold-changes) can still be estimated, and because pretreated data reveal better the hidden factors present in the raw data. For this reason, in this Thesis, for metabolomics purposes, we have decided to use relative concentrations instead of absolute concentrations.

3.6 IMPORTING AND EXPORTING NMR DATA

Most of the statistical analyses used in this Thesis has been carried out under MATLAB® environment. In order to be able to work with experimental data in this programming environment, we established a workflow protocol to import the NMR data, summarized in **Figure 2.22A**.

In this Thesis, NMR spectra were acquired using NMR instruments from the two principal NMR companies, Varian (acquired by Agilent in 2010, discontinued since 2014) and Bruker. Since the data acquisition for the two instruments is different, two different workflows were used.

In Varian, the acquired intensities of the FID are stored with the corresponding measured acquisition time values in the same file, and all the acquisition settings are stored altogether in another file. On the other hand, in Bruker, a larger number of files are generated, and the file with the FID intensities does not contain the measured times.

In addition, in this Thesis, we also established two workflows to export the data from MATLAB® to two of the most used programs for the analysis of NMR spectra: MestReNova (Mestrelab Research, S.L.) and TopSpin® (Bruker, Inc.). While TopSpin® is restricted to Bruker NMR data, MestReNova can read both Varian and Bruker NMR data. These workflows for exporting NMR data are summarized in **Figure 2.22B**.

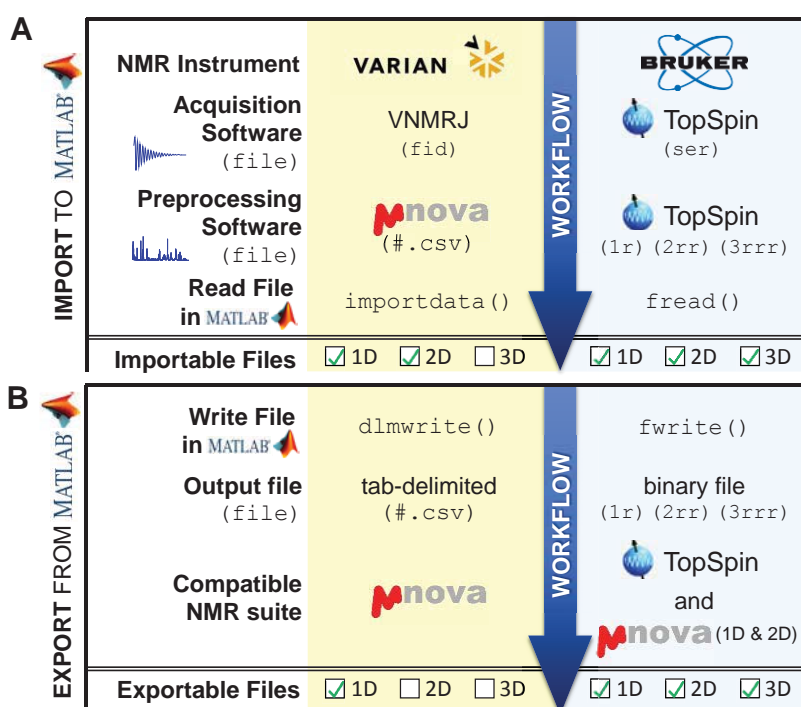


Figure 2.22. Exporting and importing NMR data.

3.6.1 Importing NMR data from Varian

For Varian NMR instruments, the FID files (named `fid`) acquired using the VNMRJ software (Varian, Inc.) are opened in MestReNova and the NMR preprocessing methods (FT, apodization, baseline correction) are applied afterwards. Then, the preprocessed NMR spectra are saved as a comma-separated value file (extension `.csv`), which are opened in MATLAB® using the `importdata()` function. For a 1D NMR spectrum (**Fig. 2.23A**), the imported data is a matrix of two columns, where the first column contains the chemical shift values, and the second column contains the intensities. For a 2D NMR spectrum (**Fig. 2.23B**), the imported data is a matrix. In the first row and column of this matrix the chemical shifts values for f_1 and f_2 dimensions are represented, while the rest of variables of the matrix contain the intensities associated to chemical shifts in f_1 and f_2 .

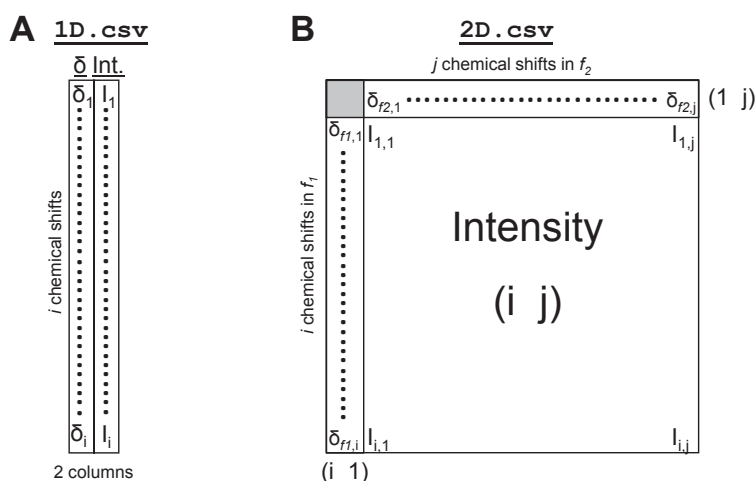


Figure 2.23. Representation of a **A**) 1D NMR spectrum or a **B**) 2D NMR spectrum saved as a `.csv` file.

For NMR data with higher dimensionality, this workflow cannot be used because these spectra are not readable with MestReNova software.

3.6.2 Importing NMR data from Bruker

In Bruker NMR instruments, FID data are stored in a 32-bit integer binary file named `ser`. After application of Fourier Transform, another binary file containing the real intensities of the FT-NMR processed spectra is automatically created (named as `1r` for 1D NMR data, `2rr` for the 2D NMR data, and `3rrr` for the 3D NMR data). However, unlike in Varian NMR instruments, the intensity data-points are generated as a vector, regardless of the dimensionality of the acquired NMR spectrum. Moreover, the chemical shifts positions associated to these intensity data-points are stored separately in a different data file.

Data from this file can be apodized and baseline corrected, and imported to MATLAB[®] afterwards using the `fread()` command. The resulting MATLAB[®] variable contains the real intensities, but not the chemical shifts. To import the two data at the same time, we used MATLAB[®] functions from the BPIO toolbox provided by Bruker. This function reads the real intensities (using the `fread()` command) as well as a large list of other values, such as the chemical shifts, the acquisition parameters, and some information regarding the pre-processing methods used in TopSpin[®]. Moreover, for 2D NMR data, the vectors of intensities are folded into a matrix format, whereas for 3D NMR data, the vectors are folded into a cube. After application of any of these functions, a structure array containing the spectral data is generated.

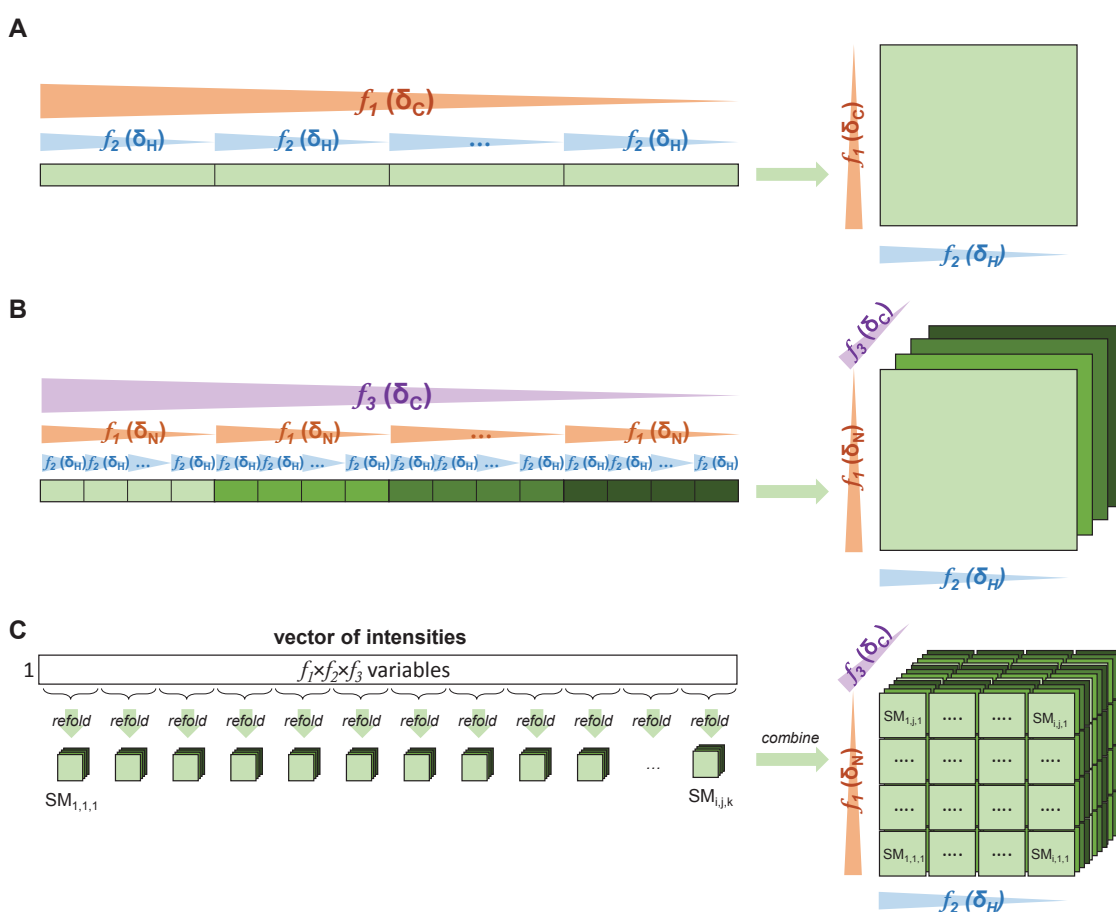


Figure 2.24. Data acquisition of Bruker NMR data. **A)** 2D NMR data. **B)** 3D NMR data. **C)** 3D NMR data acquired using submatrices (SMs).

In Bruker files containing the real intensities, these are sorted according to their corresponding chemical shifts or shieldings in decreasing order. For 2D NMR data (`2rr` files), intensities are sorted in decreasing order according first to the indirect dimension, f_1 , and then by the direct dimension, f_2 (**Fig. 2.24A**). For 3D NMR data (`3rr` files), intensities

are sorted by f_3 (indirect dimension), f_1 (indirect dimension) and f_2 (direct dimension), also in decreasing order (**Fig. 2.24B**).

Moreover, for large NMR spectra, spectral data are commonly acquired in smaller pieces, referred as submatrices (SMs). In this situation, the vector of intensities contains the vectorized data from all the SMs. SMs relative to the most shielded resonances are acquired and given first, whereas the SM relative to the most unshielded resonances are acquired last (**Fig. 2.24C**).

3.6.3 Exporting NMR data

Sometimes it may be convenient to export MATLAB[®]-processed NMR data to a particular NMR suite. For instance, this could be the case for determining *spin-spin* coupling patterns from this processed data, to perform peak picking, or to compare how the different preprocessing methods not included in the traditionally used NMR suites (*e.g.*, *icoshift*) have modified the raw data.

In order to export a MATLAB[®] variable containing the NMR data to a NMR suite, MATLAB[®] data must be converted to the same format as the originally imported NMR data file. In other words, MATLAB[®] data generated from Varian NMR files must be transformed into a tab-delimited `.csv` file, while MATLAB[®] data generated from Bruker NMR files must be transformed into a 32-bit integer binary file.

Having said this, despite it is possible to transform the 1D and 2D FT-NMR spectra into tab-delimited `.csv` files with MestReNova software, this program can only read the 1D FT-NMR spectra saved in this format, but not the 2D ones. Therefore, with this workflow, MATLAB[®] files generated from 2D NMR spectra acquired in Varian cannot be exported (**Fig. 2.22B**).

On the other hand, for exporting MATLAB[®] data to TopSpin[®], the original Bruker files containing the real intensities (`1r`, `2rr`, or the `3rrr` binary files) need to be replaced with the new binary file containing the MATLAB[®]-processed data. Since these files are equivalent to the ones generated by Bruker's TopSpin[®] software, not only TopSpin[®], but all NMR suites compatible with Bruker data are capable to import these files.

This binary file can be generated by executing the MATLAB[®] `fwrite()` function. However, for 2D and 3D NMR data, the 2D- or 3D-MATLAB[®] matrices must be reshaped into a vector before application of the MATLAB[®] `fwrite()` function. In order to export the data correctly, this reshaping operation must exactly correspond to the inverse of the refolding process applied during the import of the original binary file (**Fig. 2.24**).

Thus, with this workflow, all MATLAB® files generated from Bruker NMR data can be imported to TopSpin® regardless of their dimensionality.

4 CHEMOMETRICS

The data generated in NMR Metabolomics studies are very complex, since each NMR spectrum contains from thousands to millions of frequency data-points, from hundreds of resonances, and the number of analyzed samples is usually comprised between the several tens up to the hundreds.

In order to extract the biological information contained in these complex datasets, chemometrics methods can be used. Chemometrics is a science field focused on the extraction of knowledge from chemical systems by data-driven means [169], to address problems in chemistry [170], medicine [171], biology [172], chemical engineering [173], and more recently, in the ‘Omics’ sciences [174,175], among other fields.

In the first part of this section, a more comprehensive definition of Chemometrics is provided. In the second part of this section, the different data structures arrangements analyzed in this Thesis are introduced and the concept of bilinearity is given. Finally, the chemometric methods used in this Thesis to investigate the metabolomics datasets are described.

4.1 THE DEFINITION OF CHEMOMETRICS

The International Chemometrics Society define Chemometrics as [176]:

“The science of relating measurements made on a chemical system or process to the state of the system via application of mathematical or statistical methods”.

The scope of the Chemometrics research field covers a wide range of methods that can be applied in chemistry areas, from the generation of good quality data (optimization of experimental parameters, design of experiments, calibration and signal processing) to the extraction of meaningful information from these data (statistics, pattern recognition, modeling and structure-property-relationship estimations). Thus, Chemometrics intends to generate knowledge by building a bridge between these mathematical and statistical methods and their application in chemistry.

A more comprehensive and timely definition of Chemometrics, that refers the mentioned applications, can be found in the *Handbook of Qualimetrics and Chemometrics* [177] (**Fig. 2.25**):

“The chemical discipline that uses mathematical, logical and statistical methods: (a) to design or select optimal measurement procedures and experiments, (b) to provide maximum chemical information by analyzing chemical data, and (c) to provide knowledge from chemical processes”.

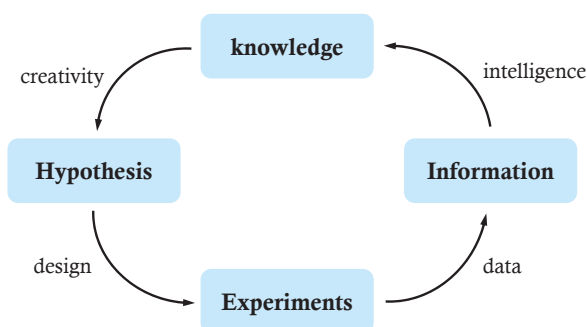


Figure 2.25. Chemometric circle of knowledge (adapted from [178]).

4.2 DATA STRUCTURES

With Chemometrics, chemical data are analyzed by multivariate methods. Within the field of Chemistry, several instrumental techniques can generate multivariate data, specially the spectroscopic and spectrometric (*e.g.*, NMR, UV, NIR, RAMAN, fluorescence, MS...) and chromatographic (*e.g.*, GC, LC, CE...) techniques, and also more recent approaches based in any of the formers, such as the ones in the transcriptomics field (*e.g.*, DNA microarray, RNA-seq). In contrast, examples of instrumental techniques that produce univariate measurements are conductimetries, calorimetries, and pH measurements.

Results for each analytical technique are expressed in many formats: intensities, concentrations, peak heights, integrals, absorbances, counts, *etc.*, where each one of these measurements represents a *variable*. Moreover, multivariate datasets can be constructed by combining data from more than one instrumental technique (*e.g.*, LC-DAD or GC-MS), either univariate or multivariate. We refer to a homogeneous dataset when all measurements have been acquired using the same instrumental approach, whereas a dataset is considered to be heterogeneous if the data come from different instruments.

In an Analytical Chemistry problem, it is important to consider what needs to be measured (variables), and also what is the analyzed entity that we are interested in (*e.g.*, a patient, a beverage, a plant, a manufactured product). We commonly refer to this entity as the samples of study. Thus, when data are generated by measuring k variables on n samples, this dataset can be arranged in a matrix format of size $n \times k$, as shown in **Figure 2.26**.

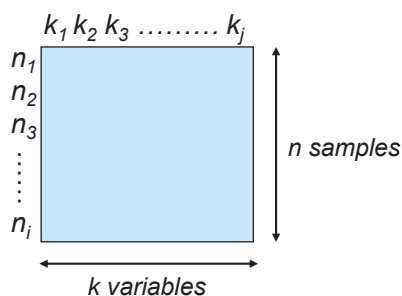


Figure 2.26. Data matrix.

In NMR metabolomics, for instance, a typical $n \times k$ data matrix is a ^1H NMR spectral datasets, having n ^1H NMR spectra in the rows, and each spectrum being composed by k chemical shift data-points. The same data structure can be found in a matrix of metabolite concentrations, with as many rows as analyzed samples, and with the same number of columns as detected metabolites. In Puig-Castellví *et al.* (2015) [30], these two dataset types were analyzed by chemometric means.

In addition, depending on the complexity of the metabolomics studies and on the nature of the data, other data structures arrangements may be recommended instead. The different data structure arrangements used in this Thesis are presented below.

For instance, a different data structure arrangement is used for 2D NMR spectral data. In this case, one sample is represented by one $n \times k$ matrix instead of a vector [119], and the dataset containing more than one sample gives a data cube (or a three-way data array, **Fig. 2.27A**). This type of structures is obtained with any analytical techniques in which the measurements are performed over two types of variables (ways, orders, directions) at the same time. In heteronuclear 2D NMR spectroscopy, for example, the two types of variables correspond to the two measured nuclei. This situation is also observed for HPLC-MS spectra, in which every measured intensity value is associated to a specific m/z and to a specific elution time.

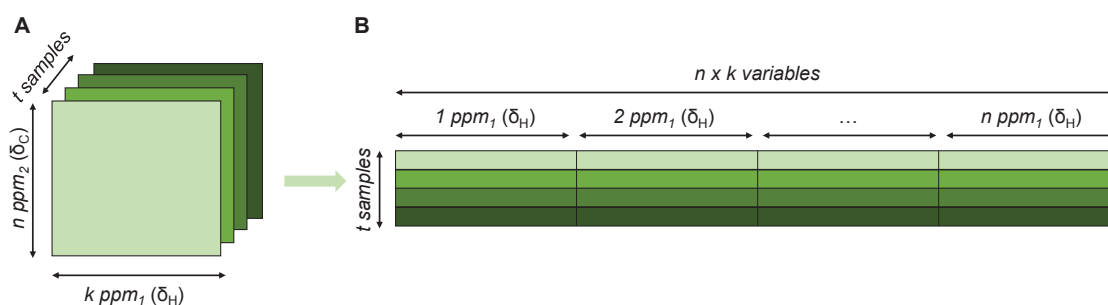


Figure 2.27. **A)** Three-way array. **B)** Unfolding of the three-way array into a two-way matrix.

In many circumstances it is convenient to convert an n by k by t three-way array into a two-way data matrix. In this Thesis, we have used two distinct approaches.

A first approach to do this dataset conversion consists in the unfolding of each sample matrix into a row vector, followed by the alignment of all the row vectors column-wisely. The result of applying this approach gives a data matrix of t rows and $n \times k$ columns, as presented in **Figure 2.27B**. This approach has been used in this Thesis for the analysis of a dataset containing several ^1H - ^{13}C HSQC NMR spectra (Puig-Castellví *et al.*, 2018 [119]).

A second approach is based on the direct column-wise data augmentation of the two-way data matrices relative to every sample. The resulting augmented data matrix has $n \times t$ rows and k columns. An illustrative example of this second augmentation is given in **Figure 2.28**. This arrangement also allows for augmenting data matrices of different number of rows. This approach has been used in this Thesis for the analysis of datasets of several samples measured with UHPLC-MS spectrometry (Puig-Castellví *et al.*, 2018 [125]).

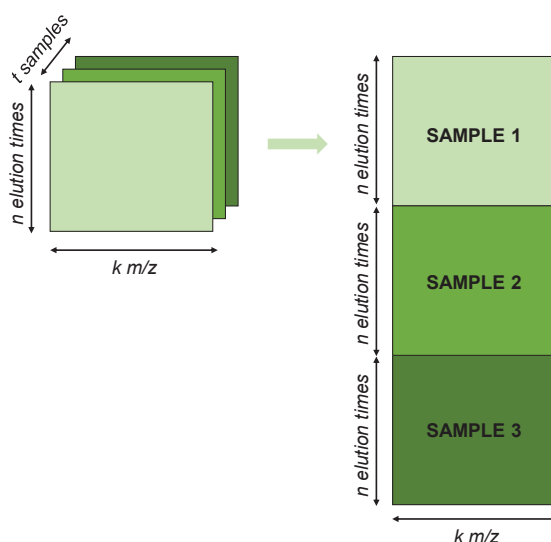


Figure 2.28. Column-wise augmentation.

A special mention needs to be done for multivariate data acquired over time, as in industrial process monitoring [179], or in time-course metabolomics experiments, the latter of special interest in our case (Puig-Castellví *et al.*, 2015 [118]; Puig-Castellví *et al.*, 2018 [119]). In these cases, the resulting data can be also considered a three-way data array, with as many data slices as screened time-points, as many rows as sampling points, and as many columns as measured variables (detected metabolites, or spectral intensity data-points). In the analysis of these datasets, a common procedure is the simultaneous analysis of the different data slices column-wisely augmented (**Fig. 2.29**). With this method, samples collected in the same

sampling point at different times are considered to give independent information, which is an assumption commonly done in the metabolomics field [118,119,180,181].

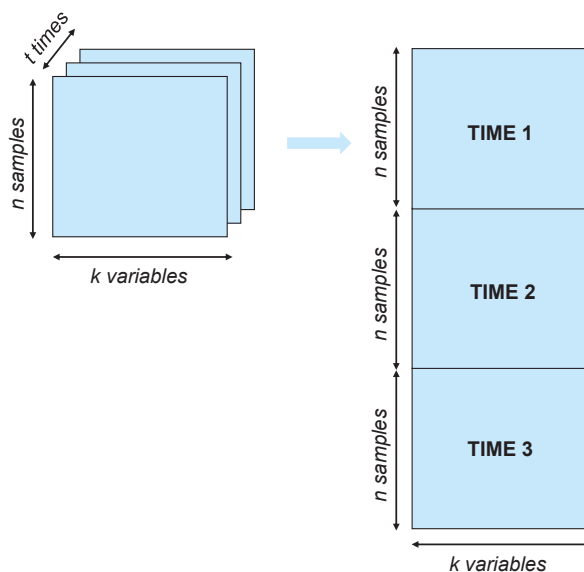


Figure 2.29. Three-way data matrix of a time-course monitoring experiment and its column-wise augmentation.

Finally, data matrices can also be augmented row-wisely. Row-wise augmented data matrices can be built when the set of samples were investigated using two (or more) different analytical platforms. Therefore, in this case, the resulting data matrix has the same number of rows as the original matrices, n , but the number of variables or columns becomes the sum of the number of variables for each one of the two datasets (**Fig. 2.30**).

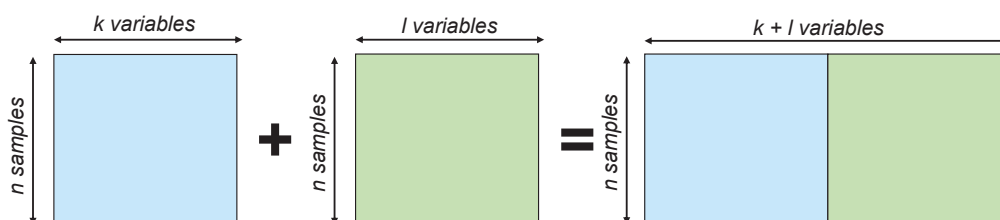


Figure 2.30. Row-wise matrix augmentation.

An example of this particular augmentation can be found in the fusion of the data from two parallel studies (a ^1H NMR metabolomics study and a UHPLC-MS lipidomic study) on the same set of samples [125], which is discussed in **Chapter 3**.

4.3 BILINEARITY

Most chemometrics methods are based on the assumption that analytical measurements contained in the analyzed \mathbf{X} datasets can be explained by a bilinear type of model similarly

to the one described by the Lambert-Beer law (eq. 2.16), in which the individual measured d_{nk} responses are the sum of the product of the s_{ak} instrumental responses of the a components, weighted by their c_{na} concentrations in the n samples, plus the corresponding e_{nk} residual part for every individual sample and response.

$$d_{nk} = \sum_{a=1}^A c_{na} s_{ak} + e_{nk} \quad \text{eq. 2.16}$$

Or in matrix notation:

$$\mathbf{D} = \mathbf{C}\mathbf{S}^T + \mathbf{E} \quad \text{eq. 2.17}$$

^1H NMR spectroscopy, as all the other spectroscopies, follows also in principle this bilinear measurement model. Thus, a \mathbf{D} matrix of a set of ^1H NMR spectra can be generated by the product between the concentration matrix of the sample constituents, \mathbf{C} , and the matrix of ^1H NMR spectra of these individual chemical constituents, \mathbf{S}^T , plus the matrix that contains the residual information not explained by the model, \mathbf{E} (Fig. 31).

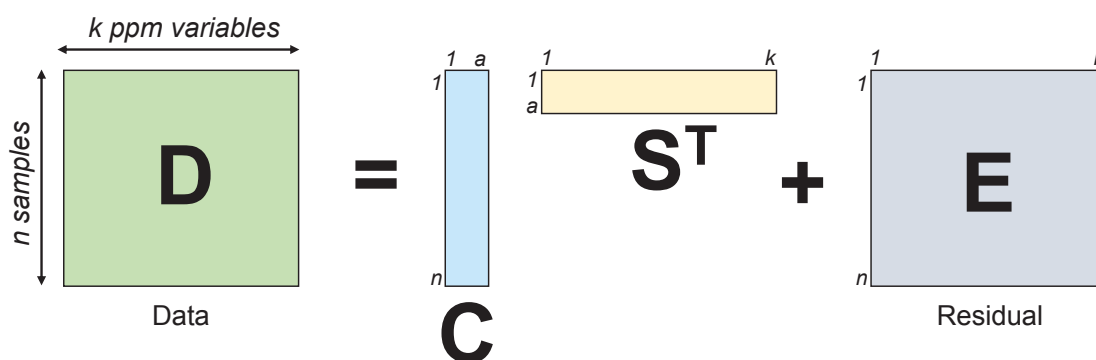


Figure 2.31. Matrix bilinear decomposition.

In agreement with this model, a ^1H NMR spectrum acquired on a mixture sample can be regarded as the sum of the set of ^1H NMR spectra acquired on each of the sample constituents, weighted by the relative concentrations of each of these constituents (Fig. 2.32).

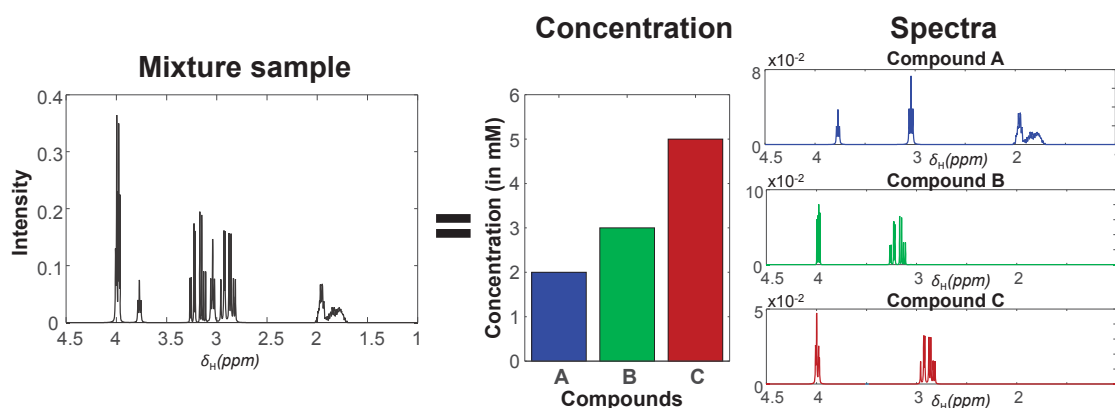


Figure 2.32. ^1H NMR spectrum of a mixture simple as the sum of ^1H NMR of single compounds.

For ^1H NMR data, this model also includes the residual term, which contains all disturbances not present in the ideal ^1H NMR spectra (*e.g.*, noise, instrumental artifacts, solvent impurities...).

For the analysis of datasets obtained in metabolomics studies, several chemometric methods are used for exploration, regression and classification purposes. Most of these chemometric methods (*e.g.*, PCA, PLS, MCR-ALS) perform the matrix decomposition of \mathbf{D} in agreement with the bilinear method of [equation 2.16](#) and some specific constraints (*e.g.*, orthogonality, no-negativity) of each particular chemometric method.

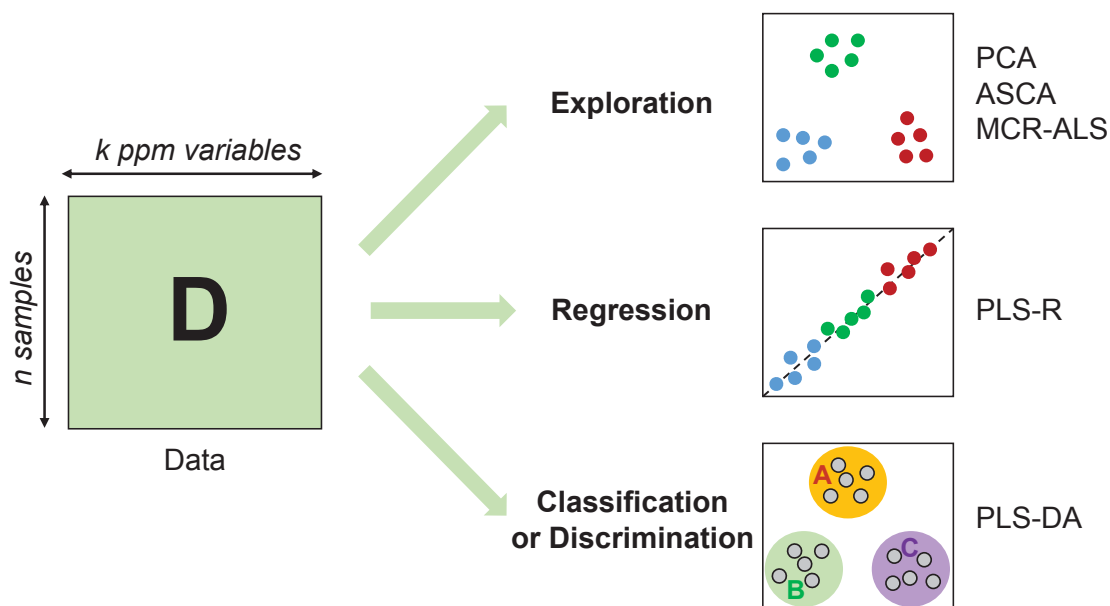


Figure 2.33. Chemometric methods based on the bilinear model, grouped by the purpose of their application (exploration, regression, and classification/discrimination).

The chemometric methods used in this Thesis are described in the next section.

4.4 CHEMOMETRIC DATA ANALYSIS METHODS

In this Thesis, the Chemometric data analysis methods used are PCA, PLS-R, PLS-DA, ASCA, and MCR-ALS.

4.4.1 Principal Component Analysis (PCA)

PCA is the most common multivariate data analysis method used to obtain a first overview of the structure of a data matrix \mathbf{X} [24,182,183].

The use of PCA is based on the idea that most of the variance observed within a dataset is caused by a few number of variance sources that affect a specific set of variables. Since these sources of variation are considered to be independent, the total observed variance can be expressed as the set of latent variances:

$$\text{Variance}(\mathbf{X}) = \text{Variance}(\text{source}_1) + \text{Variance}(\text{source}_2) + \dots + \text{Variance}(\text{source}_a) \quad \text{eq. 2.18}$$

In addition, for each set of original variables affected by the same source of variation (correlated variables), they can be linearly combined into the same principal component or PC, and each PC represents one latent variable. With this process, the resulting PCs are uncorrelated (orthogonal), and the total number of variables is substantially reduced, making the data analysis much simpler and representative of the actual variation sources.

Each PC extracted from data matrix \mathbf{X} allows the data projection onto two subspaces, coined as the 'score space' and the 'loading space'. Data projected in these two subspaces give the scores, \mathbf{T} , and the loadings, \mathbf{P}^T , which are obtained by the bilinear decomposition of the data matrix \mathbf{X} according the following equation:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \dots + \mathbf{t}_a\mathbf{p}_a^T + \mathbf{E} = \mathbf{TP}^T + \mathbf{E} \quad \text{eq. 2.19}$$

where \mathbf{t}_a and \mathbf{p}_a represents the different scores vectors and loading vectors of the principal component a obtained in the decomposition.

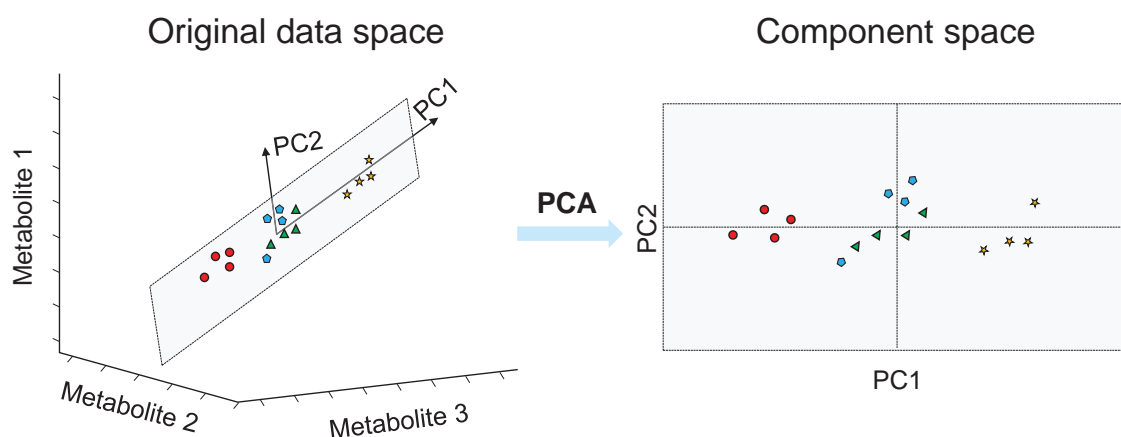


Figure 2.34. Score projection of the original samples. Adapted from [184].

From the analysis of the scores, it is possible to identify how different are the samples. On the other hand, from the analysis of the loadings, it is possible to determine which are the original variables that have stressed most the difference among these samples.

Because of all the constraints used during PCA decomposition (orthogonality, normalization and maximum variance obtained by matrix deflation) [185,186], the solution from applying PCA on a dataset is unique. There are several algorithms that can decompose the data with the given properties, such as the Singular Value Decomposition (SVD) [187] and the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm [188]. The main difference between these two methods is that SVD computes all PCs at one, while NIPALS compute each component one by one.

Even though an \mathbf{X} matrix can be always fully decomposed in the scores and loading matrices if enough components are used (\mathbf{E} is then equal to a matrix of zeros), one important aspect of the data analysis with PCA is to identify the number of PCs that describe the significant dataset variance not attributed to noise. To decide the proper number of PCs, several strategies can be used, such as the Scree test [189], the eigenvalue below one [190] method for auto-scaled data, and the Broken stick rule [191], among others.

Interpreting PCA data

It is possible to achieve an understanding of the analyzed dataset by exploring the geometrical projection of the scores and loadings on the scores plot and on the loading plot, respectively.

The projection of the scores in the space defined by the principal components allows obtaining of an overview of the similarity of the studied samples. The closer the samples are in the scores plot, the more similar they are in the considered PC plane, and vice versa. This

strategy of projecting the samples onto the scores plane, among other strategies, has been also used with the aim of detecting outliers [185].

An example of a typical PCA scores plot is presented in **Figure 2.35A**. In this example, 30 ^1H NMR spectra representative of yeast cultured at 2 different temperatures are plotted on the PCA scores plot. As observed, samples from two different classes (yeast cultures grown at two different temperatures) are separated in PC1 (54.28% of the explained variance), whereas PC2 (20.81% of the explained variance) describes the variance among samples within the same class.

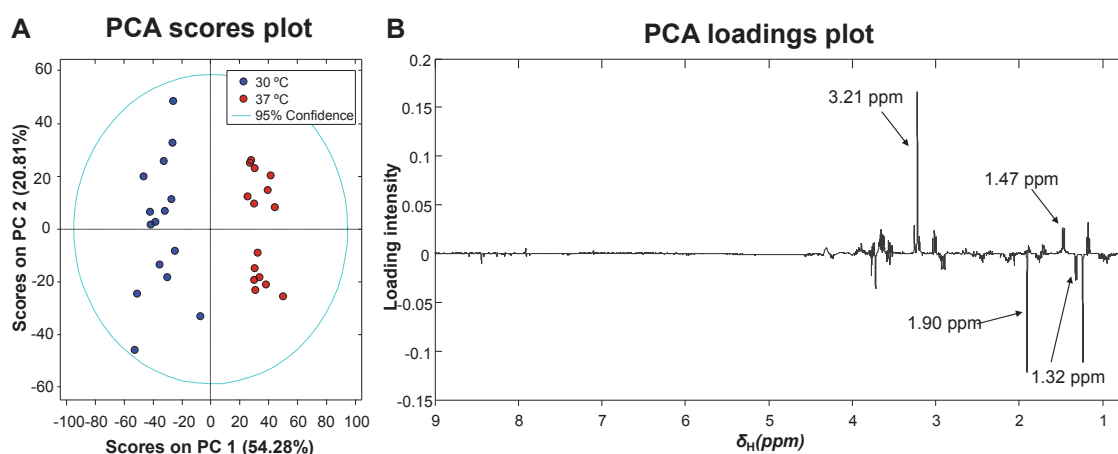


Figure 2.35. PCA analysis of a ^1H NMR dataset. **A)** PCA scores plot. **B)** PCA loadings plot.

On the other hand, the representation of the original variables in the axis defined by the loadings allows the detection of the most important variables in the given principal component, as well as to identify correlation between variables. In the example of **Figure 2.35B**, to gain a better insight, the loadings of PC1 are represented against the ppm scale. Since PCA centers the data, the resulting loading ' ^1H NMR spectrum' contains information from proton resonances with positive and negative loading values. Those with the same sign correlate in the same direction among them, and those with the opposite sign correlate inversely. Moreover, resonances associated to positive loadings in PC1 (**Fig. 2.35B**) are more important in samples associated to positive scores in the same component (**Fig. 2.35A**), whereas compounds represented with negative resonances in the loading plot are more important to describe the samples associated with negative scores in PC1. For instance, in the given example, glycerophosphorylcholine (3.21 ppm) and L-alanine (1.47 ppm) are more abundant in samples cultured at 37°C, whereas acetic acid (1.90 ppm) and 1.32 ppm (L-threonine).

Examples of PCA applications

PCA is vastly used in the metabolomics field. Recent examples of PCA applications in this field include:

- Regional discrimination of food, such as Mexican peppers [182] or Chinese rice [192].
- Evaluation of the effects of petrochemical contamination in mussels [193].
- Discrimination of bacteria strains in juice [183] and yeast strains in wine [194] based on the expressed metabolic response.
- Exploration of the different metabolic processes involved in breast cancer [195] and in plant growth development [196].
- Evaluation of extraction protocols of metabolites from *Curcuma* species [197].
- Evaluation of the metabolic effects of chemotherapeutic drugs in rats [198,199].
- Food quality control in ginseng samples [200].

In this Thesis, PCA has been primarily used as an exploratory tool to investigate the yeast metabolome when it was exposed to different environmental conditions, such as different temperature acclimation [30,125] (see **Chapter 3**), and when yeast was cultured under different media compositions [118,119,159] (consult **Chapter 3** and **Chapter 4**).

Furthermore, PCA was also applied to data generated from UHPLC-MS metabolomics analyses. Specifically, the datasets consisted of Total Ion Current (TIC) chromatograms obtained in either ESI(-) and ESI(+) modes representative of the yeast lipidome expressed under different growth temperatures [125].

Lastly, PCA was carried out on datasets of 2D NMR spectra. Since these datasets are, structurally speaking, a three-way data array, they were unfolded using the strategy of **Figure 2.27** prior PCA analysis. These datasets were an array of ^1H - ^{13}C HSQC NMR spectra of metabolomics extracts from yeast cultured using different media [159], and also to the same dataset after removal of the noise variables by the VOI-filtering strategy (see **section 3.5** in **Chapter 2**) [119,159].

4.4.2 Partial Least Squares (PLS)

In simple (univariate) linear regression, the y -intercept (β_0) and slope (β_1) coefficients of the model of **equation 2.20** are estimated by minimizing the sums of squared residuals (the difference between the measured response, y , and the corresponding predicted response, \hat{y}) term by means of a least squares algorithm.

$$y = \beta_0 + \beta_1 x + e \quad \text{eq. 2.20}$$

In the multivariate dimension, where \mathbf{X} corresponds to a data matrix of n samples and k variables (or regressors), **equation 2.20** is extended by adding one coefficient per variable, as shown in **equation 2.21** below:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + e \quad \text{eq. 2.21}$$

In matrix notation, for a set of samples, this equation is written as:

$$\mathbf{y} = \mathbf{bX} + \mathbf{e} \quad \text{eq. 2.22}$$

In PLS regression (or PLS-R) [188,201], on opposite to multivariate linear regression, the regressors used are a set of independent latent variables built by bilinear decomposition in such a way that the covariance of these latent variables to the \mathbf{y} -variables is maximized. In PLS-R, each latent variable explains part of the \mathbf{X} and \mathbf{y} data, and therefore, the percentages of the explained \mathbf{X} -variance and of the explained \mathbf{y} -variance can be estimated for every latent variable.

One of the most used algorithms to perform PLS analyses is NIPALS [188]. In PLS, similarly to PCA, a set of scores and loadings for every latent variable are obtained. However, in addition, a weight vector (\mathbf{w}) for every latent variable is used to optimally correlate the variance in \mathbf{X} and \mathbf{y} subspaces and maintain the orthogonality of \mathbf{X} [202], which can be then used to identify the original variables more linked to the \mathbf{X} variables.

PLS can also be used as a discriminant method [203]. In this case, the PLS method is referred to as PLS Discriminant Analysis or PLS-DA. This chemometric method has been commonly used in the metabolomics field. For instance, a typical metabolomics problem solved by PLS-DA is the classification of two groups of samples (*e.g.*, control and exposed samples) from their associated metabolic profiles. In this example, \mathbf{y} will define the class membership of samples, whereas the measured metabolomes for the existing samples are in the \mathbf{X} dataset.

The discriminant power of the PLS-DA method is achieved by transforming the class vector, \mathbf{y} , which has the classes or categories defining the membership of every sample, into a binary vector with as many n elements as samples. With this vector, it is denoted whether each sample belongs to a specific class (with ones) or not (with zeros), as represented in **Figure 2.36**.

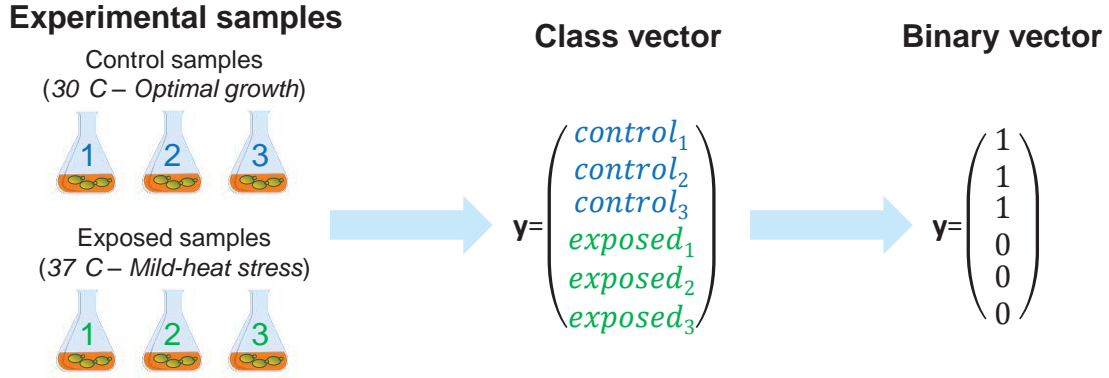


Figure 2.36. Building of the \mathbf{y} binary vector from a conceived experimental design, consisting of 3 cultures of yeast cultured at 30°C and 3 cultured at 37°C.

With PLS-DA, it is also possible to determine the variables that maximize most the discrimination between the two groups. For a metabolomics experimental design, in which the variables would be the measured concentrations for a concrete set of metabolites, these metabolites can be considered as the potential biomarkers of the studied problem.

In this Thesis, we have used the VIP (Variable Important in Projection) strategy, postulated by Wold *et al.* in 1993 [204], for selecting potential biomarkers [30,119]. In the VIP strategy, for every k variable, a VIP value is calculated by the following formula:

$$VIP_k = \sqrt{\frac{\sum_{f=1}^F w_{kf}^2 SSY_f K}{SSY_{total} F}} \quad \text{eq. 2.23}$$

where w_{kj} is the weight value of the k variable and f latent variable, SSY_f is the squared sum of the explained variance in the f latent variable, K is the total number of measured variables, SSY_{total} is the squared sum of \mathbf{Y} , and F is the total number of latent variables used in the PLS model.

For the set of calculated VIP values, a high VIP value would indicate that the corresponding variable is relevant for the discrimination, and vice versa. Thus, the VIP value allows ranking the quality of each variable to discriminate the studied classes.

Since the mean of the squared values of the VIPs is 1, this value has been established as the threshold level to consider that a variable has discriminant power or not [205]. Despite this, the threshold level of 1 can be risen in order to only select the variables with the most discriminant power, and therefore, the best biomarker metabolites [206].

Assessment of the PLS model

A practical approach to define the optimal number of latent variables in PLS is by calculating the PRESS (Predicted RESidual Sum of Squares) value for different PLS models with

increasing number of latent variables until the PRESS does not show any improvement, indicating the possibility of overfitting (fitting the noise in the PLS model) if more components are added. The same inspection can be performed from the RMSEC (Root Mean Square Error of Calibration) values.

$$PRESS = \sum_i (y_i - \hat{y}_i)^2 \quad \text{eq. 2.24}$$

$$RMSEC = \sqrt{\frac{PRESS}{n}} \quad \text{eq. 2.25}$$

The best approach to validate PLS-DA models is by using an external set of validation samples, different from the training set, not used to build the model, and evaluated from their prediction results about the class membership of every new tested validation sample.

If external validation test samples are not available, the quality of the performance of a method can be evaluated using Cross Validation (CV) strategies. In CV, the samples of the original dataset are divided into two or more different datasets, the calibration and the validation datasets. Then, the calibration set or model is used to build the PLS model, and the membership class of the samples from the validation test set are predicted. There are different ways to apply CV, based on different strategies for splitting the dataset. Examples of these methods are Venetian blinds, Leave-One-Out and random subsets, among others [203]. In this Thesis, we have used Venetian blinds CV and Leave-One-Out CV (LOOCV). In Venetian blinds CV, the dataset is split in s subsets, where each subset is determined by selecting every s^{th} sample in the dataset, starting at samples numbered 1 through s , being s the number of splits. On the other hand, in LOOCV, all samples but one are used as a calibration dataset, and the membership class of the sample left out is predicted using the PLS model based on this calibration dataset. This process is repeated n times, being n the number of samples in the dataset, and one (different) sample is excluded every time (Fig 2.37).

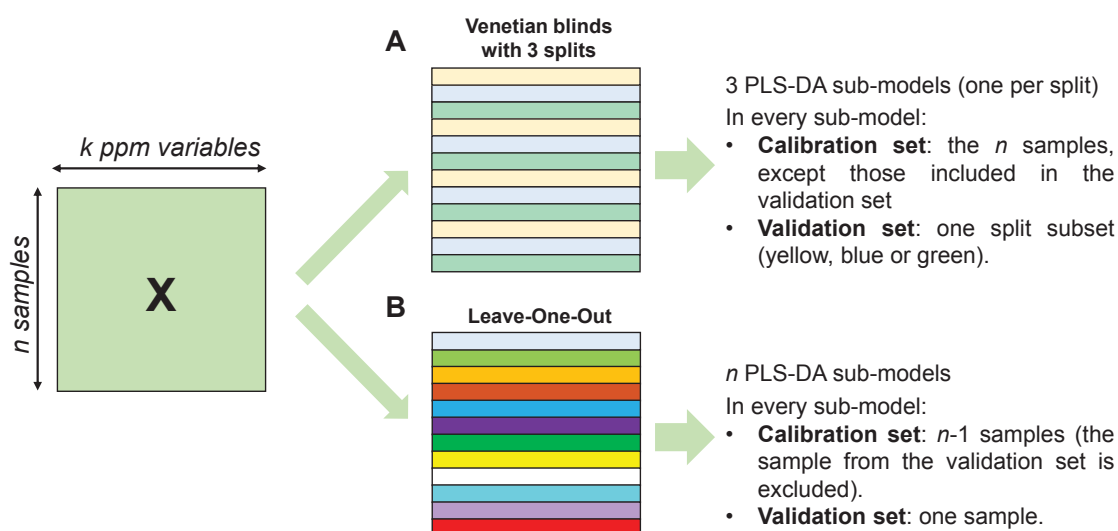


Figure 2.37. Cross-validation. **A)** Venetian blinds CV. **B)** Leave-One-Out CV.

Regardless the tested samples are from an external dataset or CV is used, it is possible to estimate the discriminant power of PLS-DA based on the number of misclassification (NMC, or the number of samples which class was correctly predicted).

Although an NMC of 0 would indicate a good PLS-DA model, it is possible that the result has been attained just by lucky random choice of the samples in the CV test, meaning that the NMC may not reflect the real discrimination power of the model. To overcome this limitation, a permutation test during PLS-DA analysis can be also applied [207]. For instance, in Puig-Castellví *et al.* (2015) [30], samples were permuted 1,000 times and PLS-DA with LOOCV was applied every time. After 1,000 permutations, it was observed that the NMC number was the lowest for the original experimental situation ($p < 0.001$), demonstrating that the sample discrimination was not just an outcome derived from pure chance.

Similar to the RMSEC mentioned before, in PLS-DA methods with CV, the RMSECV term (Root Mean Square Error of Cross Validation) can be calculated in order to help for the estimation of the optimal number of latent variables in PLS-DA modelling.

Orthogonal Signal Correction-PLS-DA (OSC-PLS-DA)

In metabolomics, it is becoming more frequent the application of PLS-DA on data previously filtered with OSC, which removes the variance (information) contained in the X matrix that is uncorrelated (orthogonal) to Y [208]. Examples of ^1H NMR metabolomics studies that take advantage of OSC-PLS-DA can be found elsewhere [30,209-211].

4.4.3 ANOVA-Simultaneous Component Analysis (ASCA)

Most metabolomics studies have an experimental design of several factors. For instance, a typical experimental design may include the cultivation of different microbial strains (factor 1: 'strain') under different conditions (factor 2: 'treatment') [24,183,212-214] (Fig. 2.38).

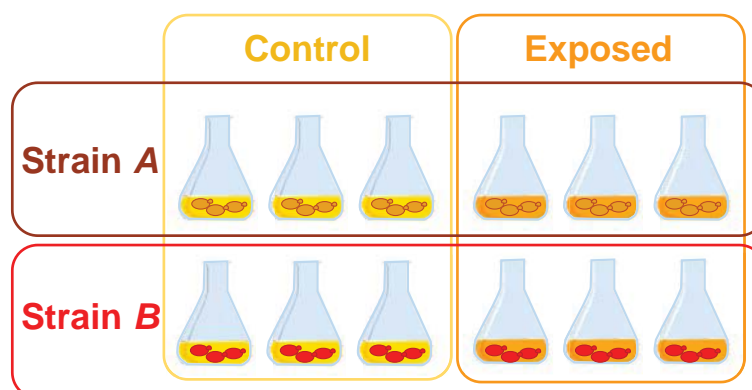


Figure 2.38. Experimental design with two factors, treatment and strain, and triplicates.

ANOVA [215] is a univariate statistical method widely used to evaluate and quantify the effect of different experimental factors on the observed outcomes of different experiments. For instance, for the example represented in Figure 2.38, the influence of the two factors can be evaluated with a two-way ANOVA model (eq. 2.26).

$$x_n = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + e_n \quad \text{eq. 2.26}$$

In this equation, μ is the offset term, α_i is the additive main effect of the factor α (treatment) on level i , β_j is the additive main effect of the factor β (strain) on level j , $\alpha\beta_{ij}$ is the interaction term between the two factors α and β on levels i and j , respectively, and e is the error.

Having said that, ANOVA is a univariate statistical method, meaning that only one variable (*e.g.*, metabolite concentration, peak area) at a time can be evaluated. Moreover, in metabolomics studies, it is also important to identify the metabolomic response (*i.e.*, the set of altered pools of metabolites) associated to the studied factors, but ANOVA, as a univariate method, does not take into account the covariance between variables, hampering the correct interpretation of the data from these metabolomic datasets.

For multivariate datasets, Multivariate ANOVA (MANOVA) can be used, which is the extension of ANOVA for the simultaneous evaluation of multiple variables. However, MANOVA (in his classical implementation) can only be applied when the number of samples is larger than the number of variables [216]. To cope with this limitation, ANOVA – Simultaneous Analysis component (ASCA) method has been proposed as an alternative method to deal with this type of multivariate datasets, typical in the metabolomics field [181,217]. In ASCA, the contributions of the different factors are disentangled into different

matrices using linear models based on ANOVA, which are then analyzed by Simultaneous Component Analysis (SCA).

Thus, for every scalar element from the two-way ANOVA model introduced above (eq. 2.26), in ASCA they are matrices instead (Fig. 2.39).

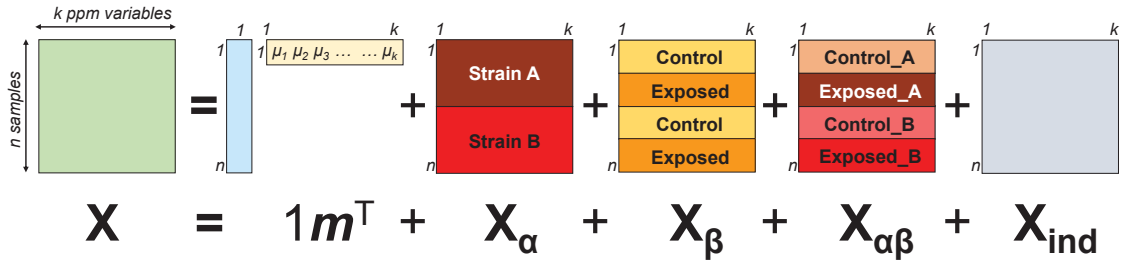


Figure 2.39. Matrix decomposition in an ASCA model of two factors and their interaction, analogous to the experimental design of Figure 2.38.

In the equation of this ASCA model (Fig. 2.39), \mathbf{m} is a row vector containing all μ estimates for every considered variable; \mathbf{X}_α , \mathbf{X}_β and $\mathbf{X}_{\alpha\beta}$ are the matrices that contain the measured response relative to the α and β effects and to the interaction between the two ($\alpha\beta$), and \mathbf{X}_{ind} is the matrix that contain the residual variations caused by differences between samples with the same treatment. For \mathbf{X}_α , \mathbf{X}_β and $\mathbf{X}_{\alpha\beta}$, all rows relative to the same factor and level are identical. In the process of this matrix decomposition, it is considered that all factor and interaction matrices (\mathbf{X}_α , \mathbf{X}_β , $\mathbf{X}_{\alpha\beta}$, and \mathbf{X}_{ind}) are mutually orthogonal.

In the next step of the SCA decomposition, multiple data matrices that contain the same number of columns (variables) are concatenated column-wisely into an augmented data matrix, and this augmented dataset is subsequently decomposed into a matrix of scores and a matrix of loadings, analogously to PCA (Eq. 2.27). These loadings and scores can be investigated to identify the source of variances of the augmented dataset [218].

Hence, the resulting ASCA model is given by the following equation:

$$\mathbf{X} = \mathbf{1}\mathbf{m}^T + \mathbf{T}_\alpha\mathbf{P}_\alpha^T + \mathbf{T}_\beta\mathbf{P}_\beta^T + \mathbf{T}_{\alpha\beta}\mathbf{P}_{\alpha\beta}^T + \mathbf{E} \quad \text{eq. 2.27}$$

Where \mathbf{E} contains all variations not contained in any factor or interaction matrix:

$$\mathbf{E} = \mathbf{X}_{\text{ind}} + \mathbf{E}_\alpha + \mathbf{E}_\beta + \mathbf{E}_{\alpha\beta} \quad \text{eq. 2.28}$$

In these two equations, α , β and $\alpha\beta$ subindices refer to the two studied factors and the interaction between these two, respectively; \mathbf{m} is a row vector containing all μ estimates for every considered variable ($\mathbf{m} = \mu_1, \dots, \mu_k$); the SCA component scores of each submodel are given by the matrices indicated by \mathbf{T}_α , \mathbf{T}_β , and $\mathbf{T}_{\alpha\beta}$; and the associated submodel loadings are given by matrices \mathbf{P}_α , \mathbf{P}_β , and $\mathbf{P}_{\alpha\beta}$ [181].

Because of the SCA application, each loading is orthogonal to the rest of loadings within a given submodel, and the components are ordered in decreasing order of explained variance.

In this Thesis, we have used ASCA to evaluate the response on yeast metabolism over time [118,119]. Thus, the first investigated ASCA models were based on the ANOVA equation of **Fig. 2.39**, and the studied factors are ‘treatment’ and ‘time’. ASCA analyses were performed on unscaled data [118] and on data scaled by the standard deviation of the reference group (samples collected at the initial time-point) [119], as recommended by Timmerman *et al.* [219].

Permutation tests in ASCA

To assess the statistical significance of every factor and of their interactions in the ASCA models, permutation tests are used [220]. In these tests, the null hypothesis (H_0) states that an experimental factor has no influence on the outcome of the experiment.

To check H_0 , the rows from the original \mathbf{X} dataset are permuted a certain number of times, and for each permuted dataset, ASCA is applied and the sum of squares for the obtained \mathbf{X}_α , \mathbf{X}_β , and $\mathbf{X}_{\alpha\beta}$ ($\|\mathbf{X}_\alpha\|^2$, $\|\mathbf{X}_\beta\|^2$ and $\|\mathbf{X}_{\alpha\beta}\|^2$) are calculated. Next, the sum of squares obtained from each permuted datasets is compared to the sum of squares obtained from the original \mathbf{X} dataset. The sum of squares is connected to the magnitude of the considered effect. This means that for a significant f factor, $\|\mathbf{X}_f\|^2$ will be also large, and vice versa.

Thus, to establish whether a factor is significant or not, the sum of squares should be larger in the original (non-permuted) situation. In a permuted dataset, samples should be more homogeneously distributed than in the original dataset, and therefore, groups of samples should be more similar among them, resulting in a lower sum of squares.

The level of significance is provided as a p-value for each evaluated factor or interaction, and it represents the fraction of the permutations where:

$$\|\mathbf{X}_f\|_{original}^2 < \|\mathbf{X}_f\|_{permuted}^2 \quad \text{eq. 2.29}$$

If the associated p-value for a considered factor is larger than a prefixed significant level, then the H_0 is accepted, and therefore the effect is not significant.

4.4.4 Multivariate Curve Resolution –Alternating Least Squares (MCR-ALS)

Multivariate Curve Resolution (MCR) is based on the same standard bilinear model ($\mathbf{D} = \mathbf{CS}^T$) introduced in **eq. 16** and depicted in **Figure 2.30**. MCR problem was first formulated by Lawton and Sylvestre in 1971 [221] and later implemented in an Alternating Least Squares (ALS) algorithm [36,222,223]. MCR-ALS has become one of the most popular MCR methods because of its simplicity and robustness.

The goal of MCR-ALS is to estimate meaningful matrices of concentration and spectra, \mathbf{C} and \mathbf{S} , from the analysis of the data matrix \mathbf{D} using chemical knowledge in the form of constraints that give information about the type of the measurements and the chemical system under study. The analysis of the \mathbf{C} and \mathbf{S} resolved matrices allow for the investigation of the underlying changes in the composition of the different samples represented in the \mathbf{D} data matrix.

Other chemometric methods based on Factor Analysis can be used to obtain concentration and spectral estimates from multivariate datasets (*e.g.*, CLS, PCR, and PLS-R) [224,225], but two main differences distinguish MCR-ALS from these other multivariate regression methods. First, in MCR-ALS, the main goal is the resolution of the pure response profiles of the components of a mixture. Therefore, for instance, the NMR resolved spectra can be directly interpreted from the chemical point of view, whereas the loadings in CLS, PCR, and PLS-R are mathematical solutions without a direct physical interpretation (for instance they are orthogonal and they can have negative values. Secondly, these regression methods are commonly used for quantitative purposes and require building a calibration curve (thus, known concentration measurements), while MCR-ALS does not (although if this information is available, it can be also used for improving resolution and perform quantitative determinations, see [226,227]).

The algorithm of MCR-ALS includes 3 main steps:

In the first step, the number of components is estimated by calculating the singular values of \mathbf{D} with the SVD method [187]. It is assumed that the larger of singular values defines the chemical rank (mathematical rank in absence of experimental noise) of \mathbf{D} , and that this coincides with the number of chemical sources of the data variance or components of the investigated system. Then, similarly to the scree test used to decide the number of components in the PLS (see PLS-DA [section 4.4](#)), the singular values can be plotted as a function of the number of components, and the chemical rank of \mathbf{D} would be defined by the maximal number of components that are not representative of noise (components of noise are associated with small singular values and decrease regularly). This step needs to be performed carefully because it defines the number of chemical patterns (*e.g.*, chemical species) to be resolved.

In the second step, an initial estimate of either \mathbf{C} or \mathbf{S}^T matrix should be defined ($\mathbf{S}_{\text{init}}^T$ or \mathbf{C}_{init} , respectively). These estimates represent a starting point of the solutions that will be obtained with the ALS iterative optimization. The selection of proper initial estimates saves computation time and minimizes convergence problems. For this reason, random estimation of \mathbf{C} or \mathbf{S}^T should be avoided. In this Thesis, we have used a method based on the selection

of the purest (the most dissimilar) samples or variables [228] from the raw \mathbf{D} dataset, but other methods exist, such as the Evolving Factor Analysis (EFA) [229] or the use of previously known spectra of pure chemical compounds.

In the third step, the iterative ALS optimization is performed. The first iteration of the ALS algorithm works in the following manner:

- First, $\hat{\mathbf{C}}$ or $\hat{\mathbf{S}}^T$ are calculated depending of the nature of the chosen initial estimate matrix obtained in the previous step ($\hat{\mathbf{C}}$ from $\mathbf{S}_{\text{init}}^T$, or $\hat{\mathbf{S}}^T$ from \mathbf{C}_{init}).
- Then, if applicable, the appropriate constraints are applied, and the obtained constrained matrix is used after to calculate the complementary matrix ($\hat{\mathbf{C}}$ from the constrained $\hat{\mathbf{S}}^T$, or $\hat{\mathbf{S}}^T$ from the constrained $\hat{\mathbf{C}}$). Some of the constraints can be embedded in the local least squares solution, like non-negativity using non-negative least squares approaches ([230,231]).
- Then, the original \mathbf{D} matrix is compared to the reconstructed $\hat{\mathbf{D}}$ matrix obtained from the reconstructed $\hat{\mathbf{C}}$ and $\hat{\mathbf{S}}^T$. If $\hat{\mathbf{D}}$ is similar enough to \mathbf{D} , then the iterative process is stopped and the current $\hat{\mathbf{C}}$ and $\hat{\mathbf{S}}^T$ matrices are defined as the solutions of the problem \mathbf{D} . Otherwise, another ALS iteration is started, using the current values of $\hat{\mathbf{C}}$ or $\hat{\mathbf{S}}^T$ as new initial estimates.

It is important to note that, due to the nature of the ALS method, the profiles from \mathbf{C} and \mathbf{S}^T matrices fulfill the natural constraints applied during the ALS optimization, like for instance non-negativity. However, in contrast to other Factor Analysis methods like PCA or PLS, these profiles are not mutually orthogonal.

Optimization through ALS is repeated until convergence, or until the maximum number of iterations predefined by the user is reached. The convergence criterion is based on the comparison of the fit obtained in two consecutive iterations. When the relative difference in fit is below a threshold value, the optimization is finished. A workflow summarizing the MCR-ALS algorithm is presented in **Figure 2.40**.

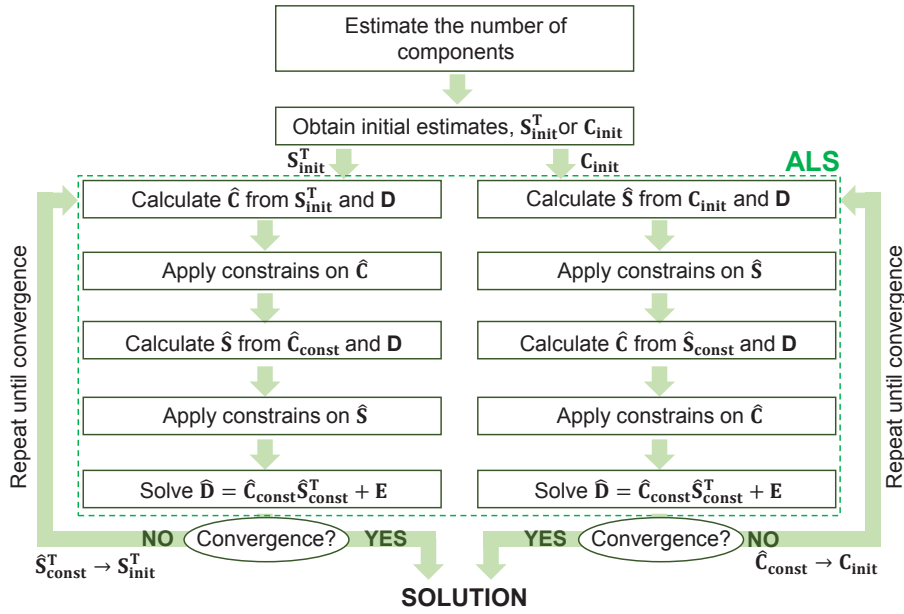


Figure 2.40. MCR-ALS algorithm. \mathbf{C} and \mathbf{S} matrices denote for concentration and spectra matrices. \mathbf{S}_{init}^T , initial estimates of \mathbf{S} ; \mathbf{C}_{init} , initial estimates of \mathbf{C} ; $\hat{\mathbf{S}}$, estimates of \mathbf{S} using ALS; $\hat{\mathbf{C}}$, estimates of \mathbf{C} using ALS; $\hat{\mathbf{S}}_{const}$, constrained $\hat{\mathbf{S}}$; $\hat{\mathbf{C}}_{const}$, constrained $\hat{\mathbf{C}}$; $\hat{\mathbf{D}}$, \mathbf{D} estimated from $\hat{\mathbf{C}}_{const}$ and $\hat{\mathbf{S}}_{const}$.

The quality of the MCR-ALS solution is evaluated from the explained variance (R^2 , eq. 2.30) and the lack-of-fit (*lof*, eq. 2.31) parameters.

$$R^2 (\text{in } \%) = \frac{\sum_{i=1}^I \sum_{j=1}^J \hat{d}_{ij}^2}{\sum_{i=1}^I \sum_{j=1}^J d_{ij}^2} 100 \quad \text{eq. 2.30}$$

$$lof (\text{in } \%) = \sqrt{\frac{\sum_{i=1}^I \sum_{j=1}^J (d_{ij} - \hat{d}_{ij})^2}{\sum_{i=1}^I \sum_{j=1}^J d_{ij}^2}} 100 \quad \text{eq. 2.31}$$

where d_{ij} are the individual data values for sample i and variable j , and \hat{d}_{ij} are the corresponding MCR-ALS calculated data values from the same sample and variable.

Hence, a good MCR-ALS solution is associated with a high R^2 value and with a low *lof* value.

Ambiguities and constraints

If no constraints are applied, there exists an infinite number of possible solutions of equation 2.17 for \mathbf{C} and \mathbf{S}^T , which multiplied with each other, give the same $\hat{\mathbf{D}}$. This is exemplified in equation 2.32, in which an infinite number of \mathbf{C} and \mathbf{S}^T solutions can be made using any nonsingular invertible matrix \mathbf{T} .

$$\hat{\mathbf{D}} = \mathbf{C}_{old} \mathbf{S}_{old}^T = (\mathbf{C}_{old} \mathbf{T})(\mathbf{T}^{-1} \mathbf{S}_{old}^T) = \mathbf{C}_{new} \mathbf{S}_{new}^T \quad \text{eq. 2.32}$$

To break this ambiguity and constraint the solutions to those with chemical meaning, a set of constraints can be applied.

There are three types of ambiguities: permutation, intensity and rotation ambiguities. Permutation ambiguities refer to the order of the components in \mathbf{C} and \mathbf{S}^T matrices. There is no prevalence and they may be ordered randomly, although the component correspondence in the two modes should be kept (*e.g.*, each column of \mathbf{C} matrix with each row of \mathbf{S}^T matrix). Intensity ambiguities only alter the magnitude of the \mathbf{C} and \mathbf{S}^T profiles, and can be avoided by closure constraints on \mathbf{C} or by normalization of the spectra in \mathbf{S}^T . Rotation ambiguities are caused by the fact that linear combination of the resolved profiles fulfilling the constraints and fitting the same data may exist. They cause a change in the profile shapes. Depending on the data nature and the applied constraints rotation ambiguities can be more or less important for the particular problem of study.

On the other hand, several types of constraints can be applied during MCR-ALS analyses. In this Thesis, we have used non-negativity, unimodality, closure, and selectivity constraints.

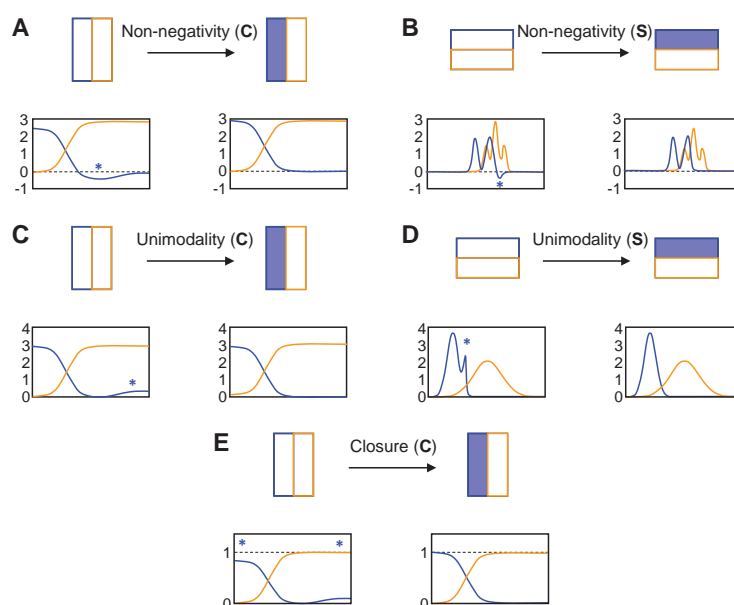


Figure 2.41. MCR-ALS constraints.

Non-negativity constraints (**Fig. 2.41A** and **Fig. 2.41B**) impose that resolved values can only be positive or zero. This constraint is very general and it can be applied on physical concentrations ($\mathbf{C} \geq 0$) and also on spectra ($\mathbf{S}^T \geq 0$) since spectral measurements are, by nature, usually (*e.g.*, ^1H NMR [232] or UV [233]).

Unimodality constraints restrict that pure components to have responses with only one maximum. This constraint has been used for chromatographic (**Fig. 2.41D**) [234], and for reaction based systems (**Fig. 2.41C**) [235], among others.

With closure constraints (**Fig. 2.41E**), the sum of the elements of each row or vector are fixed to a known constant value. This constraint results useful in closed reaction systems in which the mass balance is conserved, since it reduces the possible intensity ambiguities.

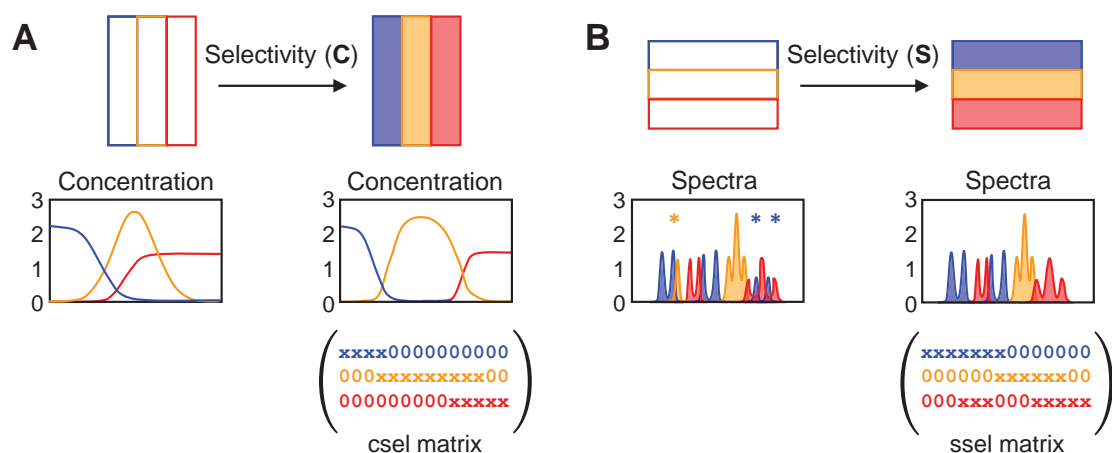


Figure 2.42. Selectivity constraints.

Selectivity and local rank constraints are very important to avoid the presence of rotation ambiguities. Then can be applied for instance when some chemical components (species) are known to exist only in certain samples or data regions (**Fig. 2.42A**), or when they give spectral signals only in a particular spectral range (**Fig. 2.42B**). In these cases, it is possible to use a constraint that defines the spectral or concentration windows where the different components contribute or not to the measured signal in the two data modes (concentration and spectra).

There are other means that contribute to the reduction of the possible rotation ambiguities. One mechanism is to perform the simultaneous analysis of complementary data [125]. For instance, the simultaneous analysis of samples containing spectra of pure samples (or just samples of reduced complexity) will reduce significantly the ambiguity problems. Also, the simultaneous analysis of the same system under different experimental conditions or different analytical methods may also result in an improvement in the resolution of the responses of the pure components in the two data modes (concentration and spectra profiles) [236].

MCR-ALS on augmented two-way datasets

MCR-ALS can be applied on matrices built from sets of samples of first-order data (analysis of one sample gives a data vector; *e.g.*, UV, ^1H NMR), and also on matrices having single samples of second-order data (analysis of one sample gives a data matrix; *e.g.*, HPLC-DAD, HPLC-MS).

For a set of K second-order data matrices, the bilinear model simultaneously applied to all of them can be described as following:

$$\begin{bmatrix} \mathbf{D}_1 \\ \vdots \\ \mathbf{D}_K \end{bmatrix} = \begin{bmatrix} \mathbf{C}_1 \\ \vdots \\ \mathbf{C}_K \end{bmatrix} \mathbf{S}^T + \begin{bmatrix} \mathbf{E}_1 \\ \vdots \\ \mathbf{E}_K \end{bmatrix} \quad \text{eq. 2.33}$$

Or summarized to:

$$\mathbf{D}_{\text{aug}} = \mathbf{C}_{\text{aug}} \mathbf{S}^T + \mathbf{E}_{\text{aug}} \quad \text{eq. 2.34}$$

In this equation, the column-wise augmented data matrix \mathbf{D}_{aug} is formed by the concatenation of the different experimental \mathbf{D}_K matrices, \mathbf{C}_{aug} matrix is formed after the concatenated \mathbf{C}_K elution profiles for all the K samples, \mathbf{S}^T matrix has the resolved mass spectra profiles of the common species in all \mathbf{D} matrices, and \mathbf{E}_{aug} matrix describes the variance not explained by the MCR-ALS model on all data sets.

The resolution of an augmented second-order data matrix (Eq. 2.33) only requires that the analyzed samples have common profiles on \mathbf{S}^T , but not in \mathbf{C} , as depicted in equation 2.34, and therefore, MCR-ALS can resolve chemical compounds that elute at different elution times for different samples and that also have different shape profiles. This results very convenient on HPLC-MS metabolomics analyses, and it makes a difference against other approaches that need a prior alignment and shape modelization of the peaks [237].

In this Thesis, this strategy has been used successfully to identify the lipid compounds characteristic of yeast samples cultured at 15°C, 30°C, 37°C and 40°C [125]. In this study, lipidomics results (obtained from UHPLC-MS measurements) were combined with metabolomics results (obtained from NMR measurements) of the same samples to achieve a better interpretation of yeast response under different acclimation temperatures. This study is discussed in detail in Chapter 3.

MCR-ALS on ^1H NMR datasets of metabolomics data

In most ^1H NMR metabolomics studies, resonances are first assigned after a careful NMR spectroscopy analysis, and then, these assigned resonances are subsequently integrated using strategies based on deconvolution [182,193,238]. Alternatively, the application of bilinear decomposition chemometric methods like MCR-ALS allow extracting directly the concentration and spectral profiles of the constituents in mixture samples of unknown composition. Nevertheless, this second procedure has been done rarely in the case of ^1H NMR studies [239-241].

This statement can be argued because there are already several user-friendly NMR platforms that allow the metabolomics characterization of the samples (see **Figure 2.2**), but they do not include approaches like MCR-ALS or Independent Component Analysis [242].

The problem of the resolution of ^1H NMR metabolomics datasets with chemometric methods is that the reliability of the resolution depends on the complexity of the data. The main difficulties are associated to rotation ambiguities, to rank deficiency problems (metabolites from the same biological pathway may be co-regulated), and to noise propagation effects (smaller resonances use to be more poorly resolved than more intense resonances).

However, in ^1H NMR data, proton resonances are narrow, sparse and broadly dispersed along the entire spectral domain, all these particularities facilitate their proper chemometric resolution. In this Thesis, these advantages are considered, and the efficiency of the MCR-ALS spectral selectivity constraint to minimize the impact of the possible rotation ambiguities and specially, of the rank deficiency limitations mentioned in the previous paragraph. In order to obtain reliable results, a new methodology to properly design spectral selectivity constraints and initial spectral estimates for ^1H NMR metabolomics datasets is proposed. This methodology, which we refer to as Decision Tree of Correlations – MCR-ALS (or DTC-MCR-ALS), can be summarized in the following steps:

1. The spectral dataset is divided into several spectral sub-regions.
2. The local rank for every sub-region is estimated.
3. MCR-ALS is applied on every sub-region using the local rank estimated as the number of components to be resolved.
4. The set of concentration vectors resolved for every MCR-ALS are combined into a column-wise augmented dataset, with n rows (samples) and m columns (concentration vectors).
5. Pair-wise correlations between all concentration vectors are calculated.
6. Resolved spectral features in step 3 are grouped if their associated concentration vectors are highly correlated.
7. Each group of spectral features is combined into a single ^1H NMR spectrum that covers all the analyzed spectral range. This ^1H NMR spectrum is representative of a compound that can be resolved from the original ^1H NMR dataset, and it will be used as a spectral initial estimate.
8. A spectral selectivity constraint complementary to the matrix of spectral initial estimates is implemented. With this constraint, it will be imposed that spectral resonances from a compound can only appear in those spectral sub-regions where a spectral feature from the same compound had been found.

9. MCR-ALS is applied to the whole spectral dataset, using as the number of components the number of groups of spectral features estimated in step 6, the spectral initial estimate obtained in step 7, and the spectral selectivity constraint implemented in step 8.

A more exhaustive explanation and implementation of this method is presented in **Chapter 4** and in [243].

MCR-ALS of metabolomics data

MCR-ALS, as to PCA, can also be applied on different metabolomics-originated data matrices to investigate the underlying common profiles that describe the considered dataset. However, the benefit of applying MCR-ALS (in combination with non-negativity constraints on \mathbf{C} and \mathbf{S}^T matrices) rather than PCA is that it produces more directly interpretable biological results.

In biological samples, where the analyzed variables are relative to biological responses (*e.g.*, ^1H NMR [180], RAMAN [244], DNA microarrays [245], metabolite concentrations [246]), the biological interdependencies that exist in the biological organism can be extracted as resolved components when they are analyzed by chemometric methods. These biological interdependencies are usually related to metabolic responses from the same pathway that are triggered by the same common stimulus.

When a metabolomics data matrix is decomposed by MCR-ALS, the two resolved factor matrices represent the relative contribution of every resolved component (sample contribution profile) for every analyzed sample (\mathbf{C}), and the set of metabolic profiles (\mathbf{S}^T). In \mathbf{S}^T , every metabolic profile is defined by the relative intensity of every measured biological signature. For example, for a metabolites concentration data matrix, MCR-ALS will resolve in \mathbf{S}^T the sets of relative metabolites concentrations that explain the underlying metabolic alterations that occurred in the studied samples.

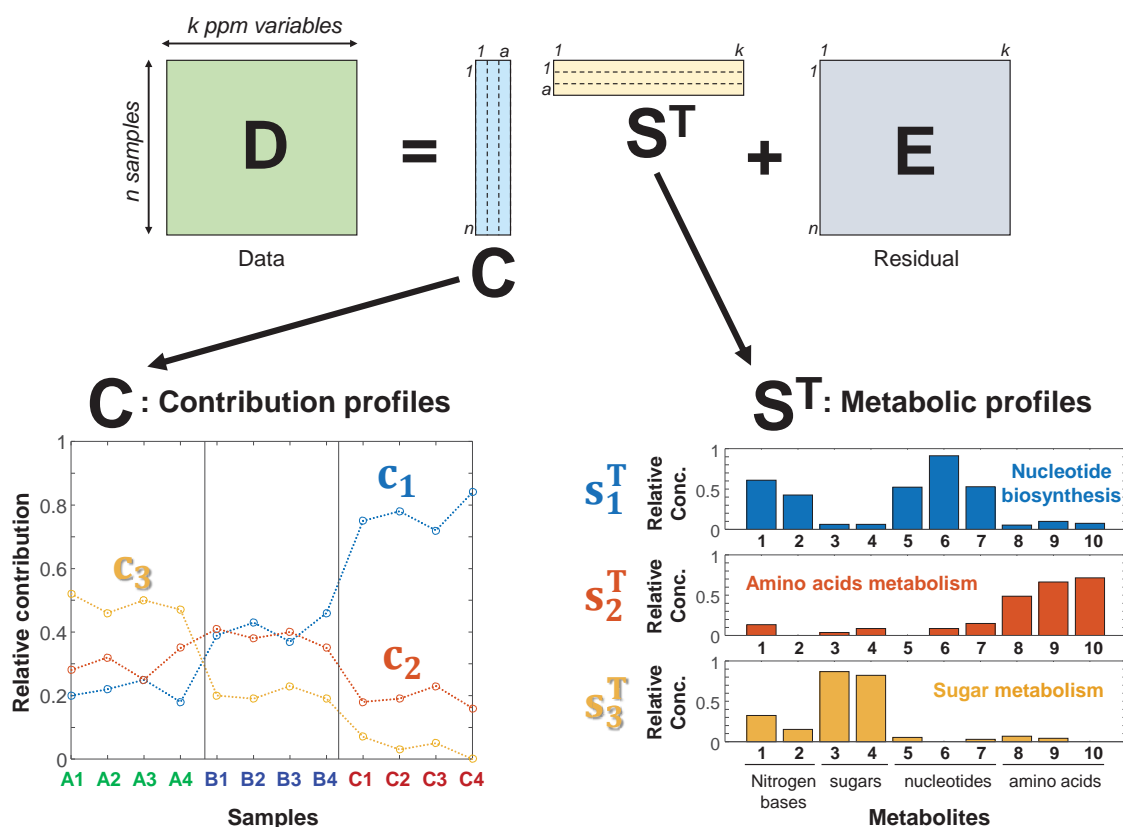


Figure 2.43. Bilinear decomposition on a concentration dataset from a metabolomics experiment. As a result of the analysis, the sample contribution (**C**) and the metabolite profiles of the different type of metabolic responses (**S^T**) are resolved.

MCR-ALS, with the aim of detecting metabolite interdependencies, has been applied in biology, food, and medicine, among others, in the last years. In fact, for example, the data analysis of imaging data (*e.g.*, hyperspectral, RAMAN, MS imaging) with MCR-ALS has become a state-of-the-art protocol, and the metabolite interdependencies detected usually describe the different biological compartments enclosed in the studied image [244,247-250].

Montoliu *et al.* in 2009 [251] also applied MCR-ALS to investigate the metabolic relationship between different biological compartments (liver, pancreas, kidney cortex, plasma, and adrenal gland samples), but using ¹H NMR data.

The biological information that can be retrieved from the MCR-ALS analysis depends on the implemented experimental design. For instance, it has been used to evaluate the ¹H NMR spectral profiles associated to nephrotoxicity from urine samples collected from rats during 9 days, and the metabolic changes in Japanese rice fish at 9 different embryogenic stages [180].

In order to obtain a better biological interpretation, MCR-ALS can be also applied to the table with already curated metabolite concentrations. With this approach, overlapping

among the directly measured variables is avoided (e.g., resonance overlapping in ^1H NMR spectra), reducing, therefore, the existence of the possible rotational ambiguities, and all metabolites are then equally represented in the dataset since every metabolite is defined by only one variable. This approach has been used to evaluate the metabolic circadian variations in rice plants [246], and in this Thesis, to evaluate the metabolic response of yeast cultured under different nutrient-limiting conditions [118] and at different temperatures [125].

Finally, metabolic signatures are encoded in the \mathbf{S}^T matrix which can be further analyzed with other chemometric methods, such as PLS-DA [247] or clustering approaches [125].

5 ENVIRONMENTAL METABOLOMICS

The potential for metabolomics to detect biomarkers of environmental stress and to delineate the modes of action of xenobiotics in other fields of science has contributed to the application of metabolomics techniques in the environmental sciences [252]. This sub-discipline of metabolomics is referred to as *environmental metabolomics* and consists in the application of metabolomics techniques to analyze the interactions of organisms with their environment [29].

A broad range of organisms are used to evaluate environmental stressors, such as microorganisms, plants, and terrestrial or aquatic organisms. In this Thesis, we have chosen the yeast *Saccharomyces cerevisiae* as the model organism for our metabolomics analyses.

5.1 *Saccharomyces cerevisiae* (YEAST)

Yeasts are unicellular fungi that have been widely used throughout human history in the fermentation industry to produce alcoholic beverages, and in the baking industry to expand dough [253]. In the research field, despite its apparent evolutionary simplicity, it has become a well-established model organism for molecular genetic research because the basic cellular mechanics are generally conserved between yeast and larger eukaryotes, including humans. Furthermore, yeast can be easily manipulated at the genetic level [254], converting it to an excellent host for conveniently analyzing and functionally dissecting gene products from other eukaryotes. The importance of yeast in research led to become the first eukaryote organism with its DNA completely sequenced (the strain S288C in 1996) [255].

Many current genetic studies are carried out with the haploid strain S288C or one of their derivative strains, although other strains with different genetic pedigrees are equally used. These strains have different properties that can influence experimental outcomes. For instance, the S288C strain contains a defective *HAP1* gene, making it incompatible with studies of mitochondrial and related systems [256].

In this Thesis, the yeast strains S288C and the S288C-derived BY4741 have been used. The BY4741 strain differs from the S288C strain because it lacks 4 genes involved in amino acid and nucleotide biosynthesis (*URA3*, *MET15*, *HIS3* and *LEU2* genes). Due to the absence of these four genes, the BY4741 strain is unable to biosynthesize (auxotroph) uracil, L-methionine, L-histidine, and L-leucine, respectively.



Figure 2.44. Yeast colonies on a YPD agar plate.

Yeast can be grown in a solid culture medium (as in a Petri Plate) or in a liquid culture medium[257]. Different culture media with different composition are used depending on the distinct desired growth conditions. Two of the most used culture media are YPD (Yeast Peptone Dextrose) and YNB (Yeast Nitrogen Base) media.

YPD is a nutritious medium preferred for the growth and propagation of yeast cultures, which contains bacteriological peptone, yeast extract, and glucose.

On the other hand, YNB is a medium used for the cultivation of yeast. YNB contains ammonium sulfate (as the nitrogen and sulfate sources), phosphate, vitamins and trace elements. If this medium is only complemented with a carbon source, such as glucose, it can be used as a selective media for culturing amino acid-auxotrophic yeast strains. On the other hand, if YNB medium is supplemented with amino acids, we can refer to this culture medium as Yeast Synthetic Complete (YSC) medium. Finally, in case not all the amino acids are added into the YNB medium, this medium can be referred as a Yeast Synthetic Drop-out Medium or just Drop-out Medium (DM).

Yeast growth can be described as a function of cell number increase and nutrient availability, and it consists of five different phases: lag phase, exponential phase, diauxic shift, post-diauxic shift and stationary phase. After inoculation, yeast is in the lag phase, adapting to the growth conditions of the fresh media. During exponential phase, yeast cells growth primarily by fermentation metabolism at full growth rate, as there is no nutrient constraint. The diauxic shift occurs when glucose becomes exhausted from the medium and cells adapt to respiratory metabolism. During the post-diauxic phase, growth resumes at a much lower rate, utilizing energy provided by respiration. Finally, the stationary phase is a result of carbon starvation, and there is no further net increase in cell number [258]. Since yeast growth depends directly on nutrient availability, different metabolic profiles are observed at different yeast growth phases and at different yeast growth culture media.

5.1.1 Environmental metabolomics studies in yeast

Yeast is not only a suitable model organism for genetic experiments, but also for addressing questions related to the metabolism. In fact, dozens of scientific articles investigating yeast metabolome have been published in the recent years [30,88,118,119,125,159,243,259-276].

In a more biological framework, yeast metabolome has been investigated to understand cell cycle regulation [262,263], aging [264], and gene regulation [265].

In the framework of environmental metabolomics, the metabolism of yeast has been deeply studied for cells cultured in sub-optimal or even adverse conditions, such as for dehydration [266], nutrient limitation [267-269], ethanol tolerance [88,270], pH [88,273], heat stress [88,272], salt stress [88,273,274], oxidative stress [88,276] and metal stress [275].

Between these two mentioned frameworks, there exists an extra in-between framework that has been overlooked and understudied in the scientific literature. This third framework consists of the study of yeast metabolism under controlled *standard* lab conditions.

Microbiology is a rather old research field [277], and the established culturing methods employed nowadays were optimized based on phenotypic responses (*e.g.*, cell density, growth rate) [257] since methods to study cellular metabolism did not exist at that moment [278].

Therefore, although yeast has been extensively investigated from the genomic point of view, it is sometimes ignored the fact that every genetic response is a dynamic event that can be altered due to the used growing conditions.

In addition, apart from the culturing conditions (*e.g.*, medium composition, temperature), the genetic background, specific of each yeast strain, may determine the metabolic responses observed.

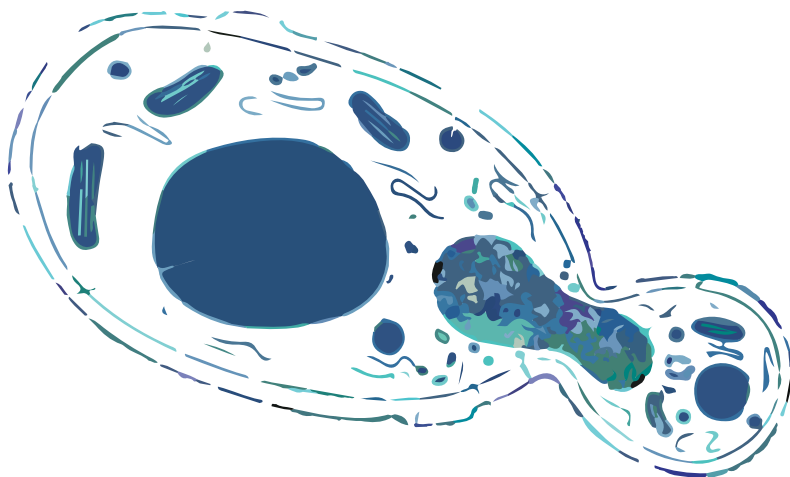
Most of the gene editing protocols in microbiology require the use of auxotroph strains, leading to the fact that most published yeast genetics experiments are performed using strains with at least one biosynthetic pathway disrupted. However, in 2012, Mülleder *et al.* affirmed that it is preferable to work with prototroph strains instead since the encoding gene deletions of the auxotroph strains can influence several physiological parameters and produce a bias in physiological and metabolic studies [279].

Therefore, despite all the efforts put into yeast metabolomics experiments that evaluate the effect of environmental conditions, basic research to improve the knowledge related to yeast metabolism and to yeast gene regulation when it is cultured under *standard* laboratory conditions is still necessary. For this reason, in one of the metabolomics studies covered in

this Thesis (**Chapter 4**), we have evaluated the impact on yeast metabolism of two different yeast growth media, YSC and YPD, over time.

Chapter 3

Current data analysis strategies
for the investigation of ^1H NMR
metabolomics datasets.
The *Saccharomyces cerevisiae*
case-study



Yeast cells are a versatile tool for investigating the metabolic effects of environmental stressors. In order to evaluate these effects, yeast cells are very convenient because they are easy to manipulate and to maintain, and because the metabolic pathways related to growth and metabolism found in yeast are also found in larger eukaryotes, such as in humans.

In this Chapter, we have studied the yeast metabolic response to two main stresses: temperature acclimation and nutrient starvation. In the first part of the introduction section, background information regarding these two stresses is presented. In the second part of the introduction section, the analytical tools used to investigate these two stresses are presented. In the scientific research section, the effects of the mild-heat acclimation stress on the metabolome of yeast cells is studied (Scientific article I). Moreover, the changes in the yeast metabolome and lipidome as a result of their cultivation at four different temperatures are investigated (Scientific article II). In addition, the metabolic response of yeast at four different nutrient starvation conditions are studied (Scientific article III). In the last discussion section, the results obtained in this research section are presented and commented. Finally, the specific conclusions drawn from these studies are listed.

1 INTRODUCTION

1.1 YEAST STRESS (TEMPERATURE AND STARVATION)

1.1.1 Temperature stress in yeast

Microorganisms, including yeast, have colonized all kind of environments, from Mediterranean countries [280] to hot springs [281], or even in the Arctic [282]. To confront these diverse conditions, microorganisms have adapted their metabolism and physiology to survive.

For instance, to cope with low temperatures, psychophilic¹ yeasts present a cell membrane rich in short and unsaturated fatty acids, resulting in an improvement of the membrane fluidity, required to maintain the appropriate physical state of the lipid bilayer and the good functionality of membranes at such low temperatures [283].

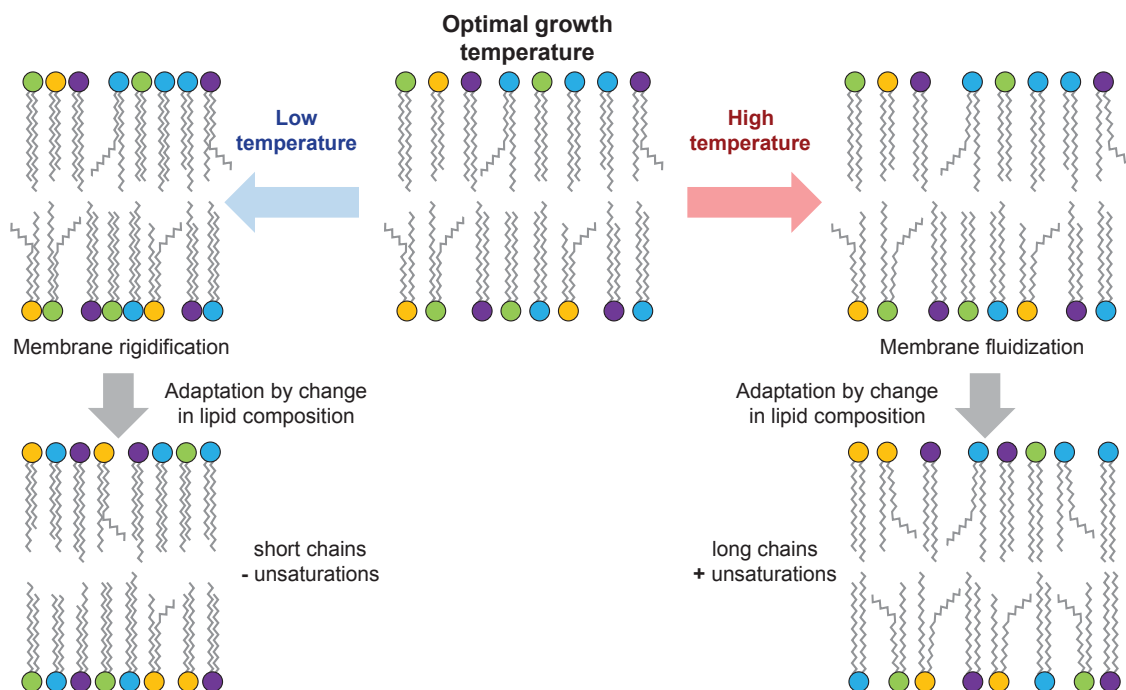


Figure 3.1. Scheme of a plasma membrane and its adaptation to changes in temperature. Colors represent different lipid families.

In addition, other metabolic adaptations observed in psychophilic yeasts are the increase of polyols (*e.g.*, glycerol, mannitol) that sustain the function and integrity of cell membranes to dehydration and osmotic level derived from the increased membrane permeability [284]; and

¹ Psychophile: organism capable of growth and reproduction in low temperatures.

the increase of trehalose, a metabolite that enhances fungi resistance to different environmental stresses [285].

Moreover, growing on a low temperature-environment produces an important impact at the transcriptional and translational levels, the so-called Cold Shock Response (CSR) [286]. Transcription levels of genes involved in the regulation of transporters of growth-limiting nutrients, glycogen metabolism and ribosome biogenesis are altered [287], and translation of antifreeze proteins is substantially enhanced in psychrophilic yeast [288].

On the other hand, yeast metabolome and transcriptome are also affected by exposition at high temperatures. Specifically, heat dynamically induces a protective transcriptional response known as the heat shock response (HSR), that alters yeast physiology, membrane composition, and overall metabolism. Mediated by HSR, genes related to energy reserves are over-expressed [289], as well as genes related to glucose transporters, gluconeogenesis and to ethanol fermentation [272,289]. As in the cold temperature stress, the trehalose production is activated [290].

Aberrant protein folding is more likely to occur at higher temperatures because they present an elevated conformational freedom. To minimize protein misfolding, HSR induces transcription of protein folding genes [291], and heat shock proteins (HSPs) represses translation of non-heat shock transcripts by modifying mRNA transport into subcellular complexes away from ribosomes [292] and by chromatin modulation [293].

Moreover, to maintain cell membrane stability at higher temperatures, cell membrane fluidity is reduced by changing its lipid composition [283]. These alterations in the cell wall activate transduction pathways such as the Cell Wall Integrity signaling pathway, a MAP² kinase pathway [294] that regulates cell wall biosynthetic enzymes and the polarization of the actin cytoskeleton [295].

Finally, under aerobic conditions, exposition to high temperatures results also on an increase of the potential oxidative stress, and the antioxidant defenses, such as catalases, peroxidases, and thioredoxins are over-expressed [293]. At the metabolome level, an increase of glutathione consumption is observed [296].

1.1.2 Starvation stress in yeast

In response to perturbations in the availability of nutrients in the environment, yeast cells modulate their gene expression levels after activation of a transcriptional reprogramming machinery specific for the exposed stress [297]. The result of this transcriptional

² MAP kinase: Mitogen-Activated Protein kinase.

reprogramming is a change in the metabolism and on the yeast physiology that minimizes the impact of this environmental stress, allowing yeast to grow and survive. However, depending on the magnitude of the stress, yeast cells might enter into a latent cell state and remain arrested until the environmental conditions become favorable [298]. Thus, nutrient limitation might modify the dynamics of yeast cell cycle.

Yeast cell cycle [299] consists of a series of events that lead to the DNA duplication and the formation of two daughter cells. This cell cycle can be divided into different phases: the first growth phase (G_1), the synthesis phase (S), the second growth phase (G_2) and the mitotic phase (M).

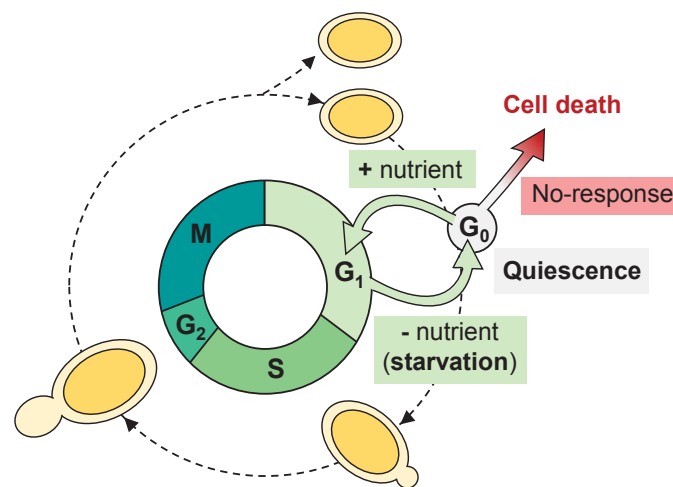


Figure 3.2. The cell cycle in yeast haploid cells and starvation.

In the G_1 phase, cells increase the number of proteins, organelles, and also their size. In the S phase, the chromosomes are replicated. In the G_2 phase, cells prepare for mitosis and grow rapidly. Then, in the M phase, the nucleus is divided by mitosis. Finally, starting during late mitosis, the cytokinesis occurs, in which the two nuclei, cytoplasm, organelles and cell membrane are divided into two functional daughter cells. These two cells will be in the G_1 phase, and each one will start a new cell cycle.

If a cell perturbation exists, such as from an environmental stress, cells will try to adapt to these adverse environmental conditions. In case this stress is prolonged over time and cells are reaching a biological point of no return due to the exhaustion of vital resources, cells found in G_1 can enter into the latent state of the G_0 phase or quiescent state [300].

In this state, cells can remain for long periods of time without proliferating. G_0 yeast cells present enhanced resistance to heat and high osmolarity, increased levels of carbohydrate storage, and a thickened cell wall [301,302]. In addition, they show a reduced metabolic activity and protein synthesis and growth is halted. Instead, resources are used for cell maintenance, survival, and repair pathways [303]. When the extracellular perturbation

disappears, cells activate their metabolism and can enter again into the cell division cycle in the G₁ phase [304].

Under starving conditions, several signaling pathways are activated. These pathways are the RAS/cAMP-dependent Protein Kinase A (PKA), AMP-dependent protein kinase (Snf2), Sch9, and Target Of Rapamycin Complex 1 (TORC1) signaling pathway [305]. The outcome resulting from the activation of these pathways is a metabolic response that alleviates the cellular stress. For instance, under glucose deprivation, genes involved in fatty acid beta-oxidation are expressed, as well as gluconeogenesis genes, and inhibits the expression of hexokinase and hexose transporters [306]. Thus, altogether, glucose catabolism is repressed, fatty acids are used instead to produce energy, and glucose anabolism is activated to improve the carbohydrate reserves.

Similarly, starvation for amino acids, purines, and glucose limitation induces the synthesis of Gcn4 protein (Gcn4p), a transcriptional activator of amino acid biosynthetic genes in multiple pathways. In amino acid-starved cells, Gcn4p represses ribosomal proteins, inhibiting protein synthesis, and activates the transcription of amino acid biosynthetic genes [307].

Another characteristic metabolic response derived from starvation is autophagy [308]. In autophagy, cytoplasmic components are delivered into vacuoles for degradation to generate an internal pool of molecules ready to be recycled. In yeast, autophagy can be activated through PKA and TORC1 signaling pathways under nitrogen starvation, carbon starvation, auxotrophic amino acids starvation and nucleic acids starvation [309].

Due to a limitation of the internal resources, autophagy, and therefore quiescence, cannot be maintained forever. Requirements for surviving starvation varies on the starvation conditions, and so the maximum lifespan at G₀ state. In general, starvation for natural nutrients such as carbon, phosphate, nitrogen or sulfate results in low death rates (half-life of > 10 days), whereas starvation for amino acids in auxotrophic mutants presents a rapid loss of viability (half-life of < 4 days) [267,269].

Unlike wild-type strains, auxotrophic mutants have not been subjected to evolutionary selection and the compensatory mechanisms activated under starvation conditions are not correctly regulated. That results in an incomplete cell cycle arrest and higher rates of glucose consumptions. On the contrary, sulfate and phosphate starvations produce a rapid and uniform cell cycle arrest and a slow glucose consumption [267].

Not all amino acid starvations show the same transcriptional and metabolic response, likely because some amino acids have additional roles in the cell apart from being used as building blocks for proteins. For instance, L-leucine, L-arginine, and L-glutamine interact with

TORC1 signaling pathway[310], while L-methionine regulates both autophagy[311] and growth [312]. Related to the latter, L-methionine starvation is the only amino acid starvation condition whose cells can enter into a quiescence state similarly as under a natural starvation condition, and yeast lifespan is also similar. It has been observed that L-methionine biosynthetic genes exhibit periodic expression, which suggests that these differences are caused because L-methionine plays an important role in the control of cell cycle regulation [313].

1.2 ANALYTICAL STRATEGIES USED

The most reported response to temperature acclimation in yeast is the variation of the lipid composition [314-322], whereas nutrient starvation in yeast has been commonly examined with transcriptomics approaches [269,304,307,323-328]. However, with only lipidomics and transcriptomics, respectively, it is not possible to completely understand yeast adaptation to these two stresses. For this reason, we have performed metabolomics experiments to try to fill the existing knowledge gap in this area.

Metabolomics analyses regarding temperature [88,272,329,330] and starvation [269,304,331] stresses, albeit less common, exist. In these studies, it is observed that these two biological processes affect drastically on the whole cell, and the two produce an impact directly on the primary metabolism.

For these metabolomics studies, researchers have used Mass Spectrometry (MS). However, for the analysis of the primary metabolism, Nuclear Magnetic Resonance (NMR) spectroscopy is a strong competitor to MS, since NMR is more robust, it has better reproducibility, metabolite identification results less ambiguous, and because (^1H) NMR is inherently quantitative. Because of all this, we have decided in this Thesis to use mainly the NMR methodology to characterize the yeast metabolome instead.

In this chapter, we present several NMR-derived strategies to assign resonances from raw ^1H NMR spectra to a set of meaningful metabolites. These approaches consisted of:

- 1) Performing spiking experiments with candidate compounds (Scientific Article I)
- 2) Analysis of the ^1H NMR dataset with Statistical TOCSY or STOCSY [332] (Scientific Article I).
- 3) Analysis of ^1H - ^1H COSY, ^1H - ^1H TOCSY, ^1H - ^{13}C HSQC and ^1H - ^{13}C HMBC NMR spectra from representative samples (Scientific Article I & III).
- 4) Acquiring selective experiments, such as the selective 1D TOCSY NMR experiment [333] (Scientific Article III).

However, to obtain relevant results, not only the studied system and the analytical methodology used are important, but also the data analyses strategies used.

Most metabolomics analyses covering the two topics are based on Principal Component Analysis [88,329,330], Partial Least Squares – Discriminant Analysis [272] or on a heat-map representation of the hierarchically clustered auto-scaled areas [269,304,330,331]. Despite these data analyses strategies result very appropriate if combined to obtain an overview of the metabolomics data, they are not powerful enough if used separately to reveal all the information hidden beneath the original data.

Moreover, in this Chapter, the chemometrics method Partial Least Squares – Discriminant Analysis (PLS-DA) has been used to identify metabolite biomarkers of temperature acclimation (Scientific Article I); and the Analysis of variance – Simultaneous Component Analysis (ASCA) was employed to assess whether the yeast metabolic response to nutrient starvation over time is statistically different to the one observed during growth at normal conditions (Scientific article III).

Finally, in this Chapter, we have proposed the use of the chemometric method Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) to resolve the metabolic profiles and their relative contributions that explain, at a biological level, the observed variations in the yeast metabolome under stress conditions (Scientific article II & III). In order to extract those metabolic profiles more descriptive of the temperature acclimation effect, the analyzed metabolomic dataset consisted of two row-wise fused datasets: one representative of the primary metabolism, generated by NMR analysis; and another one representative of the lipidome, generated by UHPLC-MS analysis from the same yeast samples (Scientific article II).

2 SCIENTIFIC RESEARCH

2.1 SCIENTIFIC ARTICLE I

A quantitative ^1H NMR approach for evaluating the metabolic response of *Saccharomyces cerevisiae* to mild heat stress.

Authors: Puig-Castellví F., Alfonso I., Piña B., Tauler R.

Citation reference: *Metabolomics* (2015), 11:1612–1625.

DOI: 10.1007/s11306-015-0812-9



A quantitative ^1H NMR approach for evaluating the metabolic response of *Saccharomyces cerevisiae* to mild heat stress

Francesc Puig-Castellví¹ · Ignacio Alfonso² · Benjamí Piña¹ · Romà Tauler¹Received: 23 January 2015 / Accepted: 20 May 2015
© Springer Science+Business Media New York 2015

Abstract Effect of growth temperature on the yeast (*Saccharomyces cerevisiae*) metabolome has been analysed by one-dimensional proton NMR spectroscopy (^1H NMR). Potential biomarkers have been first identified by a non-targeted chemometric evaluation of the spectra, followed by a comprehensive analysis of bayesian estimated concentrations of target metabolites in extracts of cells growth either at 30 or 37 °C. Tentative identification of metabolites whose concentrations were affected by this mild heat-shock stress was attempted by partial least squares-discriminant analysis (PLS-DA) on ^1H NMR data, combined with Statistical Total Correlation Spectroscopy, and further confirmed with empirical data. An extensive assignment for most of the detected NMR signals was performed, with a total number of 38 identified metabolites. Concentrations estimated using automatic BATMAN modelling revealed that bayesian integration is a sufficient approach for obtaining relevant concentration changes of metabolites and biological information of interest. In contrast to when it is applied directly on spectral data, the application of PLS-DA on BATMAN recovered metabolite concentration estimates allowed for a better overview of the investigated

samples, since more metabolites were highlighted in the discriminatory model. Observed changes in metabolite concentrations were consistent with the expected process of temperature acclimation, showing alterations in amino acid cellular pools, nucleotide metabolism and lipid composition. The strategy described in this work can thus be proposed as a powerful and easy tool to investigate complex biological processes, from biomarker screening and discovery to the study of metabolite network changes in biological processes.

Keywords ^1H NMR · BATMAN · Yeast · Temperature · OSC-PLS-DA

1 Introduction

Metabolomics is a field of ‘omics’ research that is primarily focused on the identification and characterisation of small molecule metabolites in cells, tissues, organs and organisms (German et al. 2005). Commonly, nuclear magnetic resonance (NMR) (Griffin 2003) and mass spectrometry (MS) (Dettmer et al. 2007) are used for metabolomics studies. NMR studies mostly include ^1H NMR (Mazzei et al. 2013) and bidimensional heteronuclear proton-carbon (HSQC) (Kang et al. 2012) whereas MS studies include direct infusion (Højer-Pedersen et al. 2008) and hyphenated to chromatography techniques (Farrés et al. 2015). Less frequently, LC- ^1H NMR studies are also reported, mostly in the phytochemistry area (Wolfender et al. 2005).

The choice between NMR and MS lies in the evaluation of the pros and cons for both techniques. First, MS sensitivity is higher than that of NMR, which usually allows identification of only 30–50 metabolites. Second, molecule

Electronic supplementary material The online version of this article (doi:10.1007/s11306-015-0812-9) contains supplementary material, which is available to authorized users.

✉ Romà Tauler
Roma.Tauler@idaea.csic.es

¹ Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, 08034 Barcelona, Catalonia, Spain

² Department of Biological Chemistry and Molecular Modelling, IQAC-CSIC, Jordi Girona 18-26, 08034 Barcelona, Catalonia, Spain

assignment capacity in MS depends directly on the instrument accuracy and, without MS¹⁰¹, compounds can only be tentatively assigned, as molecular isomers would have identical molecular masses. For hyphenated chromatographic methods with MS, metabolite assignment can be performed by their metabolite retention time and *m/z* values. However, the power of resolving peaks depends strictly on the capacity of the chromatographic method, usually the limiting step. This problem can be circumvented by the use of chemometric approaches like multivariate curve resolution (MCR) (Farrés et al. 2015). Traditionally, metabolite identification in NMR spectra depended heavily on the NMR-spectrometrist's previous knowledge and in the ability of identifying the spectral patterns of each metabolite. Several NMR databases (Jewison et al. 2012; Wishart et al. 2013) with ¹H and ¹³C spectra of biological compounds have been recently implemented, significantly helping in speeding-up the assignment process. Resolution of overlapping signals in NMR spectrum is not always as crucial as in (MS) chromatographic methods. If an NMR spectrum region cannot be resolved, it is possible that the resonance signals comprised in that region come from metabolites with equivalent signals in other regions of the spectrum, making the related metabolites equally estimable. And third, signals in NMR spectra have inherently absolute quantitative information, whereas MS requires a calibration curve for each compound, due to the variable ionisation rates of the different metabolites.

Saccharomyces cerevisiae is a key organism in both traditional (wine, beer, bread) and technological (bioethanol) fermentation processes. Growth temperature is a key parameter determining the yeast metabolism. For most strains, growth is optimal between 25 and 30 °C, whereas ethanol production is favoured by somewhat higher temperatures (Barnett et al. 2000; Mensonides et al. 2013). However, high temperatures affect negatively other aspects of yeast metabolism, including changes in membrane lipid composition to adapt its fluidity to the growth temperature (Arthur and Watson 1976). The compromise between ethanol production and yeast survival is solved in industrial procedures by using thermotolerant yeast strains (Nonklang et al. 2008).

In a previous study combining transcriptomics and metabolomics studies (Strassburg et al. 2010), upregulated genes during heat stress in yeast were mostly related to primary metabolism processes. Knowing in advance that the majority of compounds detectable by NMR belong to primary metabolites, and due to the fact that any metabolic approach is tightly dependent on the metabolomics strategy employed (Dunn et al. 2005), in this study we propose to use an NMR-based metabolomics approach to study the yeast thermal response to a mild heat stress.

There are two possible approaches to analyse NMR spectra and both are presented in this work. The first strategy is the non-targeted analysis of the whole set of NMR spectra through chemometric methods, like principal component analysis (PCA) (Bro and Smilde 2014) and partial least squares discriminant analysis (PLS-DA) (Wold et al. 2001), with the aim of identifying discriminatory peaks that can be related to particular metabolites, and if required, their corresponding peak areas from the raw spectrum can be compared to estimate the relative changes in their concentrations. The second approach is the assignment of the NMR spectrum peaks to a set of target metabolites (prior to know whether they are relevant or not in the studied effect) and the integration of their corresponding signals, followed by the application of chemometric methods to the resulting peak-assigned metabolite areas data matrix.

Whilst the direct non-target approach is faster and brings the information to discriminate among tested classes, the traditional targeted approach allows a better interpretation of the tested samples. However, metabolites assignment and confirmation is time-consuming and estimation of their area can be sometimes complicated due to overlapping signals. Also, manual peak integration might produce misleading results caused by human factor. To overcome with these possible problems of both type of approaches, in this work we tested the Batman approach (Hao et al. 2012), currently available in open source in R (R Core Team 2013), which uses a Bayesian approach to estimate the concentration of metabolites from previously assigned NMR peaks. This approach has already been confirmed to provide better and more accurate results than traditional manual integration approaches (Astle et al. 2012).

2 Materials and methods

2.1 Experimental

2.1.1 Yeast growth

Saccharomyces cerevisiae BY4741 (MATa; *his3Δ1*; *leu2Δ0*; *met15Δ0*; *ura3Δ0*) cells were pre-cultured in YPD medium on an orbital shaker (150 rpm) at 30 °C overnight. A 500 ml of YPD medium was inoculated with the pre-culture to an optical density at 600 nm (OD₆₀₀) of 0.1 and divided into eight aliquots of 50 ml. Aliquots were grown at either 30 or 37 °C (four aliquots each, shaking at 150 rpm). After 6 h of cultivation, at 1.4–1.6 of OD₆₀₀, cultures were arrested on ice and 4 fractions of 10 ml were collected from each culture for metabolite extraction. Growth was similar for cultures grown at both temperatures (not shown). Cell harvesting was performed by

centrifugation of each 10-ml fraction at $4000\times g$ for 5 min and discarding the supernatant. Cells were washed afterwards with 1 ml of 100 mM sodium phosphate buffer (pH 7.0). The resulting pellets were stored at -80°C and lyophilised. For both studied temperatures, 15 pellets were analysed.

2.1.2 Metabolite extraction

Metabolites were extracted by using a slight modification of the chloroform–methanol extraction protocol (Palomino-Schätzlein et al. 2013). Pellets were resuspended in 1800 μl of a cold (4°C) methanol–chloroform (1:2) solution by vigorous vortexing and submerged into liquid nitrogen for 1 min, and afterwards thawed in ice for 2 min. The process was repeated by a total of five times. Next, 400 μl of water were added to create a biphasic system and homogenised by vortexing. Organic and aqueous phases were separated by centrifugation at 16,500 rpm (3 min, 4°C). The upper aqueous phase was collected, and the process was repeated once. The combined aqueous phases were afterwards lyophilised.

2.1.3 NMR sample preparation

Extracts were dissolved in 700 μl of deuterated phosphate buffer (Na_2DPO_4 100 mM, pH 7.0) in D_2O with DSS 0.2 mM as internal standard. The resulting solution is placed into the NMR tube.

2.1.4 NMR experiments

Spectra were recorded in a 400 MHz Varian spectrometer, using a spectrometer frequency of 400.14 MHz with a OneNMR Probe and a ProTune System (Agilent). Spectral size range covered from -2 to 10 ppm, consisting of 65,536 data points. The number of scans was 512 and the relaxation delay was 5 s. For bidimensional TOCSY experiments, 8 scans and 512 t_1 increments were used.

2.1.5 NMR spectra preprocessing

Spectra were preprocessed with MestreNova v.9.0 (Mestrelab Research, Spain). Spectra preprocessing consisted in an exponential apodization of 0.5 Hz, a manual phasing and a baseline correction with Bernstein polynomial of 3rd order. After adjusting the reference to DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid), regions of water (4.7–5.1 ppm), methanol (3.30–3.37 ppm) and chloroform (7.64–7.69 ppm) were removed. Data points with chemical shift higher than 9.45 ppm or lower than 0.8 ppm were also removed. The final NMR dataset consisted on a data matrix of 30 spectra (rows) having 38,213 ppm values (columns)

each one. This data matrix was stored in an ASCII file format.

2.1.6 Metabolite identification

A preliminary 1D-STOCSY (Cloarec et al. 2005) detection of correlated peaks with significant variation between the two temperatures was performed on the most relevant proton resonances indicated by VIPs score plot from the OSC-PLS-DA (see OSC-PLS-DA in Sect. 2.2) of the raw ^1H NMR spectral data matrix (step 1 in Fig. 1). This step does not only serve as a biomarker discovery strategy, but it can be also applied to do a tentative metabolite assignment since intramolecular signals are likely to be shown as highly correlated in the 1D-STOCSY. Metabolite assignment (step 2 in Fig. 1) was performed by a detailed targeted metabolite profiling analysis of the ^1H NMR and 2D-TOCSY yeast extract samples, using the ^1H NMR spectra library developed during this work and also using the Yeast Metabolome Data Base library (Jewison et al. 2012) (YMDB). Some of the assigned metabolites were further confirmed by spiking using commercial standards. Online Resource 1 presents the list of methods used for each identification.

2.1.7 Metabolite quantification

Relative metabolite quantification was performed using BATMAN R-package (step 3 in Fig. 1). BATMAN implements a procedure based on the use of a bayesian model to deconvolute ^1H NMR peaks and automatically assigns these peaks to metabolites from a target list, whose concentration estimate expressed in arbitrary units is obtained. BATMAN uses the modelled resonances of each assigned proton to reconstruct the empirical NMR spectrum, whereas the noise signal and unassigned resonances are modelled by wavelets. Proton signals that showed first-order couplings patterns were modelled using their spectrometric parameters. In other cases, multiplets were modelled as a ‘raster’ multiplet, using in-house ^1H NMR spectra, or as a set of singlets. The list of parameters used to model each proton signal is given in Online Resource 2. BATMAN allows to interpret data directly from spectroscopic parameters of weak coupling spin systems (AX, A_2X ,...). However, for strongly coupled spin systems, alternative input parameters have to be chosen. Thus, most of the signals in Online Resource 2 with non-integer proton intensities correspond to AB or to AA'XX' proton systems modelled as two AX systems. D-Glucose proton intensities are referenced as the sum of the two existing anomeric forms. All these non-integer proton intensities, including trehalose and glycerol proton intensities as well, were calculated by manual integration

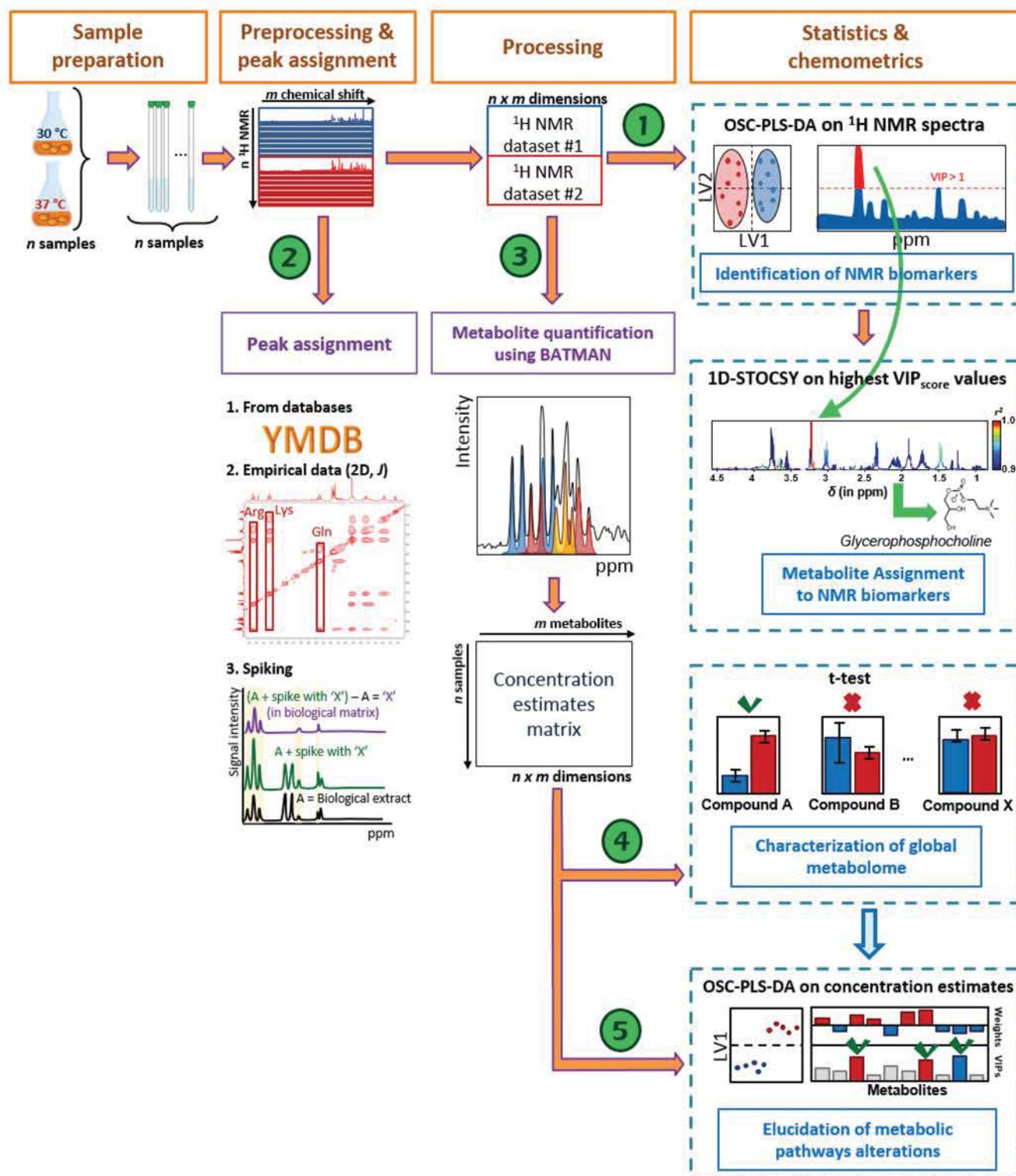


Fig. 1 Workflow of the NMR metabolomics data analysis procedure used in this work. Different steps are identified by the numbers inside the circles

with MestreNova software (Mestrelab Research, Spain). Finally, in order to let the fitting less restrictive and to obtain a better performance, Glutamine signal **23** (in

Online Resource 2) was defined as a sum of singlets, each one containing the weighted intensity of the whole multiplet. Additional parameters related to the bayesian

algorithm were also introduced. Some of these input parameters are summarised in Online Resource 3, while for the rest of input parameters their default value was used. More details of the procedure are described elsewhere (Hao et al. 2014). After introducing the input data, each NMR metabolite pattern should be correctly fitted in its corresponding ^1H NMR experimental spectrum from the yeast extract sample. By performing this step, we concluded that the modelling of every proton signal is not necessary for achieving a good estimate. For instance, the use of overlapped signals when the same metabolite has isolated signals within the NMR spectra implies that more calculations are required for obtaining the same concentration estimate.

Running time depends directly on the width of the chemical shift region studied. To shrink dataset dimension, only 1 out of each 10 values from the 38,213 spectral set points were used. In addition, instead of using the whole spectral dimension, the analysis was performed by selecting those spectral regions comprising a number of multiplets and removing regions containing only noise. In Fig. 2b, d, f and g, the individual spectral regions used are shown. Thus, although all ^1H NMR spectra were always processed simultaneously, metabolite concentrations were estimated sequentially depending on the region analysed each time.

Metabolite estimation performance from overlapped signals was evaluated carefully by observing the residual wavelet and contrasting whether non-overlapped signals from the same metabolites have been affected or not by wavelet penalisation. Therefore, additional regions containing peak signals from the same metabolite were always fitted if possible. For instance, when peaks from region of 1.15–1.35 ppm are fitted, both L-lactic acid and L-threonine signals were cross-checked by also adding into the fitting model their respective resonances from the proton at C_α and C_β , respectively, which are located above 4 ppm in the NMR spectrum. In Fig. 2a, c, e and g, a stack plot from a representative sample showing the sum of deconvoluted areas and the residual part is presented.

2.1.8 Pathway analysis and transcriptome data mining

Metabolites highlighted in the OSC-PLS-DA (see Sect. 2.2) of the concentration estimates were correlated with known yeast metabolic pathways using the KEGG database (Kanehisa et al. 2012). Transcriptomic data comparison between yeast cell growing at 29 and 36 °C was obtained from <http://genome-www5.stanford.edu/> (reference GSM1046). Metabolic maps were also obtained from the KEGG web page (<http://www.genome.jp/kegg/kegg2.html>).

2.2 Chemometric data preprocessing and analyses

2.2.1 Theory

A brief description of the different statistical and modelling methods used in this work to analyse ^1H NMR spectra is given below.

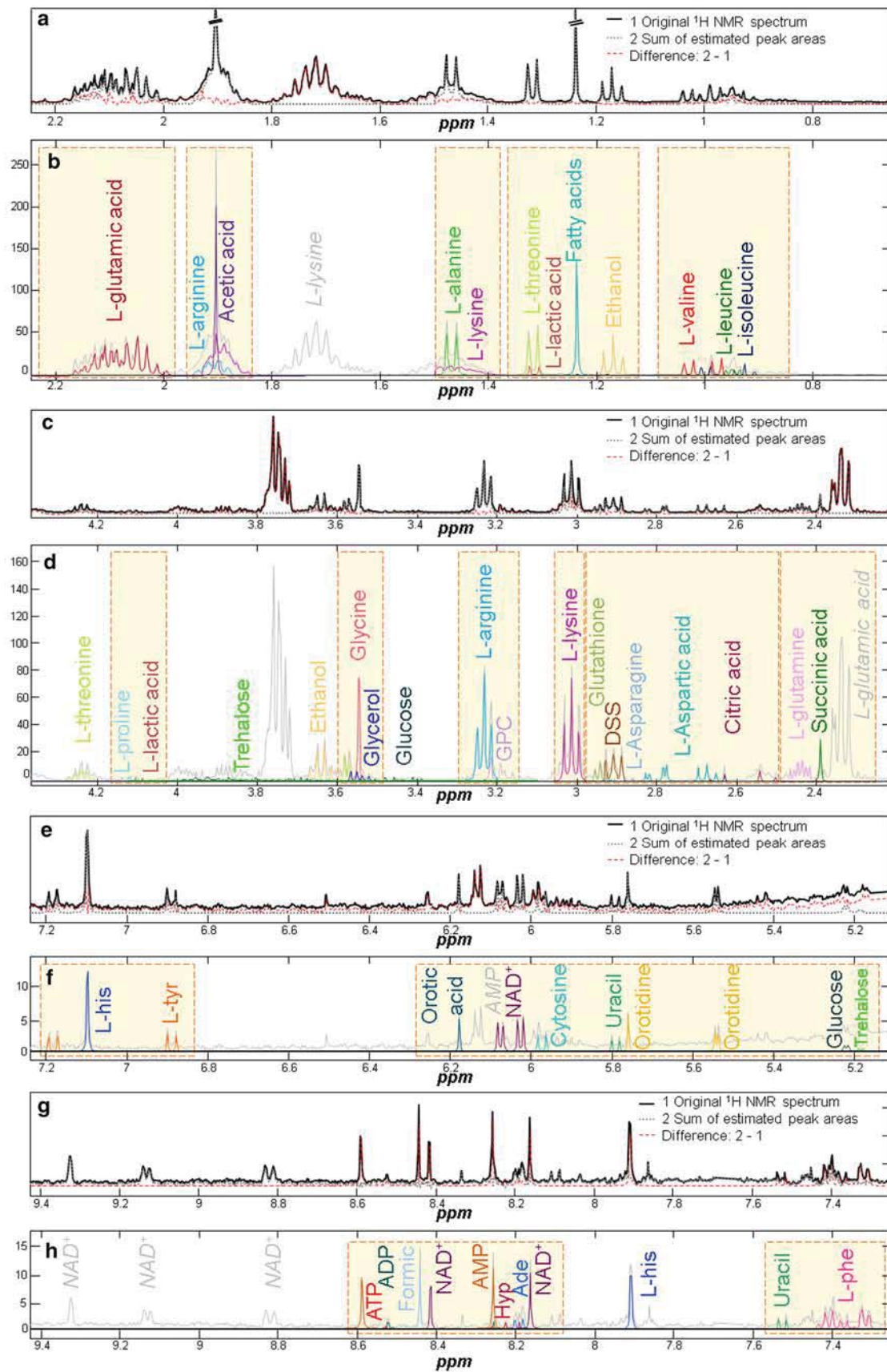
2.2.1.1 Statistical total correlation spectroscopy (STOCSY) In STOCSY (Cloarec et al. 2005), the autocorrelation matrix for a set of 1D NMR spectra of samples containing mixtures of metabolites is calculated. High correlations values might reveal which proton signals belong to the same molecule. STOCSY serves as a helpful tool for pattern recognition in complex mixtures (Li et al. 2014). In NMR spectra from biological extracts, if two or even more metabolites are strongly correlated, these correlations might be an evidence of biochemically connected metabolites. However, we have to be aware that these correlation values might be also distorted by spectral overlap and noise interferences.

2.2.1.2 OSC-PLS-DA PLS-DA (Barker and Rayens 2003) is a PLS regression method (Geladi and Kowalski 1986) variant that allows to correlate a set of discriminant response y -variables to a set of correlated X -variables (in this studied case, the temperature and the NMR spectral matrix, respectively). As a result, a reduced number of new linear combination of the independent original X -values (called latent variables, LV) is obtained. This LV optimally correlate with the variation in y -variables. For every LV, a vector of weight coefficients is acquired, showing which X -variables affect most on y .

Orthogonal signal correction (OSC) (Wold et al. 1998) is a signal pretreatment method that removes any variation within the X dataset uncorrelated (orthogonal) to y . In this work, the variable importance on projection (VIP) scores (Wold et al. 2001), which give a weighted sum of squares of PLS weights for each variable, were also calculated in order to determine the most influent variables in the model. X -variables associated with VIP scores greater than one are considered to be relevant on the definition of the PLS-DA model (Chong and Jun 2005), since the average of squared VIP score is equal to 1.

2.2.2 Multivariate analysis of the NMR spectra

The spectra from the previous preprocessing steps were imported to Matlab R2013a (The Mathworks Inc. Natick,



◀ **Fig. 2** ^1H NMR 400 MHz spectrum of the aqueous cell extracts of *S. cerevisiae* grown at 30 °C. **a, c, e** and **g** stack plots represent the fit for the modelled resonances obtained. In *black* it is shown the empirical ^1H NMR spectrum; in *dotted grey*, the sum of the modelled resonances; and in *dashed red*, the difference between the empirical and the modelled spectra. On the other hand, **b, d, f** and **h** stack plots show the individual models of each one of the used proton resonances. The *orange-dashed boxes* represent the different regions used. Coloured peak signals out of the *box* represent those signals used only as a cross-check validation of the fitting performance. Metabolite names in *grey* represent some of the identified but not used proton resonances. Abbreviated metabolites: *ADP* adenosine diphosphate, *Ade* adenine, *AMP* adenosine monophosphate, *ATP* adenosine triphosphate, *DSS* 4,4-dimethyl-4-silapentane-1-sulfonic acid, *Formic* formic acid, *GPC* glycerophosphocholine, *Hyp* hypoxanthine, *L-his* L-histidine, *L-phe* L-phenylalanine, *L-tyr* L-tyrosine (Color figure online)

MA, USA) and analysed with the PLS toolbox 7.3.1 (Eigenvector Research Inc., Wenatchee, WA, USA). Principal component analysis (PCA) was applied to the NMR spectral data matrix after standard normal variate (SNV) (Dong et al. 2011) normalisation and mean-centering pre-processing performed in this order. Although dried pellets have similar weights (6.9 ± 0.3 mg for pellets collected from samples cultured at 30 °C, and 6.7 ± 0.8 mg for 37 °C cultured samples), and according to *p* value significance level they are statistically identical (*p* value = 0.32), if spectra were normalised by dry weight, data would be biased. In order to overcome this difficulty, spectral data (with solvent and DSS regions already removed) was normalised with SNV. PLS-DA of the SNV, plus OSC-normalised spectra and mean-centered data has been also performed. In both cases, PCA and PLS-DA, cross-validation with Leave-One-Out system has been applied to check the reliability of the obtained models. Additionally, two permutation tests and a leave-one-culture-out cross-validation were executed. In the two permutation tests applied to the OSC-PLS-DA model, the evaluated parameters were the cumulative predicted residual sums of squares and the number of misclassifications. With leave-one-culture-out cross-validation test, the similarity among samples from different cultures was tested. In this test, an OSC-PLS-DA model was built using all samples except those from one specific culture as a training set, and these excluded samples were used afterward as a validation set. This process was repeated for all the cultures.

2.2.3 Analysis on relative concentration estimates

Concentration changes between classes (low and high culture temperature) were evaluated using a Mann–Whitney–Wilcoxon test and their corresponding levels of significance were calculated (step 4 in Fig. 1). In order to

contrast this information, a Kruskal–Wallis one-way analysis of variance was applied to the same dataset.

PCA and OSC-PLS-DA were also applied to these autoscaled concentration estimates (step 5 in Fig. 1). Missing concentration estimates were imputed by PCA as implemented in PLS Toolbox. The same permutation tests and leave-one-culture out cross-validation as in the NMR spectral data were performed to the OSC-PLS-DA model.

3 Results and discussions

3.1 Workflow

A scheme summarising the complete data analysis workflow of the acquired ^1H NMR data in the analysis of yeast samples at the two temperatures is shown in Fig. 1.

3.2 Preliminary study of NMR spectral data

Before performing OSC-PLS-DA, NMR spectra were analysed by PCA. PCA scores plot of the mean-centered ^1H NMR spectra from the yeast extract samples shows clearly that yeast metabolomes obtained for the two growth temperatures are distinguishable. The separation of the two sample classes was accomplished with the first component (PC1), containing 54.28 % of the explained variance (Online Resource 4a). PCA model using 3 components explained 86.58 % of the X-variance.

In PCA plot of the 2 first components, all tested samples laid inside the 95 % confidence level. Although two samples have Q residuals over the 95 % of confidence and one more is over the 95 % of confidence of Hotelling T^2 (Online Resource 4b), they were not considered outliers, as their removal did not show any significant improvement on explained variance nor contributed significantly to the residuals. Scores of the yeast culture batches were randomly distributed, implying that none of the cultures had an outlying response at the growth temperature. This lack of any distribution pattern was also observed for colour-coded scores according to their fractional sampling order (data not shown). Therefore, the fractioning step did not either constitute any source of variance nor produced any sample bias.

3.3 Initial NMR signal assignment of potential biomarkers

When OSC-PLS-DA was applied to the mean-centered ^1H NMR spectra from the yeast extract samples, one latent variable (LV) was sufficient to discriminate between the two sample classes. 53.36 % of the X-variance was already

able to explain the 99.99 % of the y-variance (Online Resource 4c). Scores were very constant within each sample class. Predictive capacity was totally achieved, with a 100 % of sensitivity and selectivity. If cross-validation is performed, correct classification is obtained for the whole sample dataset as well. Low RMSEC and RMSECV values were obtained (Online Resource 5). The permutation tests confirmed the reliability of the model with a $p < 0.001$, and the leave-one-culture-out cross-validation shown a total predictive capacity for each one of the cultures used as a validation set.

From the VIP score plot (Online Resource 4d), 83 lorentzian-shaped signals were observed whose intensity corresponds to $VIP > 1$, all located below $\delta = 5$ ppm except for one at $\delta = 8.443$ ppm. This fact explains why only half of the total X variance was needed to explain the y variance.

Since VIPs were associated to chemical shifts and not to individual NMR signals, chemical shifts associated to overlapped signals would give different VIP scores than those from isolated signals. Another appreciation from the direct examination of VIP scores plot is that while some of the VIP scores were close to 1, other reach values of some orders of magnitude higher, even greater than 1000 (Online Resource 4e), indicating significant unequal fold-changes of some resonance signals occurring during the stress situation.

A preliminary assignment of those metabolites whose concentrations were more affected by the temperature can be achieved by representing the corresponding 1D-STOCSY plots of the list of lorentzian-shaped NMR peaks selected by the OSC-PLS-DA VIPs. In STOCSY plots, spectrum intensities are differently coloured depending on their r^2 correlation with the selected chemical shift. An example of 1D-STOCSY for the peak at 3.218 ppm, which is the chemical shift where the maximum VIP value was found, is represented in Fig. 3.

As observed in Fig. 3, two ^1H NMR regions (around 3.67 and 4.31 ppm) were candidates of containing signals

from intramolecular protons connected to the one associated to the target resonance at 3.218 ppm. Using the YMDB, we tentatively assigned the peak of 3.218 ppm to be from L-glycerophosphocholine, since all three selected signals are also present in the ^1H NMR spectrum of this metabolite. Using the same strategy, ethanol, DSS, L-threonine, L-lysine, glycerol and also glycine were tentatively assigned. Also, peaks from L-leucine, L-valine and L-isoleucine are found to be highly correlated, which agrees with the fact that they are biochemically linked.

3.4 Metabolite identification in ^1H NMR spectra

Metabolite assignment is not straightforward. Although STOCSY is a useful tool for identifying compounds in NMR spectra, total assignment of the resonances cannot be achieved because correlation coefficient values are negatively affected by spectral overlapping signals and noise interferences. From the interpretation of bidimensional NMR spectra, additional assignment of correlated signals can be obtained, but the total assignment is hardly obtained due to the overlapping problem. NMR databases represent a convenient tool to assign metabolite names to resonances, but slight chemical shifts from the ones reported in library databases are expected. Even if pH is controlled, other parameters such as differences in ionic strength may produce this variation. These peaks can be confirmed from spectroscopic data, although the absolute confirmation experiment is by metabolite spiking. However, in practice, this is not always possible. Therefore, a good knowledge on the interpretation of NMR spectroscopic data is required to do a proficient metabolite assignment.

In Fig. 2b, d, f and h, the metabolite identification in a ^1H NMR 400 MHz experiment of a yeast cell extract is summarised. A wide range of biomolecules were identified, comprising α -amino acids, sugars, nucleotides, organic acids and osmolytes. Online Resource 1 presents the list of methods used for each peak assignment. The number of

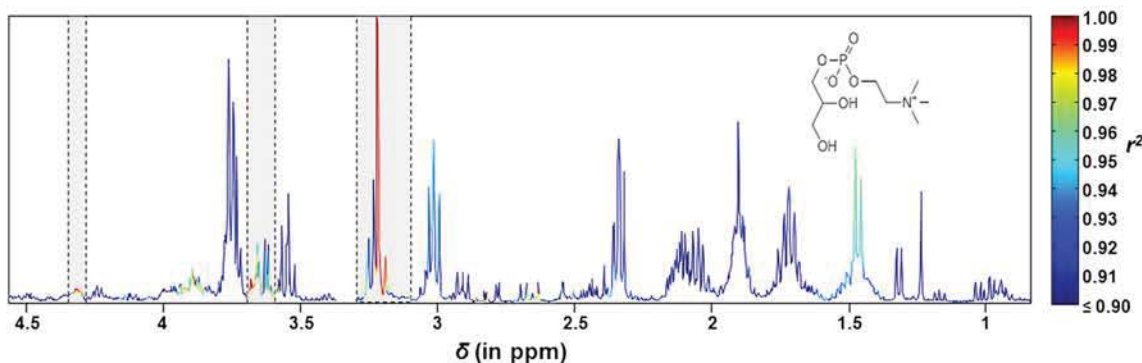


Fig. 3 1D-STOCSY plot at 3.218 ppm. Regions with segments with r^2 correlations above 0.99 are encapsulated in the grey regions (Color figure online)

assigned metabolites in this study is on the average range of most NMR metabolomics studies, despite the fact that they use NMR spectrometers of higher magnetic fields, with higher sensitivity and resolution. In our case, however, the deconvolution process with BATMAN allowed to obtain each metabolite signal, regardless of the overlapping degree. In manual integration, some of the peaks are so small that artifacts might be produced because of the signal noise. By using BATMAN, the noise is modelled by wavelets and the signals from the compounds are more precisely estimated. This principle is also applied for peaks located in the shoulder of water signal, such as trehalose or glucose.

3.5 Metabolite quantification of yeast cell extracts

In this study we propose metabolite quantification by using a procedure based on bayesian integration. NMR metabolomics software based on Bayesian algorithms have been presented in previous literature, as the R-Package BQuant (Zheng et al. 2011). However, this software does not take profit of all the spectroscopic data from each signal. Other softwares can be used alternatively, such as the NMR suite-derived Chenomx NMR suite (Chenomx Inc., Canada). It has been previously stated that BATMAN “gave very comparable results to Chenomx, with $r^2 > 0.996$ when templates were optimised” (Hao et al. 2014). In contraposition to Chenomx, BATMAN is a free and open source software. In this work the efficiency of BATMAN to obtain estimates of biologically interpretable data is tested.

Resonances used to estimate each metabolite concentration are shown in Online Resource 4, whereas their corresponding spectroscopic parameters are detailed in Online Resource 2.

Relative concentration estimates of the identified metabolites are presented in Online Resource 6. Due to deformations in peak shape, three concentration estimates of histidine were not estimated. Twenty eight metabolites showed significant concentration changes with $p < 0.05$ due to temperature stress: at 37 °C, 22 metabolites (including cytosine, glutathione, ATP, formic acid and acetic acid, among others) had their concentration diminished when compared to 30 °C cultured samples, whereas six metabolites (including glycerophosphocholine and trehalose, among others) had increased their concentration. The same list of affected metabolites are significant with $p < 0.05$ when the statistical test performed was the Kruskal–Wallis one-way analysis of variance.

In contrast with OSC-PLS-DA in the spectra matrix, more metabolites are highlighted now as being influenced by changes in temperature. Metabolite quantification revealed that glycerophosphocholine, one of the metabolites that had been pinpointed with STOCYSY (and associated

with a $\text{VIP}_{\text{score}} = 1075.7$ in the OSC-PLS-DA), has low concentration in the sample (it is the 28th in the list of most abundant estimated metabolites). In the studied system, glycerophosphocholine concentration is increased by a factor of 5.57, but due to the fact that this molecule has 9 magnetically and chemically equivalent protons, its intensity value in the ^1H NMR spectra is increased by approximately 50-fold, which explains its high VIP score obtained in the OSC-PLS-DA.

The tentative biomarker discovery strategy from the raw spectra matrix presented before is a convenient method to perform a tentative biomarker discovery, but should not be used for obtaining conclusions about general sample composition, since peak overlapping distorts the interpretation of the signals, and both OSC-PLS-DA and STOCYSY are negatively affected because of this. On the opposite, to perform a biomarker discovery strategy focused on concentration estimates is time-consuming in comparison to direct screening by PLS-DA on spectral data.

However, in our dataset, yeast metabolome is too stable within samples of the same class (Online Resource 4c), provoking that when the proposed strategy focused on the spectral data is applied, too many metabolic changes are highlighted instead of only those more intense. Since biomarkers should not be a large list of metabolites, by increasing the VIP score threshold in two orders of magnitude (dotted line in Online Resource 4d) the list of biomarkers was restricted to three (glycerophosphocholine, fatty acids and acetic acid). Since two of these metabolites are related to the lipid fraction, an alteration of this fraction due to the growth temperature is suggested.

3.6 Metabolic alterations observed by multivariate analysis of their concentration estimates

The first principal component in the preliminary PCA of the concentration estimates already explained 41.46 % of the total variance, and allowed the separation among the two classes of samples (those at 30 °C from those at 37 °C). PC2 (14.17 % of the variance) scores were homogeneously distributed within the 95 % of confidence limits, except for one sample. The residual value associated to that sample was below the limit but it had a high leverage (large hotelling T^2 value). Variance explained by the model without this sample did not improve substantially (55.92 %). Scores projection for the 2 first PCs is represented in Fig. 4a.

OSC-PLS-DA LV1 on the autoscaled concentration estimates explained a 99.92 % of the y-variance using 44.33 % of the X-variance. All samples were correctly classified when cross-validation is performed. Low values of RMSEC and RMSECV were obtained (Online Resource 5). The permutation tests confirmed the reliability of the

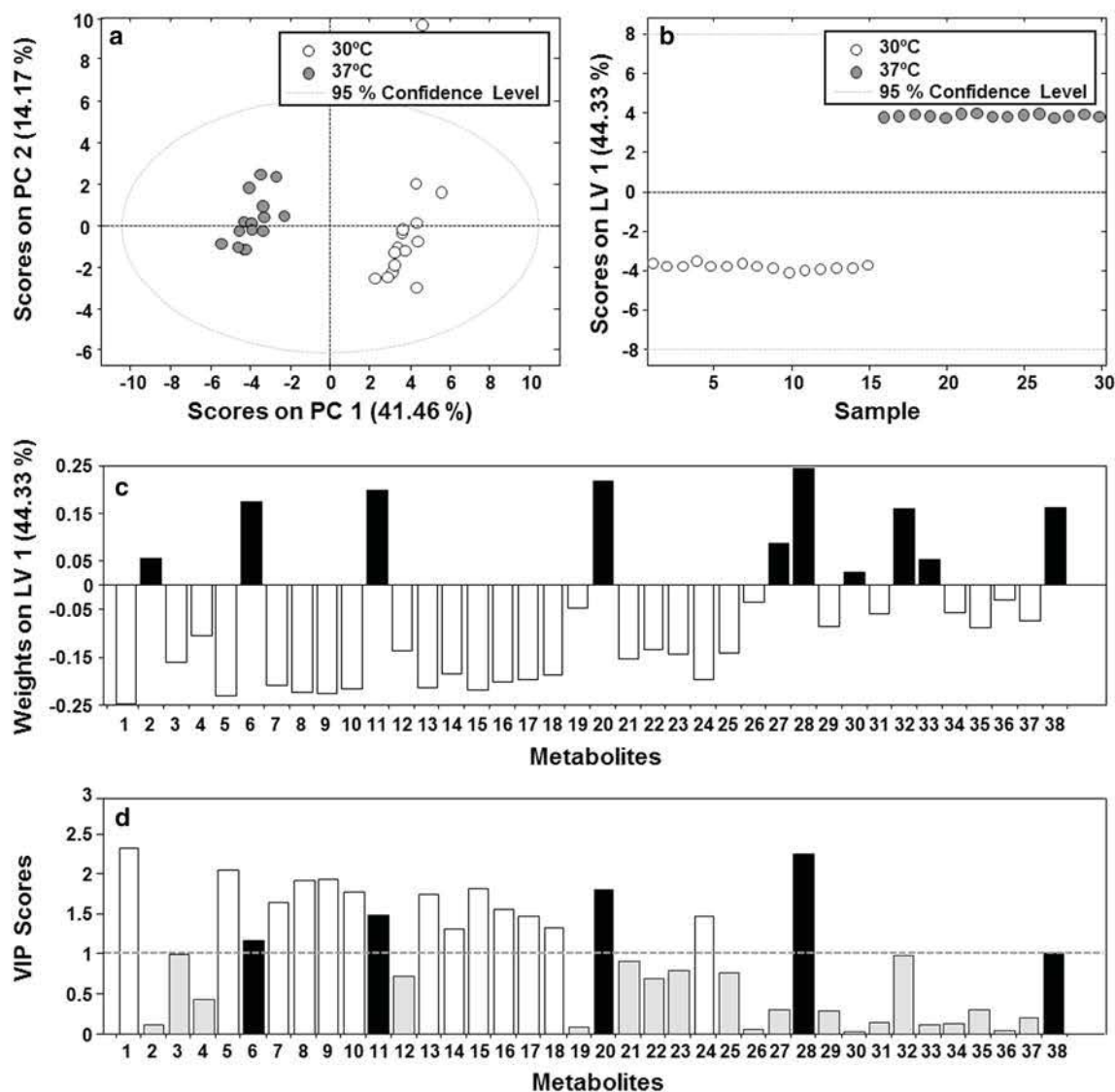


Fig. 4 **a** PCA scores projection on PC1 and PC2 subspace. *White* and *grey dots* correspond to 30 and 37 °C, respectively. **b** LV1 PLS-DA scores against each sample. **c** LV1 weights barplot. Negative weights are coloured in *white*, whereas positive weights are coloured in *black*. **d** VIP scores barplot. Bars with VIP values below threshold of 1 are coloured in *grey*. Other bars are coloured as their corresponding weight bars. Metabolites represented are the following: 1 L-threonine, 2 ethanol, 3 fatty acids, 4 L-glutamic acid, 5 L-glutathione, 6 L-histidine,

7 formic acid, 8 NAD⁺, 9 L-valine, 10 AMP, 11 L-lysine, 12 L-arginine, 13 L-isoleucine, 14 acetic acid, 15 L-leucine, 16 L-phenylalanine, 17 cytosine, 18 L-glutamine, 19 orotidine, 20 L-alanine, 21 orotic acid, 22 L-lactic acid, 23 glycine, 24 L-aspartic acid, 25 uracil, 26 L-tyrosine, 27 glycerol, 28 glycerophosphocholine, 29 ADP, 30 L-proline, 31 D-glucose, 32 citric acid, 33 L-asparagine, 34 succinic acid, 35 ATP, 36 adenine, 37 hypoxanthine, 38 trehalose.

model with a $p < 0.001$, and the leave-one-culture-out cross-validation shown a total predictive capacity for each one of the cultures used as a validation set.

Score values distribution can be observed in Fig. 4b. In Fig. 4c, LV1 weights for each metabolite are displayed. According to LV1, negative weights (in white) show those metabolites with larger concentrations at 30 °C, whereas positive weights (in black) indicate the opposite (lower concentrations at this temperature). In Fig. 4d, VIP scores are given. They indicate those metabolites whose

concentrations have changed when culture temperature changed significantly (defined by a threshold of 1). Metabolites more important in the definition of the 37 °C class were **6** (L-histidine), **11** (L-lysine), **20** (L-alanine), **28** (glycerophosphocholine) and **38** (trehalose). For the 30 °C class, the metabolites of interest are 13: **1** (L-threonine), **5** (glutathione), **7** (formic acid), **8** (NAD⁺), **9** (L-valine), **10** (AMP), **13** (L-isoleucine), **14** (acetic acid), **15** (L-leucine), **16** (L-phenylalanine), **17** (cytosine), **18** (L-glutamine) and **24** (L-aspartic acid).

Only a fraction of the metabolites identified as affected by growth temperature by the univariate statistical analysis (either at $p < 0.05$ or $p < 0.01$ confidence levels) was also detected as such by OSC-PLS-DA. This multivariate statistical analysis appeared thus to be more restrictive, probably because it models the whole set of metabolites for a given temperature at the same time, including their inter-replicate variability. Therefore, only those metabolites with concentration changes that are varying at unison within classes (but not among them) will be considered important for the model and highlighted as significant. On the contrary, in univariate statistical analysis, metabolites are evaluated one-by-one, whether or not their concentration change significantly, and coordinated metabolic variations cannot be detected.

Trehalose is one of the metabolites marked, both in the univariate and multivariate analysis, but it is not highlighted in the OSC-PLS-DA on NMR spectra. This divergence can be due to the effect of the applied pretreatment. Since NMR spectra were mean-centered instead of autoscaled, and trehalose is the identified metabolite with lowest concentration (Online Resource 6), its significance in the OSC-PLS-DA model was masked underneath other more concentrated significant metabolites. Therefore, in order to accomplish a biological interpretation from NMR data, it is preferred to perform the quantification of the effects on (autoscaled) metabolite concentrations rather than a direct interpretation from the spectrum.

3.7 Pathway analysis and biological interpretation

From a total of 28 metabolites whose concentrations varied between 30 and 37 °C and were present in the KEGG database (see below), 25 of them fell into four functional categories according to KEGG: amino acid metabolism (14 metabolites), nucleotide (purine and pyrimidine) metabolism (7 metabolites), respiration (TCA cycle and oxidative phosphorylation, 4 metabolites) and Pyruvate metabolism (4 metabolites). Some metabolites may belong to more than one category (Table 1). A comparison of these results with the transcriptomic analysis of yeast cells grown at 36 °C (GSM1046, <http://genome-www5.stanford.edu/>) revealed a rather good correlation between the metabolite changes and gene expression data (Table 1).

The figure in Online Resource 7 showed a summary of both metabolic and transcriptomic data. Note that affected metabolites and pathways are concentrated in particular areas of the total map, particularly at the top right (nucleotide metabolism), central-bottom right (amino acid metabolism) and at the centre of the map (energy gain, from sugar catabolism at the top, to oxidative phosphorylation at the bottom). In addition, most of affected metabolites (red and blue dots) are implicated in at least

one of the enzymatic steps catalyzed by a deregulated gene product, others lie shortly upstream or downstream of the same metabolic pathway (Online Resource 7).

Transcriptomic analysis showed deregulated genes in the fatty acid metabolism pathways (Online Resource 7). Although no KEGG-coded metabolite identified by NMR data analysis can be directly assigned to this pathway, it is like that the fatty acids singlet signal may be related to the alteration of the fatty acid pathway. Similarly, glycerophosphocholine concentration changes may be related to the deregulation of nine genes codifying enzymes of the Glycerophospholipid metabolism (Online Resource 7). In fact, different lipid composition changes are expected because of the type of experimental methodology applied to the analysis of yeast samples. The enhanced lipid biosynthesis allows compensating fluidity instabilities due to the changes in temperature (Sakamoto and Murata 2002; Torija et al. 2003).

Finally, glutathione showed a reduced concentration in 37 °C grown yeast cells, an effect probably related to the deregulation of different peroxidases and reductases involved in its metabolism (Table 1). Similarly, concentrations of NAD⁺, related to the redox cell machinery, are notably diminished in cells grown at 37 °C, which may be related to a decreased mitochondrial function (McConnell et al. 1990). This putative redox potential alteration may be related to the low acetate/ethanol ratio observed at 37 °C relative to cells grown at 30 °C, as a consequence of the interconversion between ethanol and acetate controlled by the NAD⁺-dependent aldehyde dehydrogenase (Racker 1949).

Trehalose, one of the few metabolites with elevated concentration at 37 °C, is not metabolically linked to the other altered metabolites, although the enzymatic reactions leading to its production and degradation are also catalyzed by deregulated gene products (Online Resource 7). Trehalose has been identified as a cellular protector to different stresses, including heat shock (Elbein et al. 2003; Estruch 2000). Our results (and others) suggest that trehalose concentration changes may be also part of the yeast acclimation process to high temperatures (Farrés et al. 2015; Strassburg et al. 2010).

Globally, our data showed a general decrease of many amino acid and nucleotide-related metabolites, whereas gene expression data of the same pathways indicate an overexpression of many enzyme-codifying genes. There is no easy way to translate changes in gene expression to changes in concentrations of metabolites, but it should be remembered that amino acids are rather strictly regulated, in part by negative feed-back mechanisms in which the absence of a given metabolite triggers transcription of the genes involved in the corresponding enzymatic pathway (Hahn and Young 2011; Hinnebusch 2005). Amino acid

Table 1 KEGG classification of metabolites and related genes affected by temperature

KEGG pathway	Metabolites		Genes ^a	
	#	KEGG name	#	Official gene name
Biosynthesis of amino acids	14	L-Glutamate, glycine, L-alanine, L-lysine, L-aspartate, L-arginine, L-glutamine, L-phenylalanine, L-leucine, L-histidine, citrate, L-valine, L-threonine, L-isoleucine	18	<i>ALT1, ARO1, BAT2, CHA1, ERR1, ERR3, GLT1, GLY1, HIS3, IDH1, IDH2, LEU4, LYS20, LYS21, LYS9, NQM1, PYC1, SHM2</i>
Purine and pyrimidine metabolism ^b	7	ADP, AMP, cytosine, glycine, L-glutamine, orotate, uracil	19	<i>ADE17, ADE2, ADE8, ADO1, APA2, DAL1, GUK1, PDE1, PGM2, RNRI, RNRI, RPA190, RPA43, RPB11, RPO31, URA1, URA10, URA4, URA8</i>
Oxidative phosphorylation and TCA cycle ^b	4	NAD ⁺ , ADP, succinate, citrate	10	<i>FUM1, IDH1, IDH2, PDA1, PYC1, SDH2, SDH4, QCR2, RIP1, VMA8</i>
Pyruvate metabolism	4	Acetate, succinate, formate, (S)-lactate	10	<i>ACCI, ALD4, FUM1, GLO2, GLO4, LEU4, LYS20, LYS21, PDA1, PYC1</i>

^a Data from dataset GSM1046, <http://genome-www5.stanford.edu>

^b Combination of two KEGG categories

pools are known to decisively determine the capacity of de novo synthesis of proteins in yeast, and their composition may vary in processes leading to changes in the cell protein composition (Onodera and Ohsumi 2005; Suzuki 2013). At this point, it is important to point out that in our study, cells were allowed to acclimate to both temperatures for two to three cell doublings, and therefore the observed changes would not reflect a typical heat shock, but rather reflect an acclimation process i.e., the compensative reaction to the actual effects of the temperature, including changes in the global content in sugars, lipids and proteins, which in turn are reflected by changes in both the transcriptome and the metabolome.

4 Concluding remarks

¹H NMR combined with statistical multivariate data analysis tools has been able to produce a robust and reliable interpretation of the effects of temperature on yeast metabolism. Detection of significant changes on the concentration of some metabolites using OSC-PLS-DA of ¹H NMR spectra, in 1D-STOCSY format, is a feasible and rapid strategy.

By using Mann–Whitney–Wilcoxon's statistical test on the comparison between metabolite concentration estimates at the two temperatures, 28 metabolites were detected as significantly affected ($p < 0.05$) by the temperature. From those, only 18 were highlighted on the corresponding PLS-DA, which appears to be more restrictive, probably because it includes the inter-replicate sample variability.

We consider that changes in glycerophosphocholine concentration may indicate more general changes in

membrane composition, likely related to the different growth temperature.

Apart from alterations on lipid membrane composition, other metabolic changes, such as the increase in the concentrations of some stress indicators like trehalose, were also detected. It is concluded that the combination of ¹H NMR metabolomics and multivariate data analysis tools are by themselves sufficient to obtain data interpretable from a biological point of view.

Acknowledgments The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement No. 320737.

Conflict of interest Francesc Puig-Castellví, Ignacio Alfonso, Benjamí Piña, and Romà Tauler declare that they have no conflict of interest.

Ethical standard This article does not contain any studies with human participants or animals performed by any of the authors.

References

- Arthur, H., & Watson, K. (1976). Thermal adaptation in yeast: Growth temperatures, membrane lipid, and cytochrome composition of psychrophilic, mesophilic, and thermophilic yeasts. *Journal of Bacteriology*, 128, 56–68.
- Astle, W., De Iorio, M., Richardson, S., Stephens, D., & Ebbels, T. (2012). A Bayesian model of NMR spectra for the deconvolution and quantification of metabolites in complex biological mixtures. *Journal of American Statistical Association*, 107, 1259–1271. doi:10.1080/01621459.2012.695661.
- Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics*, 17, 166–173. doi:10.1002/cem.785.
- Barnett, J. A., Payne, R. W., & Yarrow, D. (2000). *Yeasts: Characteristics and identification* (3rd ed.). Cambridge: Cambridge University Press.

- Bro, R., & Smilde, A. K. (2014). Principal component analysis. *Analytical Methods*, 6, 2812–2831. doi:10.1039/C3AY41907J.
- Chong, I.-G., & Jun, C.-H. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems*, 78, 103–112. doi:10.1016/j.chemolab.2004.12.011.
- Cloarec, O., Dumas, M.-E., Craig, A., et al. (2005). Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic ¹H NMR data sets. *Analytical Chemistry*, 77, 1282–1289. doi:10.1021/ac048630x.
- Core Team, R. (2013). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
- Dettmer, K., Aronov, P. A., & Hammock, B. D. (2007). Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, 26, 51–78. doi:10.1002/mas.20108.
- Dong, J., Cheng, K.-K., Xu, J., Chen, Z., & Griffin, J. L. (2011). Group aggregating normalization method for the preprocessing of NMR-based metabolomic data. *Chemometrics and Intelligent Laboratory Systems*, 108, 123–132. doi:10.1016/j.chemolab.2011.06.002.
- Dunn, W. B., Bailey, N. J. C., & Johnson, H. E. (2005). Measuring the metabolome: Current analytical technologies. *Analyst*, 130, 606–625. doi:10.1039/B418288J.
- Elbein, A. D., Pan, Y. T., Pastuszak, I., & Carroll, D. (2003). New insights on trehalose: A multifunctional molecule. *Glycobiology*, 13, 17R–27R. doi:10.1093/glycob/cwg047.
- Estruch, F. (2000). Stress-controlled transcription factors, stress-induced genes and stress tolerance in budding yeast. *FEMS Microbiology Reviews*, 24, 469–486. doi:10.1111/j.1574-6976.2000.tb00551.x.
- Farrés, M., Piña, B., & Tauler, R. (2015). Chemometric evaluation of *Saccharomyces cerevisiae* metabolic profiles using LC–MS. *Metabolomics*, 11, 210–224. doi:10.1007/s11306-014-0689-z.
- Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: A tutorial. *Analytica Chimica Acta*, 185, 1–17. doi:10.1016/0003-2670(86)80028-9.
- German, J. B., Hammock, B., & Watkins, S. (2005). Metabolomics: building on a century of biochemistry to guide human health. *Metabolomics*, 1, 3–9. doi:10.1007/s11306-005-1102-8.
- Griffin, J. L. (2003). Metabonomics: NMR spectroscopy and pattern recognition analysis of body fluids and tissues for characterisation of xenobiotic toxicity and disease diagnosis. *Current Opinion in Chemical Biology*, 7, 648–654.
- Hahn, S., & Young, E. T. (2011). Transcriptional regulation in *Saccharomyces cerevisiae*: Transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. *Genetics*, 189, 705–736. doi:10.1534/genetics.111.127019.
- Hao, J., Astle, W., De Iorio, M., & Ebbels, T. M. D. (2012). BATMAN—an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics*, 28, 2088–2090. doi:10.1093/bioinformatics/bts308.
- Hao, J., Liebeke, M., Astle, W., De Iorio, M., Bundy, J. G., & Ebbels, T. M. D. (2014). Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN. *Nature Protocols*, 9, 1416–1427. doi:10.1038/nprot.2014.090.
- Hinnebusch, A. G. (2005). Translational regulation of GCN4 and the general amino acid control of yeast. *Annual Review of Microbiology*, 59, 407–450. doi:10.1146/annurev.micro.59.031805.133833.
- Højer-Pedersen, J., Smedsgaard, J., & Nielsen, J. (2008). The yeast metabolome addressed by electrospray ionization mass spectrometry: Initiation of a mass spectral library and its applications for metabolic footprinting by direct infusion mass spectrometry. *Metabolomics*, 4, 393–405. doi:10.1007/s11306-008-0132-4.
- Jewison, T., Knox, C., Neveu, V., et al. (2012). YMDB: The yeast metabolome database. *Nucleic Acids Research*, 40, D815–D820. doi:10.1093/nar/gkr916.
- Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40, D109–D114. doi:10.1093/nar/gkr988.
- Kang, W. Y., Kim, S. H., & Chae, Y. K. (2012). Stress adaptation of *Saccharomyces cerevisiae* as monitored via metabolites using two-dimensional NMR spectroscopy. *FEMS Yeast Research*, 12, 608–616.
- Li, M., Wang, J., Lu, Z., Wei, D., Yang, M., & Kong, L. (2014). NMR-based metabolomics approach to study the toxicity of lambda-cyhalothrin to goldfish (*Carassius auratus*). *Aquatic Toxicology*, 146, 82–92. doi:10.1016/j.aquatox.2013.10.024.
- Mazzei, P., Spaccini, R., Francesca, N., Moschetti, G., & Piccolo, A. (2013). Metabolomic by ¹H NMR spectroscopy differentiates “Fiano di Avellino” white wines obtained with different yeast strains. *Journal of Agriculture and Food Chemistry*, 61, 10816–10822.
- McConnell, S. J., Stewart, L. C., Talin, A., & Yaffe, M. P. (1990). Temperature-sensitive yeast mutants defective in mitochondrial inheritance. *Journal of Cell Biology*, 111, 967–976. doi:10.1083/jcb.111.3.967.
- Mensonides, F. I. C., Hellingwerf, K. J., de Mattos, M. J. T., & Brul, S. (2013). Multiphasic adaptation of the transcriptome of *Saccharomyces cerevisiae* to heat stress. *Food Research International*, 54, 1103–1112. doi:10.1016/j.foodres.2012.12.042.
- Nonklang, S., Abdel-Banat, B. M. A., Cha-aim, K., et al. (2008). High-temperature ethanol fermentation and transformation with linear DNA in the thermotolerant yeast *Kluyveromyces marxianus* DMKU3-1042. *Applied and Environment Microbiology*, 74, 7514–7521. doi:10.1128/aem.01854-08.
- Onodera, J., & Ohsumi, Y. (2005). Autophagy is required for maintenance of amino acid levels and protein synthesis under nitrogen starvation. *Journal of Biological Chemistry*, 280, 31582–31586. doi:10.1074/jbc.M506736200.
- Palomino-Schätzlein, M., Molina-Navarro, M., Tormos-Pérez, M., Rodríguez-Navarro, S., & Pineda-Lucena, A. (2013). Optimised protocols for the metabolic profiling of *S. cerevisiae* by ¹H-NMR and HRMAS spectroscopy. *Analytical and Bioanalytical Chemistry*, 405, 8431–8441.
- Racker, E. (1949). Aldehyde dehydrogenase, a diphosphopyridine nucleotide-linked enzyme. *Journal of Biological Chemistry*, 177, 883–892.
- Sakamoto, T., & Murata, N. (2002). Regulation of the desaturation of fatty acids and its role in tolerance to cold and salt stress. *Current Opinion in Microbiology*, 5, 206–210. doi:10.1016/S1369-5274(02)00306-5.
- Strassburg, K., Walther, D., Takahashi, H., Kanaya, S., & Kopka, J. (2010). Dynamic transcriptional and metabolic responses in yeast adapting to temperature stress. *OMICS: A Journal of Integrative Biology*, 14, 249–259. doi:10.1089/omi.2009.0107.
- Suzuki, K. (2013). Selective autophagy in budding yeast. *Cell Death and Differentiation*, 20, 43–48.
- Torija, M. J., Beltran, G., Novo, M., et al. (2003). Effects of fermentation temperature and *Saccharomyces* species on the cell fatty acid composition and presence of volatile compounds in wine. *International Journal of Food Microbiology*, 85, 127–136. doi:10.1016/S0168-1605(02)00506-8.
- Wishart, D. S., Jewison, T., Guo, A. C., et al. (2013). HMDB 3.0—The human metabolome database in 2013. *Nucleic Acids Research*, 41, D801–D807. doi:10.1093/nar/gks1065.
- Wold, S., Antti, H., Lindgren, F., & Öhman, J. (1998). Orthogonal signal correction of near-infrared spectra. *Chemometrics and*

- Intelligent Laboratory Systems*, 44, 175–185. doi:10.1016/S0169-7439(98)00109-9.
- Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58, 109–130. doi:10.1016/S0169-7439(01)00155-1.
- Wolfender, J.-L., Queiroz, E. F., & Hostettmann, K. (2005). Phytochemistry in the microgram domain—a LC–NMR perspective. *Magnetic Resonance in Chemistry*, 43, 697–709. doi:10.1002/mrc.1631.
- Zheng, C., Zhang, S., Ragg, S., Raftery, D., & Vitek, O. (2011). Identification and quantification of metabolites in ^1H NMR spectra by Bayesian model selection. *Bioinformatics*, 27, 1637–1644. doi:10.1093/bioinformatics/btr118.

SUPPLEMENTARY MATERIAL FOR SCIENTIFIC ARTICLE I

A quantitative ^1H NMR approach for evaluating the metabolic response of *Saccharomyces cerevisiae* to mild heat stress.

Authors: Puig-Castellví F., Alfonso I., Piña B., Tauler R.

Citation reference: *Metabolomics* (2015), 11:1612–1625.

DOI: 10.1007/s11306-015-0812-9

Online Resource 1. Methods used for peak assignment.

#	Metabolite	YMDB or HMDB ID	KEGG C-Number	400 MHz		
				¹ H NMR	Spectral info. from extracts	Spiking
1	L-threonine	YMDB00214	C00188	NO	1D, 2D-TOCSY, J	YES
2	Ethanol	YMDB00883	C00469	NO	1D, 2D-TOCSY, J	NO
3	Fatty acids	-	-	NO	1D	NO
4	L-glutamate	YMDB00271	C00025	NO	1D, 2D-TOCSY, J	YES
5	Glutathione	YMDB00160	C00051	YES	1D, 2D-TOCSY, J	YES
6	L-histidine	YMDB00369	C00135	YES	1D, 2D-TOCSY, J	NO
7	Formic acid	YMDB00385	C00058	NO	1D, 2D-TOCSY	NO
8	NAD ⁺	YMDB00110	C00003	NO	1D, 2D-TOCSY, J	YES
9	L-valine	YMDB00152	C00183	NO	1D, 2D-TOCSY, J	NO
10	AMP	YMDB00097	C00020	YES	1D, 2D-TOCSY	YES
11	L-lysine	YMDB00330	C00047	NO	1D, 2D-TOCSY, J	YES
12	L-arginine	YMDB00592	C00062	NO	1D, 2D-TOCSY, J	YES
13	L-isoleucine	YMDB00038	C00407	NO	1D, 2D-TOCSY, J	YES
14	Acetic acid	YMDB00056	C00033	NO	1D, 2D-TOCSY	NO
15	L-leucine	YMDB00387	C00123	NO	1D, 2D-TOCSY, J	YES
16	L-phenylalanine	YMDB00304	C00079	YES	1D, 2D-TOCSY	NO
17	Cytosine	YMDB00651	C00380	NO	1D, 2D-TOCSY	NO
18	L-glutamine	YMDB00002	C00064	NO	1D, 2D-TOCSY, J	YES
19	Orotidine	HMDB00788	-	NO	1D, 2D-TOCSY	NO
20	L-alanine	YMDB00154	C00041	NO	1D, 2D-TOCSY, J	NO
21	Orotic acid	YMDB00405	C00295	NO	1D, 2D-TOCSY, J	YES
22	L-lactic acid	YMDB00247	C00186	NO	1D, 2D-TOCSY, J	YES
23	Glycine	YMDB00016	C00037	NO	1D, 2D-TOCSY	YES
24	L-aspartic acid	YMDB00896	C00049	NO	1D, 2D-TOCSY, J	NO
25	Uracil	YMDB00098	C00106	NO	1D, 2D-TOCSY	YES
26	L-tyrosine	YMDB00364	C00082	NO	1D, 2D-TOCSY, J	NO
27	Glycerol	YMDB00283	C00116	YES	1D, 2D-TOCSY, J	YES
28	Glycerophosphocholine	YMDB00309	C00670	YES	1D, 2D-TOCSY, J	YES
29	ADP	YMDB00914	C00008	YES	1D, 2D-TOCSY	YES
30	L-proline	YMDB00378	C00148	NO	1D, 2D-TOCSY, J	YES
31	D-glucose	YMDB00286	C00031	NO	1D, 2D-TOCSY, J	YES
32	Citric acid	YMDB00086	C00158	NO	1D, 2D-TOCSY, J	YES
33	L-asparagine	YMDB00226	C00152	NO	1D, 2D-TOCSY, J	YES
34	Succinic acid	YMDB00338	C00042	NO	1D, 2D-TOCSY, J	NO
35	ATP	YMDB00109	C00002	YES	1D, 2D-TOCSY	YES
36	Adenine	YMDB00887	C00147	NO	1D, 2D-TOCSY	YES
37	Hypoxanthine	YMDB00555	C00262	NO	1D, 2D-TOCSY	YES
38	Trehalose	YMDB00008	C01083	NO	1D, 2D-TOCSY	YES

Online Resource 2. Spectroscopic parameters for each assigned resonance.

#	Name	ppm	<i>J</i>	H	M	Raster
1	L-Ile	0.928	7.44	3	t	
2	L-Leu	0.946	6.10	3	d	
3	L-Leu	0.954	6.10	3	d	
4	L-Val	0.978	7.01	3	d	
5	L-Ile	1.000	7.06	3	d	
6	L-Val	1.029	7.05	3	d	
7	EtOH	1.170	7.08	3	t	
8	FA	1.237	-	4 ¹	s	
9	L-Ile	1.240	-	1	m	YES
10	L-Lac	1.313	7.00	3	d	
11	L-Thr	1.316	6.60	3	d	
12	L-Ile	1.406	-	2	m	YES
13	L-Lys	1.465	-	2	m	YES
14	L-Ala	1.468	7.14	3	d	
15	L-Arg	1.690	-	2	m	YES
16	L-Lys	1.720	-	2	m	YES
17	L-Leu	1.700	-	3	m	YES
18	DSS	1.748	-	2	m	YES
19	L-Lys	1.880	-	2	m	YES
20	HAc	1.904	-	3	s	
21	L-Arg	1.907	-	2	m	YES
22	L-Ile	1.968	-	1	m	YES
23	L-Glu	1.99-	-	2	Σ s	
24	L-Pro	2.030	-	3	m	YES
25	GSH	2.118	7.67,6.30	2	td	
26	L-Gln	2.124	-	2	m	YES
27	L-Val	2.260	-	1	m	YES
28	L-Glu	2.340	-	2	m	YES
29	L-Pro	2.341	-	1	m	YES
30	Succ.	2.390	-	4	s	
31	L-Gln	2.441	-	2	m	YES
32	Citr.	2.507	-	0.78	s	
33	Citr.	2.545	-	1.29	s	
34	GSH	2.547	7.40,3.04	2	td	
35	Citr.	2.631	-	1.22	s	
36	L-Asp	2.641	8.85	0.33	d	
37	Citr.	2.669	-	0.70	s	
38	L-Asp	2.685	8.85	0.67	d	
39	L-Asp	2.780	3.72	0.67	d	
40	L-Asp	2.825	3.72	0.33	d	
41	L-Asn	2.830	7.58	0.29	d	
42	L-Asn	2.870	7.57	0.71	d	
43	DSS	2.910	-	2	m	YES
44	L-Asn	2.922	4.34	0.71	d	
45	L-Tyr	2.933	7.75	0.67	d	
46	GSH	2.938	-	2	m	YES

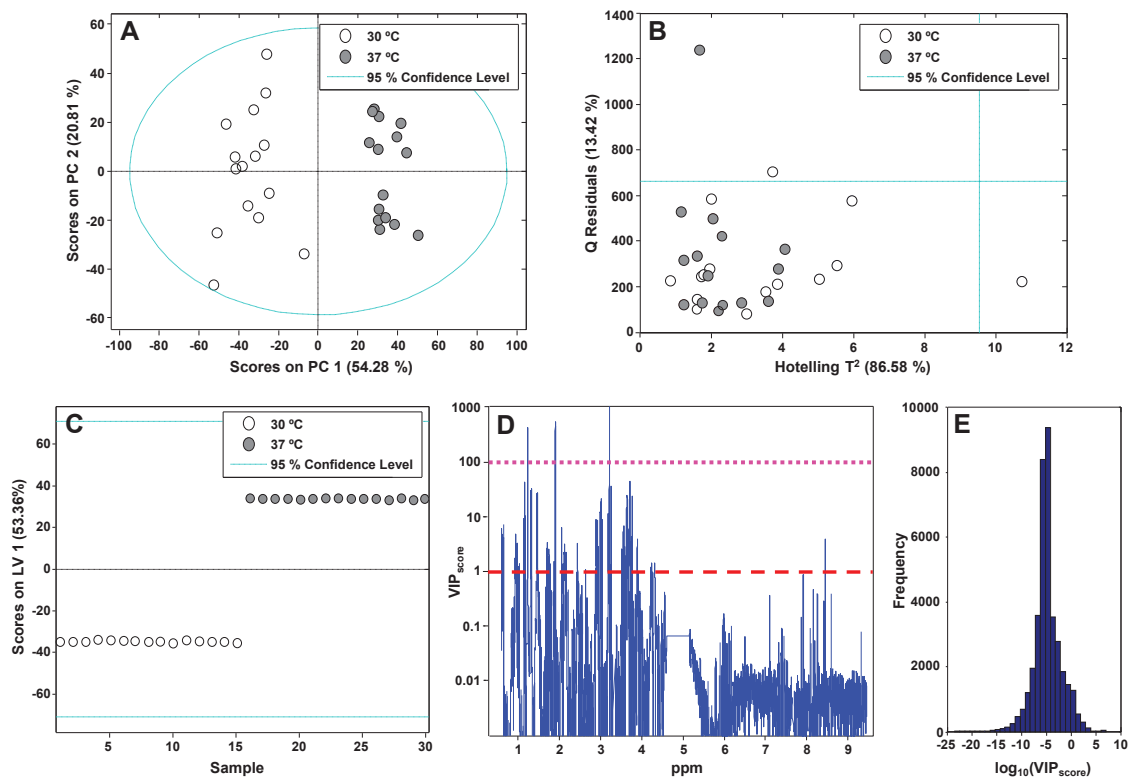
#	Name	ppm	<i>J</i>	H	M	Raster
47	L-Tyr	2.958	7.78	0.33	d	
48	L-Asn	2.965	4.37	0.29	d	
49	L-Lys	3.000	-	2	m	YES
50	L-Tyr	3.077	5.18	0.67	d	
51	L-Tyr	3.101	5.17	0.33	d	
52	L-His	3.160	1.93	0.31	d	
53	L-His	3.190	3.09	0.69	d	
54	GPC	3.217	-	9	s	
55	Glucose	3.221	7.96	0.29	d	
56	L-His	3.230	3.09	0.69	d	
57	L-Arg	3.232	6.93	2	t	
58	Glucose	3.243	7.82	0.35	d	
59	L-His	3.250	1.93	0.31	d	
60	Trehalose	3.438	9.34	2.06	t	
61	Glucose	3.460	-	2.63	m	YES
62	GlyOH	3.546	-	2.07	m	YES
63	Gly	3.547	-	2	s	
64	L-Thr	3.575	4.76	1	d	
65	Trehalose	3.624	3.90	0.9	d	
66	GPC	3.635		4	m	YES
67	EtOH	3.643	7.07	2	q	
68	GlyOH	3.645	-	2.12	m	YES
69	Trehalose	3.648	-	1.14	m	YES
70	L-Ile	3.662	3.97	1	d	
71	L-Leu	3.723	-	1	m	YES
72	L-Glu	3.746	7.19,4.72	2	dd	
73	L-Lys	3.747	6.11	1	t	
74	GSH	3.750	-	3	m	YES
75	L-Arg	3.762	6.11	1	t	
76	GlyOH	3.770	6.50, 4.37	0.79	tt	
77	L-Ala	3.787	7.20	1	q	
78	Glucose	3.790	-	2.75	m	YES
79	Trehalose	3.81	-	7.74	m	YES
80	GPC	3.895	-	3	m	YES
81	L-His	3.975	7.75, 4.92	1	dd	
82	L-Asp	4.000	7.60, 4.35	1	dd	
83	L-Lac	4.097	6.92	1	q	
84	L-Pro	4.120	8.80, 6.17	1	dd	
85	L-Thr	4.237	-	1	m	YES
86	GPC	4.315	-	2	m	YES
87	GSH	4.562	7.01,5.22	1	dd	
88	Glucose	4.634	7.97	0.61	d	
89	Trehalose	5.184	3.84	2	d	
90	Glucose	5.220	3.79	0.36	d	
91	Orotidine	5.540	3.30	1	d	
92	Orotidine	5.760	-	1	s	
93	Uracil	5.790	7.75	1	d	
94	Cyt.	5.973	8.17	1	d	

#	Name	ppm	<i>J</i>	H	M	Raster
95	NAD ⁺	6.025	5.86	1	d	
96	NAD ⁺	6.075	5.27	1	d	
97	AMP	6.130	5.96	1	d	
98	Orotate	6.177	-	1	s	
99	L-Tyr	6.888	8.68	2	d	
100	L-Hys	7.097	0.58	1	d	
101	L-Tyr	7.182	8.68	2	d	
102	L-Phe	7.370	-	5	m	YES
103	Uracil	7.525	7.65	1	d	
104	L-Hys	7.907	1.13	1	d	
105	NAD ⁺	8.160	-	1	s	
106	NAD ⁺	8.171	6.09	0.5	d	
107	Adenine	8.184	-	1	s	
108	Hyp.	8.191	7.94	2	d	
109	NAD ⁺	8.191	6.29	0.5	d	
110	Adenine	8.232	-	1	s	
111	AMP	8.256	-	1	s	
112	ADP	8.257	-	1	s	
113	ATP	8.259	-	1	s	
114	NAD ⁺	8.416	-	1	s	
115	Formate	8.442	-	1	s	
116	ADP	8.526	-	1	s	
117	ATP	8.534	-	1	s	
118	AMP	8.591	-	1	s	
119	NAD ⁺	8.822	8.22	1	d	
120	NAD ⁺	9.132	6.14	1	d	
121	NAD ⁺	9.323	-	1	s	

¹Fatty acids number of protons is defined arbitrarily to 4 in order to obtain a concentration value comprised within the concentration range for the other estimated metabolites. H: Number of protons, M: Multiplet. Raster: The signal is modelled as a raster multiplet or not.

Online Resource 3. Input parameters modified from the default *batmanoptions.txt* template.

Parameter	Value
Intensity scale factor	800000
Number of burn-in iterations	2000
Number of post-burn-in iterations	1000
Number of parallel processes	3
Number of spectra to be modelled	30
Spectrometer frequency (MHz)	400



Online Resource 4. a. PCA projection of the ^1H NMR mean-centered spectra on the two first principal components subspace. **b.** Hotelling T^2 against the Q Residuals scatterplot of each NMR spectrum. **c.** LV1 OSC-PLS-DA Scores plot. **d.** VIP scores plot represented against the chemical shift, in logarithmic y-scale. $\text{VIP}_{\text{score}}$ threshold of 1 and of 100 is represented by dashed and dotted lines, respectively. **e.** Histogram of the VIP score plot, in a logarithmic x-scale.

Online Resource 5. RMSEC and RMSECV values obtained for the different chemometric models used.

		¹ H NMR data	Concentration estimates
PCA	RMSEC	0.09	0.655
	RMSECV	8.058	0.96
PLS-DA	RMSEC	0.005	0.0139
	RMSECV	0.103	6.036

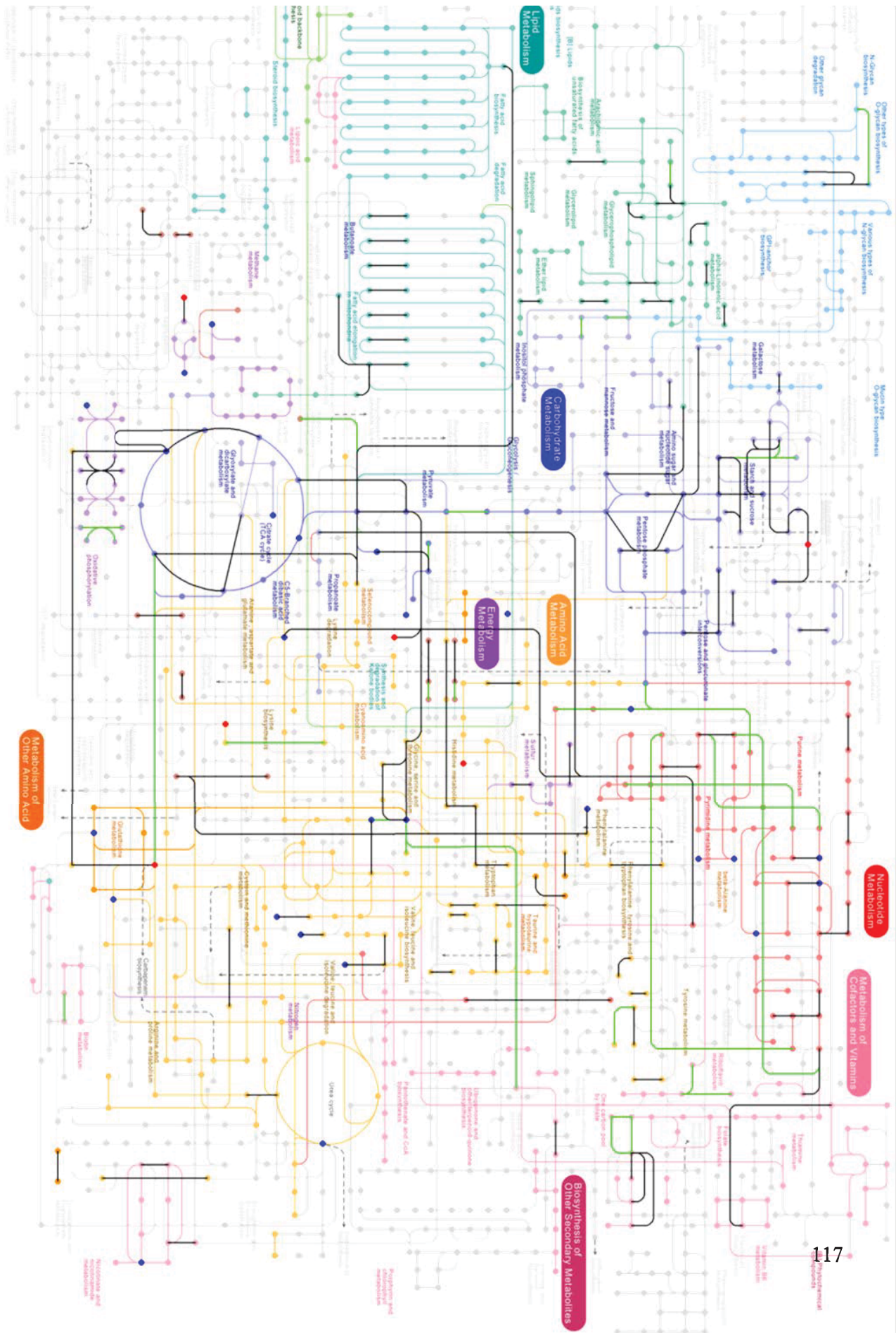
Online Resource 6. Fold changes concentration for metabolites identified in yeast extracts.

Metabolite	[] ₃₇ /[] ₃₀	p-value ¹	30°C		37°C	
			Mean ²	SD ³	Mean ²	SD ³
L-threonine	0.63	***	2.62 x 10 ⁻¹	1.65 x 10 ⁻²	1.66 x 10 ⁻¹	1.07 x 10 ⁻²
Ethanol	1.48		1.25 x 10 ⁻¹	1.19 x 10 ⁻¹	1.84 x 10 ⁻¹	1.55 x 10 ⁻¹
Fatty acids	0.39	***	1.58 x 10 ⁻¹	8.33 x 10 ⁻²	6.12 x 10 ⁻²	2.74 x 10 ⁻²
L-glutamic acid	0.94	*	1.33 x 10 ⁻¹	8.94 x 10 ⁻³	1.25 x 10 ⁻¹	8.45 x 10 ⁻³
Glutathione	0.47	***	1.16 x 10 ⁻¹	1.77 x 10 ⁻²	5.45 x 10 ⁻²	1.20 x 10 ⁻²
L-histidine	1.22	***	9.30 x 10 ⁻²	1.10 x 10 ⁻²	1.13 x 10 ⁻¹	1.18 x 10 ⁻²
Formic acid	0.52	***	8.07 x 10 ⁻²	1.90 x 10 ⁻²	4.22 x 10 ⁻²	7.23 x 10 ⁻³
NAD⁺	0.69	***	6.52 x 10 ⁻²	7.22 x 10 ⁻³	4.48 x 10 ⁻²	4.09 x 10 ⁻³
L-valine	0.66	***	6.43 x 10 ⁻²	7.09 x 10 ⁻³	4.25 x 10 ⁻²	4.80 x 10 ⁻³
AMP	0.72	***	6.12 x 10 ⁻²	6.42 x 10 ⁻³	4.42 x 10 ⁻²	4.59 x 10 ⁻³
L-lysine	1.18	***	5.07 x 10 ⁻²	3.34 x 10 ⁻³	5.96 x 10 ⁻²	4.21 x 10 ⁻³
L-arginine	0.87	**	4.99 x 10 ⁻²	5.14 x 10 ⁻³	4.36 x 10 ⁻²	5.19 x 10 ⁻³
L-isooleucine	0.59	***	4.91 x 10 ⁻²	7.85 x 10 ⁻³	2.91 x 10 ⁻²	5.66 x 10 ⁻³
Acetic acid	0.55	***	3.18 x 10 ⁻²	8.78 x 10 ⁻³	1.74 x 10 ⁻²	5.33 x 10 ⁻³
L-leucine	0.58	***	3.15 x 10 ⁻²	5.34 x 10 ⁻³	1.83 x 10 ⁻²	2.44 x 10 ⁻³
L-phenylalanine	0.68	***	3.08 x 10 ⁻²	5.24 x 10 ⁻³	2.09 x 10 ⁻²	2.12 x 10 ⁻³
Cytosine	0.44	***	2.96 x 10 ⁻²	8.82 x 10 ⁻³	1.31 x 10 ⁻²	4.89 x 10 ⁻³
L-glutamine	0.69	***	2.94 x 10 ⁻²	6.26 x 10 ⁻³	2.02 x 10 ⁻²	9.88 x 10 ⁻⁴
Orotidine	0.96		2.76 x 10 ⁻²	4.21 x 10 ⁻³	2.64 x 10 ⁻²	2.03 x 10 ⁻³
L-alanine	1.25	***	2.09 x 10 ⁻²	1.24 x 10 ⁻³	2.61 x 10 ⁻²	2.02 x 10 ⁻³
Orotic acid	0.66	***	2.52 x 10 ⁻²	7.11 x 10 ⁻³	1.67 x 10 ⁻²	4.02 x 10 ⁻³
L-lactic acid	0.7	**	2.43 x 10 ⁻²	6.05 x 10 ⁻³	1.69 x 10 ⁻²	6.10 x 10 ⁻³
Glycine	0.88	**	2.42 x 10 ⁻²	2.30 x 10 ⁻³	2.12 x 10 ⁻²	2.11 x 10 ⁻³
L-aspartic acid	0.85	***	2.38 x 10 ⁻²	1.91 x 10 ⁻³	2.02 x 10 ⁻²	1.16 x 10 ⁻³
Uracil	0.62	***	2.26 x 10 ⁻²	8.45 x 10 ⁻³	1.41 x 10 ⁻²	4.34 x 10 ⁻³
L-tyrosine	0.97		1.53 x 10 ⁻²	2.25 x 10 ⁻³	1.48 x 10 ⁻²	1.40 x 10 ⁻³
Glycerol	1.32		1.02 x 10 ⁻²	4.23 x 10 ⁻³	1.34 x 10 ⁻²	4.81 x 10 ⁻³
Glycerophosphorylcholine	5.57	***	2.39 x 10 ⁻³	2.36 x 10 ⁻³	1.33 x 10 ⁻²	1.18 x 10 ⁻³
ADP	0.63	*	1.05 x 10 ⁻²	6.78 x 10 ⁻³	6.63 x 10 ⁻³	4.30 x 10 ⁻³
L-proline	1.04		7.37 x 10 ⁻³	1.54 x 10 ⁻³	7.63 x 10 ⁻³	9.25 x 10 ⁻⁴
D-glucose	0.91		5.23 x 10 ⁻³	8.40 x 10 ⁻⁴	4.78 x 10 ⁻³	1.10 x 10 ⁻³
Citric acid	1.51	***	3.32 x 10 ⁻³	8.66 x 10 ⁻⁴	5.03 x 10 ⁻³	1.30 x 10 ⁻³

Metabolite	Π_{37}/Π_{30}	p -value ¹	30°C		37°C	
			Mean ²	SD ³	Mean ²	SD ³
L-asparagine	1.07		3.67 x 10 ⁻³	5.92 x 10 ⁻⁴	3.93 x 10 ⁻³	6.25 x 10 ⁻⁴
Succinic acid	0.91	*	3.75 x 10 ⁻³	9.31 x 10 ⁻⁴	3.41 x 10 ⁻³	5.33 x 10 ⁻⁴
ATP	0.51		3.20 x 10 ⁻³	2.93 x 10 ⁻³	1.64 x 10 ⁻³	1.06 x 10 ⁻³
Adenine	0.71		2.74 x 10 ⁻³	4.79 x 10 ⁻³	1.94 x 10 ⁻³	1.31 x 10 ⁻³
Hypoxanthine	0.71		1.22 x 10 ⁻³	5.50 x 10 ⁻⁴	8.70 x 10 ⁻⁴	6.48 x 10 ⁻⁴
Trehalose	1.89	***	3.30 x 10 ⁻⁴	1.69 x 10 ⁻⁴	6.25 x 10 ⁻⁴	2.00 x 10 ⁻⁴

¹ p -value: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Metabolites associated to $p < 0.001$ are highlighted in bold. ²Mean was calculated from the set of concentration estimates after applying *Batman()* function to the corresponding SNV preprocessed spectra. ³SD: standard deviation.

Online Resource 7. *S. cerevisiae* metabolic atlas indicating metabolites that showed higher (red dots) or lower (blue dots) concentrations at 37°C than at 30°C, and enzymatic reactions corresponding to the genetic products of genes over-(black lines) or underrepresented (green lines) in yeast cells grown at 36°C.



2.2 SCIENTIFIC ARTICLE II

Deciphering the underlying metabolomic and lipidomic patterns linked to thermal acclimation in *Saccharomyces cerevisiae*.

Authors: Puig-Castellví F., Bedia, C., Alfonso I., Piña B., Tauler R.

Citation reference: Journal of Proteome Research (2018).

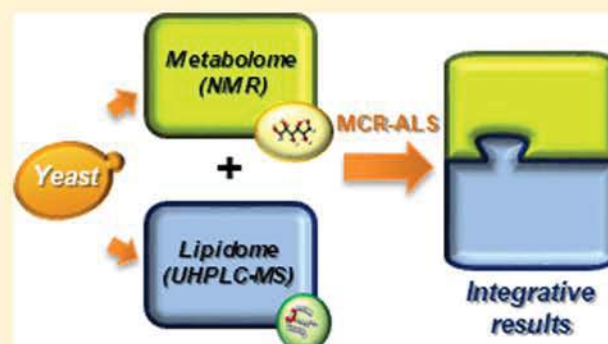
DOI: 10.1021/acs.jproteome.7b00921

Deciphering the Underlying Metabolomic and Lipidomic Patterns Linked to Thermal Acclimation in *Saccharomyces cerevisiae*

Francesc Puig-Castellví,[†] Carmen Bedia,[†] Ignacio Alfonso,[‡] Benjamin Piña,[†] and Romà Tauler^{*,†}[†]Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, Barcelona 08034, Catalonia, Spain[‡]Department of Biological Chemistry and Molecular Modelling, IQAC–CSIC, Jordi Girona 18-26, Barcelona 08034, Catalonia, Spain**S** Supporting Information

ABSTRACT: Temperature is one of the most critical parameters for yeast growth, and it has deep consequences in many industrial processes where yeast is involved. Nevertheless, the metabolic changes required to accommodate yeast cells at high or low temperatures are still poorly understood. In this work, the ultimate responses of these induced transcriptomic effects have been examined using metabolomics-derived strategies. The yeast metabolome and lipidome have been characterized by 1D proton nuclear magnetic resonance spectroscopy and ultra-high-performance liquid chromatography–mass spectrometry at four temperatures, corresponding to low, optimal, high, and extreme thermal conditions. The underlying pathways that drive the acclimation response of yeast to these nonoptimal temperatures were evaluated using multivariate curve resolution–alternating least-squares. The analysis revealed three different thermal profiles (cold, optimal, and high temperature), which include changes in the lipid composition, secondary metabolic pathways, and energy metabolism, and we propose that they reflect the acclimation strategy of yeast cells to low and high temperatures. The data suggest that yeast adjusts membrane fluidity by changing the relative proportions of the different lipid families (acylglycerides, phospholipids, and ceramides, among others) rather than modifying the average length and unsaturation levels of the corresponding fatty acids.

KEYWORDS: NMR, lipid, LC–MS, temperature, MCR-ALS, yeast



INTRODUCTION

The budding yeast *Saccharomyces cerevisiae* has been used for food processing and fermentation of alcoholic beverages over the past few millennia. Most recently, recombinant protein production and metabolic engineering are two of the many new applications of yeast.¹ Because of this historical close relationship of yeast with mankind, it is not surprising that yeast is considered as a model organism. The yeast genome was fully sequenced in 1996, and it was the first eukaryotic genome to be completely sequenced.²

The yeast metabolism has also been vastly studied because the metabolome is the closest to the phenotype.³ Examples of metabolomics research in yeast study pathway disruptions,^{4,5} exposure to external stresses,^{6,7} and acclimation metabolic responses,^{8,9} among others, are available in the literature.

The growth temperature is a key parameter that controls yeast metabolism. For most yeast strains, growth is optimal between 25 and 30 °C.¹⁰ However, in the yeast industry, fermentations are carried out at higher temperatures to obtain the highest possible ethanol yields at reduced production costs.¹¹ Conversely, wine fermentations at lower temperatures are becoming more frequent to improve their organoleptic properties, and a more pronounced aromatic profile is

obtained.¹² Thus the study of yeast performance at different temperatures is of the utmost theoretical and practical interest.

Metabolomic and lipidomic profiles of yeast cultures grown at three nonoptimal temperatures (15, 37, and 40 °C) and one at the optimal range (30 °C) were characterized by 1D proton nuclear magnetic resonance (¹H NMR) and by ultra-high-performance liquid chromatography–mass spectrometry (UHPLC–MS), respectively.

It is known that thermal stress in yeast produces changes in both the primary metabolism and lipid composition.^{13,14} Independent from the method of analysis, conventional metabolomic and lipidomic analyses are performed separately, given the different extraction strategies used for polar metabolites and for lipids.¹⁵ To combine these two “-omes”, a good data fusion strategy is crucial for a proper understanding of the studied biological system. In this article, we propose the use of the chemometric method multivariate curve resolution–alternating least-squares (MCR-ALS)¹⁶ to extract the common underlying sources (components) of metabolite variation in the two different data sets from NMR metabolomics and UHPLC–MS lipidomics. In the profiles resolved by the MCR-ALS

Received: December 27, 2017

Published: April 30, 2018

method, correlated variables (metabolites and lipids) are captured in the same single component. Preliminary examples of MCR-ALS resolution of other types of fused data sets (unrelated to “Omics”) can be found in the literature.^{17,18} Grouped metabolites in these MCR-ALS resolved components can be related to one or to a few biological pathways, thus giving informative results that are useful for biomarker discovery and interpretation.^{4,19}

This work attempts to provide some insights and a better understanding of yeast temperature regulation mechanisms. This knowledge can have potential implications, for example, in the design of thermotolerant yeast strains for biotechnology or food industry uses.

■ EXPERIMENTAL SECTION

Yeast Growth

S. cerevisiae BY4741 cells were precultured in YPD medium at 30 °C and 150 rpm overnight. Fresh YPD medium was inoculated with the preculture to an absorbance at 600 nm (A_{600}) of 0.1 and divided into volumes of 50 mL, which were grown at 15, 30, 37, or 40 °C. After reaching an A_{600} of 0.6 to 0.8, the cultures were arrested on ice. Every culture was aliquoted into two fractions of 5 and 45 mL, which were used for the lipidic and metabolic extraction, respectively. Cell harvesting was performed by centrifugation of every (5 and 45 mL) fraction at 4000g for 5 min and discarding the supernatant. The cells were washed afterward with 100 mM sodium phosphate buffer (pH 7.0). The resulting pellets were stored at -80 °C and lyophilized.

Metabolite and Lipid Extraction and Sample Preparation

The metabolites were extracted by following the protocol published in our previous work.¹⁴ Metabolic yeast extracts were finally dissolved in 700 μ L of deuterated phosphate buffer (Na_2DPO_4 100 mM, pH 7.0) in D_2O with DSS 0.2 mM as an internal standard and measured with ^1H NMR spectroscopy.

Lipids were extracted by using a slight modification of a previously published protocol, which was designed for human cell lines.²⁰ The pellets that corresponded to the 5 mL samples were resuspended in 100 μ L of deionized water. Then, 250 μ L of methanol and 500 μ L of chloroform were subsequently added. This mixture was fortified with internal standards of lipids (1,2,3-17:0 triglyceride (TG), 1,3-17:0 (d5) diglyceride (DG), 17:0 cholesteryl ester, 16:0 D31-18:1 phosphatidylethanolamine (PE), 16:0 D31-18:1 phosphatidylserine (PS), 16:0 D31-18:1 phosphatidylglycerol (PG), 16:0 D31-18:1 phosphatidylcholine (PC), 17:1 lyso PC (LPC), 17:1 lyso PE (LPE), 17:1 lyso PG (LPG), and 17:1 lyso PS (LPS)). Disruption of the cell wall was achieved by vortexing the samples with glass beads followed with sonication twice. The samples were incubated overnight at 48 °C and cooled to 37 °C afterward. Then, they were evaporated under an N_2 stream and resuspended in 500 μ L of methanol. The samples were centrifuged at 9168g for 3 min, and 130 μ L of the supernatants was transferred to UPLC vials for injection. This mixture was fortified with internal standards of sphingolipids (*N*-dodecanoylsphingosine, *N*-dodecanoylglucosyl-sphingosine, and *N*-dodecanoylsphingosylphosphorylcholine, 200 pmol each). These lipid samples were measured with UHPLC-MS spectrometry.

NMR Measurements

Spectra were recorded in a 400 MHz Varian spectrometer using a spectrometer frequency of 400.14 MHz with a OneNMR Probe and a ProTune System (Agilent). The proton spectral size range covered from -2 to 12 ppm, which consisted of 65k data points. The number of scans was 512, and the relaxation delay was 5 s.

UHPLC-MS Measurements

LC-MS analysis consisted of a Waters Acquity UPLC system connected to a Waters LCT Premier orthogonal accelerated time-of-flight mass spectrometer (Waters), operated in both positive and negative electrospray ionization modes (ESI+ and ESI-, respectively). Full-scan spectra from 50 to 1500 Da were acquired, and individual spectra were summed to produce data points, each of 0.2 s. The mass accuracy and reproducibility were maintained by using an independent reference spray via the LockSpray interference. The analytical column was a 100 \times 2.1 mm inner diameter, 1.7 mm C8 Acquity UPLC bridged ethylene hybrid (Waters). The two mobile phases were MeOH 1 mM ammonium formate (phase A) and H_2O 2 mM ammonium formate (phase B). The flow rate was 0.3 mL min^{-1} , and the gradient of A/B solvents started at 80:20 and changed to 90:2 in minute 3; from minute 3 to 6 remained at 90:10; changed to 99:1 in minute 6 until minute 15; remained 99:1 until minute 18; and finally, returned to the initial conditions until minute 20. The column was held at 30 °C.

NMR Data Preprocessing

Referencing to the NMR standard (DSS, 4,4-dimethyl-4-silapentane-1-sulfonic acid), apodization, phasing, and baseline correction of the NMR spectra were applied in MestreNova v.9.0 (Mestrelab Research, Spain) and imported to Matlab R2014b (The Mathworks, Natick, MA). Regions of water (4.66–5.16 ppm), methanol (3.30–3.37 ppm), and chloroform (7.63–7.70 ppm) were removed as well as the data points with chemical shifts >9.40 ppm or <0.75 ppm. Finally, the ^1H NMR spectra were normalized using Probabilistic Quotient Normalization (PQN)^{4,21} to correct for possible sample size effects.

UHPLC-MS Data Preprocessing

Every UHPLC-MS data file was converted to CDF format by the Databridge program of the MassLynx software (Waters, Inc.). The resulting data were imported into the MATLAB environment. The data dimensions were compressed without loss of spectral resolution by using the Regions-of-Interest (ROI) strategy.^{22,23} In this strategy, m/z traces with lower intensity signals than a defined threshold value (i.e., noise) and not measured in consecutive scans (i.e., spurious measurements) were excluded from further analysis. This procedure allows for a more than 100-fold computer storage reduction but maintains the highest possible experimental mass accuracy.²⁴ Every UHPLC-MS sample provided a data matrix with as many rows as the number of measured retention times and as many columns as the number of selected m/z ROI values. The generated ROI-MS matrices were normalized using the absorbance at 600 nm (A_{600}) of the precedent yeast culture and the peak area of the internal standards.

In a second stage, individual (one sample) ROI-MS data matrices can be joined to build up a new column-wise augmented data matrix with all ROI values for the set of studied samples using a particular MS-ESI mode. More details of this strategy can be found in Gorrochategui et al.²²

NMR Data Analysis

Metabolite assignment was performed by a detailed targeted metabolite profiling analysis of the ¹H NMR signals using a homemade ¹H NMR spectra library and the Yeast Metabolome Data Base library (YMDB).²⁵ A relative metabolite quantification of the ¹H NMR spectral matrix was performed using the BATMAN R-package.²⁶ Further information about how BATMAN works and the exact protocol can be found elsewhere.^{14,27}

UHPLC–MS Data Analysis

MCR-ALS was subsequently applied to the augmented ROI data matrices generated for the two ESI modes to resolve the elution profiles and mass spectra associated with every lipid. See the [Chemometric Data Analysis](#) section and the [Extended Methods](#) section in the [Supporting Information](#) for more information about MCR-ALS and about the integration of the MSROI and MCR-ALS methods in the ROIMCR method.²³

Tentative assignment of lipids was performed from the *m/z* values associated with the smallest delta values (difference between the query mass and adduct mass) using the Yeast Metabolome Database²⁵ and Lipid Maps²⁸ online databases and from comparing the associated retention times to those from lipids found in previous studies. Glycerophospholipids, diacylglycerol, triacylglycerol, and cholesterylestes were annotated as <lipid subclass> <total fatty acyl chain length>:<total number of unsaturated bonds>.

Each lipid was normalized to the internal standard of the same lipid family in every sample or to the average response of all of the used internal standards if no standard of the same lipid family was comprised within the list of standards.

The average number of unsaturations (*nI*) and the average length (*L*) of the fatty acyl chains were calculated using the following equations

$$nI = \frac{\sum_k^K n_k c_k i_k}{\sum_k^K n_k c_k} \quad (1)$$

$$L = \frac{\sum_k^K n_k c_k l_k}{\sum_k^K n_k c_k} \quad (2)$$

For every lipid (*k*) that contains at least one fatty acyl chain, *n_k* is the number of fatty acyl chains, *c_k* is the relative concentration for every considered lipid, *i_k* is its number of unsaturations, and, finally, *l_k* is the average carbon length of their fatty acyl chains.

Chemometric Data Analysis

Preliminary PCA Analysis. PCA was performed on the raw experimental data, either from the experimental raw ¹H NMR spectra or from the TIC chromatograms obtained with the distinct ionization modes. In all cases, data were mean-centered prior to analysis.

MCR-ALS Analysis. MCR-ALS decomposes a data matrix using the following bilinear model

$$X = CS^T + E \quad (3)$$

where **C** and **S^T** are the pure concentrations and spectral profiles of the resolved components in **X** and **E** is the residual matrix that contains the data variation not explained by **CS^T**.

MCR-ALS^{16,29} has been applied here to the analysis of two different types of data sets (data matrices **X**): (i) UHPLC–

MS–ROI data matrices and (ii) NMR-derived metabolite and UHPLC–MS-derived lipid area matrices.

ROIMCR Method: MCR-ALS Analysis of UHPLC–MS–ROI Data Matrices. In this case, MCR-ALS analysis is applied to the two column-wise augmented MSROI data matrices obtained in the analysis of the yeast samples with positive and negative electrospray ionization modes (ESI+ and ESI–, respectively).⁶ The two augmented MSROI data matrices have a total number of 15 264 rows (24 samples × 636 retention times) and as many columns as the number of detected ROI *m/z* values. 688 ROIs were detected in ESI+, and 119 ROIs were detected in ESI–.

In the MCR-ALS resolution of the UHPLC–MS–ROI data matrices (in the ROI-MCR method²³), the obtained **S^T** has the MCR-ALS resolved pure MSROI spectra of each lipid, whereas the resulting **C** has the chromatographic elution profiles of each of them.

MCR-ALS Analysis of Metabolite and Lipid Peak Area Data Matrices. Additionally, in a different data analysis stage, MCR-ALS was also applied to the set of abundances⁴ (signal/peak areas) derived from the previous metabolomic (from ¹H NMR) and lipidomic (from UHPLC–MS–ROI) studies described above. The two peak area matrices (from ¹H NMR and UHPLC–MS–ROI analyses) were joined row-wisely, producing a new lipidomic-and-metabolomic peak areas data set, especially useful to study and interpret possible correlations between these two types of compounds and their biological interrelationships.

The lipidomic-and-metabolomic fused data set gives a new data matrix **X**, which is submitted to the MCR-ALS analysis. In this case, this matrix has a number of rows equal to the number of samples (24 samples) and a number of columns equal to the number of identified compounds. (In this data example, this number was equal to 123; see the [Results](#) section.) The resolved matrix **S^T** will show, in this case, the lipid and metabolic composition (lipidomic-and-metabolomic profile) of the resolved MCR-ALS components, whereas **C** will give the contributions of these lipidomic-and-metabolomic profiles in the different samples. From these **C** matrices, the evolution of the concentrations of the different metabolites due to temperature changes can be deduced and described as metabolic thermal profiles.

In both MCR-ALS analyses, the quality of the MCR-ALS models was measured by evaluating the percent of explained variance (*R*²).²⁹

Biological interpretation of MCR-ALS resolved **S^T** profiles was attempted from the set of metabolites (or lipids) whose contribution was 50% or higher than the maximum contribution found in the given component. For every set of selected lipids and metabolites, a pathway analysis using the KEGG³⁰ pathway database was performed.

Lipid abundances for each of the resolved components were obtained after undoing the two transformation steps applied as a data pretreatment (min-max scaling and singular value scaling) of the original data on the **S^T** matrix. The lipid content of every lipid family for each **S^T** component profile was obtained by the sum of all of the reconstructed concentration values that regarded the same lipid family found in the same **S^T** profile.

Finally, the average fatty acyl chain length and the average number of unsaturations for each **S^T** profile were obtained by applying eqs 1 and 2, respectively, on the set of reconstructed concentration values associated with each **S^T** profile.

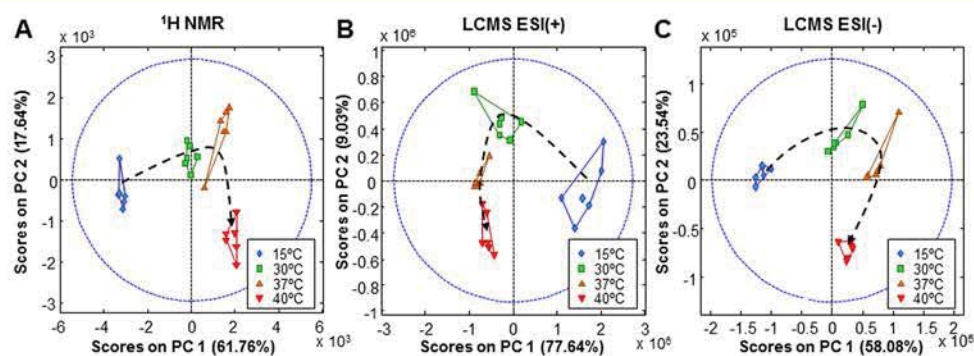


Figure 1. PC1 versus PC2 sample score plots of the mean-centered data sets: (A) ^1H NMR data set, (B) ESI+ TICs data set, and (C) ESI-TICs data set. The black dashed arrow on the PCA score plots shows how the sample scores evolve with the temperature.

A more detailed protocol of the different experimental and chemometric procedures used in this work can be found in the Extended Methods section in the Supporting Information.

RESULTS

Biological Overall Response to Acclimation

Before the metabolome and lipidome characterization, ^1H NMR raw spectra and UHPLC–MS TICs (ESI+ and ESI–) were mean-centered and analyzed separately by principal component analysis (PCA).

PCA score plots show a clear clustering of samples according to the growth temperature for both the yeast metabolomic and lipidomic data (Figure 1). In addition, a temperature-dependent gradual progression of the score distribution (denoted with a black dashed arrow) is observed in the three score plots. We concluded that these analyses reflect the acclimation of yeast cells to the different temperatures that affect both their main metabolism and their lipid profile.

A common feature in all of the score plots of Figure 1 is that the major component, PC1, which ranges from 58 to 77% of the explained variance, separates the yeast extracts from the samples cultured at 15 °C from the remaining cells, revealing that the highest metabolomic and lipidomic variations occurred between this temperature and the other temperatures. In contrast, PC2 (9 to 23% of the explained variance) separated the most extreme high temperature (40 °C) from the others, whereas the two “mild” conditions clustered relatively close to each other (Figure 1). It is remarkable that 40 °C is only 2 °C below the maximum temperature, after which our strain ceases to grow (42 °C, data not shown); in this regard, the relatively minor change in the metabolomic and lipidomic profiles that were induced under this sublethal condition compared with the variance associated with growth at low temperatures was largely unexpected.

Compound Assignment and Quantification

A total number of 42 metabolites, including mostly organic acids, amino acids, nucleobases, nucleotides, and sugars were identified by NMR according to the procedure described in the NMR Data Analysis section. Spectroscopic information for the identified metabolites is presented in Table S1 in the Supporting Information.

On the contrary, the application of the untargeted Regions-of-Interest approach (explained in the UHPLC–MS Data Preprocessing section) combined with the MCR-ALS procedure explained below in the Chemometric Data Analysis section produced 80 tentative lipid candidates in the ESI+

mode and another 50 in the ESI– mode of the MS. In both cases, the experimental data were fitted adequately (99% of the experimental data variance). After discarding the lipids of unknown identity and the coincident candidates due to the use of the two ionization methods (ESI+ and ESI–), the final list included 81 unique chemical compounds.

The list of compounds comprised basically lipids (including diacylglycerides (DG), triacylglycerides (TG), phosphatidic acids (PA), phosphatidylcholines (PC), phosphatidylglycerols (PG), phosphatidylinositols (PI), phosphatidylserines (PS), sphingolipids, inositol phosphate ceramides (PI-Cer), and cholesterol esters (CE)) but also 1D-myoinositol-4,5-biphosphate (myo-4,5-BP) and *N*-acetyl-D-glucosaminyldiphosphodolichol (NA-GADPD). Compound information that allowed their assignment, including major adducts, retention time, and mass error, can be consulted in Table S2 in the Supporting Information.

Heatmaps of the metabolite and lipid concentrations are given in Figure 2. A larger number of differences can be observed in the lipid concentrations (Figure 2A) than in the metabolite concentrations (Figure 2B). Most of the accumulated lipids (in red) at 15, 30, or 40 °C were specific for one of the studied temperatures, whereas the lipid profile of the samples grown at 37 °C was intermediate between those that grew at 30 and 40 °C. These results are in agreement with previous results obtained with PCA of the TIC chromatograms (see above). On the contrary, the NMR data set (Figure 2C) shows specific metabolite profiles for each of the tested temperatures, although the data are consistent with the previous observation that the samples grown at 15 °C were clearly differentiated from the others.

Thermal and Metabolic Profiles

Simultaneous application of MCR-ALS to the whole data set of concentrations of metabolites and lipids at the different temperatures defined three components that explained 86.9% of the total experimental variance. (See the Chemometric Data Analysis section for methodological details.) Lipidomic-and-metabolic profiles (or only “metabolic profiles”, in matrix S^T of eq 3; see the Chemometric Data Analysis section) describe the contribution of the measured lipids and metabolites in each of these three components. The thermal profiles (matrix C in eq 3 in the Chemometric Data Analysis section) for each of the MCR-ALS resolved components describe the yeast thermal response at the four investigated temperatures (Figure 3A). The metabolic profiles for the yeast cells cultured at 15 °C were predominantly described by MCR-ALS component 1 (C1, blue), whereas MCR-ALS component 3 (C3, red) described the corresponding profiles of the samples grown at 40 °C

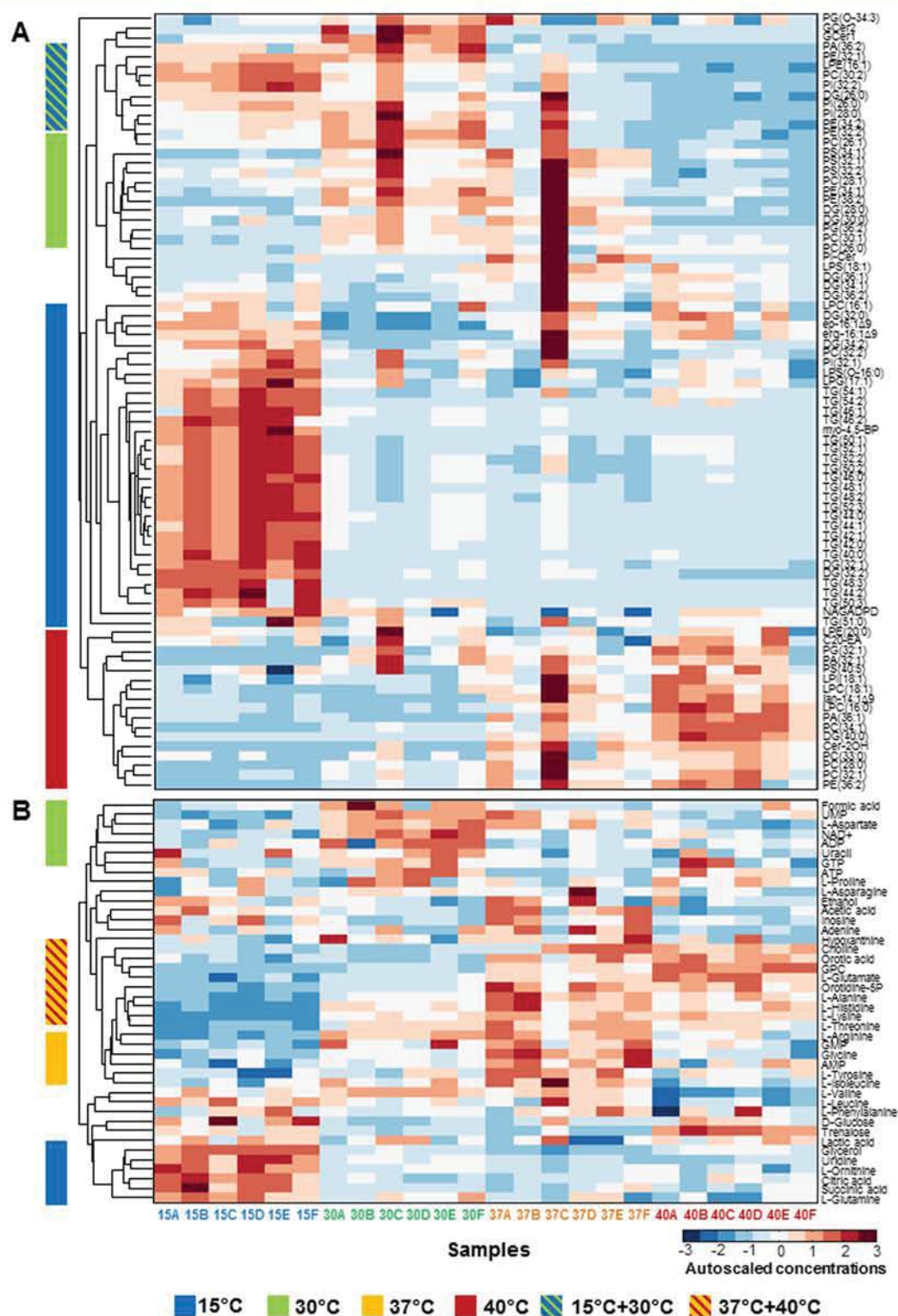


Figure 2. Heatmap of autoscaled NMR and LC–MS peak areas. (A) Heatmap of the detected compounds in the UHPLC–MS data. (B) Heatmap of the detected compounds in the ¹H NMR data. The colored bars denote the common variations among the different metabolites.

(Figure 3A). MCR-ALS component 2 (C2, green), on the contrary, is the major component that explains the metabolic profile of the cells grown at the standard 30 and 37 °C, together with a second MCR-ALS component that represents the 20–40% sample contribution (Figure 3A). This second minor component is C1 for the samples grown at 30 °C, whereas C2 is the minor component for the samples grown at 37 °C. We

can interpret these components as the metabolic profiles that correspond to the cells that grow at low, optimal, and extreme temperatures (C1, C2, and C3, respectively).

In Figure 3 and Table S3 in the Supporting Information, it is shown that a metabolite can be present in more than one metabolic profile at the same time, depending on whether it is involved in more than one biological response. To have better

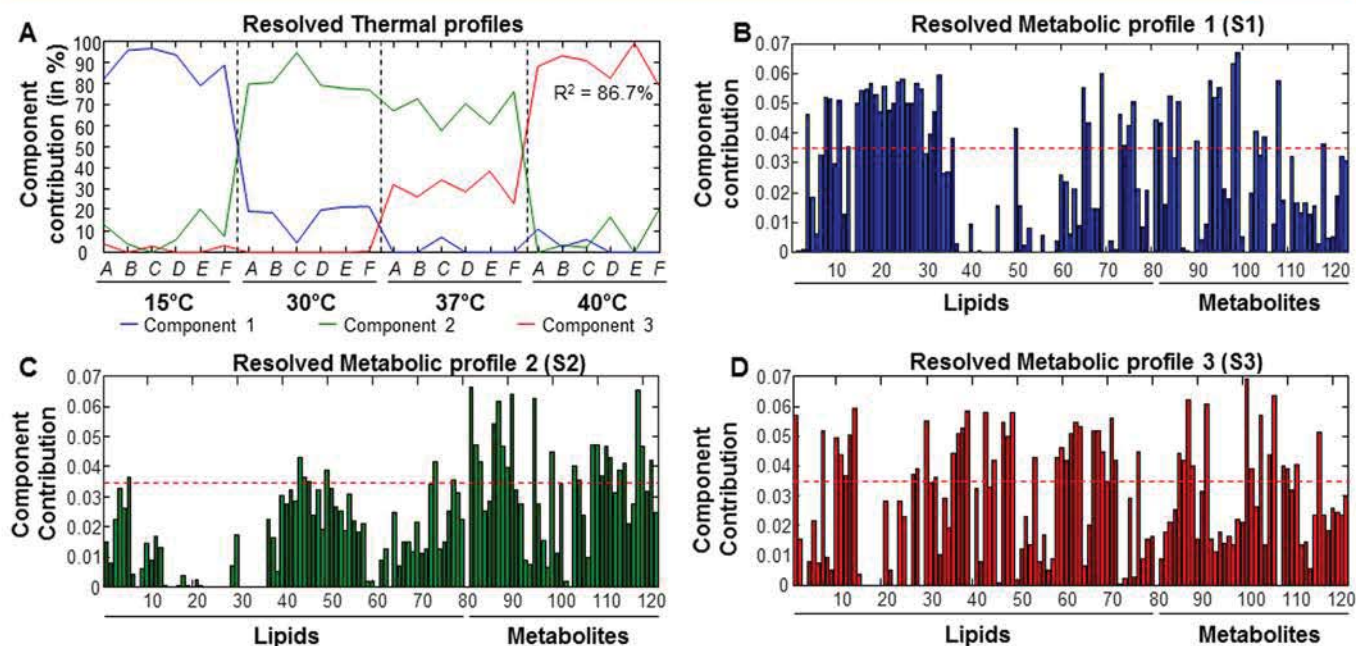


Figure 3. MCR-ALS resolution of the row-wise augmented lipidic-and-metabolic data set. (A) MCR-ALS resolved thermal (C in eq 3) profiles of yeast cells cultured at the four different temperatures. The explained variance (R^2) is also given in the plot. Sextuplicates are represented by letters A–F of the axis levels. (B–D) Resolved lipidomic-and-metabolic (S^T in eq 3) profiles for component 1 (B), component 2 (C), and component 3 (D). See Tables S1 and S2 for the identification of these metabolites and lipids.

insight into the metabolites that respond in synchrony, not only in a given metabolic profile, but in all of them, these metabolic profiles described in the MCR-ALS resolved S^T matrix of Figure 3 were analyzed with a K-medoids clustering (or PAM, partitioning around medoids) method.³¹ When this clustering method is applied, it is found that the data can be divided into three clusters that include both metabolites and lipidic species (Figure 4).

We find Cluster A to be the cluster that includes more metabolites with the highest contribution in S_1 ; therefore, these are the metabolites present at higher concentrations when the cells are grown at 15 °C (Figure 4B, left). Conversely, Cluster C includes those analytes with higher contributions in S_3 , which correspond to those present at higher amounts at cells grown at high temperatures (Figure 4B, right). Finally, Cluster B is composed of analytes that show either high contributions in S_2 or those with similar contributions to the three S^T metabolic profiles (Figure 4B, middle panel). We consider that the metabolites included in Cluster B are those that are related to conditions of optimal growth. Analysis of the analytes associated with each of these three clusters revealed that most nonlipid metabolites fall into Cluster B, whereas acylglycerides (di- and triglycerides) were overrepresented in cluster A, and phospholipids were more abundant in Cluster C than in the others (Figure 4C).

Functional Analysis of Metabolomic Profiles

KEGG³² pathway analysis of the quantified metabolites identified 25 *S. cerevisiae* functional modules that included at least three compounds (Table S4 in the Supporting Information). We added two extra pathways with only two hits to ensure that all of the studied analytes were included in the final data set (“steroid biosynthesis” and “starch and sucrose metabolism”, Table S4 in the Supporting Information). Note that lipids were introduced into the analysis as a multicongener species (in other words, diacylglycerides, triacylglycerides,

phosphatidylcholine, etc.), following the KEGG nomenclature. To better visualize the results from the pathway analysis, we performed a bipartite graph in which two metabolites are clustered if, and only if, they share at least one common functional module (represented as gray circles, Figure 5). Most nonlipidic metabolites were related to amino acids, purine, or pyrimidine metabolism modules, whereas the lipids clustered around the lipid and phospholipid metabolism functional modules (see also Table S4 in the Supporting Information). Figure 5 also includes information about the cluster that each compound was associated with, except for multicongener lipid species. As observed in Figure 4C, most nonlipidic metabolites were associated with Cluster B (green characters in Figure 5). In addition, four metabolites associated with clusters A (blue) or C (red) can be considered as precursors of nucleotides (orotidine, orotic acid, uridine) or amino acids (*L*-ornithine), which further indicates that most of the main components of the metabolome were associated with cell growth (Figure 5). Conversely, glycerol and choline were associated with Clusters A and C, respectively, which is consistent with their role as the core molecules of acylglycerides and phospholipids, respectively.

Analysis of Lipidomic Profiles

Changes in the lipidome across the different temperatures are summarized in Figure 6A. Acylglycerides, phospholipids, and phosphoserines predominate at low temperatures, whereas PI, PA, PE, LPE, LPC, and ceramides peak at 40 °C. Lysophospholipids, PC, and CE have their maximal concentrations at “growth” temperatures, 30–37 °C. On average, acylglycerides accounted for almost one-third of the total lipids in the yeast cells grown at 15 °C, whereas they were minor components of cells grown at higher temperatures. Conversely, cells that grew at 40 °C contained higher proportions of PAs and ceramides, whereas phosphocholines remained as a predominant lipid component at all temperatures (Figure 6B).

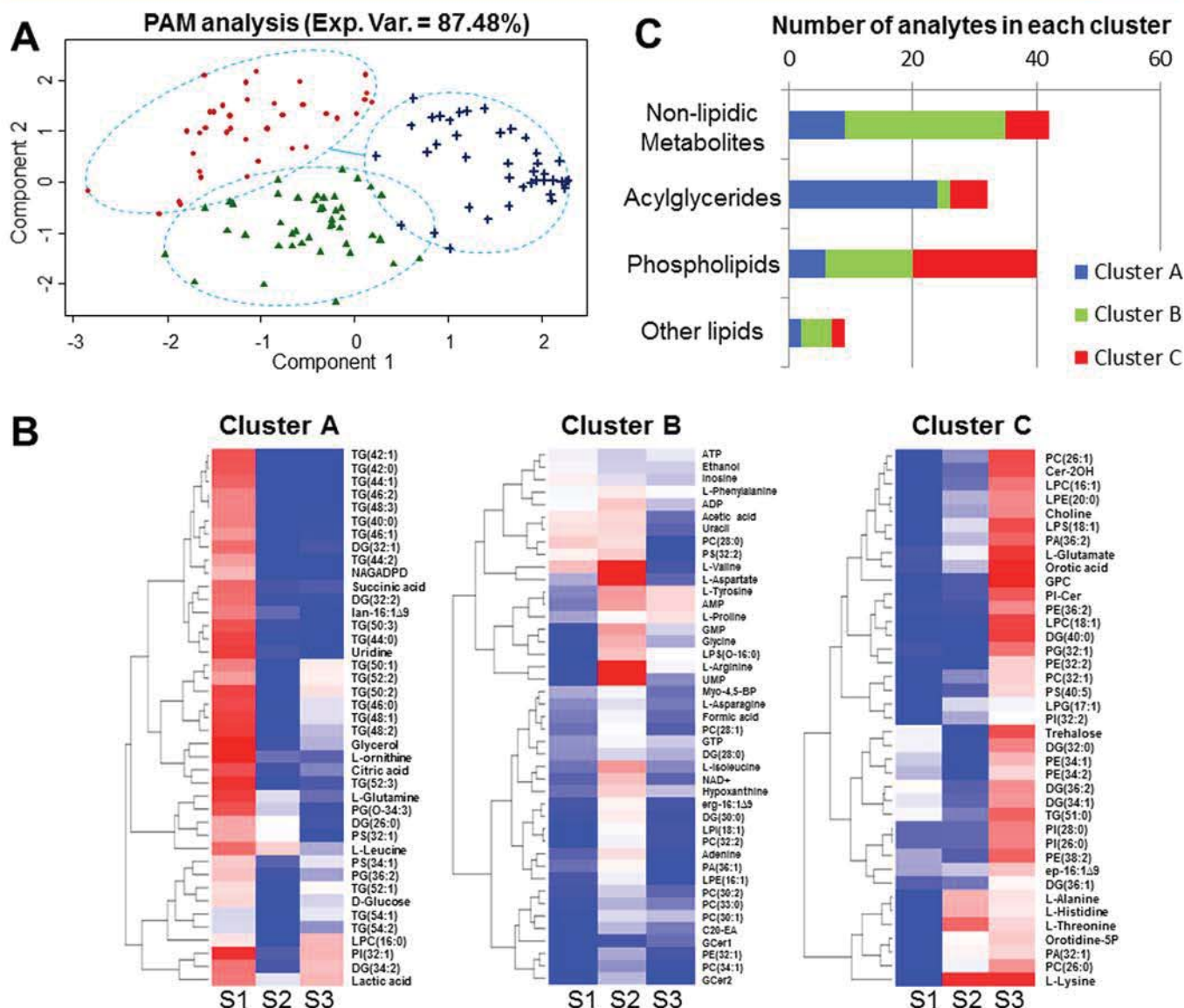


Figure 4. (A) Clustering with K-medoids of the matrix of Metabolic profiles, S. Circles, triangles, and crosses represent each one of the analytes grouped in each cluster. (B) Heatmaps with hierarchical clusterings of the S^T profiles for the analytes in Cluster A (left), Cluster B (middle), and Cluster C (right). (C) Composition of every cluster.

The average FA chain length and unsaturation level for each lipid family have been quantified from the obtained data. Most phospholipid families, such as PC, LPC, PE, and LPE, as well as the acylglycerides show a steadily subtle increase in the average FA chain length with an increase in the temperature, although some other phospholipid groups, such as PG and PA, do not show a clear trend in response to the temperature (Figure 6C). On the contrary, the average unsaturation levels also did change very slightly. The average unsaturation number for PA, PC, PG, and TG steadily decreased in response to acclimation to higher temperatures, whereas the opposite trend was found for the PE and PS lipid species (Figure 6B). These data suggest that the changes in the lipid membranes required to adapt yeast to the different temperatures were mainly brought about by changes in the lipid composition (from acylglycerides, PI, and lyso lipid forms to phospholipids, Figure 6A) and, to a lesser extent, by changing the FA length and number of double bonds.

DISCUSSION

Temperature changes are assumed to affect two main types of biological structures and processes: the enzymatic activity^{33,34} and the fluidity of the cell membranes.³⁵ Temperature changes do not equally affect all of the organisms, and thus every cell type has its own optimal growth temperature. Below this temperature, enzymatic reactions occur very slowly, whereas when this optimal temperature is surpassed, protein and membrane stability are compromised. Under the conditions of this work, the optimal temperature for yeast growth was reached at 30 °C, with a doubling time of 108 min, whereas the doubling times were 112 min at 37 °C, 150 min at 40 °C, and 652 min at 15 °C (not shown). An optimal growth temperature of 30 °C is typical for *S. cerevisiae* laboratory strains.¹⁰

When analyzed by PCA, the four groups of samples were correctly separated. However, when the same samples were analyzed by MCR-ALS, the metabolic variance among the four groups of samples only required the linear combination of three of the MCR-ALS resolved metabolic profiles.

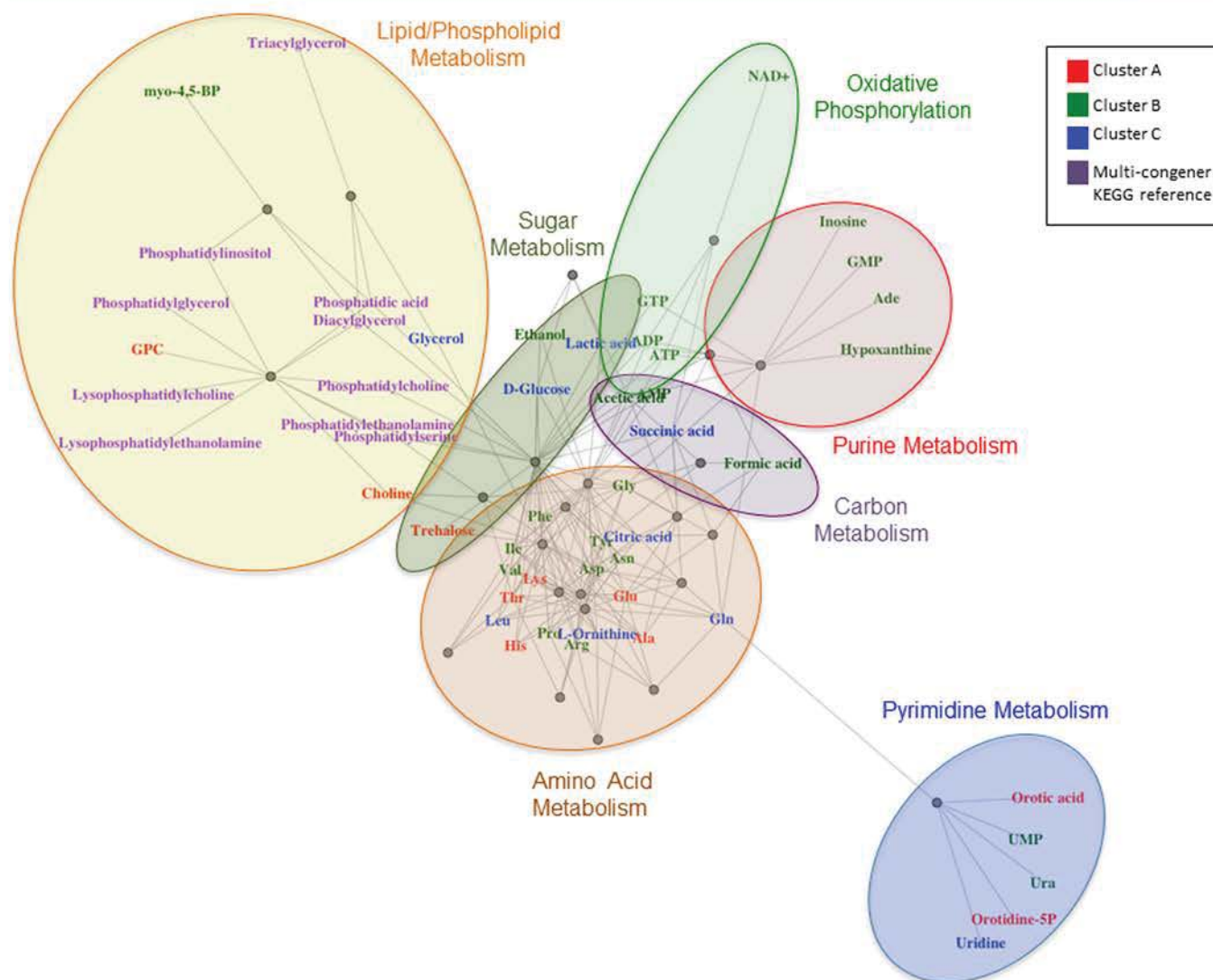


Figure 5. Network analysis showing the mutual correlations between the metabolic response of the identified compounds and their associated metabolic pathways. The compound names are colored in red, green, or blue according to their cluster (A, B, or C, respectively).

Analysis of the MCR-ALS components indicates that component 2 is strongly correlated to yeast growth because samples grown at 30 and 37 °C, which have similar doubling times, present similar C contributions for this given component (Figure 3A). Similarly, we ascribe component 1 to yeast growth at low temperatures and component 3 to yeast growth at high temperatures. In addition, a full metabolomic and lipidomic characterization has been performed for each of the components or patterns.

The observed metabolomic and lipidomic changes can be interpreted as reflecting regulatory mechanisms that act in response to different environmental conditions. Under this point of view, the observed changes suggest the existence of at least two regulatory “modules”, one of which would be activated at a less-than-optimal low temperature (component 1) and a second one that would correspond to a high-temperature growth (component 3). The evidence for these two separate modules comes from both the metabolite and lipid concentrations due to temperature changes.

Yeast cells grown at a low temperature (component 1) have lower amounts of total lipids and higher amounts of PI and TG than yeast cells grown at optimal temperatures. These results

are consistent with previously reported observations about yeast lipid content in fermentations performed at different temperatures.^{36,37} Henderson et al.³⁶ stated that under fermentative conditions the lipid composition varies in a temperature-dependent manner, but the observed changes did not significantly affect the cell membrane fluidity. In our study, we observed a similar response under nonfermentative conditions, and this response is additionally connected to the observed variations in the central metabolism. The detected variations that adapt the membrane fluidity to temperature changes are mainly in the acylglycerides. TGs, and, to a lesser extent, DGs present shorter FA chains with a higher average number of double bonds. The shortening of the fatty acyl chains decreases the van der Waals interactions among them, thus lowering the lipid viscosity and increasing their fluidity.³⁸ Therefore, shorter, unsaturated FAs produce more fluid lipids than long, saturated ones.

It is known that higher levels of unsaturated fatty acids are typical for psychrophilic yeasts, which proves that this is a typical adaptation/acclimation response to cold environments.^{39,40} Our results showed that this finding was true for *S. cerevisiae* up to a certain extent because we observed a similar

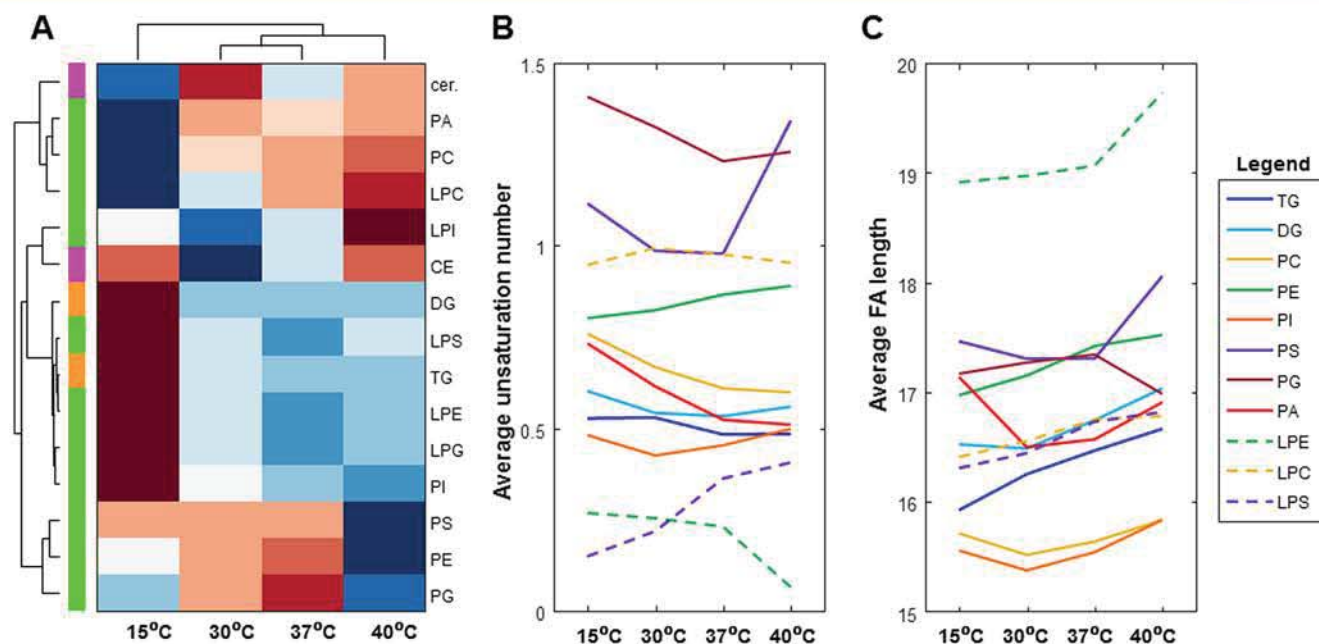


Figure 6. (A) Relative changes, in autoscaled units, in the lipid content for the different lipid families. Red indicates a higher abundance, whereas blue indicates a lower content. The side colors indicate the lipid superfamilies: green for phospholipids, orange for acylglycerides, and magenta for other lipids. (B,C) Plots of the (B) average unsaturation number and (C) the average FA chain length for the different phospholipids and acylglycerides at each studied temperature.

response for acylglycerides, even though it does not represent a significant amount of the total lipid fraction (6–15%).⁴¹ On the contrary, the small variations detected for the different phospholipid species can be explained by the fact that the yeast lipidome is not very diverse because most phospholipids in *S. cerevisiae* contain either 32 or 34 carbons with 1 or 2 unsaturations per molecule.⁴¹ Because variations in the phospholipidome are comprised within this narrow range of compounds, lipidic adaptation to changes in temperature are instead addressed by modifying the proportions of the different lipid families.

Glycerol and PI species are also characteristic of a low-temperature component (Component 1). Glycerol is protective against low temperatures and osmotic changes,^{42,43} and it is the core component of acylglycerides, whereas for PI species, it is known that phosphatidylinositol-4,5-bisphosphate (PI-4,5-BP) plays a role in cold tolerance in yeast,⁴⁴ and transcription of heat-shock genes is mediated by signal transduction of inositol-3P and PI-4,5-BP.⁴⁵ Moreover, succinic acid and citric acid have significant contributions for the C1 component (Table S3 in the Supporting Information), which can be explained as a switch to the respiratory metabolism at cold temperatures to maintain the cell viability, instead of promoting growth.

Yeast grown at high temperatures (component 3) accumulated charged amino acids and uracil precursors (Table S3 in the Supporting Information), which suggests that at higher temperatures an activation of the translation machinery occurs. In fact, one of the main responses to heat stress is the translation of genes involved in respiration and alternative carbon utilization and the induction of protein folding chaperones.⁴⁶ PI-ceramide and cholesterol esters were revealed to be significant for this component. Ceramides are known to be involved in apoptotic cell death signaling, and high levels of ceramide under cold and heat stress could reduce the cell death, hence promoting cell survival under extreme conditions.^{47,48} Alteration of sterol composition in yeast has been shown to

produce changes in thermotolerance.⁴⁹ Another response observed in these yeast cells is the accumulation of several phosphocholines (detected with UHPLC–MS) as well as glycerophosphocholine and choline (both detected with ^1H NMR). This trend reveals not only that the phosphocholine metabolism is altered due to the increase in the temperature but also that the combined use of the two different analytical platforms (NMR and LC–MS) allowed a better characterization of the overall changes in the yeast metabolism at different temperatures. While our experimental design focused on the acclimation process, rather than on a proper heat-shock (or cold-shock) response, our results are consistent with at least some characteristics of the metabolomic responses to heat-shock, which include low translation, replication, and glycolytic rates, and high levels of trehalose.⁵⁰

Finally, there are also some other chemical compounds whose concentrations were maximal under optimal growth conditions (component 2). These compounds are mostly amino acids and nucleotides, which suggests that their corresponding pathways are enhanced, promoting cell growth and viability.

CONCLUSIONS

The results described in this work have shown that the proposed application of MCR-ALS to biological metabolomics and lipidomics data sets and the interpretation of the results achieved by this method can be a convenient approach of using data analysis and integration to untangle the hidden metabolic responses of the studied organisms.

Specifically, we observed that rather than a binary metabolic switch, the thermal stress response is a fluid event. This thermal response can be defined by three components or metabolic profiles (representative of cold, optimal, and high temperature), and the contribution of these three profiles varies as a function of the temperature. From the comparison of these three metabolic profiles, we determined that yeast adjusts membrane

fluidity by changing the relative proportions of the different lipid families rather than modifying the average length and unsaturation levels of the corresponding fatty acids.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.7b00921.

Extended methods. Table S1. Metabolite assignment for the ¹H NMR data. Table S2. Lipid assignment from the UHPLC–MS data. Table S3. Classification of the relevant metabolites (including lipids) resolved by MCR-ALS analysis of the peak areas of the metabolite and lipid concentrations obtained, respectively, by the NMR and LC–MS analysis of yeast samples at different temperatures. Table S4. KEGG pathway analysis. (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Tel: +34-934006140. Fax: +34 932045904. E-mail: Roma.Tauler@idaea.csic.es.

ORCID

Francesc Puig-Castellví: 0000-0003-1064-9586

Ignacio Alfonso: 0000-0003-0678-0362

Romà Tauler: 0000-0001-8559-9670

Author Contributions

This manuscript was written through the contributions of all of the authors. All of the authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The research that produced these results was supported by funding received from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement no. 320737 and the Spanish Ministry of Economy and Competitiveness (CTQ2015-66254-C2-1-P).

■ ABBREVIATIONS

¹H NMR, 1D proton nuclear magnetic resonance; UHPLC–MS, ultra-high-performance liquid chromatography–mass spectrometry; TIC, total ion chromatogram; MCR-ALS, multivariate curve resolution-alternating least-squares; PCA, principal component analysis; ESI, electrospray ionization; DG, diacylglyceride; TG, triacylglyceride; PA, phosphatidic acid; PC, phosphatidylcholine; PG, phosphatidylglycerol; PI, phosphatidylinositol; PS, phosphatidylserine; PI-Cer, inositol phosphate ceramide; CE, cholesterol ester; myo-4,5-BP, 1D-myoinositol-4,5-biphosphate; NAGADPD, N-acetyl-D-glucosaminylidiphosphodolichol; GCer, glucoceramide; lan, lanosteryl ester; erg, ergosteryl ester; ep, episteryl ester; LPS, LPG, LPC, LPE, lyso forms of the PS, PG, PC, and PE phospholipids, respectively; ROIMCR, regions-of-interest MCR-ALS; C, concentration matrix; S^T, spectra matrix; PAM, partitioning around medoids; KEGG, Kyoto Encyclopedia of Genes and Genomes; FA, fatty acids; PI-4,5-BP, phosphatidylinositol-4,5-biphosphate; A₆₀₀, absorbance at 600 nm

■ REFERENCES

- (1) Mattanovich, D.; Sauer, M.; Gasser, B. Yeast biotechnology: teaching the old dog new tricks. *Microb. Cell Fact.* **2014**, *13* (1), 34.
- (2) Goffeau, A.; Barrell, B. G.; Bussey, H.; Davis, R. W.; Dujon, B.; Feldmann, H.; Galibert, F.; Hoheisel, J. D.; Jacq, C.; Johnston, M.; Louis, E. J.; Mewes, H. W.; Murakami, Y.; Philippsen, P.; Tettelin, H.; Oliver, S. G. Life with 6000 Genes. *Science* **1996**, *274* (5287), 546–567.
- (3) Rozpędowska, E.; Hellborg, L.; Ishchuk, O. P.; Orhan, F.; Galafassi, S.; Merico, A.; Woolfit, M.; Compagno, C.; Piskur, J. Parallel evolution of the make-accumulate-consume strategy in *Saccharomyces* and *Dekkera* yeasts. *Nat. Commun.* **2011**, *2*, 302.
- (4) Puig-Castellví, F.; Alfonso, I.; Piña, B.; Tauler, R. 1H NMR metabolomic study of auxotrophic starvation in yeast using Multivariate Curve Resolution-Alternating Least Squares for Pathway Analysis. *Sci. Rep.* **2016**, *6*, 30982.
- (5) Xu, H.; Kim, S.; Sorek, H.; Lee, Y.; Jeong, D.; Kim, J.; Oh, E. J.; Yun, E. J.; Wemmer, D. E.; Kim, K. H.; Kim, S. R.; Jin, Y. S. PHO13 deletion-induced transcriptional activation prevents sedoheptulose accumulation during xylose metabolism in engineered *Saccharomyces cerevisiae*. *Metab. Eng.* **2016**, *34*, 88–96.
- (6) Farres, M.; Pina, B.; Tauler, R. LC-MS based metabolomics and chemometrics study of the toxic effects of copper on *Saccharomyces cerevisiae*. *Metallomics* **2016**, *8* (8), 790–798.
- (7) Chen, Z.; Zheng, Z.; Yi, C.; Wang, F.; Niu, Y.; Li, H. Intracellular metabolic changes in *Saccharomyces cerevisiae* and promotion of ethanol tolerance during the bioethanol fermentation process. *RSC Adv.* **2016**, *6* (107), 105046–105055.
- (8) Ortiz-Villanueva, E.; Jaumot, J.; Benavente, F.; Piña, B.; Sanz-Nebot, V.; Tauler, R. Combination of CE-MS and advanced chemometric methods for high-throughput metabolic profiling. *Electrophoresis* **2015**, *36* (18), 2324–2335.
- (9) López-Malo, M.; Querol, A.; Guillamon, J. M. Metabolomic Comparison of *Saccharomyces cerevisiae* and the Cryotolerant Species *S. bayanus* var. *uvarum* and *S. kudriavzevii* during Wine Fermentation at Low Temperature. *PLoS One* **2013**, *8* (3), e60135.
- (10) Barnett, J. A.; Payne, R. W.; Yarrow, D. *Yeasts: Characteristics and Identification*, 3rd ed.; Cambridge University Press: Cambridge, U.K., 2000; p 1150.
- (11) Banat, I. M.; Nigam, P.; Singh, D.; Marchant, R.; McHale, A. P. Review: Ethanol production at elevated temperatures and alcohol concentrations: Part I – Yeasts in general. *World J. Microbiol. Biotechnol.* **1998**, *14* (6), 809–821.
- (12) Torija, M. a. J.; Beltran, G.; Novo, M.; Poblet, M.; Guillamón, J. M.; Mas, A.; Rozès, N. Effects of fermentation temperature and *Saccharomyces* species on the cell fatty acid composition and presence of volatile compounds in wine. *Int. J. Food Microbiol.* **2003**, *85* (1–2), 127–136.
- (13) Arthur, H.; Watson, K. Thermal adaptation in yeast: growth temperatures, membrane lipid, and cytochrome composition of psychrophilic, mesophilic, and thermophilic yeasts. *J. Bacteriol.* **1976**, *128*, 56–68.
- (14) Puig-Castellví, F.; Alfonso, I.; Piña, B.; Tauler, R. A quantitative 1H NMR approach for evaluating the metabolic response of *Saccharomyces cerevisiae* to mild heat stress. *Metabolomics* **2015**, *11* (6), 1612–1625.
- (15) Stoessel, D.; Nowell, C. J.; Jones, A. J.; Ferrins, L.; Ellis, K. M.; Riley, J.; Rahmani, R.; Read, K. D.; McConville, M. J.; Avery, V. M.; Baell, J. B.; Creek, D. J. Metabolomics and lipidomics reveal perturbation of sphingolipid metabolism by a novel anti-trypanosomal 3-(oxazololo[4,5-b]pyridine-2-yl)anilide. *Metabolomics* **2016**, *12* (7), 1–14.
- (16) Tauler, R.; Kowalski, B.; Fleming, S. Multivariate curve resolution applied to spectral data from multiple runs of an industrial process. *Anal. Chem.* **1993**, *65* (15), 2040–2047.
- (17) Peré-Trepát, E.; Tauler, R. Analysis of environmental samples by application of multivariate curve resolution on fused high-performance liquid chromatography–diode array detection mass spectrometry data. *Journal of Chromatography A* **2006**, *1131* (1–2), 85–96.

- (18) Fernández, C.; Pilar Callao, M.; Soledad Larrechi, M. UV-visible-DAD and ¹H-NMR spectroscopy data fusion for studying the photodegradation process of azo-dyes using MCR-ALS. *Talanta* **2013**, *117* (0), 75–80.
- (19) Karakach, T. K.; Knight, R.; Lenz, E. M.; Viant, M. R.; Walter, J. A. Analysis of time course ¹H NMR metabolomics data by multivariate curve resolution. *Magn. Reson. Chem.* **2009**, *47* (S1), S105–S117.
- (20) Dalmáu, N.; Jaumot, J.; Tauler, R.; Bedia, C. Epithelial-to-mesenchymal transition involves triacylglycerol accumulation in DU145 prostate cancer cells. *Mol. BioSyst.* **2015**, *11* (12), 3397–3406.
- (21) Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H. Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in ¹H NMR Metabolomics. *Anal. Chem.* **2006**, *78* (13), 4281–4290.
- (22) Gorrochategui, E.; Jaumot, J.; Tauler, R. A protocol for LC-MS metabolomic data processing using chemometric tools. *Protoc. Exch.* **2015**, DOI: 10.1038/protex.2015.102.
- (23) Gorrochategui, E.; Jaumot, J.; Lacorte, S.; Tauler, R. Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow. *TrAC, Trends Anal. Chem.* **2016**, *82*, 425–442.
- (24) Marques, A. S.; Bedia, C.; Lima, K. M. G.; Tauler, R. Assessment of the effects of As(III) treatment on cyanobacteria lipidomic profiles by LC-MS and MCR-ALS. *Anal. Bioanal. Chem.* **2016**, *408*, S829–S841.
- (25) Jewison, T.; Knox, C.; Neveu, V.; Djoumbou, Y.; Guo, A. C.; Lee, J.; Liu, P.; Mandal, R.; Krishnamurthy, R.; Sinelnikov, I.; Wilson, M.; Wishart, D. S. YMDB: the Yeast Metabolome Database. *Nucleic Acids Res.* **2012**, *40* (D1), D815–D820.
- (26) Hao, J.; Astle, W.; De Iorio, M.; Ebbels, T. M. D. BATMAN—an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics* **2012**, *28* (15), 2088–2090.
- (27) Hao, J.; Liebeke, M.; Astle, W.; De Iorio, M.; Bundy, J. G.; Ebbels, T. M. D. Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN. *Nat. Protoc.* **2014**, *9* (6), 1416–1427.
- (28) The LIPID MAPS Lipidomics Gateway. <http://www.lipidmaps.org/>.
- (29) Jaumot, J.; de Juan, A.; Tauler, R. MCR-ALS GUI 2.0: New features and applications. *Chemom. Intell. Lab. Syst.* **2015**, *140*, 1–12.
- (30) Kanehisa, M.; Goto, S.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **2014**, *42* (D1), D199–D205.
- (31) Kaufman, L.; Rousseeuw, P. J. Partitioning Around Medoids (Program PAM). In *Finding Groups in Data*; John Wiley & Sons, Inc.: 2008; pp 68–125.
- (32) Kanehisa, M.; Goto, S.; Sato, Y.; Furumichi, M.; Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **2012**, *40* (D1), D109–D114.
- (33) Razavi, B. S.; Liu, S.; Kuzyakov, Y. Hot experience for cold-adapted microorganisms: Temperature sensitivity of soil enzymes. *Soil Biol. Biochem.* **2017**, *105*, 236–243.
- (34) Singh, B.; Poças-Fonseca, M. J.; Johri, B. N.; Satyanarayana, T. Thermophilic molds: Biology and applications. *Crit. Rev. Microbiol.* **2016**, *42* (6), 985–1006.
- (35) Hassan, N.; Rafiq, M.; Hayat, M.; Shah, A. A.; Hasan, F. Psychrophilic and psychrotrophic fungi: a comprehensive review. *Rev. Environ. Sci. Bio/Technol.* **2016**, *15* (2), 147–172.
- (36) Henderson, C. M.; Zeno, W. F.; Lerno, L. A.; Longo, M. L.; Block, D. E. Fermentation Temperature Modulates Phosphatidylethanolamine and Phosphatidylinositol Levels in the Cell Membrane of *Saccharomyces cerevisiae*. *Appl. Environ. Microbiol.* **2013**, *79* (17), 5345–5356.
- (37) Redón, M.; Guillamón, J. M.; Mas, A.; Rozès, N. Effect of growth temperature on yeast lipid composition and alcoholic fermentation at low temperature. *Eur. Food Res. Technol.* **2011**, *232* (3), 517–527.
- (38) Russell, N. J. The Regulation of Membrane Fluidity in Bacteria by Acyl Chain Length Changes. In *Membrane Fluidity*; Kates, M., Manson, L. A., Eds.; Springer US: Boston, MA, 1984; pp 329–347.
- (39) Rossi, M.; Buzzini, P.; Cordisco, L.; Amaretti, A.; Sala, M.; Raimondi, S.; Ponzoni, C.; Pagnoni, U. M.; Matteuzzi, D. Growth, lipid accumulation, and fatty acid composition in obligate psychrophilic, facultative psychrophilic, and mesophilic yeasts. *FEMS Microbiol. Ecol.* **2009**, *69* (3), 363–372.
- (40) Gunde-Cimerman, N.; Plemenitaš, A.; Buzzini, P. Changes in Lipids Composition and Fluidity of Yeast Plasma Membrane as Response to Cold. In *Cold-Adapted Yeasts: Biodiversity, Adaptation Strategies and Biotechnological Significance*; Buzzini, P., Margesin, R., Eds.; Springer Berlin Heidelberg: Berlin, 2014; pp 225–242.
- (41) Klose, C.; Surma, M. A.; Gerl, M. J.; Meyenhofer, F.; Shevchenko, A.; Simons, K. Flexibility of a Eukaryotic Lipidome – Insights from Yeast Lipidomics. *PLoS One* **2012**, *7* (4), e35063.
- (42) Aguilera, J.; Rande-Gil, F.; Prieto, J. A. Cold response in *Saccharomyces cerevisiae*: new functions for old mechanisms. *FEMS Microbiology Reviews* **2007**, *31* (3), 327–341.
- (43) Hohmann, S. Osmotic Stress Signaling and Osmoadaptation in Yeasts. *Microbiology and Molecular Biology Reviews* **2002**, *66* (2), 300–372.
- (44) Córcoles-Sáez, I.; Hernández, M. L.; Martínez-Rivas, J. M.; Prieto, J. A.; Rande-Gil, F. Characterization of the *S. cerevisiae* inp51 mutant links phosphatidylinositol 4,5-bisphosphate levels with lipid content, membrane fluidity and cold growth. *Biochim. Biophys. Acta, Mol. Cell Biol. Lipids* **2016**, *1861* (3), 213–226.
- (45) Balogh, G.; Péter, M.; Glatz, A.; Gombos, I.; Török, Z.; Horváth, I.; Harwood, J. L.; Vigh, L. Key role of lipids in heat stress management. *FEBS Lett.* **2013**, *587* (13), 1970–1980.
- (46) Gasch, A. P.; Spellman, P. T.; Kao, C. M.; Carmel-Harel, O.; Eisen, M. B.; Storz, G.; Botstein, D.; Brown, P. O. Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. *Mol. Biol. Cell* **2000**, *11* (12), 4241–4257.
- (47) Yang, Z.; Khoury, C.; Jean-Baptiste, G.; Greenwood, M. T. Identification of mouse sphingomyelin synthase 1 as a suppressor of Bax-mediated cell death in yeast. *FEMS Yeast Res.* **2006**, *6* (5), 751–762.
- (48) Huang, X.; Liu, J.; Dickson, R. C. Down-Regulating Sphingolipid Synthesis Increases Yeast Lifespan. *PLoS Genet.* **2012**, *8* (2), e1002493.
- (49) Caspeta, L.; Chen, Y.; Ghiaci, P.; Feizi, A.; Buskov, S.; Hallström, B. M.; Petranovic, D.; Nielsen, J. Altered sterol composition renders yeast thermotolerant. *Science* **2014**, *346* (6205), 75–78.
- (50) Caspeta, L.; Nielsen, J. Thermotolerant Yeast Strains Adapted by Laboratory Evolution Show Trade-Off at Ancestral Temperatures and Preadaptation to Other Stresses. *mBio* **2015**, *6* (4), e00431-15.

SUPPLEMENTARY MATERIAL FOR SCIENTIFIC ARTICLE II

Deciphering the underlying metabolomic and lipidomic patterns linked to thermal acclimation in *Saccharomyces cerevisiae*.

Authors: Puig-Castellví F., Bedia, C., Alfonso I., Piña B., Tauler R.

Citation reference: Journal of Proteome Research (2018).

DOI: 10.1021/acs.jproteome.7b00921

Extended methods

Yeast Growth

S. cerevisiae BY4741 (MATa; *his3Δ1*; *leu2Δ0*; *met15Δ0*; *ura3Δ0*) cells were precultured in YPD medium on an orbital shaker (at 30°C and 150 rpm) overnight. Fresh YPD medium was inoculated with the preculture to an absorbance at 600 nm (A_{600}) of 0.1 and divided into volumes of 50 ml, which were grown at either 15°C, 30°C, 37°C or 40°C (six aliquots each, shaking at 150 rpm). After reaching an A_{600} of 0.6-0.8, cultures were arrested on ice. Every culture was aliquoted into two fractions of 5 ml and 45 ml, which were used for the lipidic and metabolic extraction, respectively. Cell harvesting was performed by centrifugation of every (5 ml and 45 ml) fraction at 4,000 g for 5 min and discarding the supernatant. The cells were washed afterward with 100 mM sodium phosphate buffer (pH 7.0). The resulting pellets were stored at -80°C and lyophilized.

Metabolite and lipid extraction and sample preparation

The metabolites were extracted by following the protocol published in our previous work (1). 1800 μ l of a solution of methanol-chloroform 1:2 (4°C) were added to the pellet from the 45-ml samples, followed by a vigorous vortexing. A cold shock using liquid nitrogen and ice was then applied to the pellets for 5 times. 400 μ l of water were added to create the biphasic system. After homogenization by vortexing, a 3 min centrifugation at 25,000 g and 4°C was carried out. The aqueous phase (upper part) was collected. This process was repeated once. The aqueous phases were afterwards combined and a N₂ stream was applied to the samples in order to evaporate chloroform traces. Afterwards, the samples were lyophilized and stored at -80°C upon sample preparation. Metabolic yeast extracts were finally dissolved in 700 μ l of deuterated phosphate buffer (Na₂DPO₄ 100 mM, pH 7.0) in D₂O with DSS 0.2 mM as internal standard.

Lipids were extracted by using a slight modification of a previously published protocol, which was designed for human cell lines(2). The pellets that corresponded to the 5-ml samples were resuspended in 100 μ l of deionized water and the suspension was transferred to borosilicate glass test tubes with Teflon caps. Then, 250 μ l of methanol and 500 μ l of chloroform were subsequently added. This mixture was fortified with internal standards of lipids (1,2,3-17:0 triglyceride (TG),

1,3-17:0 (d5) diglyceride (DG), 17:0 cholesteryl ester, 16:0 D31-18:1 phosphatidylethanolamine (PE), 16:0 D31-18:1 phosphatidylserine (PS), 16:0 D31-18:1 phosphatidylglycerol (PG), 16:0 D31-18:1 phosphatidylcholine (PC), 17:1 lyso PC (LPC), 17:1 lyso PE (LPE), 17:1 lyso PG (LPG), and 17:1 lyso PS (LPS, 200 pmol each). The samples were vortexed and sonicated until they appeared dispersed. Next, 100 μ l of glass beads were added into the samples and they were vigorously sonicated for one minute twice, followed by a second sonication. The samples were incubated overnight at 48°C and cooled down at 37°C afterwards in a heating water bath. Then, the samples were then evaporated under a N₂ stream and transferred to 1.5 ml eppendorf tubes after addition of 500 μ l of methanol. Next, the eppendorf tubes were centrifuged at 9,168 g for 3 min and 130 μ l of the supernatants were transferred to UPLC vials for injection. This mixture was fortified with internal standards of sphingolipids (N-dodecanoylsphingosine, N-dodecanoylglucosyl-sphingosine and N-dodecanoylsphingosylphosphorylcholine, 200 pmol each).

NMR measurements

Spectra were recorded in a 400 MHz Varian spectrometer, using a spectrometer frequency of 400.14 MHz with a OneNMR Probe and a ProTune System (Agilent). The proton spectral size range covered from -2 to 12 ppm, which consisted of 65k data points. The number of scans was 512 and the relaxation delay was 5 s.

UHPLC-MS measurements

LC-MS analysis consisted of a Waters Acquity UHPLC system connected to a Waters LCT Premier orthogonal accelerated time of flight mass spectrometer (Waters), operated in both positive and negative electrospray ionization modes (ESI+ and ESI-, respectively). Full scan spectra from 50 to 1,500 Da were acquired, and individual spectra were summed to produce data points each of 0.2 s. The mass accuracy and reproducibility were maintained by using an independent reference spray via the LockSpray interference. The analytical column was a 100 x 2.1 mm inner diameter, 1.7 mm C8 Acquity UPLC bridged ethylene hybrid (Waters). The two

mobile phases were MeOH 1 mM ammonium formate (phase A) and H₂O 2 mM ammonium formate (phase B). The flow rate was 0.3 ml min⁻¹ and the gradient of A/B solvents started at 80:20 and changed to 90:2 in minute 3; from minute 3 to 6 remained at 90:10; changed to 99:1 in minute 6 until minute 15; remained 99:1 until minute 18; and finally, returned to the initial conditions until minute 20. The column was held at 30°C.

NMR Data preprocessing

The NMR Spectra were preprocessed with MestreNova v.9.0 (Mestrelab Research, Spain). The NMR Spectra preprocessing consisted of an exponential apodization of 0.5 Hz, a manual phasing and a baseline correction with Bernstein polynomial of 3rd order. After adjusting the reference to DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid), regions of water (4.66 - 5.16 ppm), methanol (3.30 - 3.37 ppm) and chloroform (7.63 - 7.70 ppm) were removed, as well as the data points with chemical shifts higher than 9.40 ppm or lower than 0.75 ppm. The final NMR dataset consisted on a data matrix of 24 spectra (rows) having 37,476 ppm values (columns) each one. This data matrix was stored in an ASCII file format.

Finally, the ¹H NMR spectra were normalized using Probabilistic Quotient Normalization (PQN) (3,4) to correct for possible sample size effects. In this case, the reference spectrum used was the median spectrum obtained from the ¹H NMR spectra of yeast samples cultured at standard conditions. The quotient values were calculated using the median quotient values obtained from the intensity values of the region of 0.8 – 3.8 ppm of every ¹H NMR spectrum divided by the values of the same region of the reference spectrum.

UHPLC-MS data preprocessing

Every UHPLC-MS data file was converted to CDF format by the Databridge program of the MassLynx™ software (Waters Inc.). The resulting data were imported into MATLAB environment using the *mzcdfread* and *mzcdf2peaks* commands from the Bioinformatics Toolbox in Matlab R2014b (The Mathworks Inc. Natick, MA, USA). The data dimensions were compressed without loss of spectral resolution by using the recently developed Regions of Interest (ROI) strategy (5). In this strategy, m/z traces with lower intensity signals than a defined threshold

value (*i.e.*, noise) and not measured in consecutive scans (*i.e.*, spurious measurements) were excluded from further analysis. This procedure allows for more than 100-fold computer storage reduction maintains the highest possible experimental mass accuracy (6). Every UHPLC-MS sample provided a data matrix with as many rows as the number of measured retention times and as many columns as the number of selected m/z ROI values. The generated ROI-MS matrices were normalized using the absorbance at 600 nm (A_{600}) of the precedent yeast culture and the peak area of the internal standards.

Finally, augmented ROI data matrices were built up by column-wise augmentation of all the individual ROI data matrices obtained for the set of studied samples for each ESI mode separately. More details about this strategy can be consulted in Gorrochategui *et al.* (5).

In addition, one Total Ion Chromatogram (TIC) vector was generated for every UHPLC-MS sample and ionization mode (one TIC per ESI mode), *i.e.*, by summing all intensity values for all measured m/z in every given scan. These TICs were normalized with the PQN method previously mentioned, using as a reference TIC the median of the TICs from yeast samples cultured at standard conditions (30 °C).

NMR data analysis

Metabolite assignment was performed by a detailed targeted metabolite profiling analysis of the ^1H NMR signals using a homemade ^1H NMR spectra library and also the Yeast Metabolome Data Base library (YMDB) (7). A relative metabolite quantification of the ^1H NMR spectral matrix was performed using the BATMAN R-package (8). Further information about how Batman works and the exact protocol can be found elsewhere (1,9).

UHPLC-MS data analysis

MCR-ALS was subsequently applied to the augmented ROI data matrices generated for the two ESI modes to resolve the elution profiles and mass spectra associated with every lipid. See the **Chemometric analysis** section below for more information about MCR-ALS.

Tentative assignment of lipids was performed from the *m/z* values associated with the smallest delta values (difference between query mass and adduct mass) using the Yeast Metabolome Database (7) and Lipid Maps (10) online databases, and also from comparing the associated retention times to those from lipids found in previous studies. Glycerophospholipids, diacylglycerol, triacylglycerol and cholesterylestes were annotated as <lipid subclass> <total fatty acyl chain length>:<total number of unsaturated bonds>.

The average number of unsaturations (*nI*) and the average length (*L*) of the fatty acyl chains was calculated using the following equations:

$$nI = \frac{\sum_k^K n_k c_k i_k}{\sum_k^K n_k c_k} \quad (\text{Eq. 1})$$

$$L = \frac{\sum_k^K n_k c_k c_{lk}}{\sum_k^K n_k c_k} \quad (\text{Eq. 2})$$

For every lipid (*k*) that contains at least one fatty acyl chain, *n_k* is the number of fatty acyl chains, *c_k* is the relative concentration for every considered lipid, *i_k* is its number of unsaturations, and finally *c_{lk}* is the average carbon length of their fatty acyl chains.

Chemometric data analysis

a) Preliminary PCA analysis

PCA was performed on the raw experimental data, either from the experimental raw ¹H NMR spectra or from the TIC chromatograms obtained with the distinct ionization modes. In all three cases (¹H NMR, UHPLC-MS-ESI(+) and UHPLC-MS-ESI(-)), data was mean-centered prior analysis.

b) MCR-ALS analysis

MCR-ALS decomposes a data matrix using the following bilinear model:

$$\mathbf{X} = \mathbf{CS}^T + \mathbf{E} \quad (\text{Eq. 3})$$

where **C** and **S^T** are the pure concentrations and spectral profiles of the resolved components in **X**, and **E** is the residual matrix that contains the data variation not explained by **CS^T**.

MCR-ALS (11,12) has been applied here to the analysis of two different types of matrices: i) the column-wise augmented matrices of UHPLC-MS-ROIs (ROIMCR method), ii) NMR-derived metabolite and UHPLC-MS-derived lipid area matrices.

i) ROIMCR method: MCR-ALS analysis of UHPLC-MS-ROI data matrices

In this case, MCR-ALS analysis is applied on the column-wise augmented MSROI data matrices obtained in the analysis of the 24 yeast samples with positive and negative electrospray ionization modes (13,14). The two augmented MSROI data matrices have a total number of 15,264 rows (24 samples x 636 retention times). 688 ROIs were detected in ESI+, and 119 ROIs were detected in ESI-. A no-negativity constraint was used during the iterative process on both \mathbf{C} and \mathbf{S}^T matrices. Moreover, \mathbf{S}^T matrix was normalized to total height.

In this bilinear model, it is assumed that the measured raw UHPLC-MS data can be described as the sum of the concentration weighted MS pure spectra of the lipids present in the analyzed samples. Thus, when applied to UHPLC-MS-ROI data matrices (in the ROI-MCR method (14), \mathbf{S}^T (N, J) has the MCR-ALS resolved pure MS-ROI spectra of each lipid, whereas \mathbf{C} (I, N) contains the elution profiles of each one of them. Moreover, apart from the resolving chromatographic elution profiles of the lipids present in the analyzed samples, other components describing contributions from background, solvent and instrumental are also resolved in additional components.

ii) MCR-ALS analysis of metabolite and lipid peak area data matrices.

Additionally, in a different data analysis stage, MCR-ALS was also applied to the set of abundances (4) (signal/peak areas) derived from the previous metabolomic (from ^1H NMR) and lipidomic (from UHPLC-MS-ROI) studies described above. The two peak area matrices (from ^1H NMR and UHPLC-MS-ROI analyses) were joined row-wise, producing a new lipidomic-and-metabolomic peak areas dataset, especially useful to study and interpret possible correlations between these two type of compounds and their biological interrelationships. Before this data fusion, the data were preliminary transformed in two steps. Firstly, a min-max scaling (15) procedure was applied to each lipid and metabolite. This pretreatment allows scaling the

abundances (peak areas), for all the compounds, within the same range (between 0 and 1). Secondly, the two data subsets were block scaled by dividing all their individual elements by the first singular value of the data subset they belong to, as recommended in the multiple factor analysis method (16). This second pretreatment prevents that one subset dominates over the other in this MCR-ALS data analysis. Constraints used in this MCR-ALS were no-negativity in both **C** and **S^T** matrices, and closure in **C**.

The lipidomic-and-metabolomic fused dataset gives a new data matrix **X**, which is submitted to MCR-ALS analysis. In this case, this matrix has a number of rows equal to the number of samples (24 samples), and a number of columns equal to the number of identified compounds (in this data example this number was equal to 123; see the **Results** section). The resolved matrix **S^T** will show in this case, the lipid and metabolic composition (lipidomic-and-metabolomic profile) of the resolved MCR-ALS components, whereas **C** will give the contributions of these lipidomic-and-metabolomic profiles in the different samples. From these **C** matrices, the evolution of the concentrations of the different metabolites due to temperature changes can be deduced and be described as metabolic thermal profiles. The **E** matrix, with the same dimensions as the **X** matrix, has the residual information not explained by the model using the considered components. During the MCR-ALS analysis iterative process, closure in **C** matrix and a non-negativity constraint were also applied.

The quality of the MCR-ALS model was measured by evaluating the percent of explained variance (R^2) (11), and by comparison of the original matrix (**X**) with the one reconstructed from the resolved data (**CS^T**).

The number of components used in this bilinear decomposition can be initially estimated by using the singular value decomposition (SVD) method (17). For the estimation of **C** and **S^T** factor matrices, the MCR-ALS algorithm under non-negativity constraints on both factor matrices was used.

Biological interpretation of MCR-ALS resolved **S^T** profiles was attempted from the set of metabolites (or lipids) whose contribution was 50% or higher than the maximum contribution

found in the given component. For every set of selected lipid and metabolites, a pathway analysis using the KEGG (18) pathway database was performed.

Lipid abundances for each of the resolved components were obtained after undoing the two transformation steps applied as a data pretreatment (min-max scaling and singular value scaling) of the original data on the \mathbf{S}^T matrix. The lipid content of every lipid family for each \mathbf{S}^T component profile was obtained by the sum of all of the reconstructed concentration values that regarded the same lipid family found in the same \mathbf{S}^T profile.

Finally, the average fatty acyl chain length and the average number of unsaturations for each \mathbf{S}^T profile were obtained by applying eq.1 and eq.2, respectively, on the set of reconstructed concentration values associated with each \mathbf{S}^T profile.

Supplementary tables

Table S1. Metabolite assignment from ¹H NMR data. ¹H NMR spectroscopic data (chemical shift, multiplicity, proton integral and coupling constant) of the assigned metabolites. CAS, HMDB, YMDB and KEGG codes are provided when existing.

#	Metabolite name	¹ H signals assigned	CODE			
			CAS	YMDB	HMDB	KEGG
1	Acetic acid	1.90 ppm (s, 3 H)	64-19-7	YMDB00056	HMDB00042	C00033
2	Choline	3.19 ppm (s, 9 H)	62-49-7	YMDB00227	HMDB00097	C00114
3	Adenine	8.18 ppm (s, 1 H); 8.23 ppm (s, 1 H)	73-24-5	YMDB00887	HMDB00034	C00147
4	AMP	6.13 ppm (d, J=6.0 Hz, 1 H); 8.26 ppm (s, 1 H); 8.59 ppm (s, 1 H)	61-19-8	YMDB00097	HMDB00045	C00020
5	ADP	8.25 ppm (s, 1 H); 8.52 ppm (s, 1 H)	58-64-0	YMDB00914	HMDB01341	C00008
6	ATP	6.14 ppm (d, J=5.1 Hz, 1 H); 8.53 ppm (s, 1 H); 8.26 ppm (s, 1 H)	56-65-5	YMDB00109	HMDB00538	C00002
7	Citric acid	2.50 ppm (s, 0.66 H); 2.54 ppm (s, 1.33 H); 2.63 ppm (s, 1.22 H); 2.67 ppm (s, 0.70 H)	77-92-9	YMDB00086	HMDB00094	C00158
8	D-Glucose	3.24 ppm (dd, J=(7.8 Hz, 9.2 Hz), 0.71 H); 3.35-3.56 ppm (m, 2.63 H); 5.22 ppm (d, J=3.8 Hz, 0.36 H)	50-99-7	YMDB00286	HMDB00122	C00031
9	Ethanol	1.17 ppm (t, J=7.1 Hz, 3H); 3.65 ppm (q, J=7.1 Hz, 2 H)	64-17-5	YMDB00883	HMDB00108	C00469
10	Formic acid	8.44 ppm (s, 1 H)	64-18-6	YMDB00385	HMDB00142	C00058
11	Glycerol	3.51-3.58 ppm (m, 2.07 H); 3.61-3.67 ppm (m, 2.12 H); 3.77 ppm (tt, J=(6.5 Hz, 4.4 Hz), 0.79 H)	56-81-5	YMDB00283	HMDB00131	C00116
12	Glycerophosphocholine	3.22 ppm (s, 9 H); 3.56-3.71 ppm (m, 4 H); 4.27-4.36 ppm (m, 2 H)	28319-77-9	YMDB00309	HMDB00086	C00670
13	Glycine	3.55 ppm (s, 2 H)	56-40-6	YMDB00016	HMDB00123	C00037
14	GMP	5.93 ppm (d, J=6.2 Hz); 8.19 ppm (s, 1H)	85-32-5	YMDB00261	HMDB01397	C00144
15	GTP	5.93 ppm (d, J=6.1 Hz); 8.13 ppm (s, 1H)	86-01-1	YMDB00558	HMDB01273	C00044
16	Hypoxanthine	8.18 ppm (s, 1 H); 8.20 ppm (s, 1 H)	68-94-0	YMDB00555	HMDB00157	C00262
17	Inosine	8.22 ppm (s, 1H); 8.33 ppm (s, 1H)	75-18-3	YMDB00510	HMDB00195	C00081

#	Metabolite name	¹ H signals assigned	CODE			
			CAS	YMDB	HMDB	KEGG
18	L-alanine	1.47 ppm (d, J=7.1 Hz, 3 H); 3.76 ppm (q, J=7.2 Hz, 1 H)	56-41-7	YMDB00154	HMDB00161	C00041
19	L-arginine	1.58 - 1.79 ppm (m, 2 H); 1.80-2.00 ppm (m, 2 H); 3.23 ppm (t, J=6.9 Hz, 2 H); 3.76 ppm (t, J=6.1 Hz, 1 H)	74-79-3	YMDB00592	HMDB00517	C00062
20	L-asparagine	2.82 ppm (d, J=7.6 Hz, 0.286 H); 2.86 ppm (d, J=7.6 Hz, 0.714 H)	70-47-3	YMDB00226	HMDB00168	C00152
21	L-aspartic acid	2.64 ppm (d, J=8.9 Hz, 0.33 H); 2.69 ppm (d, J=8.9 Hz, 0.67 H); 2.78 ppm (d, 3.7 Hz, 0.67 H); 2.83 ppm (d, J=3.7 Hz, 0.33 H); 3.89 ppm (dd, J=(8.8 Hz, 3.8 Hz), 1 H)	56-84-8	YMDB00896	HMDB00191	C00049
22	L-glutamic acid	1.99 - 2.17 ppm (m, 2 H); 2.26-2.42 ppm (m, 2 H); 3.75 ppm (dd, J=(7.2 Hz, 4.7 Hz), 1 H)	56-86-0	YMDB00271	HMDB00148	C00025
23	L-glutamine	2.44 (td, J = (7.5, 3.7 Hz), 1H)	56-85-9	YMDB00002	HMDB00641	C00064
24	L-histidine	3.11 ppm (d, J=7.4 Hz, 0.35 H); 3.15 ppm (d, J=7.7 Hz, 0.65 H); 7.07 ppm (s, 1 H); 7.84 ppm (s, 1 H)	71-00-1	YMDB00369	HMDB00177	C00135
25	L-isoleucine	0.93 ppm (t, J=7.4 Hz, 3 H); 1.00 ppm (d, J=7.1 Hz, 3 H)	73-32-5	YMDB00038	HMDB00172	C00407
26	L-lactic acid	1.31 ppm (d, J=7.0 Hz, 3 H); 4.10 ppm (q, J=6.9 Hz, 1 H)	79-33-4	YMDB00247	HMDB00190	C00186
27	L-leucine	0.95 ppm (d, J=6.1 Hz, 3 H); 0.95 ppm (d, J=6.1 Hz, 3 H)	61-90-5	YMDB00387	HMDB00687	C00123
28	L-lysine	1.35-1.60 ppm (m, 2 H); 1.65-1.80 ppm (m, 2 H); 1.81-1.94 ppm (m, 2 H); 2.95-3.1 ppm (m, 2 H); 3.75 ppm (t, J=6.1 Hz, 1 H)	56-87-1	YMDB00330	HMDB00182	C00047
29	L-ornithine	1.65-2.00 ppm (m, 4 H); 3.04 ppm (t, J=7.6 Hz, 2 H)	70-26-8	YMDB00353	HMDB00214	C00077
30	L-phenylalanine	7.37 ppm (m, 5 H)	63-91-2	YMDB00304	HMDB00159	C00079
31	L-proline	1.90-2.12 ppm (m, 3 H); 2.27-2.40 ppm (m, 1 H); 4.12 ppm (dd, J=(8.6 Hz, 6.4 Hz), 1 H)	344-25-2	YMDB00378	HMDB00162	C00148
32	L-threonine	1.32 ppm (d, J=6.6 Hz, 3 H); 3.57 ppm (d, J=4.9 Hz, 1 H); 4.19-4.28 ppm (m, 1 H)	72-19-5	YMDB00214	HMDB00167	C00188
33	L-tyrosine	6.89 ppm (d, J=8.4 Hz, 2 H); 7.18 ppm (d, J=8.4 Hz, 2 H)	60-18-4	YMDB00364	HMDB00158	C00082
34	L-valine	0.98 ppm (d, J=7.1 Hz, 3 H); 1.03 ppm (d, J=7.1 Hz, 3 H); 2.26 ppm (m, 1 H)	72-18-4	YMDB00152	HMDB00883	C00183

#	Metabolite name	¹ H signals assigned	CODE			
			CAS	YMDB	HMDB	KEGG
35	NAD ⁺	6.03 ppm (d, J=5.9 Hz, 1 H); 6.08 ppm (d, J= 5.3 Hz, 1 H); 8.16 ppm (s, 1 H); 8.17 ppm (d, J=6.1 Hz); 8.15-8.21 ppm (m, 2 Hz); 8.42 ppm (s, 1H); 8.82 ppm (d, J=7.6 Hz, 1 H); 9.13 ppm (d, J=7.6 Hz, 1 H); 9.33 ppm (s, 1 H)	53-84-9	YMDB00110	HMDB00902	C00003
36	Orotic acid	6.18 ppm (s, 1 H)	65-86-1	YMDB00405	HMDB00226	C00295
37	Orotidine-5P	5.54 ppm (d, J=3.3 Hz, 1 H); 5.76 ppm (s, 1 H);	2149-82-8	YMDB00025	HMDB00218	C01103
38	Succinic acid	2.39 ppm (s, 4 H)	110-15-6	YMDB00338	HMDB00254	C00042
39	Trehalose	3.44 ppm (t, J=9.3 Hz, 2 H); 3.59 - 3.92 ppm (m, 10 H); 5.18 ppm (d, J=3.8 Hz, 2 H)	99-20-7	YMDB00008	HMDB00975	C01083
40	UMP	5.98 ppm (m, 2 H); 8.10 ppm (J=7.9 Hz, 1H)	58-97-9	YMDB000049	HMDB00288	C00105
41	Uracil	5.79 ppm (d, J=7.8 Hz, 1 H); 7.53 ppm (d, J=7.7 Hz, 1 H)	66-22-8	YMDB00098	HMDB00300	C00106
42	Uridine	5.87-5.92 ppm (m, 2H); 7.87 ppm (d. J=8.5 Hz, 1 H)	58-96-8	YMDB00127	HMDB00296	C00299

Note: Spectroscopic constants relative to strong coupled protons are described as various set of regular coupled proton (ex: a *dd* is expressed as two *d* with different proton integrals).

Table S2. Metabolite assignment from UHPLC-MS data. Summary of the molecules found using the UHPLC-MS-ROI-MCR-ALS strategy. Lipidmaps codes are provided when existing. Otherwise, YMDB code names are provided.

#	RT	Measured <i>m/z</i>	Identified compound	ESI	Adduct	Calculated <i>m/z</i>	Mass error (ppm)	CODE									
1	12.86	712.6797471	Cer(t18:0/h26:0)(2OH)	positive	[M+H] ⁺	712.6814	2.3	LMSP02030003									
									2	9.44	663.4554962	Gcer1: Gal α 1-3(Fuca1-2)Gal β 1-4Glc β -Cer(d18:1/26:1(17Z))	positive	[M+H+NH ₄] ²⁺	663.4552	0.4	LMSP0505DI08
4	12.86	938.7054047	PI-Cer(t18:0/26:0)	positive	[M+H] ⁺	938.7056	0.2	LMSP03030017									
									5	5.74	502.4465307 467.4087997	DG(26:0)	positive	[M+NH ₄] ⁺	502.4466	0.1	LMGL02010324
6	6.82	530.4751844	DG(28:0)	positive	[M+H-H ₂ O] ⁺	467.4089	0.2	LMGL02010331									
									7	7.45	558.5103603	DG(30:0)	positive	[M+NH ₄] ⁺	558.5092	2.1	LMGL02010335
8	9.59	586.5384633	DG(32:0)	positive	[M+NH ₄] ⁺	86.5405	3.5	LMGL02010009									
									9	8.66	584.5255124 549.489371	DG(32:1)	positive	[M+NH ₄] ⁺	584.5248	1.1	LMGL02010452
10	7.66	582.5103803 582.5103803	DG(32:2)	positive	[M+H-H ₂ O] ⁺	549.4872	4.0	LMGL02010408									
									11	10.00	612.5529074 577.518048	DG(34:1)	positive	[M+NH ₄] ⁺	582.5092	2.0	LMGL02010004
12	9.06	610.5412377 575.5018286	DG(34:2)	positive	[M+H-H ₂ O] ⁺	612.5561	5.3	LMGL02010004									
									13	11.24	640.5849535	DG(36:1)	positive	[M+H-H ₂ O] ⁺	577.5185	0.7	LMGL02010021
				positive	[M+NH ₄] ⁺	610.5405	1.2	LMGL02010021									
													positive	[M+H-H ₂ O] ⁺	575.5028	1.7	LMGL02010043
				positive	[M+NH ₄] ⁺	640.5874	3.9	LMGL02010043									

#	RT	Measured <i>m/z</i>	Identified compound	ESI	Adduct	Calculated <i>m/z</i>	Mass error (ppm)	CODE
DG	14	638.5760179	DG(36:2)	positive	[M+NH ₄] ⁺	638.5718	6.6	LMGL02010049
	15	698.6615414	DG(40:0)	positive	[M+NH ₄] ⁺	698.6657	6.0	LMGL02010117
TG	16	712.6282877	TG(40:0)	positive	[M+NH ₄] ⁺	712.6250	4.6	LMGL03012632
	17	740.6694547	TG(42:0)	positive	[M+NH ₄] ⁺	740.6763	9.2	LMGL03012616
	18	738.6587476	TG(42:1)	positive	[M+NH ₄] ⁺	738.6606	2.5	LMGL03012638
	19	768.7034582	TG(44:0)	positive	[M+NH ₄] ⁺	768.7076	5.3	LMGL03012645
	20	766.6887759	TG(44:1)	positive	[M+NH ₄] ⁺	766.6919	4.1	LMGL03012646
	21	764.6728214	TG(44:2)	positive	[M+NH ₄] ⁺	764.6763	4.5	LMGL03012647
	22	796.7386579	TG(46:0)	positive	[M+NH ₄] ⁺	796.7389	0.3	LMGL03010007
	23	794.7204028	TG(46:1)	positive	[M+NH ₄] ⁺	794.7232	3.5	LMGL03012653
	24	792.7118417	TG(46:2)	positive	[M+NH ₄] ⁺	792.7076	5.4	LMGL03012654
	25	822.7493862	TG(48:1)	positive	[M+NH ₄] ⁺	822.7545	6.2	LMGL03010017
	26	820.7341149	TG(48:2)	positive	[M+NH ₄] ⁺	820.7389	5.8	LMGL03010018
27	818.7171252	TG(48:3)	positive	[M+NH ₄] ⁺	818.7232	7.4	LMGL03010020	
28	850.7810769	TG(50:1)	positive	[M+NH ₄] ⁺	850.7858	5.6	LMGL03010005	
29	848.7666918	TG(50:2)	positive	[M+NH ₄] ⁺	848.7702	4.1	LMGL03010032	
30	846.750456	TG(50:3)	positive	[M+NH ₄] ⁺	846.7545	4.8	LMGL03010037	
31	866.8115699	TG(51:0)	positive	[M+NH ₄] ⁺	866.8171	6.4	LMGL03010031	
32	878.8118558	TG(52:1)	positive	[M+NH ₄] ⁺	878.8171	6.0	LMGL03010071	
33	876.7971274	TG(52:2)	positive	[M+NH ₄] ⁺	876.8015	4.9	LMGL03010083	

#	RT	Measured m/z	Identified compound	ESI	Adduct	Calculated m/z	Mass error (ppm)	CODE	
TG	34	15.53	821.7372159	TG(52:3)	positive	[M+H-2H ₂ O] ⁺	821.7370	0.3	LMGL03010099
	35	18.24	906.8421157	TG(54:1)	positive	[M+NH ₄] ⁺	906.8484	6.9	LMGL03010177
PA	36	17.65	904.8269621	TG(54:2)	positive	[M+NH ₄] ⁺	904.8328	6.4	LMGL03010203
	37	4.82	691.454192	PA(32:1)	negative	[M+HCOO] ⁻	691.4550	1.2	LMGP10010019
	38	7.07	735.5584558	PA(36:1)	positive	[M+CH ₃ OH+H] ⁺	735.5534	6.8	LMGP10010037
	39	6.05	662.4724516	PA(36:2)	positive	[M+NH ₄] ⁺	662.4755	4.6	LMGP10010058
PC	40	2.33	540.3271047	Lyso-PC(16:0)	negative	[M+HCOO] ⁻	540.3301	5.6	LMGP01050018
	41	2.16	494.3226224	Lyso-PC(16:1)	positive	[M+H] ⁺	494.3241	3.0	LMGP01050021
538.3139782			negative		[M+HCOO] ⁻	538.3145	1.0		
42	2.72	522.3548285	Lyso-PC(18:1)	positive	[M+H] ⁺	522.3554	1.1	LMGP01050029	
		566.3453712		negative	[M+HCOO] ⁻	566.3458	0.8		
43	4.90	650.4766101	PC(26:0)	positive	[M+H] ⁺	650.4755	1.7	LMGP01010388	
		648.4596363		positive	[M+H] ⁺	648.4599	0.4		
44	4.46	692.4527771	PC(26:1)	negative	[M+HCOO] ⁻	692.4503	3.6	LMGP01011316	
		678.502869		positive	[M+H] ⁺	678.5068	5.8		
45	5.80	722.4959197	PC(28:0)	negative	[M+HCOO] ⁻	722.4972	1.8	LMGP01010390	
		676.4906912		positive	[M+H] ⁺	676.4912	0.7		
46	5.15	720.4854479	PC(28:1)	negative	[M+HCOO] ⁻	720.4816	5.4	LMGP01010392	
		704.5218728		positive	[M+H] ⁺	704.5225	0.9		
48	5.39	702.5061362	PC(30:2)	positive	[M+H] ⁺	702.5068	1.0	LMGP01011323	
		746.4971292		negative	[M+HCOO] ⁻	746.4972	0.1		

#	RT	Measured m/z	Identified compound	ESI	Adduct	Calculated m/z	Mass error (ppm)	CODE
49	7.32	732.5495272	PC(32:1)	positive	[M+H] ⁺	732.5538	5.8	LMGP01010490
		776.5392322		negative	[M+HCOO] ⁻	776.5442	6.3	
50	6.42	730.5384368	PC(32:2)	positive	[M+H] ⁺	730.5381	0.4	LMGP01010494
		774.5284523		negative	[M+HCOO] ⁻	774.5285	0.1	
		1460.077265		positive	[2M+H] ⁺	1460.069	5.7	
51	7.73	1518.13629	PC(33:0)	positive	[2M+Na] ⁺	1518.1448	5.6	LMGP01010399
52	7.71	804.5717026	PC(34:1)	negative	[M+HCOO] ⁻	804.5755	4.7	LMGP01010005
53	2.14	450.2612786	Lyso-PE(16:1)	negative	[M-H] ⁻	450.2621	1.8	LMGP02050010
54	2.84	510.3549121	Lyso-PE(20:0)	positive	[M+NH ₄] ⁺	510.3554	1.0	LMGP02050012
55	7.32	690.506157	PE(32:1)	positive	[M+H] ⁺	690.5068	1.0	LMGP02010395
56	6.48	688.4905073	PE(32:2)	positive	[M+H] ⁺	688.4912	1.0	LMGP02010354
		686.4756215		negative	[M-H] ⁻	686.4761	0.7	
57	8.66	718.5336018	PE(34:1)	positive	[M+H] ⁺	718.5381	6.3	LMGP02010009
		716.5195158		negative	[M-H] ⁻	716.523	4.9	
58	7.76	716.5205389	PE(34:2)	positive	[M+H] ⁺	716.5225	2.7	LMGP02011220
		714.5071909		negative	[M-H] ⁻	714.5079	1.0	
59	8.69	761.5784066	PE(36:2)	positive	[M+NH ₄] ⁺	761.5803	2.5	LMGP02010510
		742.5365784		negative	[M-H] ⁻	742.5387	2.8	
60	10.00	772.5857109	PE(38:2)	positive	[M+H] ⁺	772.5851	0.8	LMGP02010513
61	2.67	495.2717184	Lyso-PG(17:1)	negative	[M-H] ⁻	495.2723	1.2	LMGP04050036
62	8.77	719.4848965	PG(32:1)	negative	[M-H] ⁻	719.4863	2.0	LMGP04010059

#	RT	Measured m/z	Identified compound	ESI	Adduct	Calculated m/z	Mass error (ppm)	CODE
PG	63	763.5465134	PG(O-34:3)	positive	$[M+CH_3OH+H]^+$	763.5484	2.4	LMGP04020077
	64	792.572821	PG(36:2)	positive	$[M+NH_4]^+$	792.5749	2.6	LMGP04010107
PI	65	597.3032118	Lyso-PI(18:1)	negative	$[M-H]^-$	597.304	1.3	LMGP06050005
	66	725.4219129	PI(26:0)	negative	$[M-H]^-$	725.4241	3.0	LMGP06010949
	67	753.4542539	PI(28:0)	negative	$[M-H]^-$	753.4554	1.5	LMGP06010890
	68	807.5009848	PI(32:1)	negative	$[M-H]^-$	807.5024	1.7	LMGP06010027
		826.5433157		positive	$[M+NH_4]^+$	826.544	0.8	
	69	805.4878915	PI(32:2)	negative	$[M-H]^-$	805.4867	1.5	LMGP06010028
70	2.39	Lyso-PS(O-16:0)	negative	$[M-H_2O-H]^-$	464.2777	1.3	LMGP03060003	
PS	71	522.2828151	Lyso-PS(18:1)	negative	$[M-H]^-$	522.2832	0.7	LMGP03050001
	72	1674.098597	PS(18:1(9Z)/22:4(7Z,10Z,13Z,16Z))	negative	$[2M-H]^-$	1674.0962	1.5	LMGP03010339
	73	732.4810489	PS(32:1)	negative	$[M-H]^-$	732.4816	0.7	LMGP03010059
	74	730.4675723	PS(32:2)	negative	$[M-H]^-$	730.4659	2.3	LMGP03010060
	75	760.5104737	PS(34:1)	negative	$[M-H]^-$	760.5129	3.1	LMGP03010007
	76	15.60	lanosteryl palmitoleate	positive	$[M+NH_4]^+$	680.634	5.9	LMST01031008
CF	77	652.599399	episteryl palmitoleate	positive	$[M+NH_4]^+$	652.6027	5.1	LMST01031009
	78	650.5845811	ergosteryl palmitoleate	positive	$[M+NH_4]^+$	650.5871	3.8	LMST01031011
79	0.99	1Dmyoinositol 4,5bisphosphate	negative	$[M+Cl]^-$	374.9655	1.3	YMDB00764	
80	4.24	Eicosanoyl-EA	positive	$[M+H-H_2O]^+$	338.3412	1.4	LMFA08040038	
81	1.22	N-Acetyl-D-glucosaminyldiphosphodolichol	positive	$[M+K]^+$	626.1892	7.9	YMDB00934	

Table S3. Classification of relevant metabolites, including lipids, resolved by MCR-ALS analysis of the peak areas of the metabolite and lipid concentrations obtained respectively by NMR and LC-MS analysis of yeast samples at different temperatures). Metabolic contribution for every metabolite, expressed as the fraction of the maximum metabolic contribution found in the same component, is given inside parenthesis.

KEGG Pathway	Component 1 (cold)		Component 2 (growth)		Component 3 (heat)	
	#	Metabolites	#	Metabolites	#	Metabolites
Biosynthesis of amino acids	5	L-valine (0.65), L-leucine (0.78), L-glutamine (0.86), L-ornithine (0.95), citric acid (0.82)	13	L-valine (1), L-isoleucine (0.72), L-leucine (0.62), L-threonine (0.82), L-lysine (0.93), L-alanine (0.70), L-arginine (0.97), L-aspartic acid (0.95), L-tyrosine (0.71), L-histidine (0.71), L-phenylalanine (0.56), L-proline (0.52)	8	L-Glutamic acid (0.87), L-alanine (0.57), L-lysine (0.89), L-phenylalanine (0.57), L-tyrosine (0.58), L-histidine (0.56), L-threonine (0.60), L-proline (0.56)
			4	L-glutamine (0.86), uracil (0.58), uridine (0.86), inosine (0.54)	10	AMP (0.70), NAD+ (0.65) ADP (0.63), glycine (0.67), GMP (0.71), adenine (0.58), hypoxanthine (0.62), UMP (0.99), uracil (0.60), orotidine-5P (0.53)
Purine and Pyrimidine	4	L-glutamine (0.86), uracil (0.58), uridine (0.86), inosine (0.54)	10	AMP (0.70), NAD+ (0.65) ADP (0.63), glycine (0.67), GMP (0.71), adenine (0.58), hypoxanthine (0.62), UMP (0.99), uracil (0.60), orotidine-5P (0.53)	3	AMP (0.58), orotidine-5P (0.63), orotic acid (0.91)

KEGG Pathway	Component 1 (cold)		Component 2 (growth)		Component 3 (heat)	
	#	Metabolites	#	Metabolites	#	Metabolites
Carbon metabolism	3	Citric acid (0.82), succinic acid (0.77), acetic acid (0.56) Phosphatidyl-myo-inositol-4,5-BP (0.67), glycerol (1), DG(26:0) (0.69), DG(32:1) (0.78), DG(32:2) (0.77), DG(34:2) (0.76), DG(36:2) (0.53), TG(40:0) (0.75), TG(42:0) (0.81), TG(42:1) (0.82), TG(44:0) (0.85), TG(44:1) (0.79), TG(44:2) (0.70), TG(46:0) (0.84), TG(46:1) (0.71), TG(46:2) (0.75), TG(48:1) (0.86), TG(48:2) (0.87), TG(48:3) (0.74), TG(50:1) (0.75), TG(50:2) (0.85), TG(50:3) (0.82), TG(52:1) (0.59), TG(52:2) (0.71), TG(52:3) (0.89), LPE(16:1) (0.82), LPE(20:0) (0.57), PE(32:1) (0.62), LPG(17:1) (0.65), LPS(O-16:0) (0.90), PI(26:0) (0.969), PI(28:0) (0.53), PI(32:1) (0.63), PI(32:2) (0.75)	4	Acetic acid (0.60), Glycine (0.67), L-alanine (0.70), L-aspartic acid (0.95)	2	L-glutamic acid (0.87), L-alanine (0.57) Glycerophosphocholine (1), choline (0.73), Cer-2OH (0.82), DG(32:0) (0.74), DG(34:1) (0.71), DG(36:1) (0.53), DG(36:2) (0.73), DG(40:0) (0.86), TG(50:1) (0.54), TG(50:2) (0.56), TG(51:0) (0.79), TG(52:2) (0.52), LPC(16:0) (0.77), LPC(16:1) (0.73), LPC(18:1) (0.76), PA(36:1) (0.84), PC(28:0) (0.83), PC(30:1) (0.61), PC(32:1) (0.79), PC(32:2) (0.72), PC(33:0) (0.84), PC(34:1) (0.81), PG(32:1) (0.60), LPE(20:0) (0.64), PE(36:2) (0.62), LPI(18:1) (0.75), LPS(18:1) (0.75), LPS(O-16:0) (0.65), PS(40:5) (0.65), lan-16:1Δ9 (0.62), ep-16:1Δ9 (0.67), erg-16:1Δ9 (0.61), eicosanoyl-EA (0.79), NAGADPD (0.79)
	Lipid metabolism	34		8	DG(30:0) (0.55), PC(28:1) (0.65), PC(30:1) (0.55), PC(30:2) (0.53), PE(32:1) (0.59) PS(32:1) (0.54), PI(26:0) (0.51), PI(28:0) (0.63)	34
Starch, sucrose and glycolysis	3	D-glucose (0.60), acetic acid (0.56), lactic acid(0.76)	1	Acetic acid (0.60)	2	L-lactic acid (0.64), trehalose (0.82)

Table S4. KEGG Pathway analysis.

KEGG pathway	Description	# of Hits	Metabolites
sce01100	Metabolic pathways	51	C00002, C00003, C00008, C00020, C00025, C00031, C00033, C00037, C00041, C00042, C00044, C00047, C00049, C00058, C00062, C00064, C00077, C00079, C00082, C00105, C00106, C00114, C00116, C00123, C00135, C00144, C00147, C00148, C00152, C00157, C00158, C00183, C00186, C00188, C00195, C00262, C00295, C00299, C00344, C00350, C00407, C00416, C00422, C00469, C00641, C01083, C01103, C01190, C01194, C01220, C02737
sce01110	Biosynthesis of secondary metabolites	31	C00002, C00008, C00020, C00025, C00031, C00033, C00037, C00042, C00044, C00047, C00049, C00062, C00077, C00079, C00082, C00123, C00135, C00148, C00152, C00157, C00158, C00183, C00186, C00188, C00350, C00407, C00416, C00469, C00641, C01083, C02737
sce01130	Biosynthesis of antibiotics	21	C00002, C00008, C00020, C00025, C00031, C00033, C00037, C00042, C00047, C00049, C00062, C00077, C00079, C00082, C00148, C00158, C00183, C00186, C00188, C00407, C00469
sce02010	ABC transporters	19	C00025, C00031, C00037, C00041, C00047, C00049, C00062, C00064, C00077, C00079, C00114, C00116, C00123, C00135, C00148, C00183, C00188, C00407, C01083
sce01230	Biosynthesis of amino acids	18	C00025, C00037, C00041, C00047, C00049, C00062, C00064, C00077, C00079, C00082, C00123, C00135, C00148, C00152, C00158, C00183, C00188, C00407
sce00970	Aminoacyl-tRNA biosynthesis	16	C00025, C00037, C00041, C00047, C00049, C00062, C00064, C00079, C00082, C00123, C00135, C00148, C00152, C00183, C00188, C00407
sce00564	Glycerophospholipid metabolism	11	C00114, C00157, C00344, C00350, C00416, C00641, C00670, C01194, C02737, C04230, C04438
sce00230	Purine metabolism	10	C00002, C00008, C00020, C00037, C00044, C00064, C00081, C00144, C00147, C00262
sce01210	2-Oxocarboxylic acid metabolism	10	C00025, C00047, C00049, C00077, C00079, C00082, C00123, C00158, C00183, C00407
sce01200	Carbon metabolism	8	C00025, C00033, C00037, C00041, C00042, C00049, C00058, C00158
sce00460	Cyanoamino acid metabolism	7	C00037, C00049, C00079, C00082, C00152, C00183, C00407

KEGG pathway	Description	# of Hits	Metabolites
sce00250	Alanine, aspartate and glutamate metabolism	7	C00025, C00041, C00042, C00049, C00064, C00152, C00158
sce00240	Pyrimidine metabolism	6	C00064, C00105, C00106, C00295, C00299, C01103
sce00630	Glyoxylate and dicarboxylate metabolism	6	C00025, C00037, C00042, C00058, C00064, C00158
sce00260	Glycine, serine and threonine metabolism	5	C00037, C00049, C00114, C00188, C02737
sce00220	Arginine biosynthesis	5	C00025, C00049, C00062, C00064, C00077
sce04070	Phosphatidylinositol signaling system	4	C00416, C00641, C01194, C01220
sce00290	Valine, leucine and isoleucine biosynthesis	4	C00123, C00183, C00188, C00407
sce00261	Monobactam biosynthesis	4	C00049, C00062, C00082, C00188
sce00010	Glycolysis / Gluconeogenesis	4	C00031, C00033, C00186, C00469
sce00330	Arginine and proline metabolism	4	C00025, C00062, C00077, C00148
sce00620	Pyruvate metabolism	4	C00033, C00042, C00058, C00186
sce00680	Methane metabolism	4	C00033, C00037, C00058, C00082
sce00561	Glycerolipid metabolism	4	C00116, C00416, C00422, C00641
sce00190	Oxidative phosphorylation	4	C00002, C00003, C00008, C00042

2.3 SCIENTIFIC ARTICLE III

¹H NMR metabolomic study of auxotrophic starvation in yeast using Multivariate Curve Resolution-Alternating Least Squares for Pathway Analysis.

Authors: Puig-Castellví F., Alfonso I., Piña B., Tauler R.

Citation reference: *Scientific Reports* (2016), 6:30982.

DOI: 10.1038/srep30982

SCIENTIFIC REPORTS

OPEN

¹H NMR metabolomic study of auxotrophic starvation in yeast using Multivariate Curve Resolution-Alternating Least Squares for Pathway Analysis

Received: 21 March 2016
Accepted: 12 July 2016
Published: 03 August 2016

Francesc Puig-Castellví¹, Ignacio Alfonso², Benjamin Piña¹ & Romà Tauler¹

Disruption of specific metabolic pathways constitutes the mode of action of many known toxicants and it is responsible for the adverse phenotypes associated to human genetic defects. Conversely, many industrial applications rely on metabolic alterations of diverse microorganisms, whereas many therapeutic drugs aim to selectively disrupt pathogens' metabolism. In this work we analyzed metabolic changes induced by auxotrophic starvation conditions in yeast in a non-targeted approach, using one-dimensional proton Nuclear Magnetic Resonance spectroscopy (¹H NMR) and chemometric analyses. Analysis of the raw spectral datasets showed specific changes linked to the different stages during unrestricted yeast growth, as well as specific changes linked to each of the four tested starvation conditions (L-methionine, L-histidine, L-leucine and uracil). Analysis of changes in concentrations of more than 40 metabolites by Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS) showed the normal progression of key metabolites during lag, exponential and stationary unrestricted growth phases, while reflecting the metabolic blockage induced by the starvation conditions. In this case, different metabolic intermediates accumulated over time, allowing identification of the different metabolic pathways specifically affected by each gene disruption. This synergy between NMR metabolomics and molecular biology may have clear implications for both genetic diagnostics and drug development.

Metabolomics aims to identify the specific cellular processes undergoing in biological organisms by the identification and quantitation of dozens to thousands metabolites with high-throughput techniques, by using a non-aprioristic approach¹. Metabolomic analyses have been performed in many organisms, including human and mammalian tissues^{2,3}, different animal species, both vertebrates⁴ and invertebrates⁵, plants⁶, and microorganisms, both Eukaryotes (yeasts⁷, protists⁸) and Prokaryotes (bacteria⁹, archaea¹⁰).

Among the eukaryotic microorganisms, the yeast *Saccharomyces cerevisiae* is widely used in many biological fields, such as biotechnology¹¹ or food industry¹², and it constitutes an excellent model organism for metabolomics¹³ and other “omic” approaches¹⁴. We present here an NMR analysis of the metabolome variations induced by auxotrophic starvation in yeast, which occurs when a strain lacking specific genes (in this case, *HIS3*, *LEU2*, *MET15* and *URA3*, also called genetic markers) is confronted with a medium devoid of one or more of the essential metabolites it can no longer synthesize (L-histidine, L-leucine, L-methionine and/or uracil, respectively). Reports of metabolic disruption in yeast have previously focused on the triggered proteins¹⁵ or on the transcriptomic¹⁶ evidences. The general consensus is that starvation is controlled through the RAS/protein kinase A (PKA) and TOR pathways¹⁷ that mediate the transcriptional, translational, and metabolic state of the cell. However, most studies analyzed the effects of the depletion in a requiring nutrient, like carbon, nitrogen, phosphate or

¹Department of Environmental Chemistry, Institute of Environmental Assessment and Water Research, (IDAEA-CSIC), Jordi Girona 18-26, 08034 Barcelona, Catalonia, Spain. ²Department of Biological Chemistry and Molecular Modelling, Institute of Advanced Chemistry of Catalonia (IQAC-CSIC), Jordi Girona 18-26, 08034 Barcelona, Catalonia, Spain. Correspondence and requests for materials should be addressed to R.T. (email: roma.tauler@idaea.csic.es)

sulphur sources^{14,18}. Our analysis is focused on the final downstream product of the whole biological system (the metabolites), which are closer to the final phenotype than RNA or proteins¹⁹. In addition, the study of different auxotrophic starvations allows a more detailed and specific analysis of the metabolic changes induced by characterized disruptions of the endogenous metabolic pathways.

Conventional technologies for metabolomic analyses include Nuclear Magnetic Resonance (NMR)²⁰ or hyphenated techniques of Liquid Chromatography¹³, Gas Chromatography⁹ or Capillary Electrophoresis⁷ coupled to Mass Spectrometry (MS). Despite the fact that NMR allows to identify a quite lower number of metabolites than the maximal capacity of MS²¹, the set of identified metabolites with NMR comprises a broad range of molecules, resulting advantageous for explaining the changes occurring in the cell, whereas for MS a preliminary selection of the peaks (using targeted or non-targeted approaches) has to be performed and the final metabolic overview might result biased.

A previous study has proven the capacity of ¹H NMR to identify metabolic variations related to distinct genetic backgrounds²². The most relevant limitation of NMR, in comparison to MS-derived techniques, is that most of the detectable metabolites are related to primary metabolism, whereas the less abundant ones may be left unobserved due to the relatively low sensitivity of the technique. However, if the studied condition is drastic enough, target metabolites might be raised to detectable concentrations. In addition, and because of the intrinsic properties of NMR, known resonances can be used for quantifying the corresponding metabolites, whereas chemical structures can be deduced from the not-yet-assigned resonances. In this article we applied ¹H NMR to the study of local disruptions of metabolic pathways, an approach that can be extrapolated to determine the specific triggers for other locally disrupted processes.

By combining the information obtained from the previous metabolic profiling with advanced data analyses, distinct metabolic profiles can be obtained rather than only specific metabolic markers for each studied condition. Here we took advantage of the chemometric approach MCR-ALS²³ (Multivariate Curve Resolution Alternating Least Squares) to capture the synergistic metabolic patterns for every studied condition. In fluxomics, MCR-ALS has been applied to study reaction contributions of known pathways²⁴. In metabolomics, MCR-ALS has been already applied to resolve most of the metabolites from complex HPLC-MS samples¹³, as well as to identify directly the biological profiles from a raw ¹H NMR dataset²⁵. While the use of raw data may allow a more holistic approach to detect unknown changes on the metabolome profile, the use of the concentration areas estimates obtained from peak integration, as it is presented here, can simplify the interpretation of the results.

Results

Kinetic analysis of raw ¹H NMR data. Lack of each one of different auxotrophic markers resulted in a reduction of cell growth, probably reflecting the depletion of the respective internal pools (Fig. 1a). In this context, the drop-out medium lacking L-leucine (Leu-DM) appeared as the most restrictive condition, as it reached only 140% of the original OD after 24 h of culture (Fig. 1a, cyan line). Conversely, the medium lacking L-methionine (Met-DM) allowed growth to reach almost 200% of the original OD after 24 h (Fig. 1a, green line). In contrast, growing in non-restrictive conditions (YSC medium) allowed a vigorous growth for more than 10 hours, reaching more than 250% of the original OD before entering in the stationary phase, likely imposed by the consumption of the available fermentable carbon source (Fig. 1a, purple line).

Metabolic changes during the 24 h-incubation periods in all five conditions were first analyzed by a global overview of the ¹H NMR spectral dataset (Fig. 1b). The data show examples of proton resonances increasing (red arrows) or decreasing (green arrows) during yeast growth that are characteristic for each physiological condition tested.

¹H NMR spectra were also analyzed by Principal Component Analysis (PCA). Two components explained 61.6% of the total variability, most of it associated to metabolomic changes in yeast cultured in His-DM and Ura-DM (Fig. 1c, blue and red symbols), but only for samples taken at least after four hours of incubation. We interpret this as indicating accumulation of specific metabolites over time for these two particular auxotrophic starvation conditions. In contrast, a similar analysis performed only for samples from the YSC medium showed a quasi-cyclic variation of the yeast metabolome, in which samples taken at the late stationary phase (24 h) were more similar to the initial inoculum (0 h) than samples taken during the exponentially growing (log) phase (2–10 h) (Supplementary Fig. S1, see also purple symbols in Fig. 1c).

The temporal changes of yeast metabolome under the different starving conditions can be better observed on the four score plots in Fig. 2, corresponding to PCA analyses combining data from each starving condition and the control samples. These plots show the relatively small temporal variability of Leu-DM or Met-DM samples, particularly when compared to their Ura-DM or His-DM counterparts.

Permutation tests using the ASCA method (see *Material and Methods* section), confirmed a significant interaction ($p \leq 0.018$) between *time* and *yeast medium*, either for the complete dataset or when using partial datasets, including any of the drop-out media versus control samples, confirming the temporal metabolome variability associated to the different growth conditions (Supplementary Fig. S2).

Metabolite assignment and quantification. An exhaustive assignment process (see *Material and Methods* and Supplementary Methods) for the resonances from the NMR spectra allowed for the identification and determination of a total of 47 metabolites. In addition, concentrations from three additional peak resonances were estimated but not unequivocally assigned. Tentative candidates for these three metabolites were deduced from their respective chemical shifts (2.10 ppm, 8.03 ppm and 8.37 ppm) and multiplicities (singlet for all the cases). We propose that the first signal corresponds to a methyl donor of structure R-S-CH₃, whereas the remaining two correspond to modified purine rings with only one detectable proton, such as isoguanine or xanthine. A table containing the list of metabolites with the identified features in the spectrum is presented in Supplementary

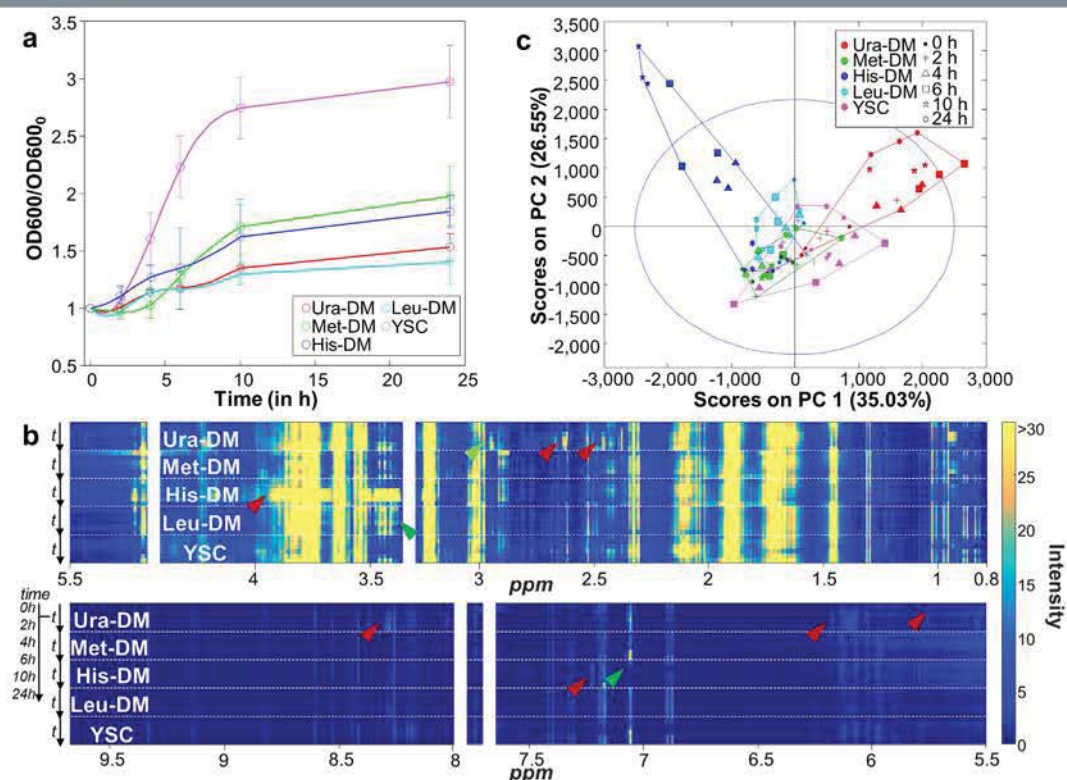


Figure 1. Yeast growth. (a) Graphic representation of cell growth at the different media (normalized for the initial OD_{600}). (b) Heatmap representation of the ^1H NMR spectra for the five different time-courses. Each row corresponds to the average of 3 spectra. Within each time-course studied, data are presented row-wise following an increasing time order. The red and green arrows point to peak signals of some metabolites accumulated (increasing intensity) or consumed (decreasing intensity) over time. (c) PCA scores projection of the mean-centered spectral dataset on PC1 and PC2 subspace. DM, Drop-out Media; YSC, Yeast Nitrogen Base Synthetic Complete medium.

Table S1, whereas relative concentration plots are presented in Supplementary Fig. S3. A biological overview of the main interconnections for these metabolites in yeast can be found in Fig. 3.

Hierarchical clustering of the auto-scaled concentration estimates defined three clusters: one corresponding to metabolites accumulated in the lack of uracil (Ura-DM), a second, less defined one, including metabolites accumulated in the lack of L-histidine (His-DM), and the last one including the remaining metabolites (Fig. 4). Close inspection of the individual profiles shows the non-consumption of metabolites in Leu-DM medium and quasi-cyclic variations for some metabolites (see for example L-methionine, 2-isopropylmalate and L-Tyrosine) in YSC and also for some of the auxotrophic starvation conditions tested.

Metabolome variations during growth. Estimated concentration changes from proton resonances were analyzed using MCR-ALS (see *Materials and Methods* and Supplementary Methods). Four temporal components, \mathbf{t}_1 – \mathbf{t}_4 , associated to four metabolic profiles, \mathbf{m}_1 – \mathbf{m}_4 , were obtained from this analysis, with an explained data variance of 85.7%. \mathbf{t}_1 – \mathbf{t}_4 temporal components for each experimental condition are presented in Fig. 5a–e, whereas the \mathbf{m}_1 – \mathbf{m}_4 metabolic profiles associated to each temporal profile are represented in the heatmap of Fig. 5f.

Most of the metabolic variability of the yeast metabolome during unrestricted growth (YSC, Fig. 5a) could be explained by only two MCR-ALS components (YSC, Fig. 5a). In addition, as observed in this figure, \mathbf{t}_1 and \mathbf{t}_2 temporal components practically mirror one each other: Component \mathbf{t}_1 (blue dots and lines in Fig. 5a) peaked after 2–6 h of incubation, coinciding with the period of maximal growth, precisely the same time point at which component \mathbf{t}_2 (red dots and lines in Fig. 5a) showed a minimum. We thus assign the corresponding metabolic profiles (\mathbf{m}_1 and \mathbf{m}_2) to exponential and lag growth phases, respectively. Analysis of the metabolites associated to each of these two components revealed that \mathbf{m}_1 has strong contribution of L-methionine, L-leucine, fatty acids, uracil precursors (orotate and orotidine-5P), and AMP. On the other side, \mathbf{m}_2 shows strong contributions of amino acids and amino acid precursors (2-isopropylmalate and 3-hydroxyisobutyrate), citrate and trehalose, among others (Fig. 5f).

MCR-ALS results show completely different metabolic dynamic processes for each of the tested conditions, suggesting different patterns of arrest of cell growth depending on the missing auxotrophic marker. Samples grown in Leu-DM hardly showed any change in metabolic concentrations over the measured time, and their main contribution was for the temporal component \mathbf{t}_2 , related to the stationary phase (Fig. 5e). In Ura-DM conditions (Fig. 5b), components \mathbf{t}_1 and \mathbf{t}_2 showed significant contributions only during the first two hours of

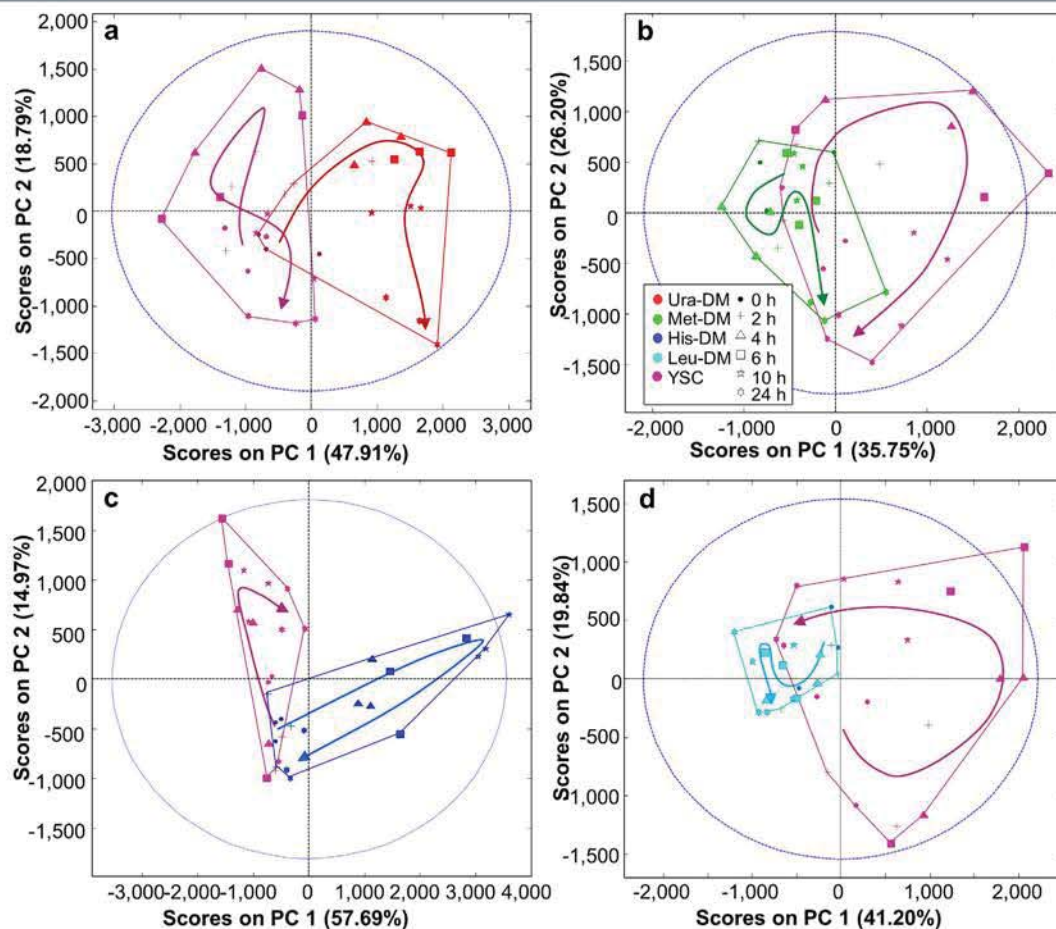


Figure 2. PCA analyses of the metabolomic variance for each starved condition. PC1 and PC2 scores projection of control samples and of samples from yeast cultured in (a) Ura-DM, (b) Met-DM, (c) His-DM, and in (d) Leu-DM.

incubation, whereas the specific temporal component t_3 increased over time, peaking at 10 h and remaining as the major component for the rest of the analyzed period. A similar pattern was observed for His-DM samples (Fig. 5d), in which the specific component t_4 , peaked at 4–10 h, slightly decreasing afterwards. Analyses of the major metabolite contributors to the corresponding temporal profiles revealed a high contribution of precursors of uracil and also of other purine-related molecules to m_3 and an equally strong contribution of D-glucose, erythro-imidazole-glycerol phosphate (EIGP), L-ornithine, L-proline and L-lactic acid to m_4 (Fig. 5d). The contribution of the different components to the variability of Met-DM samples was more complex, with relevant contribution of t_1 , t_2 and t_4 during the first 10 h, followed by a strong increase of the stationary phase-related component t_2 , which became predominant at the end of the incubation (Fig. 5c).

Starvation-induced metabolomic changes can be interpreted under the point of view of the biochemical pathways interrupted by the corresponding gene disruptions (Fig. 6). This analysis reveals a decrease in concentration of metabolites downstream the disrupted gene under all four starvation conditions and a parallel increase of upstream genes in at least three of them (Uracil, L-histidine and L-leucine). Note that these changes were condition-specific, that is, they only affected the metabolic pathway related to each particular starvation condition. Therefore, these data suggest a specific regulation for each metabolic pathway, at least at the metabolic level. Other condition-specific changes indicate far-ranging effects of the auxotrophic starvation. For example, different purine-related metabolites (i.e. AMP, hypoxanthine and N⁶-methyladenosine) were also accumulated in Ura-DM, suggesting that both purine and pyrimidine pools were affected due to the lack of uracil in the medium. N⁶-methyladenosine, which is the most common internal mRNA modification in eukaryotes²⁶, was first detected in yeast under sporulation conditions²⁷. The increasing presence of this compound in the Ura-DM extracts suggests that the lack of uracil induces mRNA degradation, likely to increase the diminishing uracil reservoirs to enhance cell survival under this limiting condition. Another example of the effects of auxotrophic starvation upon apparently unrelated metabolic pathways is the increase of intracellular D-glucose in histidine-starving cells, the meaning of which is unclear at the present.

Discussion

There are two characteristics of ¹H NMR analysis that limit its use for metabolomic studies. On one hand, it is assumed that in terms of sensitivity (limit of detection), NMR clearly lags behind chromatographic/MS-based

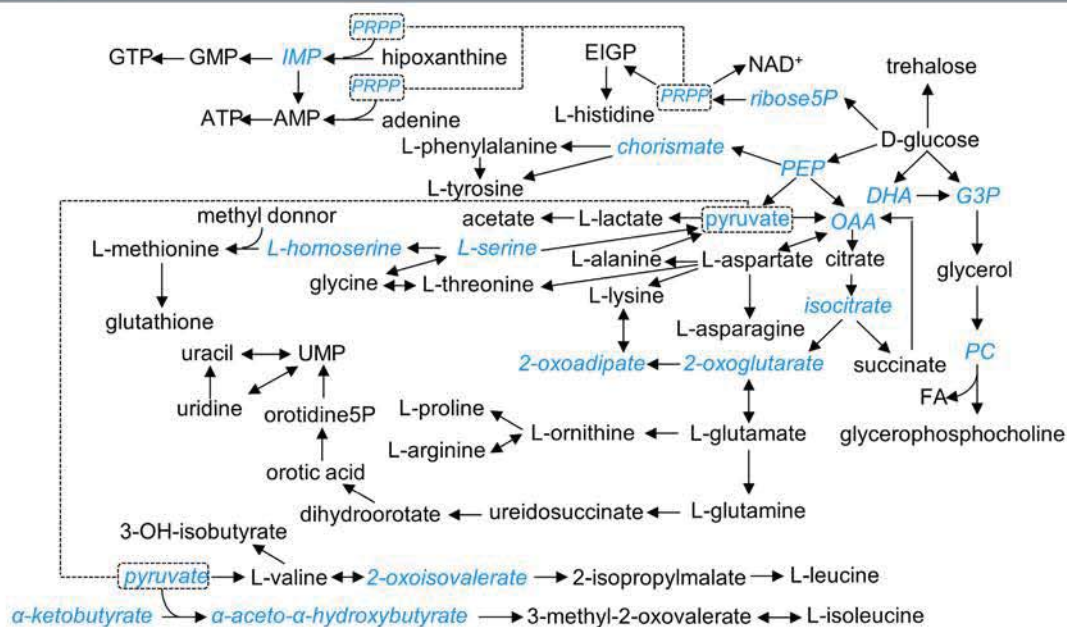


Figure 3. Pathway diagram representing the main interconnections for the assigned metabolites. Assigned metabolites are written in black, whereas non-assigned ones are written in blue italic letters. Solid arrows connect metabolites from a same metabolic pathway, showed here in a simplified way. Dashed arrows connect different pathways sharing a same metabolite.

methods. Second, the intrinsic complexity of the ¹H NMR spectra due to overlapping makes very complicated to evaluate complex mixtures of a wide variety of molecules present at very different molar concentrations, such as in metabolite extracts. Although the sensitivity and the resolution aspect is an instrumental problem that can be tackled by many technical improvements (higher magnetic fields, low temperature probes, etc.), the dynamic range problem²⁸ in NMR still prevents the detection of trace compounds. Despite these limitations, NMR is a convenient choice for performing screening metabolomics-based studies, since the compounds present in the studied samples can be unambiguously identified and robustly quantified due to the inherent particularities of this technique.

In this manuscript, we applied PCA and ASCA to first analyze the raw ¹H NMR spectral dataset, and MCR-ALS on the metabolite concentrations. With these three chemometric methods, the metabolic responses of yeast at four starving conditions have been characterized (summarized in Table 1). Finally, we used all this information to interpret the observed variations in a biochemical context.

Chemometric analysis of the complete ¹H NMR spectral dataset detected specific variations in yeast metabolomic profiles as a response to different auxotrophic starvation conditions, without requiring the identification of the metabolites implicated in these changes. The analysis allows drawing some a priori unexpected conclusions, like that single amino acid starvation (e.g., DM-Leu and DM-His samples) may trigger very different responses in the yeast metabolome, or that yeast growth efficiency is not correlated to the metabolic variance. These conclusions can be easily deduced from Figs 1 and 2, simply by comparing growth curves with PCA score plots for the different growing conditions. The same analysis reveals that metabolomic differences between dropout and control cultures tended to increase over time, and that this divergence is maximal in the case of Ura-DM samples. In contrast, Met-DM and Leu-DM samples showed relatively little changes during the 24-hour incubation period, compared to the other two drop-out media and even to the control samples (Fig. 3). On the other hand, the interaction between the two factors, *time* and *yeast medium*, confirmed by ASCA, indicates that the lack of auxotrophic markers alters the velocity and duration of the different yeast growth events.

Therefore, either by using PCA or ASCA, the analysis of ¹H NMR data allowed the characterization and evaluation of the physiological conditions of a cultured organism. Qualitative metabolite concentration changes can be extracted from the loading plots associated to PCA analysis, as shown elsewhere²⁹, although they can be also observed by a simpler heatmap representation, as in Fig. 1.

We interpret the metabolic changes observed in yeast under unrestricted conditions (components \mathbf{m}_1 and \mathbf{m}_2) as reflecting the alternation between the stationary and the exponential growth phases, induced by the presence of high concentrations of glucose in the medium and regulated by the PP2A/TOR signalling pathway³⁰. These results are also consistent with the known relative decrease of oxidative metabolism (including respiration) in yeast at high glucose conditions³¹. The enhanced growth observed in the exponential phase is characterized by a decrease in the amino acid pools (\mathbf{m}_2 in Fig. 5f) and with an increase of the transcription machinery (\mathbf{m}_1 in Fig. 5f). On the other side, the metabolic transition at the beginning of the stationary phase (10 h and 24 h) is marked by the increase of the levels of trehalose, a resistance metabolite, and of some amino acids (\mathbf{m}_2 in Fig. 5f). This is probably caused by a decrease in the translation activity and linked to the restoration of the internal amino acid pools^{32,33}.

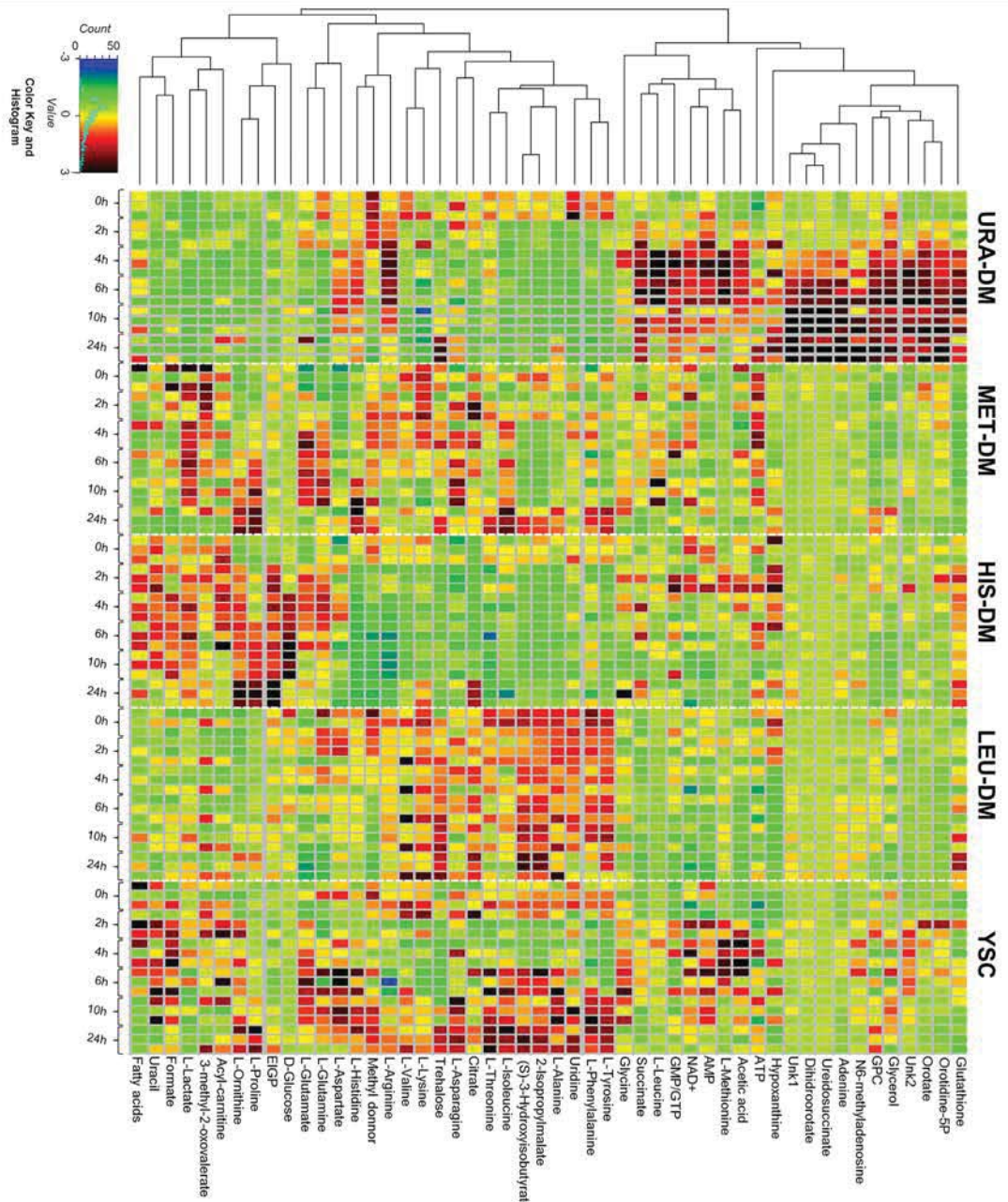


Figure 4. Heatmap of the auto-scaled concentration estimates for all assigned metabolites. Metabolites were clustered using the Pearson method. All individual samples (including replicates) were included.

Similarly, the metabolic variations associated to the corresponding auxotrophic starvation (components m_3 and m_4) can be easily interpreted as the primary metabolic response from depleting the missing metabolites (i.e. uracil in Ura-DM), consisting on the accumulation of the precursors generated at the metabolic steps immediately upstream them (i.e. EIGP to His3p and Orotidine-5P to Ura3p), (Fig. 6). It is important to note that this analysis allowed as well the detection of a priori unexpected metabolome variations, like the purines accumulation after inducing pyrimidine (uracil) starvation, or the accumulation of D-glucose in histidine-starving conditions.

It is important to note that the use of MCR-ALS for studying -omic data should be applicable to essentially any other biological sample aside of the model organism *Saccharomyces cerevisiae*. This includes population mixtures or non-model biological organisms in which conventional pathway analyses approaches (i.e. 'fluxomics') might be difficult to implement.

Perhaps the most striking observation of the metabolome analysis was the radical differences between the effects of L-histidine starvation on one side and of L-leucine and L-methionine starvation on the other side. Irrespectively of the effects for the global growth, Leu-DM and Met-DM cultures showed minimal metabolome

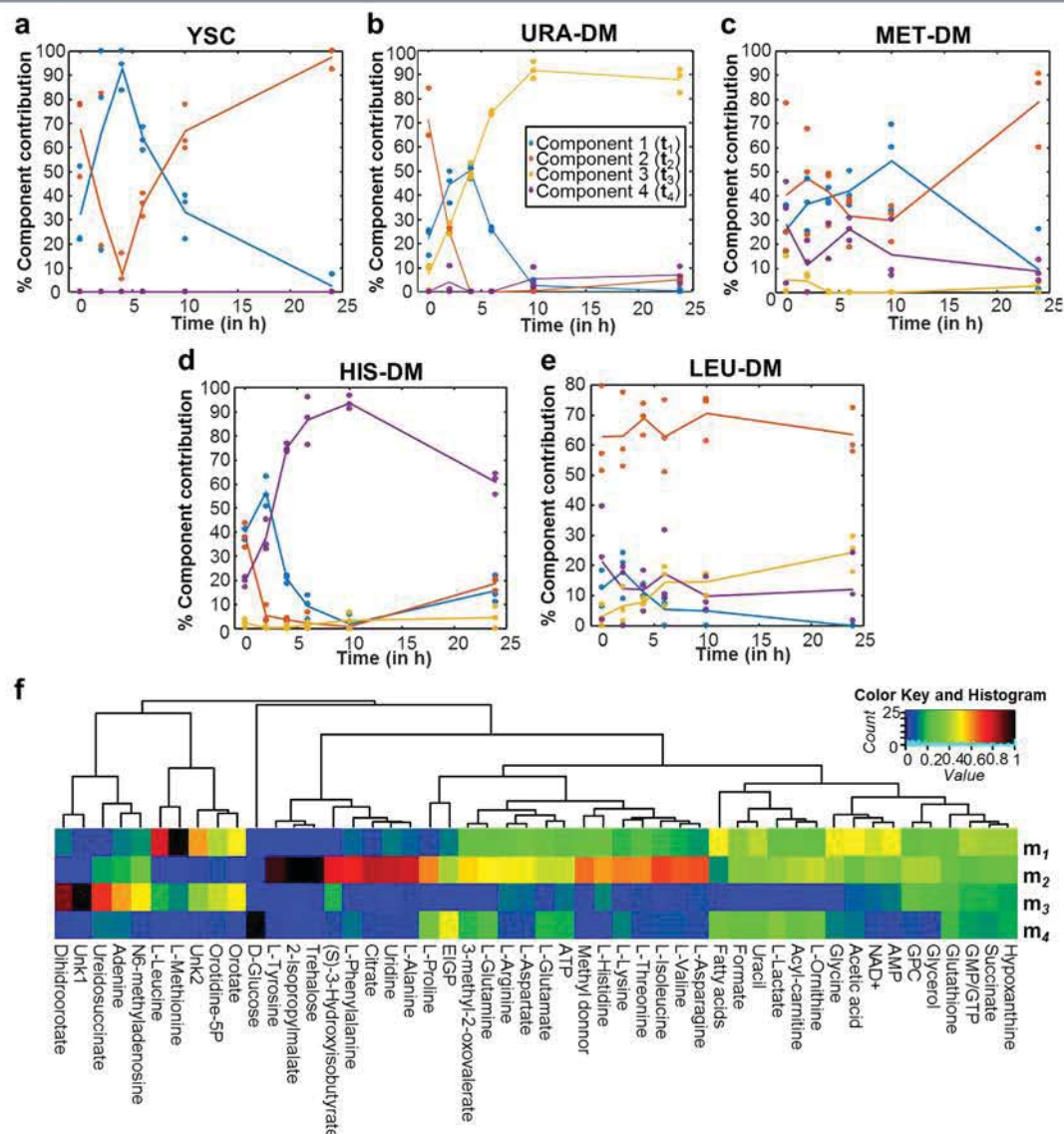


Figure 5. Growth pattern of yeast metabolism resolved by MCR-ALS using 4 components. (a–e) Temporal growth pattern (in %) of yeast cells cultured in YSC (a), URA-DM (b), Met-DM (c), His-DM (d) and Leu-DM (e) medium described by each MCR-ALS component. (f) Hierarchical clustering of the relative contribution of every metabolite in the 4 MCR-ALS resolved components given in (a–e) to every metabolite.

variations during the 24 h-incubation period (actually, even less variations than cells cultured under unrestricted conditions), whereas His-DM cells' metabolome steadily diverged from its initial composition upon time. This effect, observed both for the complete ¹H NMR spectral profiles and when identified metabolites were individually analyzed, suggests different regulatory pathways activated as a response to the different amino acids. Both L-leucine and L-methionine starvation have been linked to G0/G1 arrest through the PP2A/TOR signaling pathway, a mechanism fully compatible with our observations at the metabolome levels^{34–38}.

In this work, we used a genetically well-known model system to test the power of ¹H NMR to analyze metabolome effects of enzymatic pathway disruptions without any previous hypothesis nor anticipation of the nature of the possible metabolic pathway blockage. Analysis of the complete ¹H NMR spectral profiles allowed us to distinguish between those conditions blocking the entry into the cell cycle (Leu-DM and Met-DM) from those that allow the progression of at least some metabolic pathways (e.g., Ura-DM and His-DM), and this without the need of identifying the specific metabolic concentration changes. Identification of the altered metabolites under each starvation condition could in principle allow the identification of the affected metabolic pathway in all four cases, and pinpoint the disrupted enzymatic steps in at least two of them. We conclude that the proposed NMR metabolomics strategy can be useful for studying models of metabolomic disruption either with genetic defects or with enzymatic inhibitors in many biological systems, including the study of the molecular target of biocides or the metabolic response of malignant cells to antitumorals.

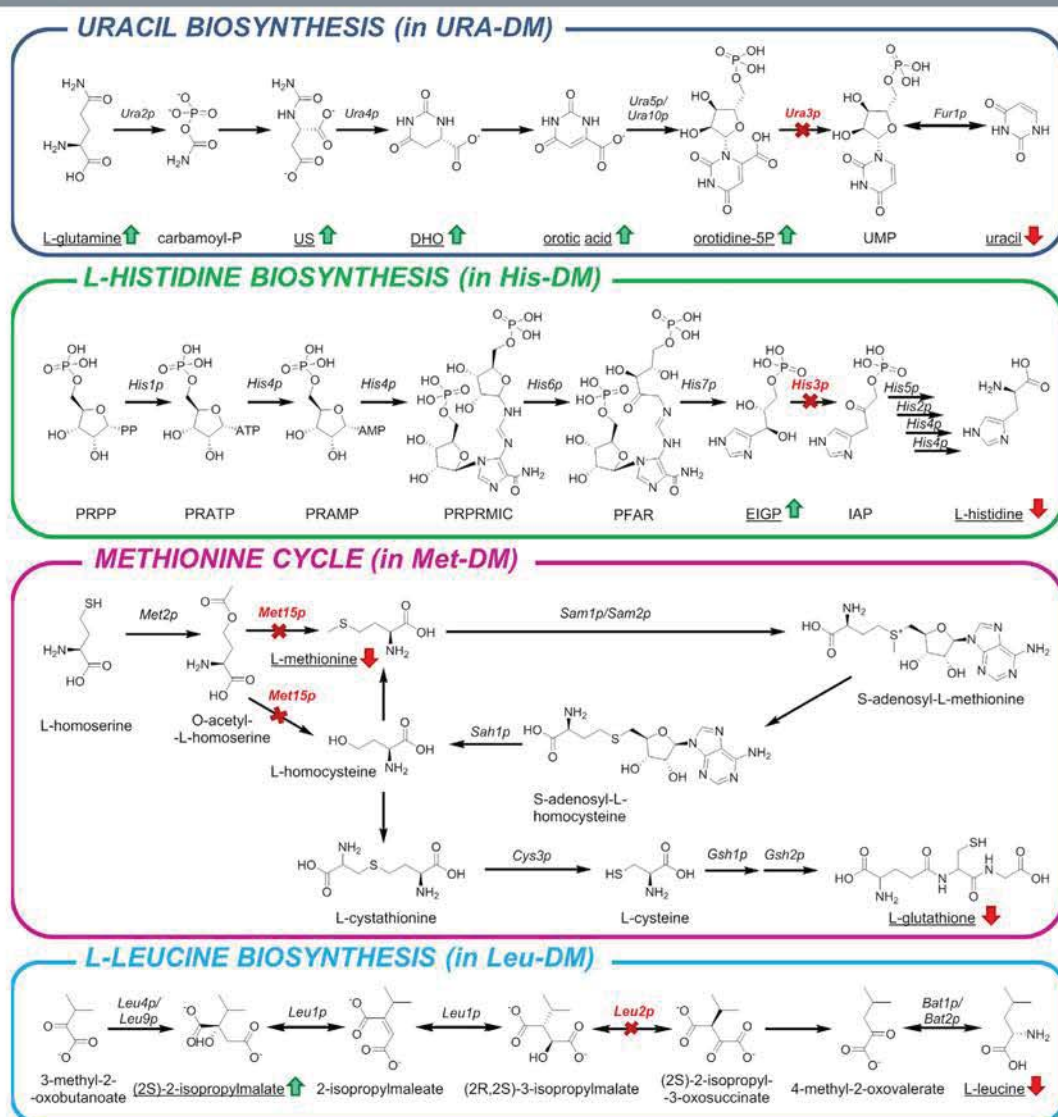


Figure 6. Disrupted metabolic pathways of the used auxotrophic yeast strain. Coloured arrows denote the accumulation (green) or consumption (red) of the detected metabolites in the associated DM medium. Red enzyme names and red crosses denote the depleted genes. Underlined metabolites were those detected and quantified by ^1H NMR. US: ureidosuccinate; DHO: dihydroorotate; UMP: uridine monophosphate; PRPP: 5-phospho- α -D-ribose PP; PRATP: 5-phospho- α -D-ribose ATP; PRAMP: 5-phospho- α -D-ribose AMP; PRPRMIC: 1-(5-phospho- β -D-ribose)-5-[(5-phosphoribosylamino) methylideneamino]imidazole-4-carboxamide; PFAR: phosphoribulosylformimino-AICAR-P; EIGP: D-erythro-imidazole-glycerol-phosphate; IAP: imidazole-acetol-phosphate.

Therefore, we consider that the synergy between NMR metabolomics and molecular biology can be very useful in genetic studies, such as transcriptomics, but may be also applicable in genetic diagnostics and drug development.

Methods

Yeast Growth. *S. cerevisiae* BY4741 (MATa; his3 Δ 1; leu2 Δ 0; met15 Δ 0; ura3 Δ 0) cells were pre-cultured in YPD (1% yeast extract, 1% peptone, 2% glucose) medium on an orbital shaker (150 rpm) at 30 °C overnight. All following cultures were cultured with these shaking and temperature conditions. 2 L of YNB Synthetic Complete medium (YSC, 1.7 g/L Yeast Nitrogen Base without amino acids and sulphate (Difco), 5 g/L $(\text{NH}_4)_2\text{SO}_4$) supplemented with the appropriated auxotrophic markers (4 mg/L uracil, 1 mg/L leucine, 1 mg/L histidine, and 1 mg/L methionine, as requested) were inoculated with 200 μl of the pre-culture sample and left at the same temperature and shaking conditions until the culture reached an optical density at 600 nm (OD_{600}) of approximately 0.8–1. Pellets from these resulting cultures were collected by centrifuging the cultures, but not washed, at 2000 rpm for 3 min and 4 °C. Pellets were used right after for inoculating erlenmeyers containing either YSC medium (control)

Starving condition	Related gene	Cell growth (OD ₆₀₀)	Metabolic variation over time (PCA)	Effect in cell growth timing (ASCA)	Metabolic response (MCR-ALS)
Uracil	URA3	153%	↑↑	YES (p ≤ 0.0001)	Enhanced biosynthesis of uracil precursors and purine-related compounds (m ₃ metabolic profile)
L-methionine	MET15	197%	↑	YES (p = 0.0180)	Prolongation of log phase
L-histidine	HIS3	184%	↑↑	YES (p ≤ 0.0001)	Enhanced biosynthesis of L-histidine precursors (m ₄ metabolic profile)
L-leucine	LEU3	140%	—	YES (p ≤ 0.0001)	Cell cycle arrest

Table 1. Summary of the responses observed in starved yeast cultures.

or drop-out media missing either L-leucine, L-histidine, uracil or L-methionine, up to a final OD₆₀₀ of 0.4–0.5. Resulting cultures were grown at 30 °C and 150 rpm.

Sample collection. 100 ml aliquots of every culture were collected six times during one day (0 h, 2 h, 4 h, 6 h, 10 h and 24 h). Samples were arrested with a cold shock in ice and cell were harvested by centrifugation at 4000 g for 3 min, discarding the supernatant. Cells were washed twice in Na₂HPO₄ 100 mM pH 7.0 followed by a centrifugation at 4700 g for 3 min. Resulting pellets were stored at –80 °C and lyophilized. Cell density was calculated as OD₆₀₀, and viable cell counted by plating culture dilutions in YPD agar plates (1% yeast extract, 1% peptone, 2% agar, 2% glucose) for each sample.

Metabolite extraction. Metabolites were extracted by following the protocol published in a previous work²⁰. 1800 μl of a solution of methanol-chloroform 1:2 (4 °C) were added to the pellet, followed by a vigorous vortexing. A cold shock is then applied to the pellets for 5 times using the following procedure: the pellets are submerged in liquid nitrogen for 1 minute and consequently thawing in ice for 2 minutes. 400 μl of water are added to create the biphasic system. After homogenization by vortexing, a 3 min centrifugation at 16500 rpm and 4 °C is carried out. The aqueous phase (upper part) is collected. This process is repeated and water and methanol are removed from the aqueous phase in a speedvac.

NMR sample preparation. Aqueous samples were dissolved in 700 μl of deuterated phosphate buffer (Na₂DPO₄ 100 mM, pH 7.0) in D₂O with DSS 0.2 mM as internal standard.

¹H NMR experiments. Spectra were recorded in a 400 MHz Varian spectrometer, using a spectrometer frequency of 400.14 MHz with a OneNMR Probe and a ProTune System (Agilent). Spectral size range covered from –2 to 10 ppm. Receiver gain was fixed to 34. Also, 512 scans were used with a relaxation delay of 5 seconds. Spectral size contained 65 k data points, and the acquired size was made of 32 k complex data points.

NMR spectra preprocessing. Spectra were preprocessed with MestreNova v.10.0 (Mestrelab Research, Spain). Spectra preprocessing consisted in an exponential apodization of 0.5 Hz, a manual phasing and a baseline correction with Bernstein polynomial of 3rd order. After adjusting the reference to DSS, water (4.41–5.16 ppm), methanol (3.30–3.37 ppm), chloroform (7.64–7.69 ppm) and DSS (<0.8 ppm) regions were removed. Data points which chemical shifts were higher than 10.3 ppm were also removed. The final NMR dataset consisted on a data matrix of 90 spectra (rows) having 35,342 ppm values (columns) each one. This data matrix was stored in ASCII file format.

Metabolite identification. Metabolite assignment was performed by a detailed targeted metabolite profiling analysis of the ¹H NMR signals using a home-made ¹H NMR spectra library²⁰ and also the Yeast Metabolome Data Base library³⁹ (YMDB). Proton correlations were checked on gCOSY spectra. Additional metabolite confirmations were performed using complementary NMR pulse sequences (see Supplementary Methods). Pathway diagram data was obtained from *BioCyc* database⁴⁰.

Metabolite quantification. Relative metabolite quantifications of the ¹H NMR spectral matrix were performed using BATMAN R⁴¹ package. Further information about how Batman works and the exact protocol can be found elsewhere^{20,42}.

Chemometric data analysis of the NMR dataset. ¹H NMR spectra from the previous preprocessing steps were imported to Matlab R2014a (The Mathworks Inc. Natick, MA, USA) and analysed with the PLS toolbox 7.8.0 (Eigenvector Research Inc., Wenatchee, WA, USA). In order to eliminate sample size effects, ¹H NMR spectra were normalized using the Probabilistic Quotient Normalization (PQN)⁴³ method, taking only into account the region of 0.8–3.8 ppm. A reference spectrum was used for every time and tested condition, consisting on the average of all spectra measured at the same conditions in the previous sample collection time, except for time 0 h spectra, where the reference spectrum used was obtained using all samples measured at 0 h, regardless of the tested condition studied.

Chemical shift corrections were performed using the icoshift⁴⁴ algorithm in the 7.02–7.14, 7.72–7.77, 8.51–8.55, 8.56–8.63 and 9.09–9.17 ppm regions. For Principal Component Analysis (PCA) and Analysis of variance of Simultaneous Component Analysis (ASCA), the 7.78–8.00 region was ignored.

PCA was applied to the mean-centered PQN-normalized NMR spectral data matrix (dimensions of 90×34443 data points), using Cross-Validation with Venetian Blinds to find the most reliable number of components to be included in the models.

ASCA^{45,46} was applied to the NMR spectral dataset to evaluate the effects of each experimental factor (*time* and *medium* in this study). In order to check the statistical significance of the effects of the investigated factors and of their possible interactions, a permutation test was performed^{47,48}. In this study, the number of permutations was set at 10000. Before performing ASCA, matrix dimensions were reduced by taking every 1 out of 10 values, thus obtaining a reduced size data matrix with 90×3445 values. In contrast to PCA, data used were not mean-centered.

Chemometric data analysis of the concentration profiles. In order to resolve the metabolic patterns that evolve over time, the Multivariate Curve Resolution by Alternating Least Squares⁴⁹ (MCR-ALS) method has been used. Data analysis was performed using the MCR-ALS GUI 2.0²³ under Matlab 2014b (The Mathworks Inc. Natick, MA, USA) environment.

MCR-ALS is a chemometric method which decomposes a data matrix using the following bilinear model:

$$\mathbf{X} = \mathbf{TM}^T + \mathbf{E} \quad (1)$$

In this particular case, the data matrix \mathbf{X} (size $I \times J$) has the concentrations of the J metabolites obtained by integrating their corresponding proton resonances, in the I yeast samples, cultured in a particular medium during a particular time period. This matrix bilinear decomposition gives two factor matrices, \mathbf{M}^T and \mathbf{T} , the matrix of metabolic profiles, \mathbf{M}^T ($N \times J$), and the matrix of temporal profiles, \mathbf{T} ($I \times N$). Each metabolic profile in \mathbf{M} shows its metabolite composition, whereas \mathbf{T} shows the contribution of each individual metabolic profile in every sample at different collection times. N represents the number of components used in the decomposition generated in the MCR-ALS analysis. N can be initially estimated by using a singular value decomposition (SVD)⁵⁰. \mathbf{E} matrix (size $I \times J$) contains the residual information not explained by the model using the N considered components. A more detailed description of this method can be found in Supplementary Methods.

The quality of the MCR-ALS model was measured evaluating the lack-of-fit parameter, which is expressed by the percent of explained variance (R^2)²³.

On the other hand, hierarchical cluster analysis of concentration estimates and of the \mathbf{M} matrices was performed. Clustering was performed using the heatmap.2 function from the gplots R package⁵¹ using the complete agglomeration method for clustering. Metabolite concentration estimates were auto-scaled before analysis, whereas heatmap representations of \mathbf{M} matrices had the MCR-ALS results without any data pretreatment.

References

- German, J. B., Hammock, B. & Watkins, S. Metabolomics: building on a century of biochemistry to guide human health. *Metabolomics* **1**, 3–9, doi: 10.1007/s11306-005-1102-8 (2005).
- Bedia, C., Dalmau, N., Jaumot, J. & Tauler, R. Phenotypic malignant changes and untargeted lipidomic analysis of long-term exposed prostate cancer cells to endocrine disruptors. *Environ. Res.* **140**, 18–31, doi: http://dx.doi.org/10.1016/j.envres.2015.03.014 (2015).
- Gorochategui, E., Casas, J., Porte, C., Lacorte, S. & Tauler, R. Chemometric strategy for untargeted lipidomics: Biomarker detection and identification in stressed human placental cells. *Anal. Chim. Acta* **854**, 20–33, doi: http://dx.doi.org/10.1016/j.aca.2014.11.010 (2015).
- Huang, S.-M., Xu, F., Lam, S. H., Gong, Z. & Ong, C. N. Metabolomics of developing zebrafish embryos using gas chromatography-and liquid chromatography-mass spectrometry. *Mol. Biosyst.* **9**, 1372–1380, doi: 10.1039/C3MB25450J (2013).
- Nagato, E. G. *et al.* 1H NMR-based metabolomics investigation of Daphnia magna responses to sub-lethal exposure to arsenic, copper and lithium. *Chemosphere* **93**, 331–337 (2013).
- 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.*, 1–13, doi: 10.1007/s00216-015-9042-2 (2015).
- Ortiz-Villanueva, E. *et al.* Combination of CE-MS and advanced chemometric methods for high-throughput metabolic profiling. *Electrophoresis* **36**, 2324–2335, doi: 10.1002/elps.201500027 (2015).
- Halter, D. *et al.* *In situ* proteo-metabolomics reveals metabolite secretion by the acid mine drainage bio-indicator, Euglena mutabilis. *ISME J.* **6**, 1391–1402, doi: 10.1038/ismej.2011.198 (2012).
- Hossain, S. M. Z., Bojko, B. & Pawliszyn, J. Automated SPME-GC-MS monitoring of headspace metabolomic responses of *E. coli* to biologically active components extracted by the coating. *Anal. Chim. Acta* **776**, 41–49, doi: http://dx.doi.org/10.1016/j.aca.2013.03.018 (2013).
- Hamerly, T. *et al.* Untargeted metabolomics studies employing NMR and LC-MS reveal metabolic coupling between Nanoarchaeum equitans and its archaeal host Ignicoccus hospitalis. *Metabolomics* **11**, 895–907, doi: 10.1007/s11306-014-0747-6 (2015).
- Ro, D.-K. *et al.* Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* **440**, 940–943, doi: http://www.nature.com/nature/journal/v440/n7086/supinfo/nature04640_S1.html (2006).
- Torija, M. a. J. *et al.* Effects of fermentation temperature and Saccharomyces species on the cell fatty acid composition and presence of volatile compounds in wine. *Int. J. Food Microbiol.* **85**, 127–136, doi: http://dx.doi.org/10.1016/S0168-1605(02)00506-8 (2003).
- Farrés, M., Piña, B. & Tauler, R. Chemometric evaluation of Saccharomyces cerevisiae metabolic profiles using LC-MS. *Metabolomics* **11**, 210–224, doi: 10.1007/s11306-014-0689-z (2015).
- Klosinska, M. M., Crutchfield, C. A., Bradley, P. H., Rabinowitz, J. D. & Broach, J. R. Yeast cells can access distinct quiescent states. *Genes Dev.* **25**, 336–349, doi: 10.1101/gad.2011311 (2011).
- Rodkaer, S. V. *et al.* Quantitative proteomics identifies unanticipated regulators of nitrogen- and glucose starvation. *Mol. Biosyst.* **10**, 2176–2188, doi: 10.1039/C4MB00207E (2014).
- Natarajan, K. *et al.* Transcriptional Profiling Shows that Gcn4p Is a Master Regulator of Gene Expression during Amino Acid Starvation in Yeast. *Mol. Cell Biol.* **21**, 4347–4368, doi: 10.1128/mcb.21.13.4347-4368.2001 (2001).
- Cebollero, E. & Reggiori, F. Regulation of autophagy in yeast Saccharomyces cerevisiae. *BBA-Mol. Cell Res.* **1793**, 1413–1421, doi: http://dx.doi.org/10.1016/j.bbamer.2009.01.008 (2009).
- Lafaye, A. *et al.* Combined Proteome and Metabolite-profiling Analyses Reveal Surprising Insights into Yeast Sulfur Metabolism. *J. Biol. Chem.* **280**, 24723–24730, doi: 10.1074/jbc.M502285200 (2005).

19. Urbanczyk-Wochniak, E. *et al.* Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Rep.* **4**, 989–993, doi: 10.1038/sj.embor.embor944 (2003).
20. Puig-Castellví, F., Alfonso, I., Piña, B. & Tauler, R. A quantitative ¹H NMR approach for evaluating the metabolic response of *Saccharomyces cerevisiae* to mild heat stress. *Metabolomics* **11**, 1612–1625, doi: 10.1007/s11306-015-0812-9 (2015).
21. Patti, G. J., Yanes, O. & Siuzdak, G. Innovation: Metabolomics: the apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.* **13**, 263–269 (2012).
22. Szeto, S. S. W., Reinke, S. N., Sykes, B. D. & Lemire, B. D. Mutations in the *Saccharomyces cerevisiae* Succinate Dehydrogenase Result in Distinct Metabolic Phenotypes Revealed Through ¹H NMR-Based Metabolic Footprinting. *J. Proteome Res.* **9**, 6729–6739 (2010).
23. Jaumot, J., de Juan, A. & Tauler, R. MCR-ALS GUI 2.0: New features and applications. *Chemometrics Intell. Lab. Syst.* **140**, 1–12, doi: http://dx.doi.org/10.1016/j.chemolab.2014.10.003 (2015).
24. Folch-Fortuny, A. *et al.* MCR-ALS on metabolic networks: Obtaining more meaningful pathways. *Chemometrics and Intelligent Laboratory Systems* **142**, 293–303, doi: http://dx.doi.org/10.1016/j.chemolab.2014.10.004 (2015).
25. Karakach, T. K., Knight, R., Lenz, E. M., Viant, M. R. & Walter, J. A. Analysis of time course ¹H NMR metabolomics data by multivariate curve resolution. *Magn. Reson. Chem.* **47**, S105–S117, doi: 10.1002/mrc.2535 (2009).
26. Desrosiers, R., Friderici, K. & Rottman, F. Identification of Methylated Nucleosides in Messenger RNA from Novikoff Hepatoma Cells. *Proceedings of the National Academy of Sciences of the United States of America* **71**, 3971–3975 (1974).
27. Flancy, M. J., Shambaugh, M. E., Timpte, C. S. & Bokar, J. A. Induction of sporulation in *Saccharomyces cerevisiae* leads to the formation of N(6)-methyladenosine in mRNA: a potential mechanism for the activity of the IME4 gene. *Nucleic Acids Research* **30**, 4509–4518 (2002).
28. Davies, S., Bauer, C., Barker, P. & Freeman, R. The dynamic range problem in NMR. *Journal of Magnetic Resonance (1969)* **64**, 155–159, doi: http://dx.doi.org/10.1016/0022-2364(85)90045-9 (1985).
29. Griffin, J. L. Metabonomics: NMR spectroscopy and pattern recognition analysis of body fluids and tissues for characterisation of xenobiotic toxicity and disease diagnosis. *Curr. Opin. Chem. Biol.* **7**, 648–654 (2003).
30. Galdieri, L., Mehrotra, S., Yu, S. & Vancura, A. Transcriptional Regulation in Yeast during Diauxic Shift and Stationary Phase. *OMICS* **14**, 629–638, doi: 10.1089/omi.2010.0069 (2010).
31. Verduyn, C., Zomerdijk, T. L., van Dijken, J. & Scheffers, W. A. Continuous measurement of ethanol production by aerobic yeast suspensions with an enzyme electrode. *Appl. Microbiol. Biotechnol.* **19**, 181–185, doi: 10.1007/BF00256451 (1984).
32. Hans, M. A., Heinzle, E. & Wittmann, C. Free intracellular amino acid pools during autonomous oscillations in *Saccharomyces cerevisiae*. *Biotechnol. Bioeng.* **82**, 143–151, doi: 10.1002/bit.10553 (2003).
33. Fuge, E. K., Braun, E. L. & Werner-Washburne, M. Protein synthesis in long-term stationary-phase cultures of *Saccharomyces cerevisiae*. *J. Bacteriol.* **176**, 5802–5813 (1994).
34. Binda, M. *et al.* The Vam6 GEF Controls TORC1 by Activating the EGO Complex. *Molecular Cell* **35**, 563–573, doi: http://dx.doi.org/10.1016/j.molcel.2009.06.033 (2009).
35. Saldanha, A. J., Brauer, M. J. & Botstein, D. Nutritional Homeostasis in Batch and Steady-State Culture of Yeast. *Mol. Biol. Cell* **15**, 4089–4104, doi: 10.1091/mbc.E04-04-0306 (2004).
36. Boer, V. M., Amini, S. & Botstein, D. Influence of genotype and nutrition on survival and metabolism of starving yeast. *Proc. Natl. Acad. Sci. USA* **105**, 6930–6935, doi: 10.1073/pnas.0802601105 (2008).
37. Laxman, S., Sutter, B. M. & Tu, B. P. Methionine is a signal of amino acid sufficiency that inhibits autophagy through the methylation of PP2A. *Autophagy* **10**, 386–387, doi: 10.4161/auto.27485 (2013).
38. Sutter, B. M., Wu, X., Laxman, S. & Tu, B. P. Methionine Inhibits Autophagy and Promotes Growth by Inducing the SAM-Responsive Methylation of PP2A. *Cell* **154**, 403–415, doi: 10.1016/j.cell.2013.06.041 (2013).
39. Jewison, T. *et al.* YMDB: the Yeast Metabolome Database. *Nucleic Acids Res.* **40**, D815–D820, doi: 10.1093/nar/gkr916 (2012).
40. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* **42**, D459–D471, doi: 10.1093/nar/gkt1103 (2014).
41. R Core Team. *R: A language and environment for statistical computing*. (R Foundation for Statistical Computing, 2013).
42. Hao, J. *et al.* Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN. *Nat. Protoc.* **9**, 1416–1427, doi: 10.1038/nprot.2014.090 (2014).
43. Dieterle, F., Ross, A., Schlotterbeck, G. & Senn, H. Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in ¹H NMR Metabonomics. *Anal. Chem.* **78**, 4281–4290, doi: 10.1021/ac051632c (2006).
44. Savorani, F., Tomasi, G. & Engelsen, S. B. icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *J. Magn. Reson.* **202**, 190–202, doi: http://dx.doi.org/10.1016/j.jmr.2009.11.012 (2010).
45. Smilde, A. K. *et al.* ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics* **21**, 3043–3048, doi: 10.1093/bioinformatics/bti476 (2005).
46. Jansen, J. J. *et al.* ASCA: analysis of multivariate data obtained from an experimental design. *J. Chemometr.* **19**, 469–481, doi: 10.1002/cem.952 (2005).
47. Vis, D., Westerhuis, J., Smilde, A. & van der Greef, J. Statistical validation of megavariable effects in ASCA. *BMC Bioinformatics* **8**, 322 (2007).
48. 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. *Journal of Chemometrics* **25**, 561–567, doi: 10.1002/cem.1400 (2011).
49. Tauler, R., Kowalski, B. & Fleming, S. Multivariate curve resolution applied to spectral data from multiple runs of an industrial process. *Anal. Chem.* **65**, 2040–2047, doi: 10.1021/ac00063a019 (1993).
50. Golub, G. H. & Van Loan, C. F. *Matrix Computations*. (Johns Hopkins University Press, 1996).
51. Warnes, G. R. *et al.* Gplots: Various R Programming Tools for Plotting Data. R packages version 3.0.1. URL <https://CRAN.R-project.org/package=gplots/> (2016).

Acknowledgements

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement n. 320737. We also thank Dr. Yolanda Pérez for her helpful recommendations on setting up the acquisition parameters for some of the NMR experiments.

Author Contributions

F.P.-C. and B.P. designed experiments, F.P.-C. performed experiments, F.P.-C. analysed data, I.A. contributed to the NMR analysis and to the assignment process, R.T. contributed to the MCR-ALS analysis and statistic tools, and F.P.-C. and B.P. performed the biological discussion. F.P.-C., I.A., B.P. and R.T. wrote, read and approved the manuscript.

Additional Information

Supplementary information accompanies this paper at <http://www.nature.com/srep>

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Puig-Castellví, F. *et al.* ^1H NMR metabolomic study of auxotrophic starvation in yeast using Multivariate Curve Resolution-Alternating Least Squares for Pathway Analysis. *Sci. Rep.* **6**, 30982; doi: 10.1038/srep30982 (2016).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2016

SUPPLEMENTARY MATERIAL FOR SCIENTIFIC ARTICLE III

^1H NMR metabolomic study of auxotrophic starvation in yeast using Multivariate Curve Resolution-Alternating Least Squares for Pathway Analysis.

Authors: Puig-Castellví F., Alfonso I., Piña B., Tauler R.

Citation reference: *Scientific Reports* (2016), 6:30982.

DOI: 10.1038/srep30982

Supplementary Methods

Yeast Growth. For assignment purposes, *S. cerevisiae* BY4741 (MATa; his3 Δ 1; leu2 Δ 0; met15 Δ 0; ura3 Δ 0) cells were cultured in 1-L of His-DM medium on an orbital shaker (150 rpm) at 30 °C for 24 h.

Metabolite extraction on the extract from 1 L culture of His-DM. The same method as in the 100-ml samples case was applied, but the solvent volumes were up-scaled accordingly.

NMR experiments for assignment confirmation. In order to check proton correlations and to provide a more robust assignment, additional homonuclear (gCOSY, zTOCSY and zTOCSY1D) and heteronuclear experiments ($^1\text{H}/^{13}\text{C}$ gHMBCAD and $^1\text{H}/^{13}\text{C}$ gHSQCAD) were performed. Unless stated, receiver gain was fixed to 34. In all bidimensional experiments, 512 t1 increments were used. Both zTOCSY and gCOSY experiments were acquired with 8 scans. zTOCSY1D experiments were acquired with 256 scans. In zTOCSY and zTOCSY1D, spinlock time used was 80 ms. Heteronuclear experiments were recorded using a relaxation delay of 1 second and an automatic detection of the optimum value for the receiver gain. $^1\text{H}/^{13}\text{C}$ gHMBCAD and $^1\text{H}/^{13}\text{C}$ gHSQCAD were recorded using 16 scans. The carbon spectral size covered from -10 to 190 ppm.

Metabolite identification. Proton correlations were checked on gCOSY spectra. 2-isopropylmalate assignment was confirmed using zTOCSY1D, irradiating individually the two resonances from each methyl, with a window width of 18 Hz for each doublet. Results from these experiments are in agreement with spectroscopic values of the isopropyl branch of the 2-isopropylmalate, as seen in YMDB. On the other hand, D-eritro-imidazole-glycerol-phosphate was confirmed by contrasting spectroscopic data from $^1\text{H}/^{13}\text{C}$ gHMBCAD and $^1\text{H}/^{13}\text{C}$ gHSQCAD of a metabolic yeast extract from a 1L culture for 24 h in His-DM.

Chemometric data analysis of the concentration profiles. MCR-ALS was applied to the whole dataset (\mathbf{X} , size dimension of 90 x 50), and to the subset containing only samples cultured in YSC medium (\mathbf{X}_{YSC} , size dimension of 18 x 50). An initial estimation of either \mathbf{T} or \mathbf{M} factor matrices is also needed to start the MCR-ALS analysis. For the MCR-ALS decomposition of \mathbf{X}_{YSC} , these initial estimations were selected from the purest samples (rows)^{1,2}. MCR-ALS results of this

subset, using two components, are given in the **Supplementary Fig. S3**. In this analysis, R^2 value was 84.6.

In the MCR-ALS analysis of the whole dataset, \mathbf{X} , metabolic profiles obtained in previous MCR-ALS analysis (\mathbf{M}_{VSC}) were used as the initial estimates of the growth components. Initial estimations of component profiles related to starving conditions were normalized concentration estimates of original samples containing an important contribution due to stress. In this case, these two samples corresponded to samples cultured in Ura-DM and His-DM, and collected after 10 h. Only additional components related to Uracil- and Histidine-starving conditions were added in the MCR-ALS analysis of the whole dataset since it had been observed in the PCA analysis that the major variability derived from these two cultures, and due to the fact that only these two starving conditions triggered the synthesis of the specific metabolite precursors, as showed in the heatmap of **Figure 4**.

MCR-ALS resolution using five components (three components related to growth at normal conditions and two related to growth at starving conditions) was also evaluated. However, when five components were used, the exponential phase (t_1) was divided in two peaks and we considered that explaining the exponential phase with two components was less meaningful. In addition, when five components were used instead of four, lack-of fit and R^2 do not considerably improve. These values were 34.6 % and 88.0%, respectively, for five components used. For four components, lack-of-fit and explained variance corresponded to 37.5 % and 85.7 %, respectively.

Then, the estimation of \mathbf{T} and \mathbf{M} factor matrices was performed by means of the alternating least squares optimization under constraints. In this work, we applied non-negativity constraints on both factor matrices and an equal height constraint on \mathbf{M} matrix. Metabolic profiles related to starvation metabolism were constrained to not be included in the resolution of control samples³.

Since \mathbf{T} and \mathbf{M}^T matrices are obtained using an iterative least squares process with the goal of explaining the maximum possible variance, metabolites at higher concentration would have a higher weight in the model. In order to avoid that, \mathbf{X} raw data matrix was scaled by the total sum of every column (metabolite concentrations) before MCR-ALS analysis. With this approach, low

concentrated metabolites (for all the samples) become equally represented than the high concentrated ones.

References:

1. Windig W, Stephenson DA (1992) Self-modeling mixture analysis of second-derivative near-infrared spectral data using the SIMPLISMA approach. *Analytical Chemistry* 64: 2735-2742.
2. Windig W, Guilment J (1991) Interactive self-modeling mixture analysis. *Analytical Chemistry* 63: 1425-1432.
3. Tauler R, Smilde A, Kowalski B (1995) Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. *Journal of Chemometrics* 9: 31-58.

Supplementary Figures and Tables with captions

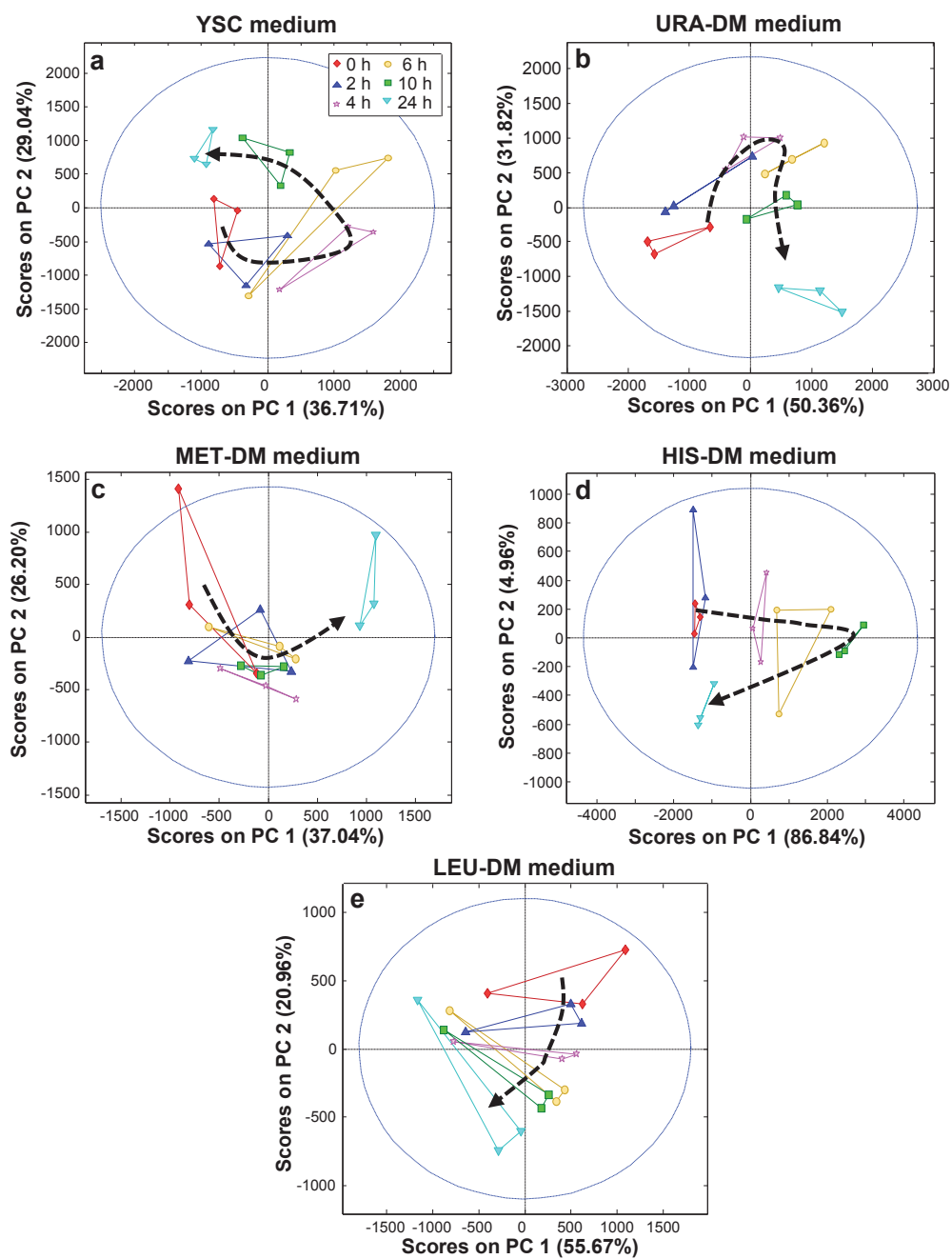


Figure S1. PCA analysis of the internal metabolic variance of starved cultures over time. a-e) PCA scores projection on PC1-PC2 subspace of samples cultured in YSC (a), Ura-DM (b), Met-DM (c), His-DM (d) and Leu-DM (e).

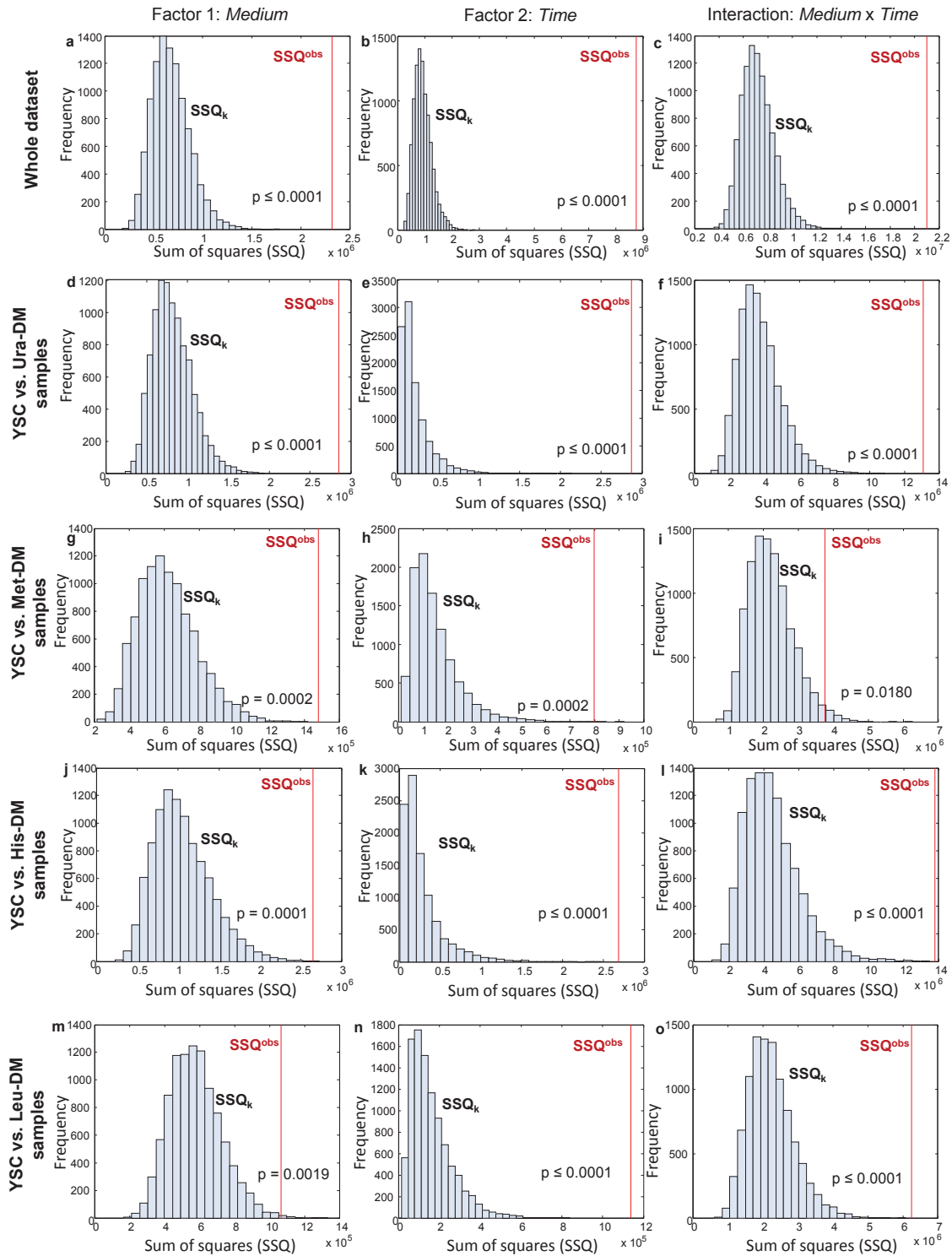


Figure S2. Histogram of the sum of squares (SSQ) obtained during permutation test in ASCA analyses. a-c) All samples. d-f) YSC vs. Ura-DM samples. g-i) YSC vs. Met-DM samples. j-l) YSC vs. His-DM samples. m-o) YSC vs. Leu-DM samples. a), d), g), j) and m) histograms show the significance of the *Medium* factor; b), e), h), k) and n) histograms show the significance for the *Time* factor; and c), f), i), l) and o) histograms show the significance for the *Medium* x *Time* interaction. The number of

permutations, k , used was 10000. Factors (**a** and **b**) or their interaction (**c**) are considered significant, as the sum of squares for each observed value (SSQ^{obs}) are larger than 95 % of the SSQ values obtained when the corresponding levels are randomized. For more information about the permutation test, see Zwanenburg *et al.* (2011).

Table S1. Metabolite assignment. ^1H NMR spectroscopic data (chemical shift, multiplicity, proton integral, and coupling constant) of the assigned metabolites. ^{13}C spectroscopic data is also provided for the erythro-imidazole-glycerol phosphate compound. CAS, HMDB, YMDB and KEGG codes are provided when existing.

#	Metabolite name	^1H and ^{13}C NMR signals assigned	CODE			
			CAS	YMDB	HMDB	KEGG
1	Acetic acid	1.90 ppm (s, 3 H)	64-19-7	YMDB000056	HMDB000042	C00033
2	Acyl-carnitine	3.19 ppm (s, 9 H)	25518-54-1	YMDB01529*	HMDB02250*	C02301
3	Adenine	8.18 ppm (s, 1 H); 8.23 ppm (s, 1 H)	73-24-5	YMDB000887	HMDB000034	C00147
4	AMP	6.13 ppm (d, J=6.0 Hz, 1 H); 8.26 ppm (s, 1 H); 8.59 ppm (s, 1 H)	61-19-8	YMDB000097	HMDB000045	C00020
5	ATP	6.14 ppm (d, J=5.1 Hz, 1 H); 8.53 ppm (s, 1 H); 8.26 ppm (s, 1 H)	56-65-5	YMDB00109	HMDB00538	C00002
6	Citric acid	2.50 ppm (s, 0.66 H); 2.54 ppm (s, 1.33 H); 2.63 ppm (s, 1.22 H); 2.67 ppm (s, 0.70 H)	77-92-9	YMDB000086	HMDB000094	C00158
7	EIGP	δ_{H} : 7.17-7.25 ppm (s, 1 H); 7.83-8.08 ppm (s, 1 H); 4.78 ppm (nd, 1 H); 4.00 ppm (nd, 1 H); 3.78 ppm (nd, 1 H); 3.58 ppm (nd, 1 H)	36244-87-8	YMDB000089	HMDB12208	C04666
		δ_{C} : 65.1 ppm (CH_2), 69.5 ppm (CH), 76.3 ppm (CH), 119.8 ppm (CH), 138.2 ppm (CH)				
8	D-Glucose	3.24 ppm (dd, J=(7.8 Hz, 9.2 Hz), 0.71 H); 3.35-3.56 ppm (mm, 2.63 H); 5.22 ppm (d, J=3.8 Hz, 0.36 H)	50-99-7	YMDB00286	HMDB00122	C00031
9	L-dihydroorotic acid	2.76 ppm (d, J=6.4 Hz, 0.43 H); 2.81 ppm (d, J=6.4 Hz, 0.57 H)	5988-19-2	YMDB00396	HMDB03349	C00337
10	Methyl donor (R-S- CH_3)	2.10 ppm (s, 3 H)	75-18-3	-	HMDB02303	-
11	Fatty acid singlet	1.24 ppm (s, 4 H)	143-07-7	YMDB00678**	HMDB00638**	C00162
12	Formic acid	8.44 ppm (s, 1 H)	64-18-6	YMDB00385	HMDB00142	C00058
13	Glutathione	2.15 ppm (q, J=7.6 Hz, 2 H); 2.87-3.00 ppm (mm, 2 H); 4.56 ppm (dd, J=(7.0 Hz, 5.2 Hz), 1 H)	70-18-8	YMDB00160	HMDB00125	C00051
14	Glycerol	3.51-3.58 ppm (mm, 2.07 H); 3.61-3.67 ppm (mm, 2.12 H); 3.77 ppm (tt, J=(6.5 Hz, 4.4 Hz), 0.79 H)	56-81-5	YMDB00283	HMDB00131	C00116

#	Metabolite name	¹ H and ¹³ C NMR signals assigned	CODE			
			CAS	YMDB	HMDB	KEGG
15	Glycerophosphocholine	3.22 ppm (s, 9 H); 3.56-3.71 ppm (mm, 4 H); 4.27-4.36 ppm (mm, 2 H)	28319-77-9	YMDB00309	HMDB00086	C00670
16	Glycine	3.55 ppm (s, 2 H)	56-40-6	YMDB00016	HMDB00123	C00037
17	GMP/GTP	5.93 ppm (d, J=5.3 Hz)	85-32-5/ 86-01-1	YMDB00261/ YMDB00558	HMDB01397/ HMDB01273	C00144/ C00044
18	S-3-Hydroxyisobutyric acid	1.06 ppm (d, J=6.9 Hz)	2068-83-9	YMDB00337	HMDB00442	C06001
19	Hypoxanthine	8.18 ppm (s, 1 H); 8.20 ppm (s, 1 H)	68-94-0	YMDB00555	HMDB00157	C00262
20	2-Isopropylmalic acid	0.84 ppm (d, J=6.9 Hz, 3 H); 0.89 ppm (d, J=6.9 Hz, 3 H)	49601-06-1	YMDB00106	HMDB00402	C02504
21	L-alanine	1.47 ppm (d, J=7.1 Hz, 3 H); 3.76 ppm (q, J=7.2 Hz, 1 H)	56-41-7	YMDB00154	HMDB00161	C00041
22	L-arginine	1.58 - 1.79 ppm (ms, 2 H); 1.80-2.00 ppm (mm, 2 H); 3.23 ppm (t, J=6.9 Hz, 2 H); 3.76 ppm (t, J=6.1 Hz, 1 H)	74-79-3	YMDB00592	HMDB00517	C00062
23	L-asparagine	2.82 ppm (d, J=7.6 Hz, 0.286 H); 2.86 ppm (d, J=7.6 Hz, 0.714 H)	70-47-3	YMDB00226	HMDB00168	C00152
24	L-aspartic acid	2.64 ppm (d, J=8.9 Hz, 0.33 H); 2.69 ppm (d, J=8.9 Hz, 0.67 H); 2.78 ppm (d, 3.7 Hz, 0.67 H); 2.83 ppm (d, J=3.7 Hz, 0.33 H); 3.89 ppm (dd, J=(8.8 Hz, 3.8 Hz), 1 H)	56-84-8	YMDB00896	HMDB00191	C00049
25	L-glutamic acid	1.99 - 2.17 ppm (ms, 2 H); 2.26-2.42 ppm (mm, 2 H); 3.75 ppm (dd, J=(7.2 Hz, 4.7 Hz), 1 H)	56-86-0	YMDB00271	HMDB00148	C00025
26	L-glutamine	2.44 (td, J = 7.5, 3.7 Hz, 1H)	56-85-9	YMDB00002	HMDB00641	C00064
27	L-histidine	3.11 ppm (d, J=7.4 Hz, 0.35 H); 3.15 ppm (d, J=7.7 Hz, 0.65 H); 7.06 ppm (s, 1 H); 7.81-7.92 ppm (s, 1 H);	71-00-1	YMDB00369	HMDB00177	C00135
28	L-isoleucine	0.93 ppm (t, J=7.4 Hz, 3 H); 1.000 ppm (d, J=7.1 Hz, 3 H)	73-32-5	YMDB00038	HMDB00172	C00407
29	L-lactic acid	1.31 ppm (d, J=7.0 Hz, 3 H); 4.10 ppm (q, J=6.9 Hz, 1 H)	79-33-4	YMDB00247	HMDB00190	C00186
30	L-leucine	0.95 ppm (d, J=6.1 Hz, 3 H); 0.95 ppm (d, J=6.1 Hz, 3 H)	61-90-5	YMDB00387	HMDB00687	C00123
31	L-lysine	1.35-1.60 ppm (mm, 2 H); 1.65-1.80 ppm (mm, 2 H); 1.81-1.94 ppm (ms, 2 H); 2.95-3.1 ppm (mm, 2 H); 3.75 ppm (t, J=6.1 Hz, 1 H)	56-87-1	YMDB00330	HMDB00182	C00047

#	Metabolite name	¹ H and ¹³ C NMR signals assigned	CODE			
			CAS	YMDB	HMDB	KEGG
32	L-methionine	2.13 ppm (s, 3 H); 2.63 ppm (t, J=7.5 Hz, 3 H)	63-68-3	YMDB00318	HMDB00696	C00073
33	L-ornithine	1.65-2.00 ppm (mm, 4 H); 3.04 ppm (t, J=7.6 Hz, 2 H)	70-26-8	YMDB00353	HMDB00214	C00077
34	L-phenylalanine	7.37 ppm (mm, 5 H)	63-91-2	YMDB00304	HMDB00159	C00079
35	L-proline	1.90-2.12 ppm (mm, 3 H); 2.27-2.40 ppm (mm, 1 H); 4.12 ppm (dd, J=(8.6 Hz, 6.4 Hz), 1 H)	344-25-2	YMDB00378	HMDB00162	C00148
36	L-threonine	1.32 ppm (d, J=6.6 Hz, 3 H); 3.57 ppm (d, J=4.9 Hz, 1 H); 4.19-4.28 ppm (mm, 1 H)	72-19-5	YMDB00214	HMDB00167	C00188
37	L-tyrosine	6.89 ppm (d, J=8.4 Hz, 2 H); 7.18 ppm (d, J=8.4 Hz, 2 H)	60-18-4	YMDB00364	HMDB00158	C00082
38	L-valine	0.98 ppm (d, J=7.1 Hz, 3 H); 1.03 ppm (d, J=7.1 Hz, 3 H); 2.26 ppm (mm, 1 H)	72-18-4	YMDB00152	HMDB00883	C00183
39	3-Methyl-2-oxovaleric acid	1.09 ppm (d, J=6.6 Hz, 3 H)	816-66-0	YMDB00168	HMDB00491	C00671
40	N ⁵ -methyl-adenosine	2.96 ppm (s, 3 H); 6.11 ppm (d, J=4.4 Hz, 1H); 8.28 ppm (s, 1H); 8.29 ppm (s, 1H)	1867-73-8	-	HMDB04044	-
41	NAD ⁺	6.03 ppm (d, J=5.9 Hz, 1 H); 6.08 ppm (d, J= 5.3 Hz, 1 H); 8.16 ppm (s, 1 H); 8.17 ppm (d, J=6.1 Hz); 8.15-8.21 ppm (mm, 2 Hz); 8.42 ppm (s, 1H); 8.82 ppm (d, J=7.6 Hz, 1 H); 9.13 ppm (d, J=7.6 Hz, 1 H); 9.33 ppm (s, 1 H)	53-84-9	YMDB00110	HMDB00902	C00003
42	Orotic acid	6.18 ppm (s, 1 H)	65-86-1	YMDB00405	HMDB00226	C00295
43	Orotidine-5P	5.54 ppm (d, J=3.3 Hz, 1 H); 5.76 ppm (s, 1 H);	2149-82-8	YMDB00025	HMDB00218	C01103
44	Succinic acid	2.39 ppm (s, 4 H)	110-15-6	YMDB00338	HMDB00254	C00042
45	Trehalose	3.44 ppm (t, J=9.3 Hz, 2 H); 3.59 - 3.92 ppm (mm, 10 H); 5.18 ppm (d, J=3.8 Hz, 2 H)	99-20-7	YMDB00008	HMDB00975	C01083
46	Uracil	5.79 ppm (d, J=7.8 Hz, 1 H); 7.53 ppm (d, J=7.7 Hz, 1 H)	66-22-8	YMDB00098	HMDB00300	C00106
47	Ureidosuccinic acid	2.42 pm (d, J=9.7 Hz, 0.44 H); 2.46 ppm (d, J=9.7 Hz, 0.56 H); 2.64 ppm (J=3.9 Hz, 0.57 H); 2.68 ppm (d, J=4.0 Hz, 0.43 H)	13184-27-5	YMDB00027	HMDB00828	C00438
48	Uridine	5.87-5.92 ppm (mm, 2H); 7.87 ppm (d, J=8.5 Hz, 1 H)	58-96-8	YMDB00127	HMDB00296	C00299

#	Metabolite name	¹ H and ¹³ C NMR signals assigned	CODE			
			CAS	YMDB	HMDB	KEGG
49	Unknown-1	8.37 ppm (s, 1 H)	-	-	-	-
50	Unknown-2	8.03 ppm (s, 1 H)	-	-	-	-

*Code for acyl-carnitine is from Acyl(C12:0)-carnitine

**Code for fatty acid is from C12:0

mm = modelled multiplet

ms = multiplet modelled as deconvoluted singlet

nd = assignment from TOCSY/HMBC/HSQC spectra (too overlapped in the ¹H NMR spectrum).

Note: Spectroscopic constants relative to strong coupled protons are described as various set of regular coupled proton (ex: a *dd* is expressed as two *d* with different proton integrals).

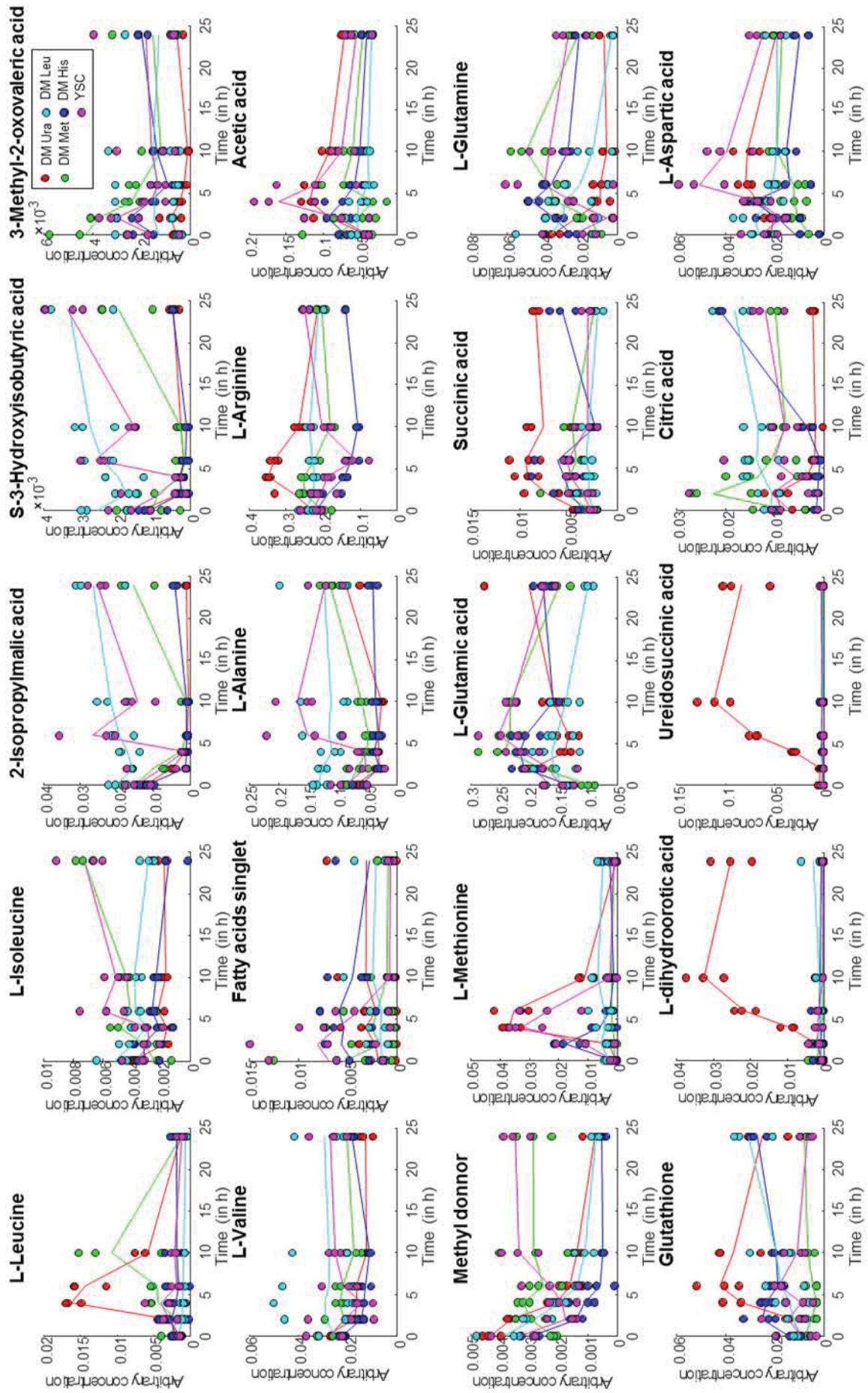


Figure S3A. Concentration estimates over time of 20 assigned metabolites.

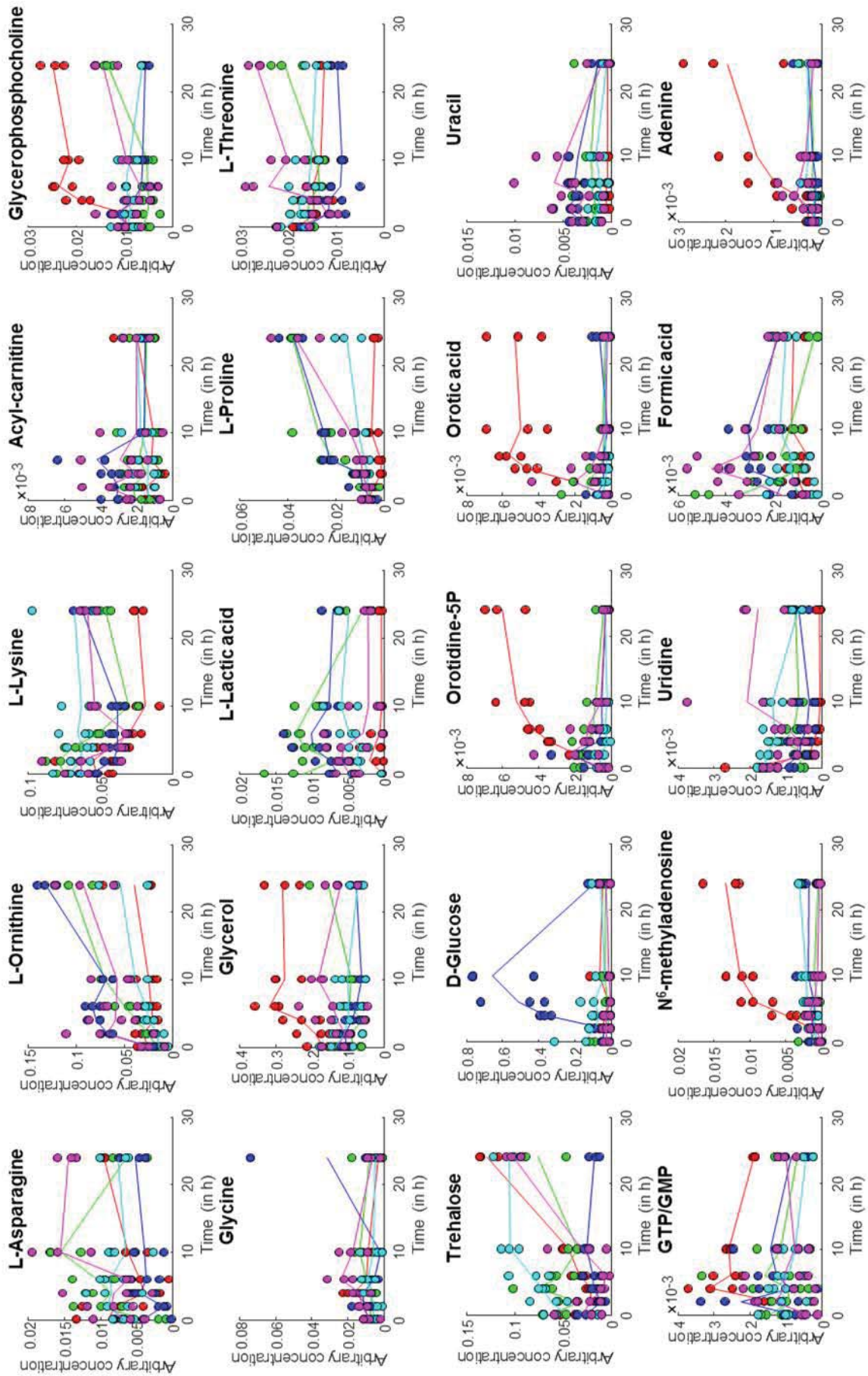


Figure S3B. Concentration estimates over time of 20 assigned metabolites. Continuation from Fig S2A.

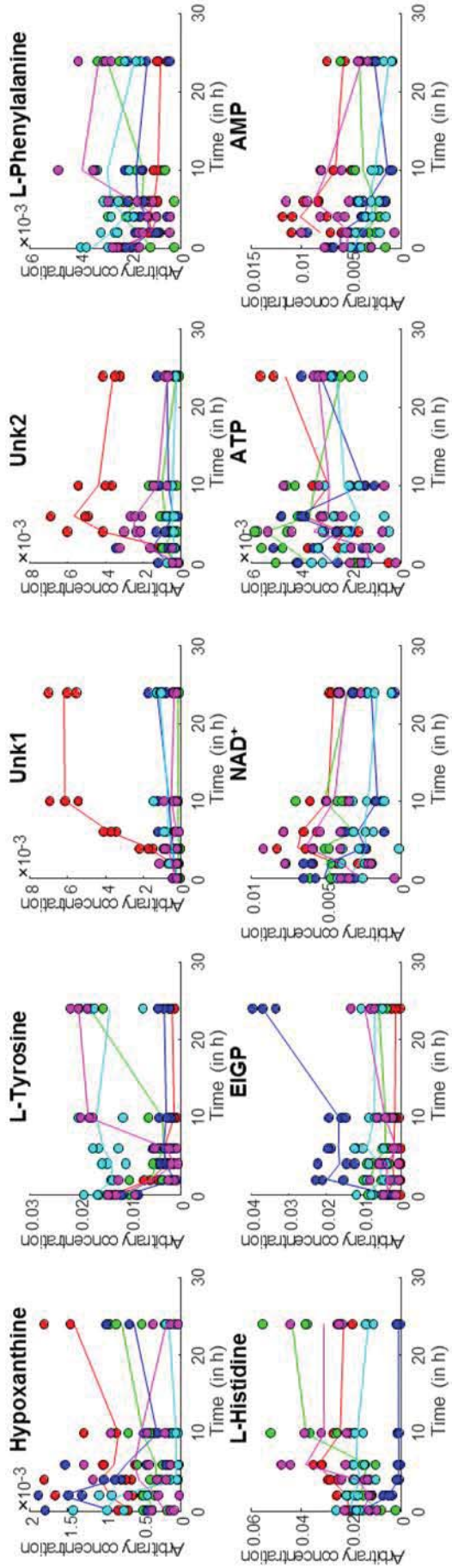


Figure S3C. Concentration estimates over time of 10 assigned metabolites. Continuation from Fig S2B.

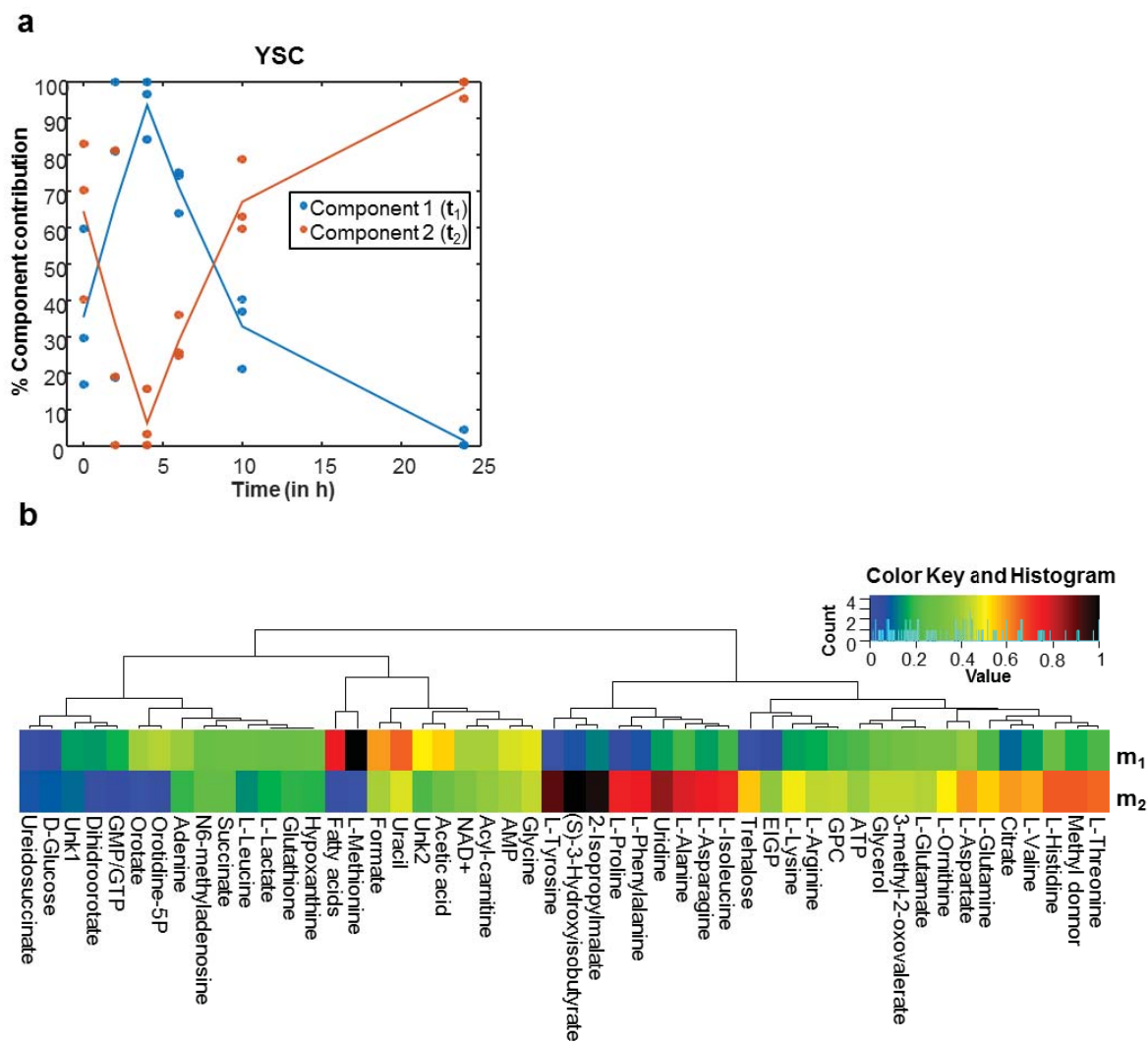


Figure S4. Growth pattern of yeast metabolism cultured in YSC using 2 components.

(a) Temporal growth pattern (in %). (b) Hierarchical clustering of the relative contribution of every metabolite in the MCR-ALS resolved components from X_{YSC} data matrix.

3 DISCUSSION OF THE RESULTS

In this section, the suitability of NMR to identify biomarkers is discussed. In addition, MCR-ALS is presented as a powerful choice to unravel the yeast metabolic responses derived from environmental stress conditions. Finally, the new findings obtained from yeast metabolomics analyses are exposed and compared to those existing in the literature.

3.1 NMR IS A POWERFUL ANALYTICAL TECHNIQUE TO IDENTIFY BIOMARKERS

In ^1H NMR, every proton³ produces a resonance signal in the spectrum, and therefore, most likely several proton resonances per metabolite can be detected.

Although this signal redundancy causes that the resulting ^1H NMR metabolomics spectrum becomes overcrowded and very complex due to signal overlapping, if this overlapping is meticulously disentangled, the complete NMR assignment for the detected compounds will provide non-refutable proofs of their presence in the sample. This identification power is only comparable to the one obtained for MS-fragmentation (or MS^n) techniques. However, while, two NMR instruments with the same external magnetic field will acquire identical spectra for the same molecules, this is unlikely to happen with two different MS^n instruments. MS^n fragments depend on several parameters (*e.g.*, ionization source, ionization mode, cone voltage, m/z detector), implying that the obtained MS spectra cannot easily be extrapolated to other MS^n instruments. For this reason, when NMR sensitivity is not a limitation, NMR spectroscopy is the best choice to confirm biomarkers, as the spectroscopic signatures that define these biomarkers can be straight away transferred and used by the scientific community.

To disentangle the resonance constituents of an overlapping region, the most unveiling option is by adding or spiking an NMR sample with a tentative constituent. If the added metabolite was already in the sample but in a lower concentration, the total number of resonances will be the same, and the resonances from this metabolite will be now more intense (**Fig. 3.3B**). On the opposed situation, the NMR spectrum after the addition will contain more resonances than before (**Fig. 3.3A**).

³ Except exchangeable protons in a deuterated solvent.

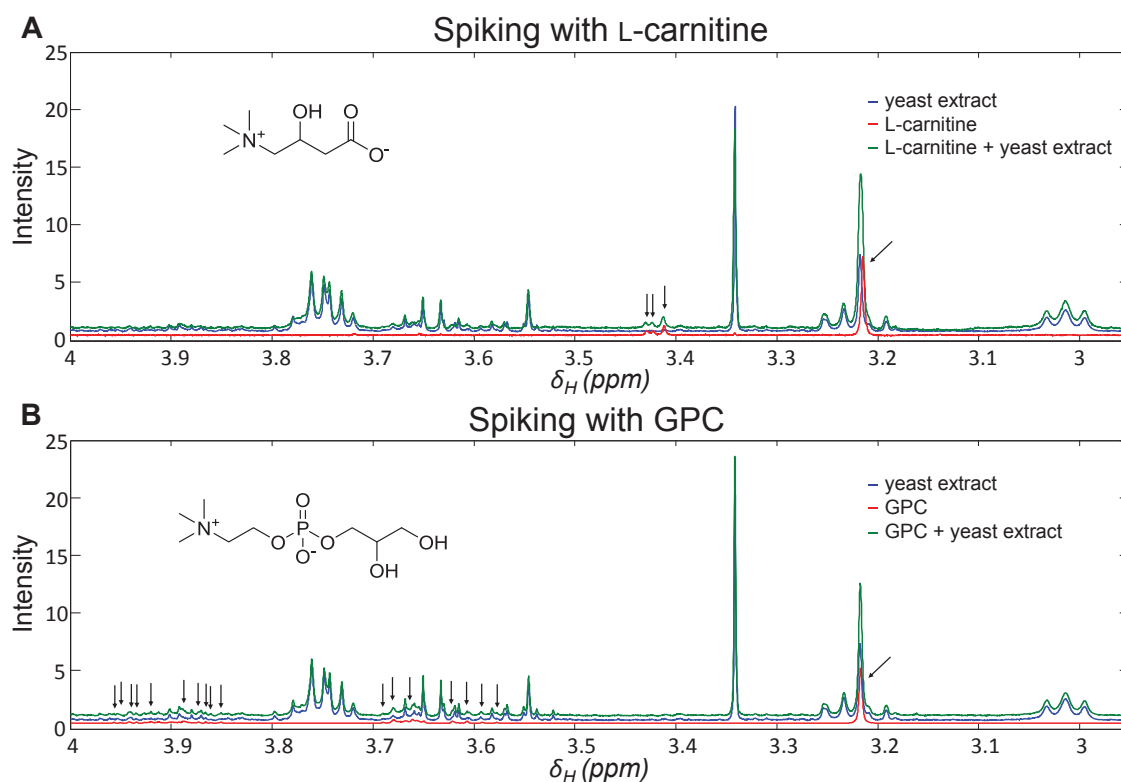


Figure 3.3. Spiking with **(A)** L-carnitine and **(B)** glycerophosphorylcholine (GPC). The singlet resonance from L-carnitine at 3.21 ppm is not exactly at the same position as the singlet in the yeast extract sample. This chemical shift position of the singlet coincides in the case for GPC.

The trickiest part of the spiking approach is to decide which metabolites need to be checked. Since *Saccharomyces cerevisiae* has a database (YMDB) [142] containing its metabolites constituents and the NMR spectrum of these metabolites can be consulted, we used this resource as a starting point. However, at the beginning of this Thesis, YMDB contained 2,007 metabolites and, from those, 930 metabolites entries included NMR data. Since, at most, only 50 metabolites are usually detected by NMR of a yeast sample, the first attempt of consulting this database produced too many hits per queried resonance. After refining this search using NMR data from ^1H - ^1H COSY and ^1H - ^1H TOCSY NMR experiments, we reduced the list of detectable metabolites to around 70. Then, we performed spiking experiments using all the metabolites from that list that were available. A total of 53 metabolites were tested and, from those, the actual presence of 23 was confirmed. This became the beginning of our home-made NMR dataset [30].

Other metabolites that were overlooked in the first NMR analysis but were highlighted in the posterior chemometric analysis were also tested by spiking. For instance, this was the case for the singlet resonance at $\delta_{\text{H}} = 3.21$ ppm, associated to a very high discriminant power by

PLS-DA (VIP > 1,000) [30]. With the spiking strategy, the resonance was satisfactorily assigned to glycerophosphorylcholine or GPC (**Fig. 3.3**).

In the long run, spiking is very demanding and expensive because all the tentative compounds need to be bought and experimentally checked.

A second strategy used in this Thesis was to acquire ^1H - ^{13}C NMR spectra from our target samples or, in case the sensitivity was not enough, to prepare NMR samples of concentrated extracts. For instance, in 2016 [118], the acquisition of ^1H - ^{13}C HSQC (data not shown) and ^1H - ^{13}C HMBC NMR (**Figure 3.4**) from an NMR sample containing yeast extracts from 1-liter cultures was performed. These NMR spectra, for instance, confirmed the presence of the erythro-imidazole-glycerol-phosphate (EIGP) compound, an L-histidine precursor (**Fig. 3.4**).

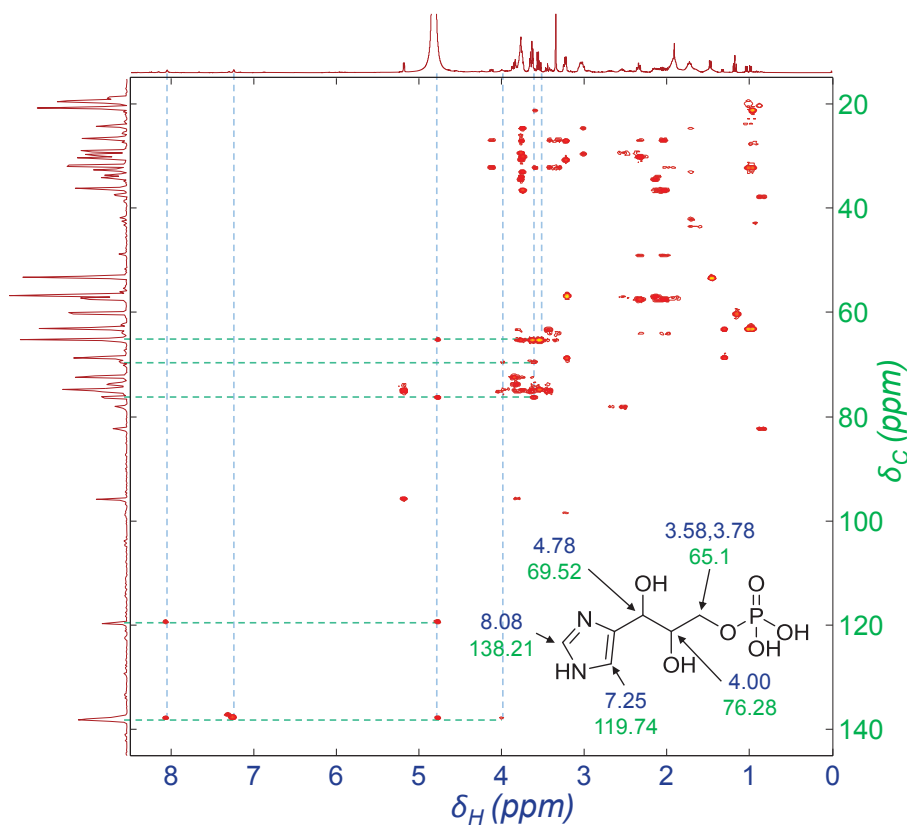


Figure 3.4. ^1H - ^{13}C HMBC of a yeast extract. Correlations from erythro-imidazole-glycerol-phosphate are shown with dashed lines and indicated in the drawn molecule.

It is also possible to assign some of the unknown resonances guided by the interpretation of the observed changes in the known metabolome. For instance, also in 2016 [118], two doublets resonances (at $\delta_{\text{H}} = 0.84$ and $\delta_{\text{H}} = 0.89$ ppm) from an unknown compound were attributed to a metabolite related to L-leucine biosynthesis pathway in yeast. This deduction was made because the evolution of the intensity of these two resonances within the sample

set was inversely proportional to the intensity of L-leucine resonances, and because their resonance pattern reminds to an isopropyl group, which is also present in L-leucine carbon backbone.

Seven compounds from L-Leucine biosynthesis pathway contain an isopropyl group (**Fig. 3.5**) and one of them may actually be the compound associated to these two resonances. From these seven compounds, L-leucine and 3-methyl-2-oxobutanoate compounds were directly discarded because their isopropyl groups are associated with resonances with different chemical shifts. The other five compounds cannot be discarded with this criteria, because they have similar resonances, or their ^1H NMR spectra could not be consulted in any NMR metabolomics databases.

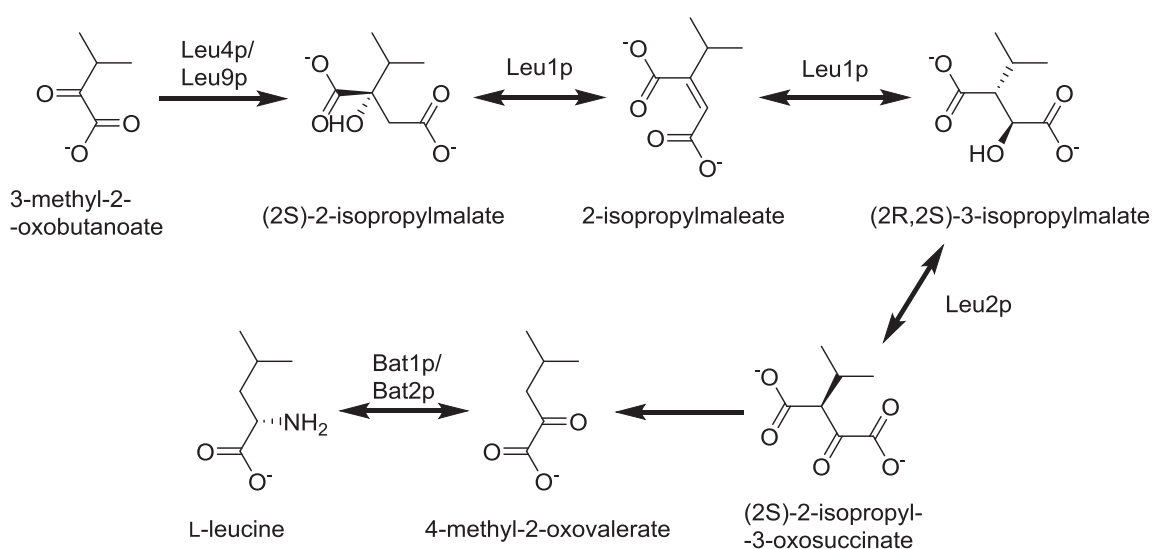


Figure 3.5. L-leucine biosynthesis in yeast.

For compounds without experimental ^1H NMR spectra available in the databases, it is possible to generate a theoretical ^1H NMR spectrum. In this case, the ^1H NMR spectra of the 5 remaining metabolites were calculated using the ^1H NMR prediction tool from MestReNova (**Fig. 3.6**). From these 5 predictions, 2 metabolites were discarded because their ^1H NMR spectra only contain one doublet (blue and red spectra in **Figure 3.6**) and not two because the two methyl groups are magnetically equivalents.

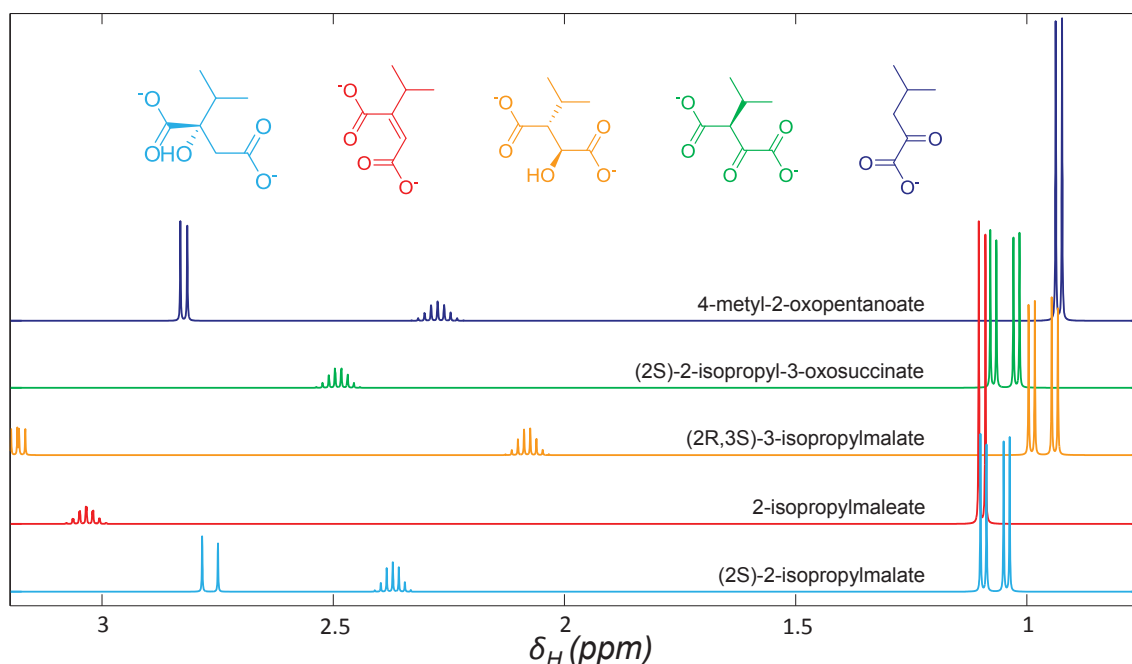


Figure 3.6. Simulated spectra of five compounds from the L-leucine biosynthetic pathway in yeast. In MestReNova, spectra are simulated without taking into account the sample conditions, the external magnetic field nor any other acquisition parameter. For all these reasons, predicted chemical shifts and coupling constants may diverge from the real ones.

Finally, by acquiring two selective 1D TOCSY NMR spectra [333], using as the irradiating frequency the ones from the two doublets (**Fig. 3.7**), it was observed that the protons from the isopropyl group formed an isolated spin system of three resonances. This implies that the carbon next to the isopropyl group does not have any proton bound and, therefore, the only possible option left is (2S)-2-isopropylmalate.

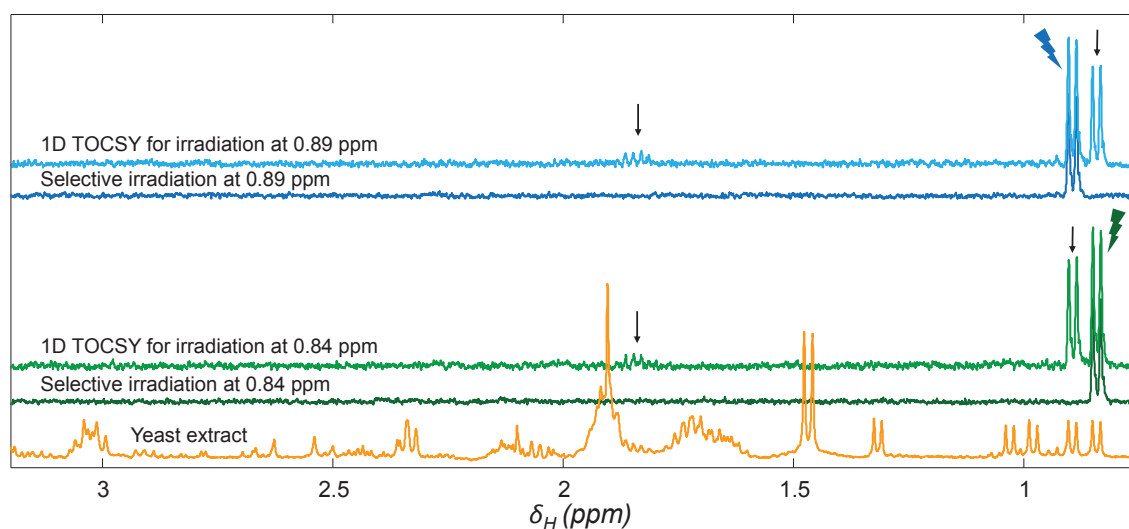


Figure 3.7. Selective 1D TOCSY at 0.84 and at 0.89 ppm in a yeast extract sample.

The last strategy used to assign NMR resonances is the application of STOCSY [332]. This approach consists in calculating, from a 1D NMR dataset, the statistical pair-wise correlations between the intensities measured in a given chemical shift for all the samples and all the other measured intensities in different chemical shifts. Then, a reference NMR spectrum is colored using a color scheme that represents the calculated correlation (red denotes high correlations). In **Figure 3.8**, the STOCSY showing the correlations to a signal at $\delta_{\text{H}} = 2.615$ is given.

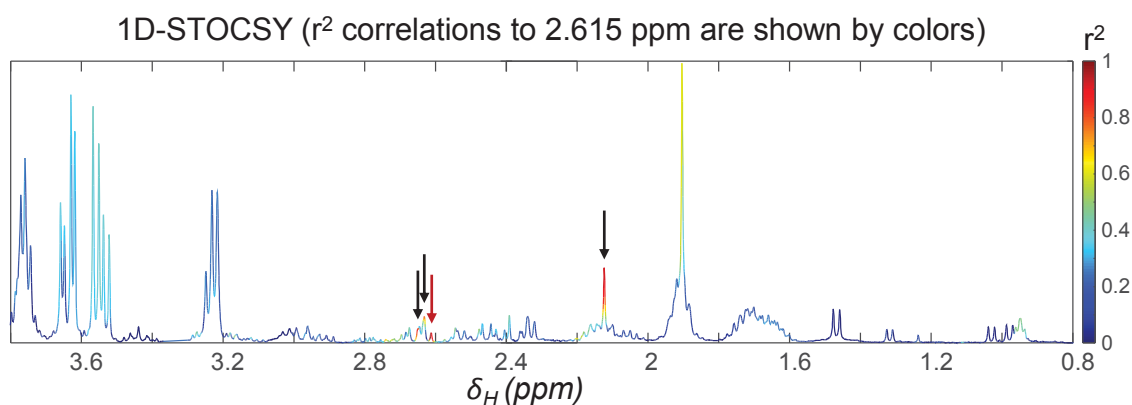


Figure 3.8. 1D STOCSY. The red arrow points out the target chemical shift ($\delta_{\text{H}} = 2.615$ ppm), while the three black arrows point the chemical shifts which intensities best correlate to intensities at 2.615 ppm.

From this statistical analysis, it is revealed that the resonance at $\delta_{\text{H}} = 2.615$ ppm (red arrow in **Figure 3.8**) was one of the three constituents of the triplet at $\delta_{\text{H}} = 2.63$ ppm (indicated by the same red arrow and two more black arrows in **Figure 3.8**). Moreover, this triplet resonance is found to be highly correlated to a singlet at 2.10 ppm. After consulting YMDB database [142], these two resonances were satisfactorily assigned to L-methionine. It is worth to mention that the knowledge information generated with STOCSY cannot be obtained with ^1H - ^1H TOCSY since these two resonances belong to different spin systems (their protons are not coupled).

Therefore, different strategies based on NMR spectroscopy can be used to characterize and assign the unknown compounds within a metabolomics sample.

3.2 MCR-ALS HIGHLIGHTS THE AFFECTED METABOLIC PATHWAYS UNDER DIFFERENT STRESSES

For any living organism, metabolism can be regarded as the summation of all the metabolic processes occurring at the same time to sustain life and to promote growth. Since sets of these

metabolic reactions are synchronically orchestrated as a result of the needs of this organism at every moment (all reactions from glucose catabolism reactions are activated under glucose availability, or all processes needed to replicate DNA are activated during the S phase of the cell cycle), the metabolome can be mathematically expressed as the combination of the metabolic profiles characteristic of each one of these metabolic pathways, \mathbf{m}_i , weighted by their relative contribution, t_i (**Eq. 3.1**).

$$\text{Metabolism} = \sum t_i \mathbf{m}_i, \quad \text{eq. 3.1}$$

Since metabolism can be explained using an equation analogous to the bilinear model of **equation 2.17** (page 55), chemometric approaches based on this model can be used to untangle the metabolic pathways and their relative contribution for every measured sample. In this Thesis, we used MCR-ALS to extract the metabolic pathways (or metabolic profiles) that describe the yeast metabolic state at the two explored conditions, temperature [125] and starvation [118] stresses. The application of this chemometric method to extract these profiles has been detailed in **Section 5.4 of Chapter 2** and in the methods section in the Scientific articles II and III.

From all possible chemometric methods based on the bilinear model, MCR-ALS is the most suitable approach because it allows the use of non-negativity constraints. With this constraints, we ensure that the metabolic profiles will have real meaning: metabolic concentrations will be positive, on the relative contribution of each pathway will be either 0 (metabolic pathway not activated) or positive (metabolic pathway activated).

In order to maximize the metabolic changes, even for the smallest concentrated metabolites, data was scaled prior MCR-ALS analysis. Without scaling, the smaller metabolic changes may be captured on the residual matrix and the MCR-ALS components will only be descriptive of the more abundant metabolites.

The used data-scaling methods transformed the original concentration estimates for every metabolite to relative concentrations within the range of 0 and 1. These scaling transforms were performed by dividing every metabolite concentration by the maximum concentration value [118] or by using min-max scaling [125].

Results described in these works have shown that the proposed application of MCR-ALS to investigate metabolomics datasets results in a convenient approach for untangling the underlying metabolic responses in the different studied biological systems.

3.3 BIOLOGICAL INTERPRETATION OF THE TEMPERATURE ADAPTATION

In the MCR-ALS analysis of the metabolomics dataset descriptive of the thermal stress (Scientific article II, [125]), three metabolic profiles were obtained, representative of the metabolism at low, optimal and high temperatures.

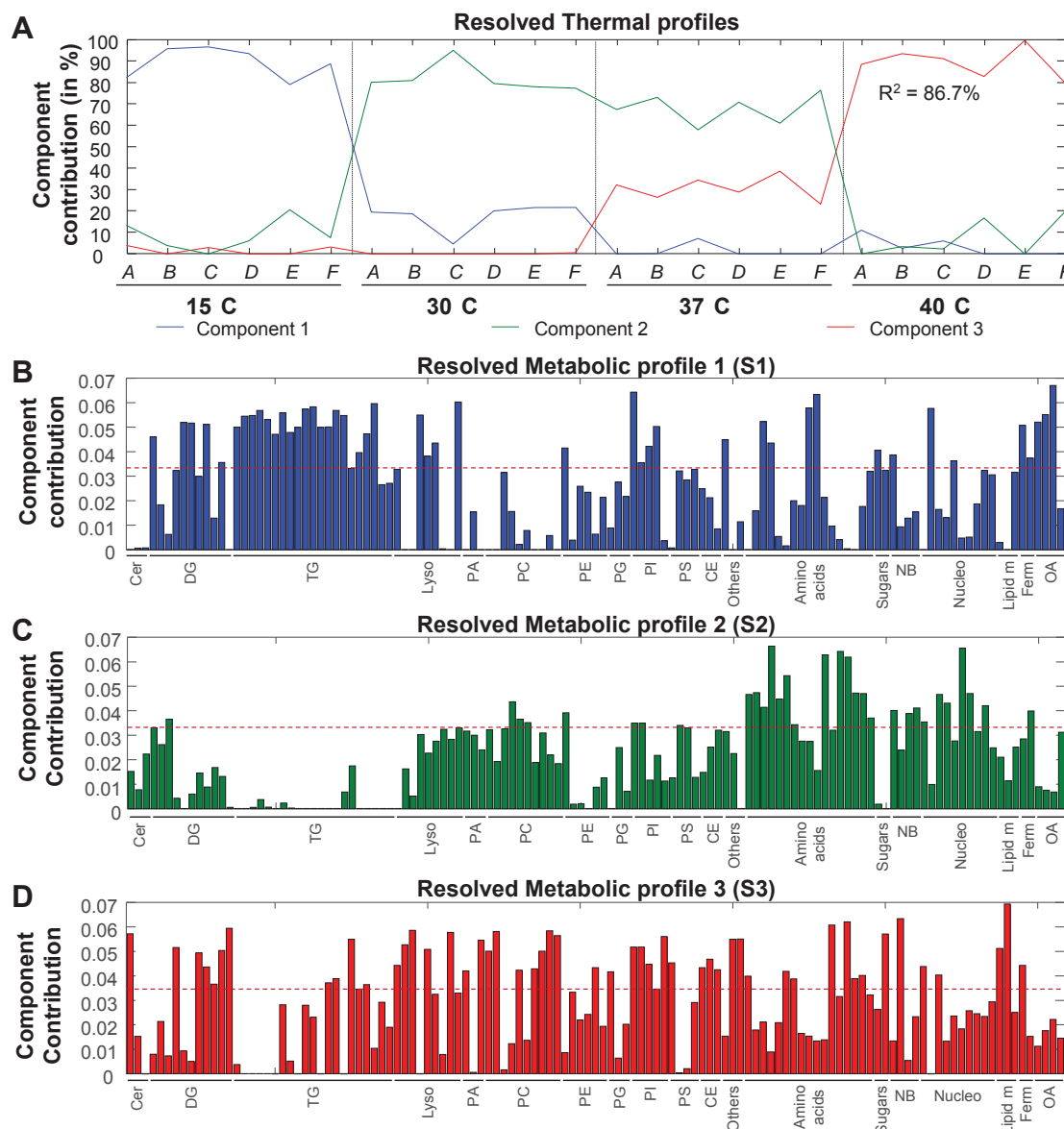


Figure 3.9. MCR-ALS resolution lipidic-and-metabolic dataset. **A)** MCR-ALS resolved thermal (**C** in Eq. 2.17) profiles of yeast cells cultured at the four different temperatures. Explained variance (R^2) is also given in the plot. Sextuplicates are represented by *A-F* letters of the axis levels. **B-D)** Resolved lipidomic-and-metabolic (S^T in Eq. 2.17) profiles for component 1 (**B**), component 2 (**C**) and component 3 (**D**). Lipid families: *Cer*, ceramides; *DG*, diacylglycerides; *TG*, triacylglycerides; *Lyso*, lyso-phospholipids; *PA*, phosphatidic acid; *PC*, phosphatidylcholines; *PE*, phosphatidylethanolamines; *PG*,

phosphatidylglycerol; *PI*, phosphatidylinositol; *PS*, phosphatidylserines; *CE*, cholesterol esters. Polar metabolite families: *amino acids*; *sugars*; *NB*, nitrogen bases; *Nucleo*, nucleosides and nucleotides; *Lipid m*, lipid metabolites (glycerophosphorylcholine, choline, and glycerol); *Ferm*, fermentation metabolites (ethanol, lactic acid); *OA*, organic acids.

We observed that the contribution of these profiles change gradually when the growth temperature is increased or decreased (**Figure 3** in Scientific Article II), pointing out that the degree of the metabolic adaptation response is proportional to the intensity of the stress temperature (*i.e.*, the metabolic profile of growth at low temperatures is more prevalent at low temperatures than at high temperatures).

This gradual response suggests that, rather than a binary metabolic switch regulated by the activation or inactivation of stress response-related genes, the thermal stress response can be regarded as a fluid event, where the activation of these genes is also fluid.

The analyzed dataset in Scientific Article II contains two different types of metabolomics data: data relative to the lipidic fraction, characterized by UHPLC-MS analysis; and data relative to the primary metabolism (*e.g.*, mostly sugars, amino acids and nucleotides), characterized by ¹H NMR analysis. After MCR-ALS analysis, it was revealed that combining more than one data type in the same analysis is more insightful than the two separated analyses because the MCR-ALS resolved components will be descriptive of the coordinated response of both lipids and primary metabolites. In addition, the combined analysis is preferable because yeast adaptation to low or hot temperatures uses different metabolic strategies, and therefore the two types of metabolites have different importance in the three resolved components. For instance, in the temperature experiment we observed that the predominant component at low temperatures is mainly described by triglycerides (**Fig. 3.9B**), obtained by the UHPLC-MS analysis, while the metabolism at optimal conditions was described by amino acids and other primary metabolites (**Fig. 3.9C**), obtained by the NMR analysis.

The different evaluated thermal stresses cause changes or adaptation responses in the lipid fraction and in the primary metabolism.

Low temperatures alter the lipid membranes by promoting the accumulation of short triglycerides (TGs) and diglycerides (DGs) species with a low number of unsaturations, as well as accumulating more phosphatidylinositol (PIs) lipid species (**Fig. 3.9B**). The observed stress response agrees with the response observed in previous studies of yeast cultured at low temperatures [314], and with the phenotype observed in psychrophilic yeasts [282,284].

On the other hand, at optimal growth, DGs and TGs are less abundant and phospholipid species, such as PE and PS, are more prominent (**Fig 3.9C**).

Finally, at higher temperatures, DG, phosphatidylcholines (PCs) and long, poly-unsaturated TG, plays an important role in the cell membrane modification (**Fig 3.9D**).

Therefore, from the MCR-ALS analysis and from the fatty acid composition analysis, we conclude that lipid membrane adapts mostly by changing the relative fraction of the different lipid families (*i.e.*, more PC lipids at high temperatures, more PI and TG at low temperatures), and by altering the unsaturation number and carbon length of the fatty acids from TG species.

The primary metabolism is also affected by changes in the growth temperature.

At low temperatures (**Fig 3.9B**), organic acids are accumulated, reflecting that yeast cells are focused on survival because resources are invested to maintain vital pathways such as the Krebs cycle. Glycerol is also found abundant at low temperatures, acting as osmoprotectant to palliate the osmotic stress derived from the increased membrane permeability [334,335].

At optimal growth, pathways used to build cell structures and to promote growth are activated. Because of this, metabolites characteristic of these conditions are amino acids and nucleotides (**Fig 3.9C**).

At higher than optimal temperatures, derived from the accumulation of PCs lipid species, glycerophosphorylcholine (a PCs precursor) is also accumulated under this growing conditions (**Fig 3.9D**). In addition, trehalose, a stress biomarker metabolite, and L-lactic acid were found significant. The accumulation of L-lactic acid pointed out an increase of the fermentative metabolism. This agrees with the fact that, at higher temperatures, genes involved in alternative carbon utilization are expressed [323]. Finally, another typical gene response is the expression of genes from protein folding chaperones [323], which can be connected with the observed accumulation of uracil and some amino acids.

In Scientific Article I, we studied the metabolic response of yeast at a mild-heat temperature (37°C). Despite using a different chemometric method in this study (OSC-PLS-DA instead of MCR-ALS), obtained results are in agreement with the metabolic profile observed for yeast cells cultured at high temperatures in Scientific Article II. Specifically, all metabolites detected up-regulated at a mild-heat temperature in the OSC-PLS-DA in Scientific Article I (trehalose, L-lysine, L-histidine, L-alanine, and glycerophosphocholine) were also characteristic for the metabolic response at high temperatures described in Scientific Article II.

In addition, in this first study, some metabolic pools, such as the pools of NAD^+ and glutathione, were reduced. This can be connected to the increased consumption of glutathione [296] to attenuate the effect of the increased oxidative stress at higher temperatures [293].

3.4 BIOLOGICAL INTERPRETATION OF THE STARVATION STRESS

The metabolomics dataset from Scientific article III[118], descriptive of four different starvation stresses (L-methionine, L-histidine, L-leucine and uracil deprivation), was first investigated with ASCA. This analysis revealed that, at a metabolic level, yeast responded differently at every studied medium. In order to identify the underlying metabolic pathways that drove to this observed response, MCR-ALS was used on the same dataset.

In the MCR-ALS analysis, four components were resolved. For every component, a temporal profile, \mathbf{t} , and a metabolic profile, \mathbf{m} , were obtained.

Two of the four MCR-ALS components resolved in this analysis were descriptive of the basal metabolism (\mathbf{t}_1 and \mathbf{m}_1 , and \mathbf{t}_2 and \mathbf{m}_2 in **Figure 5** in Scientific article III), while the two other resolved components were descriptive of de-regulated metabolic processes (\mathbf{t}_3 and \mathbf{m}_3 , and \mathbf{t}_4 and \mathbf{m}_4 in **Figure 5** in Scientific article III).

The first MCR-ALS component (\mathbf{t}_1 in **Figure 3.10**) represents the metabolic response associated to the exponential growth phase in yeast, while second MCR-ALS component (\mathbf{t}_2 in **Figure 3.10**) represents the metabolic response associated to the lag and stationary growth phases. On the other hand, the third MCR-ALS component (\mathbf{t}_3 in **Figure 3.10**) corresponds to the metabolic response derived from the de-regulation of uracil biosynthetic pathway, and the fourth MCR-ALS component (\mathbf{t}_4 in **Figure 3.10**) is the equivalent metabolic response derived from the de-regulation of L-histidine biosynthetic pathway.

Since yeast metabolism was explored at 6 different time-points, the rate of the basal metabolism under starvation conditions can be evaluated when compared with the control conditions.

Under normal conditions (**Fig. 3.10A**), \mathbf{t}_1 peaked at maximal growth, while \mathbf{t}_2 was the predominant component when yeast growth rate was at a much lower rate.

For uracil and L-histidine starvation cultures (**Fig. 3.10B** and **Fig. 3.10D**), due to the de-regulation of the biosynthetic pathways of uracil and L-histidine, respectively, the stationary growth phase was never reached and all resources were employed towards these two de-regulated biosynthetic pathways. Because of this, \mathbf{t}_2 disappeared, while \mathbf{t}_3 (for uracil

starvation) and t_4 (for L-histidine starvations) became the predominant component at late time-points.

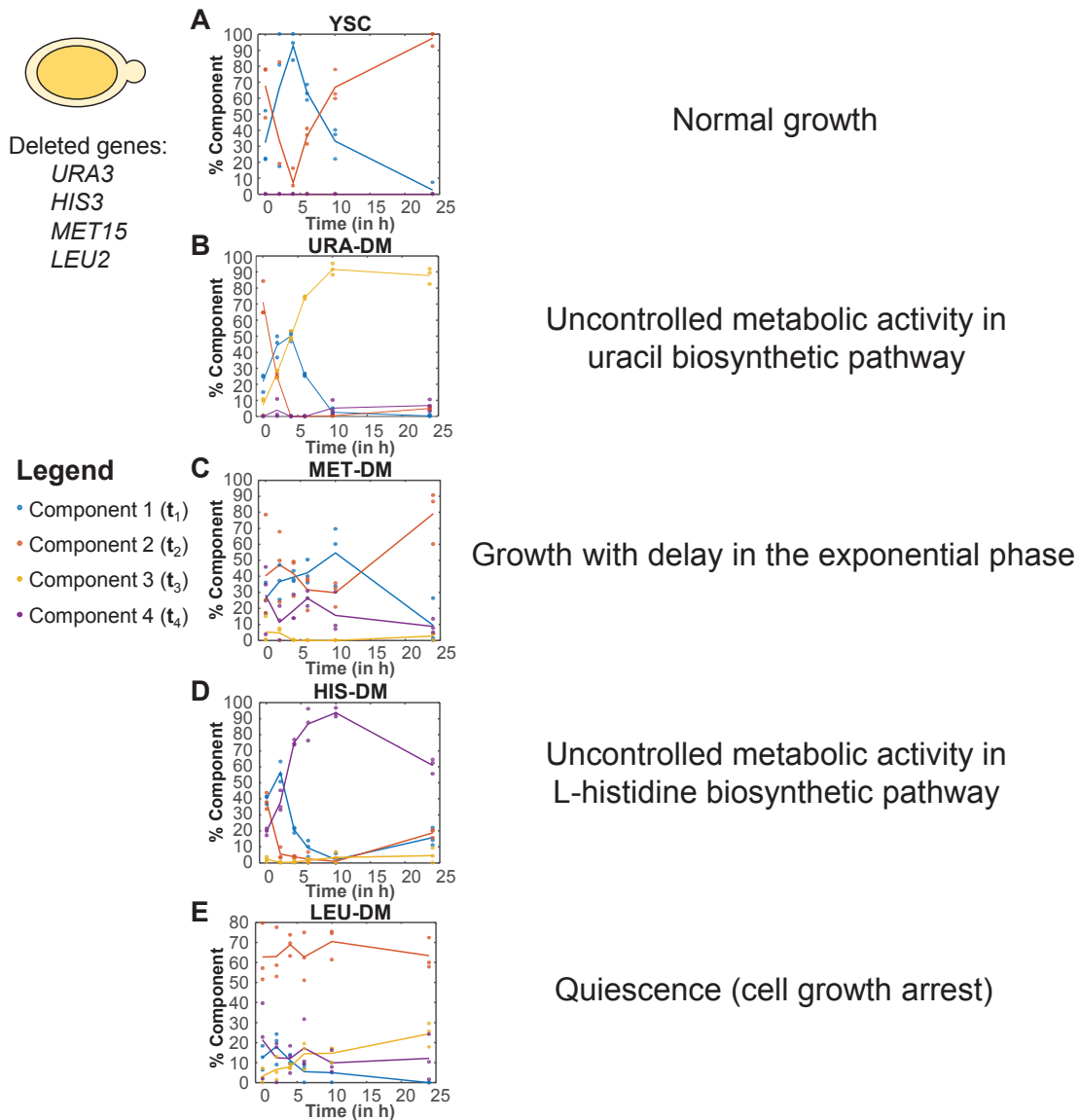


Figure 3.10. Summary of the observed metabolic responses under the different studied growth conditions. **A)** Normal growth. **B)** Absence of uracil. **C)** Absence of L-methionine. **D)** Absence of L-histidine. **E)** Absence of L-leucine. DM: Dropout-medium.

For L-methionine starvation, the exponential phase was substantially delayed. This is observed in **Figure 3.10C** as a larger span of t_1 when compared to the growth at normal conditions (**Fig. 3.10A**).

Finally, for L-leucine starvation conditions, L-leucine deprivation produced an arrest of the basal metabolism at the beginning of the experiment. This response is observed as the metabolic contributions from the four metabolic profiles in **Figure 3.10E** did not change over the 24 hours of study.

Characteristic metabolites for the metabolic profile of the first component (m_1 in **Figure 5** in Scientific Article III) are L-leucine, L-methionine, fatty acids, AMP and uracil precursors. It is known that L-methionine regulates growth [311,312] and it may play a role in the control of cell cycle regulation [313]. This agrees with the fact that, without L-methionine, yeast growth was delayed. On the other hand, L-leucine controls growth through activation of TORC1 signaling pathway [310]. Thus, in L-leucine starvation conditions, without TORC1 activation, yeast cells cannot produce any metabolic response to alleviate the cellular state derived from the starvation condition. The absence of any appreciable metabolic response and the lack of growth (**Figure 3a** in Scientific Article III) suggests that L-leucine starvation caused an entry into a quiescence state on yeast cells.

The metabolic profile for the second component (m_2 in **Figure 5** in Scientific Article III), relative to the lag and stationary growth phases, shows strong contributions of amino acids, amino acid precursors, citric acid, and trehalose, among others. This result reflects the restoration of amino acid pools after the intense metabolic activity during the exponential phase [336]. On the other hand, since most glucose from the liquid medium has already been consumed, the remaining glucose was stored as trehalose to confront the anticipated adverse carbon-limiting conditions [337].

The two observed de-regulated pathways were described with the two remaining components, m_3 and m_4 (in **Figure 5** in Scientific Article III). m_3 metabolic profile, relative to the metabolic response associated to uracil starvation, is mostly defined by uracil precursors (ureidosuccinic acid, dihydroorotic acid, orotic acid, orotidine-5-phosphate), while m_4 metabolic profile, relative to the metabolic response associated to L-histidine deprivation, includes the histidine precursor EIGP.

As stated in [267], auxotrophic mutants may present incomplete cell cycle arrest under starvation conditions. This was observed for uracil-starved cells, since they did not show a substantial growth (**Figure 3a** in Scientific Article III) because, at the metabolic level, they were metabolically active trying to produce uracil without success. This uncontrolled metabolic activity may explain the short half-life observed for these auxotrophic strains [267] when subjected to starvation for the metabolite they cannot produce.

Finally, as for uracil-starved cells, a similar uncontrolled metabolic activity was detected for L-histidine-starved cells. However, due to the different nature of the starvation, yeast growth was slightly better in the latter growth conditions (**Figure 3a** in Scientific Article III).

4 CONCLUSIONS

From the scientific research included in this Chapter, the following specific conclusions can be extracted:

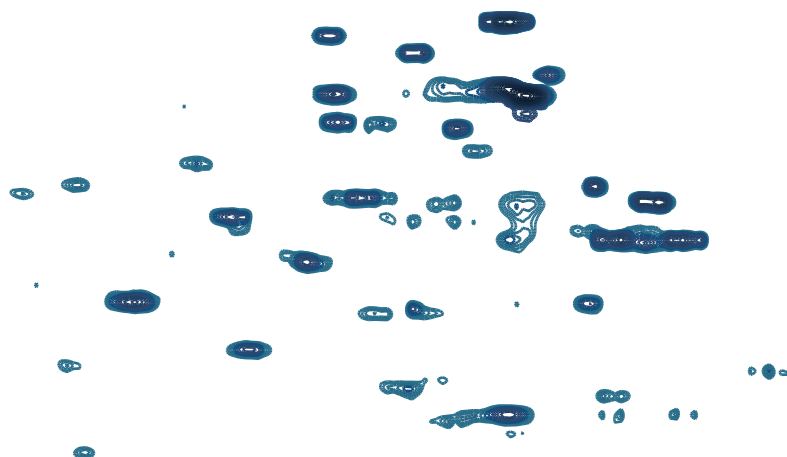
- NMR spectroscopy is a very versatile tool to identify the constituents of complex mixtures, such as metabolomics samples. Stress biomarkers in yeast relative to temperature and starvation could be confirmed using a battery of NMR-based strategies that includes the resonance assignment from investigating conventional 1D and 2D NMR spectra, the acquisition of selective TOCSY NMR experiments to highlight proton resonances from the same spin system, the consultation of NMR databases to retrieve a list of potential compound candidates, the prediction of ¹H NMR spectra of metabolites not listed in the NMR databases, the use of STOCSY to identify statistically correlated resonances, and the spiking of NMR samples with candidate metabolites as the ultimate approach to confirm the presence of these compounds in the mixture.
- The chemometric method MCR-ALS is an effective tool to reveal metabolite relationships derived from an environmental stress or any other biological perturbation. Since these perturbations produce an orchestrated metabolic response in the organism, metabolites affected by the same stimuli are represented within the same component in the MCR-ALS analysis.
- Combining ‘-omics’ data from two high-throughput analytical platforms, such as data from the yeast metabolome (obtained with ¹H NMR) and data from the yeast lipidome (obtained with UHPLC-MS), improves the characterization of the cellular response to the exposed stress.
- *Saccharomyces cerevisiae* (strain BY4741) adapts to perturbations in the growth temperature by changing the composition of the lipidome and of the metabolome level.
 - Regarding the lipidome, biomarkers of yeast growth at low temperatures are short DGs and TGs with a low number of saturations, as well as PIs. For optimal growth conditions, several phospholipid species including PEs and PSs are more abundant than in the other conditions. At higher temperatures, the lipidome accumulates some DGs, PCs, and long poly-unsaturated TGs in a larger amount than in the other screened conditions.
 - Regarding the primary metabolism, organic acids from the Krebs cycle are accumulated at lower temperatures, suggesting that resources are mainly invested pathways required for maintaining life (such as the Krebs cycle) working, which is the main priority of the yeast cells in these undermined conditions. At optimal growth conditions, amino acids and nucleotides are abundant because they are used

to build cell structures and to promote growth. Finally, at higher temperatures, metabolites detected to be descriptive of this condition suggest an increase of the fermentative metabolism, as well as the activation of adaptation mechanisms related to heat stress (trehalose accumulation).

- In *Saccharomyces cerevisiae* (strain BY4741), the ASCA analyses confirmed that the removal of essential nutrients from the media causes a complete metabolic, nutrient-specific, de-regulation.
 - Monitoring of cell growth by means of measurements of OD600 pointed out that the growths of L-methionine- and L-histidine-starved cells were softly repressed, while the growths of L-leucine- and uracil-starved cells were more severely repressed. This apparent growth inhibition for L-leucine- and uracil-starved cells suggested that these cells may have entered into a quiescence state.
 - The chemometric method MCR-ALS exposed that normal growth can be explained by the linear combination of two components, one descriptive of the exponential growth phase, and another descriptive of the lag and late growth phases. In addition, uracil- and L-histidine- starved cells showed an increase of a third MCR-ALS component descriptive of the biosynthesis of uracil and L-histidine precursors, respectively.
 - Despite being the cell growth of uracil-starved cells very limited, their metabolic activity was considerable and mainly focused on the biosynthesis of these uracil precursors. Thus, the apparent quiescence state detected for these starved cells was not a true quiescence state. Since uracil precursors could not be converted to uracil because the yeast strain lacks the required *URA3* gene, this enhanced metabolic activity could not be used to stimulate cell growth and development, and for this reason, it was not detected from the OD600 measurements.
 - The metabolism of L-methionine- and L-leucine- starved cells were explained by the same MCR-ALS components as the metabolism of yeast cells cultured under normal conditions. Nevertheless, the metabolic growth response over time of the starved cells was different to the one observed for the same cells cultured under normal conditions. For L-methionine-starved cells, the exponential growth was delayed, implying that the absence of L-methionine de-regulates the yeast growth cycle. On the other hand, for L-leucine-starved cells, metabolism was completely arrested, indicating an entry into a quiescence state.

Chapter 4

Development and application of data analysis strategies for the investigation of 1D NMR and 2D NMR metabolomics datasets



Different types of data can be acquired using NMR spectroscopy. These data types differ in their dimensionality (*e.g.*, 1D NMR, 2D NMR), the measured nuclei (*e.g.*, ^1H , ^{13}C), connectivity (short-range and long-range), relaxation (T_1 and T_2), phase-sensitivity, and others.

In this Chapter, we have evaluated the intrinsic differences between 1D NMR data and 2D NMR metabolomics data, with a special focus on ^1H NMR and ^1H - ^{13}C HSQC NMR metabolomics data. Apart from the different number of dimensions, major differences were found in resonance overlapping, spectral resolution, sensitivity, and noise intensity. Due to these differences, different analysis strategies should be considered. This is not only true for manually-driven analyses, but also for chemometric analyses.

In the scientific research section of this Chapter, two different analytical tools to investigate 1D NMR and 2D NMR metabolomics datasets, respectively, are proposed. First, in the Scientific Article IV, the MCR-ALS chemometrics method is proposed to resolve ^1H NMR metabolomics datasets. A new noise filtering strategy with applicability for 2D NMR datasets is presented in Scientific Article V. In Scientific Article VI, a metabolomics experiment is performed in parallel with both 1D NMR and 2D NMR spectroscopies. Both datasets have been analyzed with chemometric methods, and the outcomes from these two experiments are presented and compared, providing an insight of the strengths and weaknesses of each type of NMR data used. Finally, in the last section, the results obtained in the research section are discussed.

1 INTRODUCTION

1.1 NUCLEAR RELAXATION AND RESONANCE WIDTH

In NMR, nuclear relaxation describes how the magnetic spin of the measured nuclei evolve over time.

Nuclear relaxation includes two different phenomena, *spin-lattice* relaxation (T_1 , also known as longitudinal relaxation, or relaxation in the z-direction) and *spin-spin* relaxation (T_2 , also known as transverse relaxation, or relaxation in the x-y plane).

On one hand, T_1 relaxation corresponds to the time needed for re-establishing the normal Gaussian population distribution of α and β spin states in the magnetic field. In order to obtain the highest possible sensitivity in the acquisition, it is important to know the T_1 relaxation of the measured compounds. For quantitative purposes, the relaxation delay used should be set long enough to allow the full T_1 relaxation of the measured nuclei.

On the other hand, T_2 relaxation corresponds to the loss of phase coherence among the differently measured nuclei, meaning that the distribution of the magnetic spin vectors disperses over the ideal situation during the decay. This led to a more extended distribution of the resonances frequencies representing all equivalent nuclei. After application of FT, this results in broader resonances.

T_1 and T_2 relaxation depend on molecular size and motion (**Fig. 4.1**). Small and rapidly rotating molecules (such as water) have long T_1 (1-2 s) and long T_2 relaxation times (30-90 ms). For large molecules (such as proteins), molecular motion slows, T_2 shortens and T_1 increases. For small metabolites commonly detected in NMR metabolomics, T_1 and T_2 relaxation times are similar to T_1 and T_2 relaxation times in water, although T_2 can be larger due to chemical processes that affect molecular motion (*e.g.*, interconversion of conformations, chemical exchange) [338].

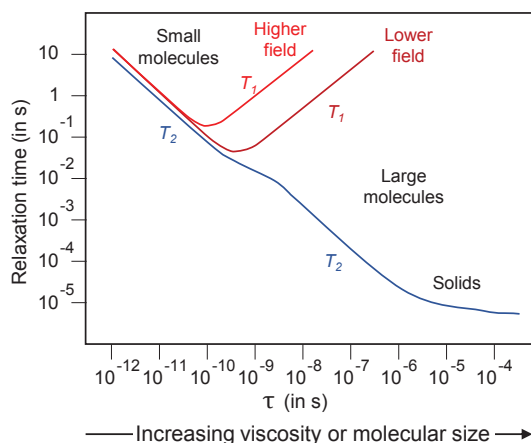


Figure 4.1. Behavior of T_1 and T_2 as a function of correlation time from spin $\frac{1}{2}$ nuclei relaxing by the Dipole-Dipole mechanism. τ = Molecular correlation time: the time it takes the average molecule to rotate one radian (adapted from [339]).

Resonance width can be precisely calculated by means of the effective T_2 relaxation (T_2^*) and the strength of the external magnetic field as follows:

$$v_{1/2} = \frac{1}{\pi T_2^*} = \frac{1}{\pi T_2} + \gamma \Delta B_0 \quad \text{eq. 4.1}$$

where $v_{1/2}$ is the resonance width at half height (in Hz units), γ is the gyromagnetic ratio of the measured nuclei, and ΔB_0 represents the residual macroscopic magnetic field inhomogeneities [338].

T_2^* for proton nuclei is found between 1 and 10 seconds, giving a width at half height of 0.3 or 0.03 Hz, respectively. For other nuclei of spin= $\frac{1}{2}$, such as ^{13}C , T_2^* is around 0.2 and 50 seconds, giving a resonance width at half height of 1.6 Hz and 0.006 Hz, respectively.

In the FT-NMR spectra, resonance width is also dependent on the external magnetic field. When the external magnetic field is increased, the signal resolution is improved, since resonances cover a smaller range of the total measured frequency domain, resulting in sharper peaks when converted to ppm units. For instance, a resonance width of 0.3 Hz corresponds in a 300 MHz magnetic field to 0.001 ppm (0.3/300) width, while the same resonance covers a width of 0.0005 ppm (0.3/600) in a 600 MHz magnetic field.

Apart from field inhomogeneities, long T_2 , and the use of lower external magnetic fields, resonances are also broadened due to the inherent resonance patterns derived from the *spin-spin* coupling (resonance multiplicity), and to chemical exchange dynamics [131]. Then, if low intense resonances are broadened due to any (or combination) of these five factors, their signal-to-noise ratio will worsen, to the extent that the resonances may not be even detected.

1.2 NMR DATA

A single 1D NMR spectrum can be regarded, in the mathematical sense, as a vector of intensities. A single 2D NMR spectrum can be stored as a two-dimensional data matrix (f_1 measurements in rows and f_2 measurements in columns). A single 3D NMR spectrum can be stored in the three dimensions of a data cube, and so on.

Typically, one 1D NMR spectrum consists of thousands of data points. For instance, ^1H NMR spectra have normally 16k, 32k, or 64k acquired data points. Multi-dimensional NMR spectra, because of the vast amount of data measured, are usually acquired with a lower digital resolution (number of data-points in the frequency domain) per screened dimension. For example, a typical data matrix from a 2D NMR spectrum has 2,048 columns and 1,024 rows (~2 million data values in total), while the corresponding data cube from a 3D NMR spectrum have around 256 columns, 256 rows and 128 slices (2D planes), giving a total of circa 8 million of intensity data points [340].

Even though the digital resolution per screened dimension is lower in multi-dimensional NMR spectra, the acquisition time used is drastically increased when compared to 1D NMR spectroscopy because the total number of data points is considerably larger.

Having said this, in NMR metabolomics, either for high-resolution ^1H NMR spectra or for low-resolution multi-dimensional NMR spectra, some signals will still appear overlapped.

Resonances are found at a fixed chemical shift predetermined by the chemical environment of each measured nuclei. This means that resonances from nuclei of similar structures will have similar chemical shifts, and therefore, these regions will be crowded with multiple resonances (**Fig. 4.2B**), while other regions will be mostly empty (**Fig. 4.2A**). Moreover, within these crowded regions, it is likely that more resonances will be found near the center and fewer will be near the edges [340].

Technically, in a ^1H NMR spectrum, within one ppm unit width, 33 resonances of width 0.03 ppm (average width for a singlet signal at 500 MHz) could be found without overlap. However, due to the fact that *spin-spin* coupling constants are larger than resonance widths, resonances become commonly wider, and the possibility of finding two overlapped resonances becomes very high.

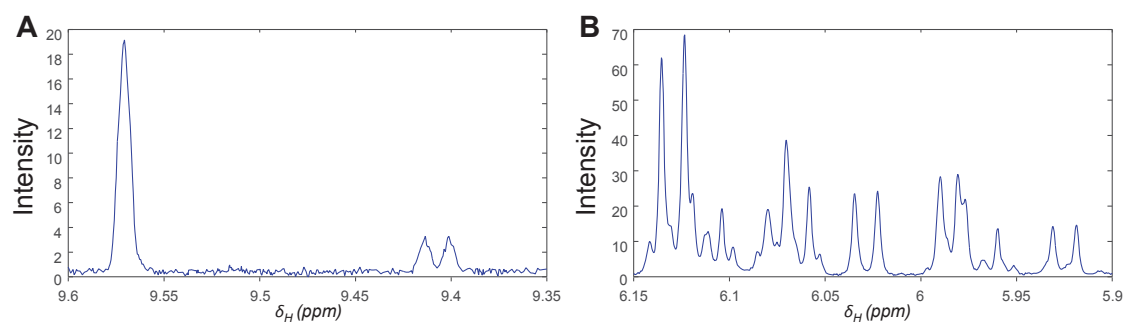


Figure 4.2. Zoomed regions of a ^1H NMR spectrum acquired in a 500 MHz NMR spectrometer. **A)** Isolated peaks. **B)** Overlapped peaks. Note that in **A** and **B** figures, the two zoomed regions have the same spectral width.

In 2D NMR, the second dimension favors the separation of some overlapping peaks, while some others remain together. The quality of the separation depends on the used NMR pulse sequence. For 2D NMR pulse sequences that measure resonances from only short-range correlations, where only one (^1H - ^{13}C HSQC NMR) or a few (^1H - ^1H COSY NMR) signals are expected per measured nuclei, overlapping will be really low. On the other hand, for 2D NMR measuring resonances from long-range correlations (*e.g.*, ^1H - ^{13}C HMBC NMR, ^1H - ^{13}C HMQC NMR, ^1H - ^1H TOCSY NMR), several cross-peak correlations per measured nuclei are expected, and signal overlapping will be more common. In addition, as in ^1H NMR spectra, resonances from 2D NMR spectra appear at a predetermined spectral region that depends on the molecular structure. This causes, for instance, that ^1H - ^{13}C HSQC NMR cross-peaks arise over the diagonal of the spectrum (**Fig. 4.3**), or that ^1H - ^1H TOCSY cross-peaks appear concentrated in the aliphatic ($\delta_{\text{H}1} = 0.8\text{-}2.5$ ppm, $\delta_{\text{H}2} = 0.8\text{-}2.5$ ppm), sugar ($\delta_{\text{H}1} = 3.0\text{-}4.5$ ppm, $\delta_{\text{H}2} = 3.0\text{-}4.5$ ppm), and aromatic ($\delta_{\text{H}1} = 7.0\text{-}9.0$ ppm, $\delta_{\text{H}2} = 7.0\text{-}9.0$ ppm) regions (**Fig. 4.4**).

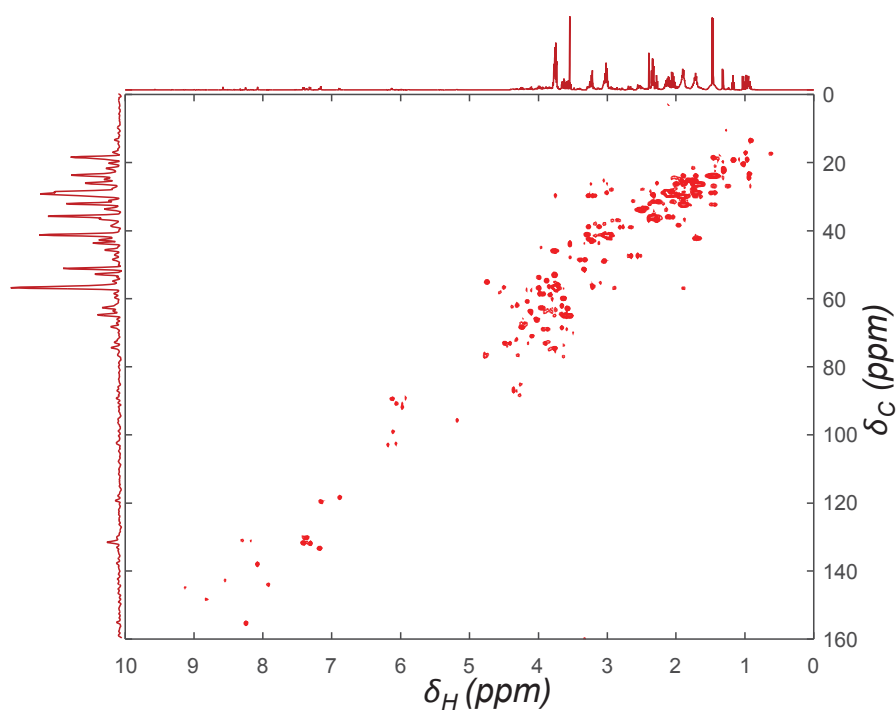


Figure 4.3. ^1H - ^{13}C HSQC NMR spectrum of a metabolomics sample (yeast extract).

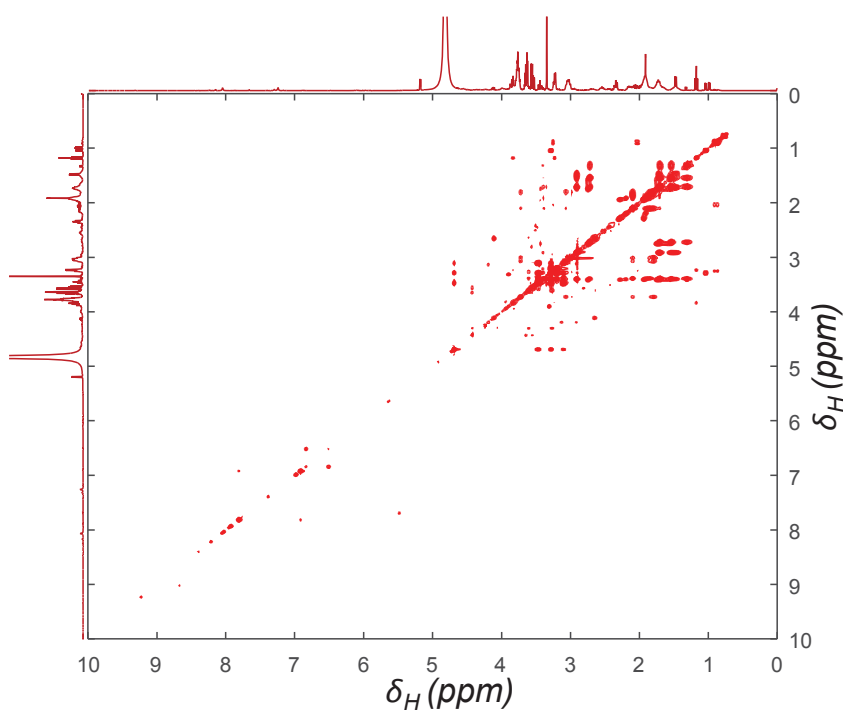


Figure 4.4. ^1H - ^1H TOCSY NMR spectrum of a metabolomics sample (yeast extract).

Thus, in NMR, signal overlapping is not caused by a low digital resolution, but due to a low signal resolution. In a 500 MHz NMR spectrometer, the width of an isolated resonance in a ^1H NMR is between 0.03 (singlet resonance) and 0.1 (multiplet resonance) ppm. That means that for a 64k data points spectrum covering a spectral width of 12 ppm, the resonance is defined by between 164 and 546 data points.

On the other hand, in a ^1H - ^{13}C HSQC NMR, the width of a cross-peak resonance presents the same width in the proton dimension than in the ^1H NMR, and between 0.5 and 1.0 ppm in the carbon dimension. Thus, in the same NMR instrument, for a ^1H - ^{13}C HSQC NMR spectrum with 2,048 δ_{H} (spectral width of 12 ppm) and 1,024 δ_{C} (spectral width of 180 ppm), the digital size of the cross-peak is between 5 and 17 data points in the proton dimension and between 3 and 6 data points in the carbon dimension (between 15 and 102 data points if the two dimensions are considered).

To improve signal resolution by instrumental means and minimize overlapping, the only feasible way is by using a stronger external magnetic field. However, the most powerful NMR instruments (>1 GHz) cost over a 1M \$ nowadays and are not affordable for most research institutions.

In this Thesis, we have presented a strategy to resolve resonance overlapping by means of the chemometric method MCR-ALS instead.

In the early days of Chemometrics, several chemometric methods were used to solve signal overlapping problems in spectrophotometric analysis [179,341,342] and, in the last decades, these methods have expanded into other analytical areas, such as in NMR. In these chemometric-based NMR analyses, NMR spectra of pure compounds were resolved from their mixtures (for example, [160,343-345]). These studies can be considered to be the previous background works to the methodology presented in this Thesis and for this reason, they are briefly introduced in the following section.

1.3 RESOLUTION OF NMR DATA BY CHEMOMETRICS: PREVIOUS WORK

The decomposition by chemometric means of an NMR spectrum from a mixture of compounds into the set of NMR spectra of their pure constituents has been always considered a compelling challenge.

The first study pursuing this goal appeared in 1994 [343], and it has been recurrently investigated by chemometricians (*e.g.*, Willem Windig [344,346] and Rasmus Bro [347]) and by NMR spectroscopists (*e.g.*, Gareth A. Morris [348-351] and Rafael Brüschweiler [162]).

In this first reported example, cross-peak resonances from a 2D NMR spectrum were decomposed as a set of 1D NMR spectra [343]. However, these 1D NMR spectra were not representative of pure compounds, but of pure resonances, since each spectrum contained only one resonance, and therefore a chemical compound was represented by as many components as detected resonances.

In order to improve the relevance of the resolution, most of the following studies analyzed diffusion NMR spectra (*e.g.*, PGSE NMR [344,346,352,353], DOSY NMR [347]). Diffusion is a physical property specific for each compound, since it depends on the molecular size and shape. Thus, with diffusion NMR spectroscopy, resonances from the same compound present the same diffusion profile and can be therefore resolved in the same component.

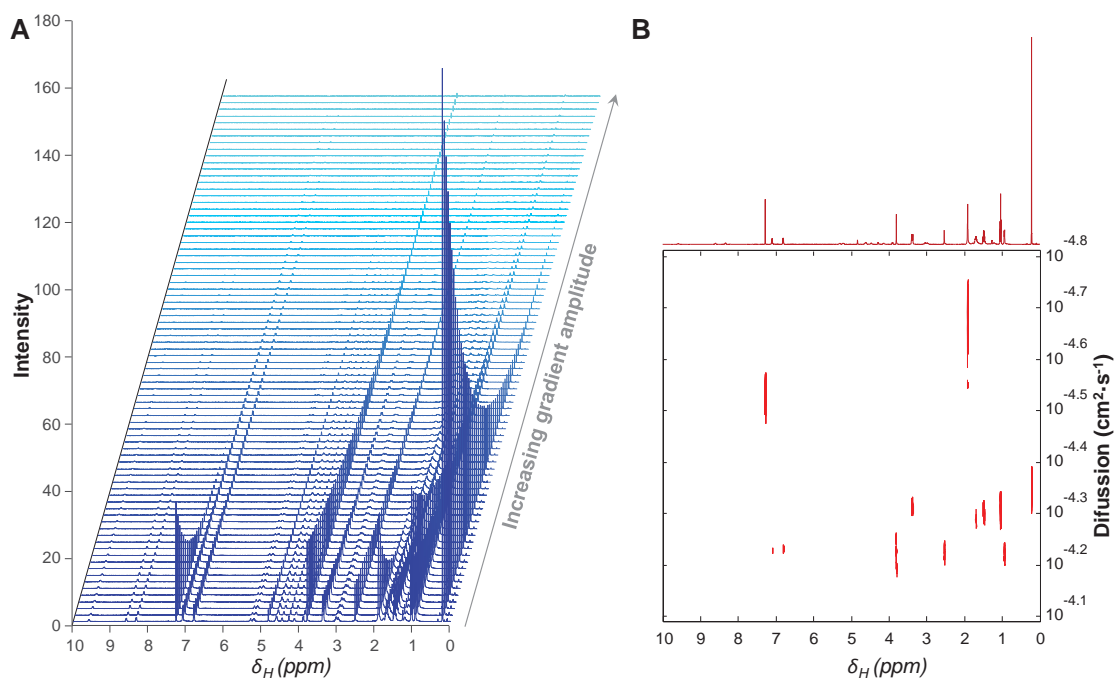


Figure 4.5. Diffusion NMR spectra **A)** PGSE **B)** DOSY. Data provided by [354].

Despite this approach produces satisfactory results, it is limited to mixtures of compounds with different self-diffusion rates, such as mixtures of surfactants [344], polymers [346,353] and lipoproteins [347]. For crowded spectra of chemical compounds of low molecular weight, such as in NMR metabolomics experiments, this approach is not practical because the diffusion dimension is not sufficient to separate the different resonances. Having said this, simple mixtures of 2-3 small metabolites have been satisfactorily analyzed with High-Resolution DOSY [348-351].

Thus, for the analysis of many complex mixtures, spectra different from diffusion NMR needed to be used.

Promising results were obtained in the decomposition of 2D ^1H - ^1H TOCSY NMR spectra, used in metabolomics for structural elucidation. With the chemometric decomposition of a single 2D TOCSY NMR spectrum, one ^1H NMR spectrum for each proton spin system was obtained [161,162], and therefore, the resolved ^1H NMR spectra only coincided with the real ^1H NMR spectra in those cases where all intramolecular protons belong to the same proton

spin system. Although the resolved ^1H NMR spectra were not complete in some cases, they were sufficient enough to identify metabolites.

A substantial improvement in the resolution can be obtained by performing a simultaneous resolution analysis of several NMR spectra. For instance, 23 2D ^1H - ^{15}N HSQC NMR spectra of 5 chemical species were resolved from the corresponding mixture dataset [160]. To resolve all resonances from the same species in the same component, the 23 spectra were vectorized and appended column-wisely before chemometric analysis as shown in **Figure 2.27** (see page 52) [160].

Since a vectorized 2D NMR spectrum can be regarded as a 1D NMR spectrum containing resonances with a low degree of overlapping, the study in [160] can be considered as one of the first studies of resolution of 1D NMR datasets.

Results from this work [160] and other subsequent works [239,241,355] demonstrated that the best resolutions of 1D NMR datasets were obtained when the differences in metabolite concentrations among samples were large. Yet, this resolution is hindered if the dataset contains strongly overlapped signals.

For moderately overlapped ^1H NMR data, such as LC- ^1H NMR data, the impact of resonance overlapping can be diminished by using spectral constraints (limiting the NMR regions where the resonance from a given metabolite may be found) [356,357].

In ^1H NMR metabolomics datasets, the differences in metabolite concentrations among samples is limited because, in a living organism, metabolites are co-regulated. This causes that components resolved by chemometric methods contain resonances from co-regulated metabolites, where every set of co-regulated metabolites is descriptive of a metabolic event in the analyzed samples [180,251]. In order to separate the different co-regulated metabolites into different components, as suggested by the previous work with LC- ^1H NMR data [356,357], spectral constraints must be used during chemometric multivariate resolution methods. Within the framework of this Thesis, we have explored this strategy, as it has not been investigated in detail before. The strengths and weaknesses of this chemometrics strategy are presented in the second part of this chapter.

1.4 PROPOSED CHEMOMETRIC STRATEGIES

For 1D NMR (specifically, ^1H NMR) metabolomics datasets, we propose the multivariate resolution method based on the MCR-ALS method combined with selective constraints that allow the separation of the ^1H NMR spectra of the pure metabolites (Scientific article IV).

For 2D NMR (specifically, ^1H - ^{13}C HSQC NMR) metabolomics datasets, we propose a noise-filtering approach which keeps only those variables containing information from meaningful resonances. After filtering noise, samples become much easier to analyze and resonances can be much accurately integrated (Scientific article V).

2 SCIENTIFIC RESEARCH

2.1 SCIENTIFIC ARTICLE IV

Untargeted Assignment and Automatic Integration of ^1H NMR metabolomic datasets using a Multivariate Curve Resolution Approach.

Authors: Puig-Castellví F., Alfonso I., Tauler R.

Citation reference: *Anal. Chim. Acta* (2017), 964: 55-66.

DOI: 10.1016/j.aca.2017.02.010



Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Untargeted assignment and automatic integration of ^1H NMR metabolomic datasets using a multivariate curve resolution approach[☆]



Francesc Puig-Castellví^a, Ignacio Alfonso^b, Romà Tauler^{a,*}

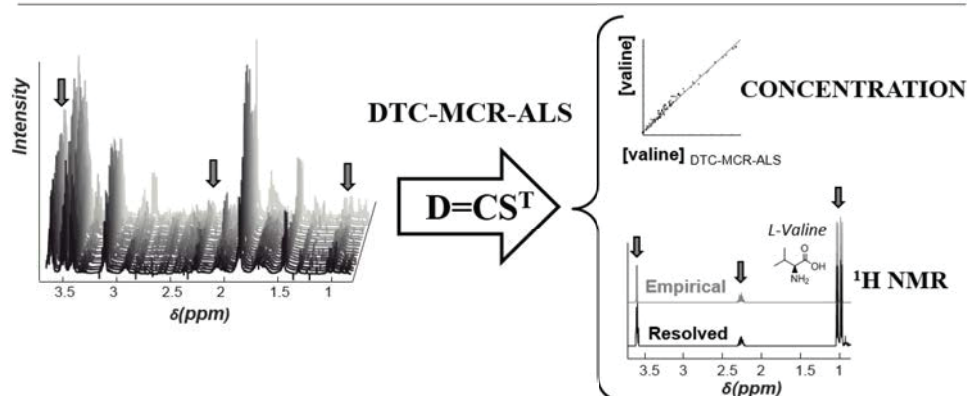
^a Department of Environmental Chemistry, Institute of Environmental Assessment and Water Research (IDAEA-CSIC), Barcelona, Spain

^b Department of Biological Chemistry and Molecular Modelling, Institute of Advanced Chemistry of Catalonia (IQAC-CSIC), Barcelona, Spain

HIGHLIGHTS

- An untargeted resolution approach of ^1H NMR metabolomics datasets is proposed.
- The approach uses MCR-ALS combined with the application of equality constraints.
- Equality constraints were designed based on observed proton inter-correlations.
- This strategy was validated with simulated and real ^1H NMR metabolomics datasets.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:

Received 19 October 2016

Received in revised form

9 February 2017

Accepted 10 February 2017

Available online 20 February 2017

Keywords:

Metabolomics

Nuclear magnetic resonance

Multivariate curve resolution

In this article, we propose the use of the Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS) chemometrics method to resolve the ^1H NMR spectra and concentration of the individual metabolites in their mixtures in untargeted metabolomics studies. A decision tree-based strategy is presented to optimally select and implement spectra estimates and equality constraints during MCR-ALS optimization.

The proposed method has been satisfactorily evaluated using different ^1H NMR metabolomics datasets. In a first study, ^1H NMR spectra of the metabolites in a simulated mixture were successfully recovered and assigned. In a second study, more than 30 metabolites were characterized and quantified from an experimental unknown mixture analyzed by ^1H NMR. In this work, MCR-ALS is shown to be a convenient tool for metabolite investigation and sample screening using ^1H NMR, and it opens a new path for performing metabolomics studies with this chemometric technique.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

High-throughput analytical techniques can be used at present in metabolomics studies to simultaneously characterize a large number of small molecules from biologically relevant samples [1]. Among most used analytical techniques are Nuclear Magnetic Resonance (NMR) spectroscopy and Mass Spectrometry (MS) hyphenated techniques, such as Gas Chromatography (GC-MS), Liquid

[☆] Selected paper from XVI Chemometrics in Analytical Chemistry, 6–10 June 2016, Barcelona, Spain.

* Corresponding author.

E-mail address: Roma.Tauler@idaea.csic.es (R. Tauler).

Chromatography (LC-MS) and Capillary Electrophoresis (CE-MS) [2–5]. In the last years, untargeted Mass Spectrometry (MS) metabolomics has been presented as a promising tool for studying metabolome changes of living organisms without assuming any prior metabolic knowledge of the studied samples [6]. However, untargeted curation methods are still scarce in Nuclear Magnetic Resonance (NMR) spectroscopy metabolomics experiments, and metabolite assignment depends mostly on the NMR analyst's skills.

NMR and MS spectra generated in metabolomics studies usually contain hundreds to thousands of signals, whose identification and interpretation require significant effort. In order to facilitate their analysis, different strategies have been proposed in the literature. Most of MS-based studies use methods that reduce the original data dimensions by extracting significant signals from raw data and assigning them to metabolites [7,8].

In NMR metabolomics, most of the methods are focused on the identification of metabolite NMR profiles of known compounds. Despite the fact that prior knowledge about sample composition is not required to identify these metabolite profiles (since proton resonances usually provide enough structural information by themselves), most of the recent methods rely on independent NMR spectral libraries (i.e. HMDB [9] and MMCD [10]) and also on NMR spectra fitting software approaches that already have their own reference library, such as Bayesil [11] or Chenomx [12]. However, using these methods, it is still difficult to obtain quantitative data for a new biomarker if no prior spectroscopic data from this compound is available.

Other data analysis approaches that can be used in NMR metabolomic studies are those based on the application of chemometric strategies such as Principal Component Analysis (PCA) or Partial Least Squares-Discriminant Analysis (PLS-DA), which are directly applied to the raw ^1H NMR spectral matrix to identify possible significant variables (or chemical shifts). However, these significant variables have no additional spectroscopic information (multiplicity, coupling constant) and further spectroscopic analysis is required for a proper peak assignment.

Therefore, in order to complete reference NMR spectral libraries or to improve biomarker characterization, the acquisition of complementary spectra is usually required. Although NMR metabolomics using bidimensional NMR spectra provides more structural information and reduces signal overlapping, this strategy is usually not preferred (with some exceptions [13,14]) because spectral acquisition time is substantially increased and the associated post-processing step would also need sophisticated and complex data analysis approaches.

In order to get structural information directly from ^1H NMR spectra datasets, several data analysis chemometric tools have been proposed. One of the best known approaches is the Statistical Total Correlation Spectroscopy (STOCSY) [15], which is based in the principle that resonances from the same molecule will present the same variation among different samples. Therefore, it is possible to identify which signals belong to the same molecule since they should show strong correlation. However, in datasets where signal overlapping changes drastically among samples, as in time-course experiments, correlation values can be misleading and pattern recognition becomes more difficult. Other chemometric methods have been proposed, such as Independent Component Analysis (ICA) [16] or Multivariate Curve Resolution [17]. ICA has shown promising results for relatively simple mixtures, resolving at most 9 compounds under the constraint of signal independence [18]. On the other hand, MCR-ALS has been shown to be a reliable method for the resolution of metabolite mixtures [19,20]. However, in all these methods, correct signal resolution will depend on how complex the investigated system is, especially in relation to the changes in the concentrations of the different metabolites in the

analyzed sample datasets. When metabolite concentrations change independently in the samples, the resolution of metabolic spectral profiles will be better [20].

In ^1H NMR metabolomics datasets, metabolite signals distribution does not usually accomplish the full rank conditions (chemical rank equals to the number of metabolites) nor the conditions for elimination of rotation ambiguities, since metabolites are always present at a fixed concentration range (varying their concentration not independently and only a little) for ensuring life survival of the studied organism, and usually concentrations of biologically-linked metabolites are strongly correlated. For this reason, when multivariate resolution methods are applied, like MCR-ALS or PARAFAC [21,22] to these datasets, the resolved components are in fact a mixture of metabolites describing some relevant biological processes.

In this work we demonstrate how MCR-ALS can be adapted to achieve an improved resolution of individual metabolite ^1H NMR profiles, without the need of knowing in advance their pure spectral profiles. Using this strategy, the proposed method provides a more suitable strategy for assisting metabolite assignment and for automatizing metabolite quantification in untargeted ^1H NMR metabolomics studies.

2. Material and methods

Three datasets were investigated in this work: two simulated datasets (\mathbf{X}_1 and \mathbf{X}_2) and one experimental dataset (\mathbf{X}_3).

2.1. Simulated ^1H NMR datasets (\mathbf{X}_1 and \mathbf{X}_2)

Two simulated ^1H NMR datasets have been analyzed in the current article, \mathbf{X}_1 and \mathbf{X}_2 , both of them with a set of 60 ^1H NMR spectra from different samples (rows) containing 10,801 ppm values (columns) each.

Individual ^1H NMR metabolite spectrum profiles, \mathbf{S}^0 , used to build the two simulated datasets was identical, comprising the same 10 metabolites (Fig. 1A), but different metabolite concentrations were used in each case. On dataset \mathbf{X}_1 , metabolite concentrations followed a random uniform distribution (Fig. 1B), whereas on dataset \mathbf{X}_2 , metabolite concentrations followed a normal distribution, whose mean and standard deviations were calculated from a previous experimental study of metabolic extracts of yeast cells cultured at 15 °C, 30 °C, 37 °C and 40 °C (see Supplementary material Appendix A), as represented in Fig. 1C. Mean abundances for every metabolite on \mathbf{X}_2 followed the order Glu > Lys >> Cit > Asn > Gly > His > Leu > Orn > AMP >> Tyr. Individual ^1H NMR spectra, \mathbf{S}^0 , were generated by BATMAN R-package [23] during the previous analysis of the experimental data, normalized to the total integral, and weighted by the metabolite proton number (see Supplementary material Appendix A).

Therefore, one ^1H NMR spectrum in one sample i , \mathbf{x}_i , with the contributions from the 10 metabolites, was modelled according to Eqn. (1) below:

$$\mathbf{x}_i = \sum_{j=1}^{J=10} c_{i,j}^0 \mathbf{s}_j^0 + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma) \quad (1)$$

Thus, \mathbf{x}_i is calculated as the sum of all metabolic contributions, defined by the product of each individual ^1H NMR metabolite spectrum profile, \mathbf{s}_j^0 , weighted by the concentration of this metabolite j in sample i , $c_{i,j}^0$. In addition, normally distributed random noise, with standard deviation σ , defined as in Eqn. (2), was also generated and added to each \mathbf{x}_i .

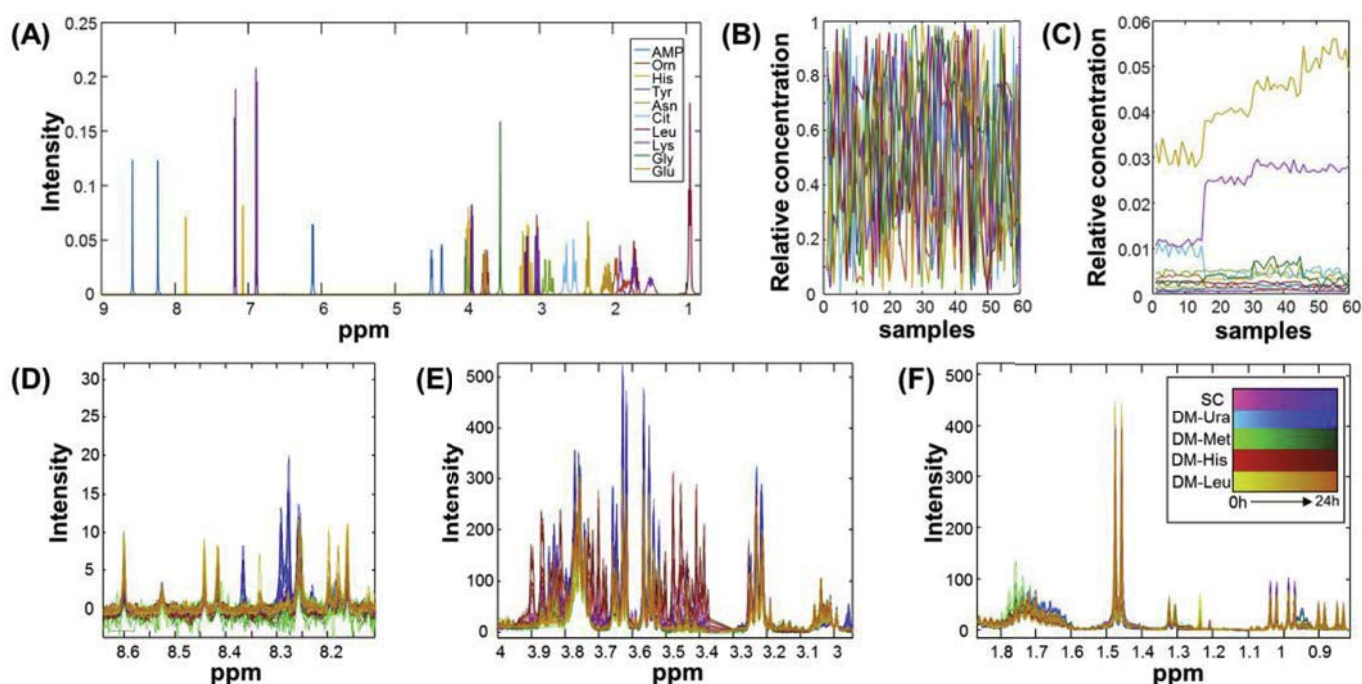


Fig. 1. ^1H NMR datasets. (A) ^1H NMR metabolite spectra (S^0) used for the simulated mixture datasets (X_1 and X_2 , see section 2.1). (B) Randomly distributed concentrations of the different metabolites in X_1 dataset. (C) Uniformly distributed concentrations of the different metabolites in X_2 dataset. (D)–(F) Different ^1H NMR regions of the experimental data set, X_3 (see section 2.2). Metabolites abbreviation: AMP, adenosine monophosphate; Orn, L-ornithine; His, L-histidine; Tyr, L-tyrosine; Asn, L-asparagine; Cit, citric acid; Leu, L-leucine; Lys, L-lysine; Gly, glycine; Glu, L-glutamate.

$$\sigma = \frac{\max(\mathbf{X})}{\text{SNR}} \quad (2)$$

σ was estimated by dividing the maximum peak intensity in the given datasets by the signal to noise ratio (SNR) defined by the user, as in Ref. [24]. The SNR used varied between 50 and 1000.

2.2. Experimental ^1H NMR dataset (X_3)

The experimental dataset X_3 consisted of 90 spectra from yeast cells cultured at 5 different growth conditions, whose samples were taken at 6 different time-points during a day period.

^1H NMR spectra were recorded in a 400 MHz Varian spectrometer, using a spectrometer frequency of 400.14 MHz with a OneNMR Probe and a ProTune System (Agilent), using the s2pul pulse sequence available in the vendors' software. Spectral size range covered from -2 to 10 ppm. 512 scans were used with a relaxation delay of 5 s and the receiver gain was fixed to 34.

65 k spectral data points were obtained for every sample. After excluding for the analysis the spectral regions of 4.41–5.16 ppm, 3.30–3.37 ppm, 7.64–7.69 ppm, below 0.8 ppm and above 10.3 ppm, the spectral domain of the dataset consisted of 35,342 data points. Highlighted ^1H NMR regions of these spectra are shown in Fig. 1D–F. For more information about sample preparation, see Puig-Castellví et al. (2016) [25].

2.3. Chemometrics methodology

2.3.1. MCR-ALS method

Chemometric data analysis was performed using the MCR-ALS GUI 2.0 [26] under Matlab 2014b (The Mathworks Inc. Natick, MA, USA) environment. MCR-ALS [17] is a chemometric method which decomposes a given data matrix using the following bilinear model:

$$\mathbf{X} = \mathbf{C}\mathbf{S}^T + \mathbf{E}, \quad (3)$$

Eqn. (3) is solved iteratively by an Alternating Least Squares algorithm which calculates concentration \mathbf{C} and pure spectra \mathbf{S}^T matrices which optimally fit the experimental data matrix \mathbf{X} . N represents the number of components used in the decomposition generated in the MCR-ALS analysis. N can be initially estimated by using the singular value decomposition (SVD) [27]. \mathbf{E} matrix contains the residual information not explained by the model using the N considered components.

MCR-ALS iterative process begins with an initial estimation of the concentrations (\mathbf{C}) or spectra (\mathbf{S}^T). These initial estimates are usually found using the most dissimilar rows (the purest ^1H NMR spectra) or columns (the purest concentrations) from the analyzed data [28]. In this study, initial estimates used were the purest \mathbf{S}^T or the purest \mathbf{C} estimated from the matrix to be resolved, or the ^1H NMR \mathbf{S}^T reconstructed profiles obtained using the Decision Tree of Correlations (DTC) methodology proposed in this work (see section 2.3.2).

Concentration and spectral profiles obtained by MCR-ALS may not be the correct ones due to the existence of rotational ambiguities in the solution of Eqn. (3), [29,30]. For instance, in Fig. 2A, the optimal solution found in the third component (green) presented wrong spectral regions (in red, pointed out with an arrow in the given figure) due to these ambiguities. Constraints are added to drive the iterative process to a better solution, more in agreement with the true one. For instance, in this work, non-negativity constraints on both concentration and spectral profiles were applied during their ALS optimization. A different type of possible constraint is the equality constraint. This constraint can be used in the spectral domain to fix the values of the signal of some component (Fig. 2B). For example, in this figure, the undesired signal found in the third component can be removed without loss

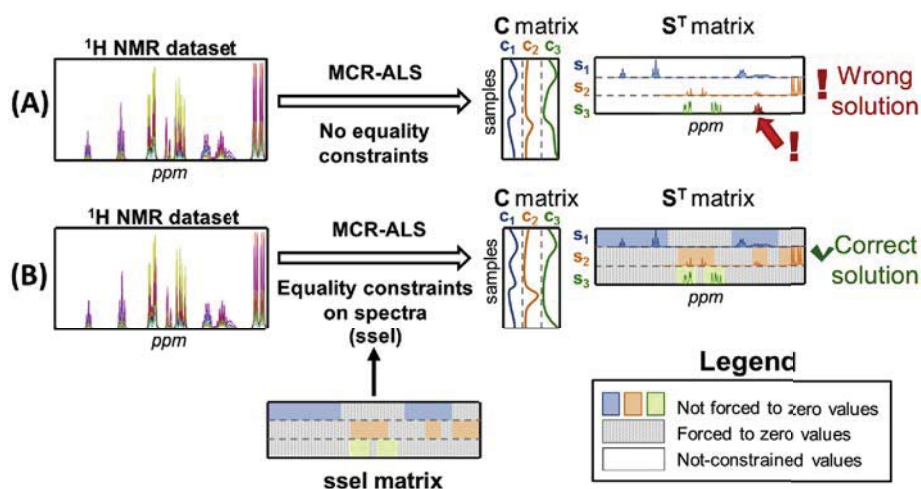


Fig. 2. MCR-ALS analysis with constraints. (A) Only non-negativity constraints. The interferent signal in red was resolved in the third component (in green) due to rotation ambiguities. (B) Application of spectrum equality constraints (sael). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

of fit by forcing this signal not to be present in this region. See Refs. [17,26,30] for more details about the MCR-ALS method and the possible constraints to be applied. The quality of the MCR-ALS model was measured evaluating the lack-of-fit (lof in Eqn. (4)) parameter, and the percent of explained variance (R^2 in Eqn. (5)).

$$\text{lof}(\%) = 100 \sqrt{\frac{\sum \sum e_{ij}^2}{\sum \sum x_{ij}^2}} \quad (4)$$

$$R^2(\%) = 100 \frac{\sum \sum x_{ij}^2 - \sum \sum e_{ij}^2}{\sum \sum x_{ij}^2} \quad (5)$$

In addition, on the simulated datasets, correlation coefficients (Eqn. (6) and Eqn. (7)) and similarity angles (Eqn. (8)) between the resolved (spectral or concentration) profiles, \mathbf{s}_n^r or \mathbf{c}_n^r , and the original ones, \mathbf{s}_j^o or \mathbf{c}_j^o , respectively, were calculated as shown below:

$$r_{\text{so},\text{sr}}^2 = \frac{\mathbf{s}_j^o \widehat{\mathbf{s}}_n^r}{\|\mathbf{s}_j^o\| \|\widehat{\mathbf{s}}_n^r\|} \quad (6)$$

$$r_{\text{co},\text{cr}}^2 = \frac{\mathbf{c}_j^o \widehat{\mathbf{c}}_n^r}{\|\mathbf{c}_j^o\| \|\widehat{\mathbf{c}}_n^r\|} \quad (7)$$

$$\alpha = \cos^{-1}(r^2) \quad (8)$$

where a small α and a high r^2 values between the two vector profiles implies a better agreement between the resolved and the original profiles.

2.3.2. DTC-MCR-ALS method

Proton resonances from ^1H NMR spectra are highly selective, broadly dispersed along the entire spectral domain, and highly correlated with other resonances belonging to the same molecule. All these specific particularities of NMR can be used to deduce the resonance distribution of every metabolite, which can be then used to create MCR-ALS equality constraints for every one of them.

The methodology proposed in this work, which will be referred here as Decision Tree of Correlations-MCR-ALS (or DTC-MCR-ALS), is based on the division of the dataset on different spectral subregions, and in the application of MCR-ALS to each one of them (Fig. 3A). Limits (beginning and end) of these spectral subregions are selected from those chemical shifts whose intensities are at the noise level, and they should contain at least one intense resonance.

MCR-ALS is then applied to every spectral subregion, and a small set of metabolite features (\mathbf{s}_k) are resolved together with a set of concentration profiles, \mathbf{c}_k . Subindex k represents the number of resolved metabolite features (in Fig. 3, each metabolite features is represented by a different roman number, I-X).

All \mathbf{c}_k profiles are joined in a single concentration data matrix and the pair-wise correlation coefficients between them (or $\mathbf{c}_k\mathbf{c}_k'$ correlations) are calculated. Since \mathbf{c}_k profiles from the same molecule will be highly correlated, this information can be used to relate them to the same individual metabolite. In addition, after this grouping strategy is performed, metabolite assignment to every component can be performed by evaluating their corresponding spectral features, or \mathbf{s}_k , altogether, since they will provide structural information about the sample metabolites.

Since signal resolution is negatively affected by signal overlapping and noise, correlation coefficients may vary depending on the influence of these two parameters and, therefore, fixing the same correlation threshold for all cases is not a suitable approach for grouping \mathbf{c}_k . To circumvent this problem, a decision tree-based approach is proposed to group the different metabolic spectral features instead of using a single correlation threshold value. In this decision tree, metabolic spectral features are sequentially grouped from the most correlated ones to the least ones, until the next spectral feature \mathbf{s}_k does not correspond to the same metabolite.

The simplest approach to check whether the picked \mathbf{s}_k is from the same metabolite is by visual inspection. For instance, if the \mathbf{s}_k is noisy (it is very poorly resolved), then we will ignore the matching suggested by DTC. Therefore, noisy \mathbf{s}_k can be removed from the pool of features when they are identified during DTC analysis.

Also, since the concentration range of metabolites in ^1H NMR metabolomics datasets may vary between 2 or 3 orders of magnitude, if the picked feature is found in a different magnitude than the previously selected features, then it is not considered to be from the same metabolite. To visually compare the magnitude of two or more features, we need to plot the reconstructed features to

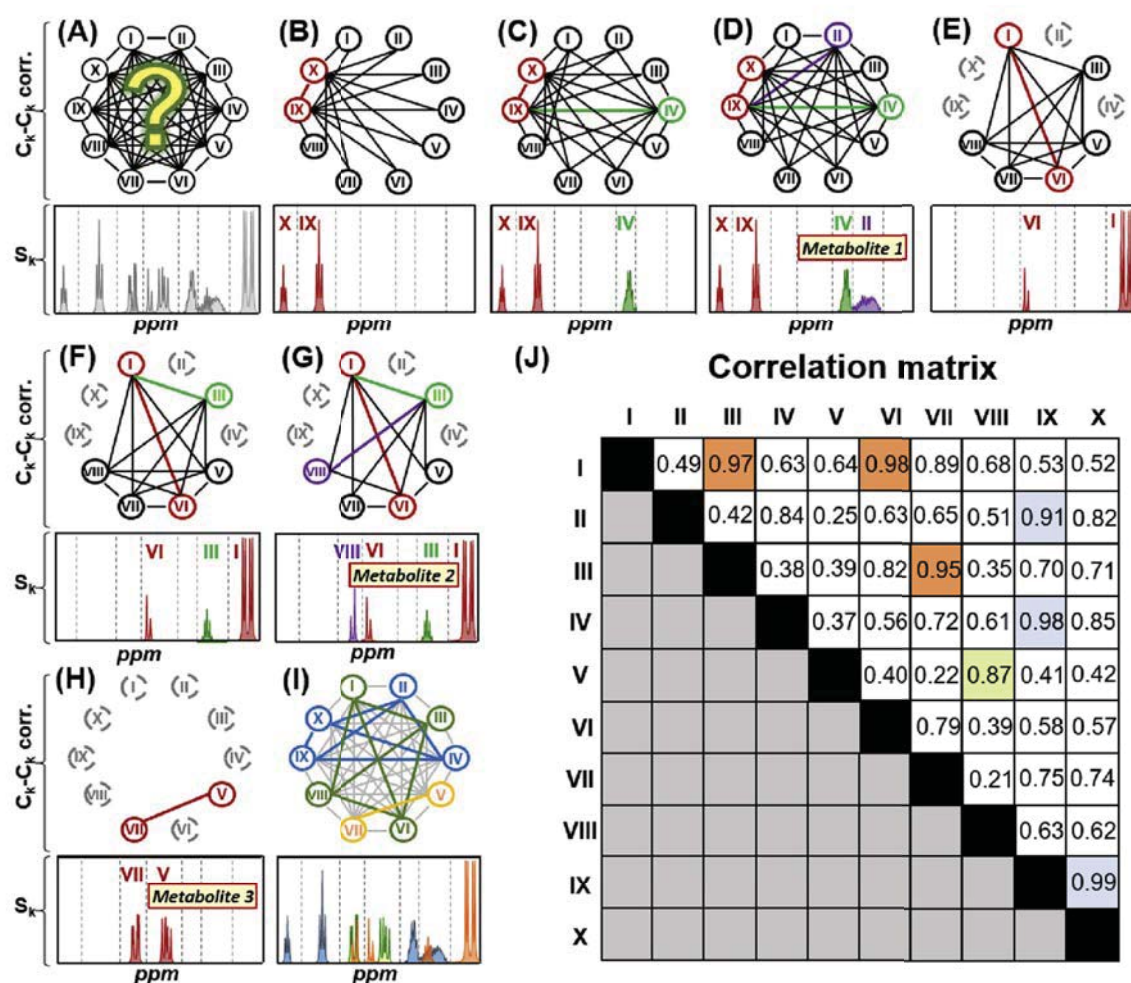


Fig. 3. Workflow for the reconstruction of the ^1H NMR metabolite profiles using the DTC-MCR-ALS method proposed in this work. (A) Calculation of pair-wise correlation coefficients between all \mathbf{c}_k (above). Spectral features, \mathbf{s}_k , in grey, are assigned to metabolites deduced using the grouping strategy of searching the highest \mathbf{c}_k - \mathbf{c}_k correlations sequentially (see section 2.3.2 and below). (B) Selection of the two highest pair-wise correlated \mathbf{c}_k , the \mathbf{c}_{IX} - \mathbf{c}_X pair, which are represented in red (above). The corresponding \mathbf{s}_k are initially assigned to one of the non-yet identified metabolites (below). (C)–(D) Looking for the maximum correlation coefficient value between one already assigned and one unassigned \mathbf{c}_k . (E)–(H) Repeat the steps (B)–(D) until all \mathbf{c}_k are assigned. (I) Reconstruction of three different ^1H NMR metabolite spectra profiles from \mathbf{s}_1 - \mathbf{s}_{III} , \mathbf{s}_{IV} - \mathbf{s}_{VIII} , \mathbf{s}_{II} - \mathbf{s}_{IV} - \mathbf{s}_{IX} - \mathbf{s}_X and \mathbf{s}_{V} - \mathbf{s}_{VII} . See section 2.3.2 for further details. (J). Correlation matrix of \mathbf{c}_k . The highest \mathbf{c}_k - \mathbf{c}_k correlation found at every (A)–(H) step is colored in blue (metabolite 1), red (metabolite 2) or green (metabolite 3). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

compare the magnitude of the peaks (i.e. comparing $c_{ix}s_{ix}^T$ with $c_x s_x^T$) using the same sample as a reference.

The DTC strategy is represented in Fig. 3. For 10 \mathbf{c}_k profiles, 45 correlations were calculated (Fig. 3A and J), and the highest \mathbf{c}_k - \mathbf{c}_k correlation was found for the \mathbf{c}_{IX} - \mathbf{c}_X pair (Fig. 3B and J). Then, the next highest pair-wise correlation should involve one of the already previously selected \mathbf{c}_{IX} or \mathbf{c}_X profiles. As shown in Fig. 3C, this condition was met for \mathbf{c}_{IV} , since the highest correlation was found between \mathbf{c}_{IX} and \mathbf{c}_{IV} (Fig. 3J). This process was repeated again and, as shown in Fig. 3D, the finally selected \mathbf{c}_k was \mathbf{c}_{II} , since its pair-wise correlation with \mathbf{c}_{IX} was the highest from the remaining possible options (Fig. 3J). Next correlations were not linked to this group of \mathbf{c}_k profiles (\mathbf{c}_{II} - \mathbf{c}_{IX} - \mathbf{c}_X). Therefore, the metabolite features grouped up to this point were assumed to correspond to the same metabolite 'metabolite 1', and the strategy is then started again with the remaining \mathbf{c}_k (Fig. 3E). Steps from Fig. 3E–G show the grouping of \mathbf{c}_k that compose the next 'metabolite 2'. This process is continued until all \mathbf{c}_k were grouped, situation that was achieved at Fig. 3H. The reconstructed ^1H NMR spectra of the proposed metabolites after grouping all \mathbf{c}_k are presented in Fig. 3I.

With the information deduced from this correlation decision

tree, an equality constraint on the resolved spectra can be implemented on the resolved spectra during MCR-ALS simultaneous analysis of the whole spectral dataset. In this MCR-ALS procedure, each resolved component is assigned to a metabolite ($N = J$). Using this spectrum equality constraint, intensity values in those regions known not to contain signals from a given metabolite are forced to be zero. In addition, the finally reconstructed ^1H NMR \mathbf{S}^T spectra profiles of the proposed metabolites obtained after grouping all \mathbf{c}_k are used as initial spectral estimates of the ALS optimization. Improved concentration and spectral profiles for every metabolite will be finally obtained next using this constrained DTC (Decision Tree of Correlations) MCR-ALS approach. Resolved spectral profiles were finally exported from Matlab environment to MestreNova v.9.0 (Mestrelab Research, Spain) for their further ^1H NMR assignment.

It is worth mentioning that the DTC strategy is implementable as a semiautomatic grouping tool, since the division of the dataset into different spectral subregions and the search of the highest \mathbf{c}_k - \mathbf{c}_k correlations can be automatized, and the user should only confirm that the DTC matching was correct. Different strategies for dividing the dataset can be used (i.e. a variable selection strategy

based on the comparison of the standard deviation [31] or the morphological score [32] associated to every chemical shift variable to a threshold value fixed by the user). Since all the required inputs for the final MCR-ALS analysis (spectrum equality constraints, initial estimates and number of components) are obtained directly from the DTC analysis, this single semiautomatic tool can be also programmed to perform directly the MCR-ALS analysis.

3. Results and discussion

The previously described DTC-MCR-ALS method is a convenient way for extracting pure metabolite ^1H NMR spectral profiles using an untargeted approach. Although it is more laborious than other more traditional soft-modelling bilinear decomposition methods (MCR-ALS, ICA, ...), it helps to circumvent the intrinsic difficulties of these methods for resolving these dataset types. These difficulties are mostly related to rank deficiency problems (not all signals vary independently [20]), and to rotation ambiguities and noise propagation effects. The impact of these limitations is discussed in section 3.1 below.

3.1. MCR-ALS analysis of the simulated datasets (X_1 and X_2)

In order to resolve any dataset using the bilinear decomposition model given in Equation (3) (section 2.3.1), a number of factors or components, N , should be proposed first. SVD is a good approach to estimate N (chemical rank or number of singular values different to zero in absence of noise) when metabolite spectra and composition among samples vary independently [33]. However, in complex biological datasets, where various metabolites may evolve synchronously because different metabolic responses are equally triggered by the same factor, SVD will detect a lower number of components than the real number of different metabolites (chemical rank deficiency). This premise can be perceived when the eigenvalues of X_1 (Fig. 4A) and X_2 (Fig. 4B) are examined. In Fig. 4B,

despite the fact that the ^1H NMR spectra had signals from ten metabolites, a lower number of eigenvalues different from zero were detected.

It is also important to note that, in MCR-ALS analyses, the more concentrated metabolites are usually more easily resolved than the ones at lower concentrations, since MCR-ALS algorithm tries to maximize the explained variance. Therefore, metabolites with concentration variability close to the noise level may be poorly resolved or even considered to be noise. Indeed, the increase of noise level in a given dataset causes a more difficult detection of the number of components estimated by SVD (Fig. 4A–B), and therefore, it penalizes MCR-ALS resolution (Fig. 4C–D).

In DTC-MCR-ALS, the use of the decision tree strategy described previously in section 2.3.2 helps for the selection of the best N . In DTC-MCR-ALS, N is the number of groups of metabolite features selected after application of the DTC strategy. In addition, metabolites with lower contributions will be better resolved when the DTC strategy is used than when the standard MCR-ALS bilinear decomposition is used. Final MCR-ALS solutions may not be unique due to a) rotation ambiguities associated to the bilinear decomposition method [29], b) the fact that the ^1H NMR of biological samples may contain hundreds of resonances from dozens of metabolites and c) the fact that NMR signals can be also highly overlapped. Different solutions fit equally well the data and fulfill the applied constraints. Therefore, MCR-ALS solution will be dependent on the initial estimates used. This situation is shown in Fig. 5, where the performance of the different MCR-ALS resolutions are obtained depending on the initial estimates and on the equality constraints used.

In this Fig. 5, similarity angles between each one of the original profiles, s_j^0 or c_j^0 , and every profile resolved by MCR-ALS in dataset $X_{2,\text{SNR}=500}$ are given. The corresponding boxplots are shown in Appendix B. Values of α and r^2 coefficients for these two studied cases are given in more detail in the Supplementary Material Appendix C. For most of the compared situations in Fig. 5,

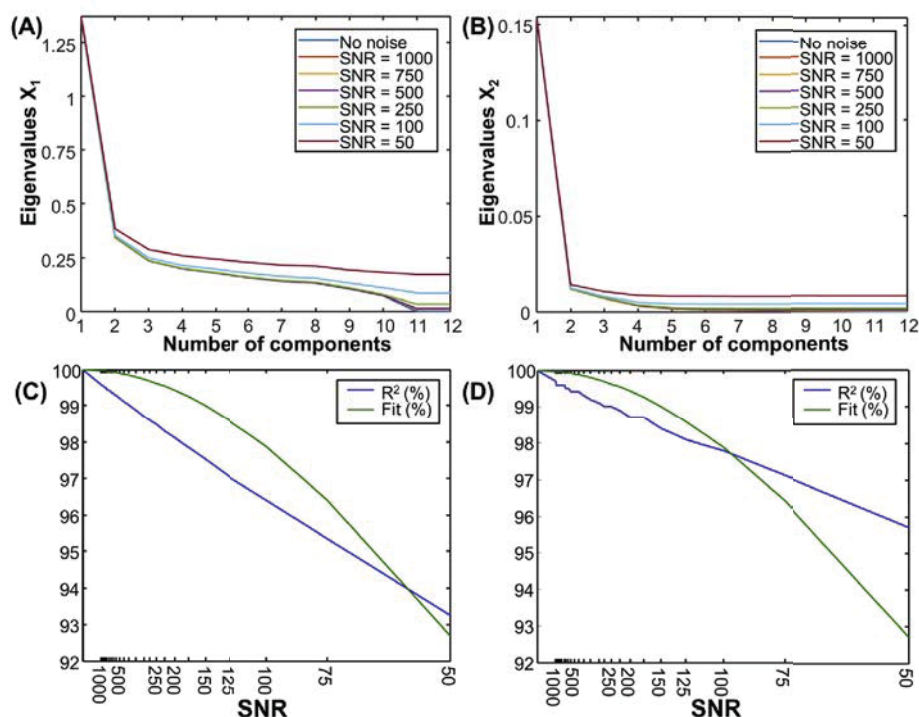


Fig. 4. SVD and MCR-ALS analysis of datasets with different signal to noise ratios (SNR). SVD of X_1 (A) and SVD of X_2 (B) at different SNR values. MCR-ALS explained data variances in the analysis of X_1 (C) and of X_2 (D) using 10 components at different SNR (initialization was performed using purest concentration profiles as in Ref. [28]).

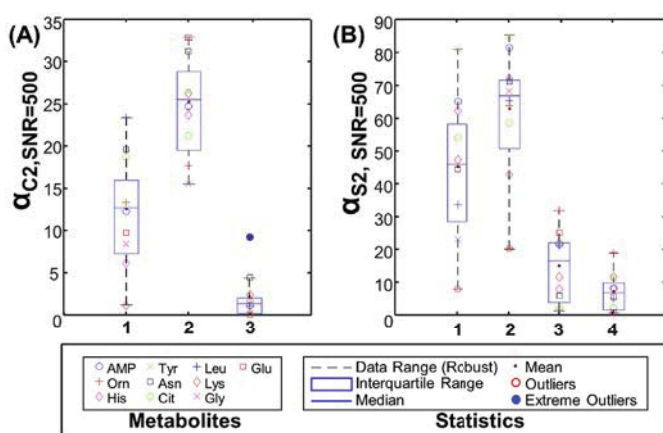


Fig. 5. Comparison of MCR-ALS results depending on initial estimates and on spectrum equality constraints in $X_{2,SNR=500}$ dataset. (A). Similarity angle between the original concentrations, c_i^0 , and the MCR-ALS resolved ones, c_i^r . (B). Similarity angle between the metabolic 1H NMR known spectra (s_i^0) and the MCR-ALS resolved ones (s_i^r). Methods used: 1, MCR-ALS with purest C as initial estimates [28]; 2, MCR-ALS with purest S^T as initial estimates [28]; 3, DTC-MCR-ALS; 4, DTC-MCR-ALS, but only variables from the unconstrained regions were used to calculate similarity angles. Numerical values shown in boxplots are given in more detail in Appendix C of Supplementary material.

concentration profiles (Fig. 5A) were better resolved than spectra profiles (Fig. 5B), due to that fact that there was a higher selectivity in the spectral direction than in the concentration direction [30].

The effect of using either the purest C or the purest S^T as initial estimates is shown in the first two boxes in the boxplots of Fig. 5, respectively. For these two cases, none of the obtained solutions was completely satisfactory. The reason for this was that the profiles were not correctly resolved due to presence of remaining rotation ambiguities and to noise propagation effects. One examples of this resolution is presented in Fig. 6.

Similar interpretations were obtained in the case of the analysis of the more noisy data, $X_{2,SNR=125}$ (see Supplementary material Appendix B and Appendix C), although resolution performance was lower than for the $X_{2,SNR=500}$ dataset due to the influence of noise.

To improve the results, the DTC methodology is proposed next.

3.2. DTC-MCR-ALS analysis of the simulated dataset (X_2)

From previous sections, it is concluded that MCR-ALS resolution of complex 1H NMR datasets is not feasible in the general case due to rotation ambiguities and rank deficiency problems. Untargeted quantitative 1H NMR DTC-MCR-ALS analyses is therefore proposed as a means to generate good initial estimates and spectrum equality constraints (ssel), and to limit significantly the number of feasible solutions and get improved ones.

This idea is demonstrated with the resolution of the $X_{2,SNR=500}$ and $X_{2,SNR=125}$ datasets. Datasets were divided in 26 spectral regions, and MCR-ALS (using as initial estimates the most dissimilar variables [28]) was applied to them. The number of components varied from 1 to 3 for each spectral region, and the total number of metabolic features (resolved spectra for each region) was 33 ($SNR = 500$) and 31 ($SNR = 125$).

The decision tree-based methodology presented in section 2. *Material and Methods* showed that $X_{2,SNR=500}$ could be properly analyzed using 13 components, whereas $X_{2,SNR=125}$ needed 12 components. Since only 10 metabolites were present in the metabolite mixtures in both cases, the additional components are either due to noise or to a combination of resonances from different metabolites that were not correctly resolved (combined effects of noise and rotational ambiguities). Metabolites were then assigned from the group of signals resolved in s_k . All metabolites (even the least concentrated ones) were correctly assigned, demonstrating that DTC-MCR-ALS strategy resulted to be very appropriate for untangling mixtures of unknown composition in 1H NMR datasets. Examples of DTC-MCR-ALS performance in the resolution of 1H NMR spectra are shown in Fig. 7.

However, it is important to state that if two or more resonances from distinct metabolites were included in the same spectral region and have always the same evolving pattern (concentration changes) in different samples, they will be considered to belong to the same metabolic feature, which therefore implies that fewer features will be attributed to a particular metabolite. For instance, in the analyzed dataset, at around 4.00 ppm, proton resonances from AMP, L-histidine and L-asparagine were highly overlapped, and only 2 components could be distinguished (i.e. the addition of a third component did not improve the ALS fitting nor their spectral resolution). As a consequence, in both datasets $X_{2,SNR=500}$ and $X_{2,SNR=125}$, these two components were finally assigned to 1H NMR profiles of L-histidine and L-asparagine, because they were more abundant than AMP, and consequently, the resolved 1H NMR profile of AMP lacked this resonance (Fig. 7C).

Noise propagation also affected the correct grouping of signal resonances. For instance, as shown in Fig. 7C, using the decision tree-based strategy previously described, only 5 out of the 6 possible s_k were related to AMP in $X_{2,SNR=500}$, and only 4 s_k in $X_{2,SNR=125}$ (the noisiest dataset) were grouped. Nevertheless, only with these groups of s_k , the assignment of AMP could be still performed correctly.

MCR-ALS simultaneous analysis of the whole datasets using the DTC-generated initial estimates and derived equality constraints explained 92.2% of the total data variance and gave a lack-of-fit of 27.8% for $X_{2,SNR=500}$, and of 95.2%, and 19.9% respectively for $X_{2,SNR=125}$. When all components were examined in more detail, the extra components were related to some poorly resolved spectral features and to residual contributions of the more concentrated metabolites.

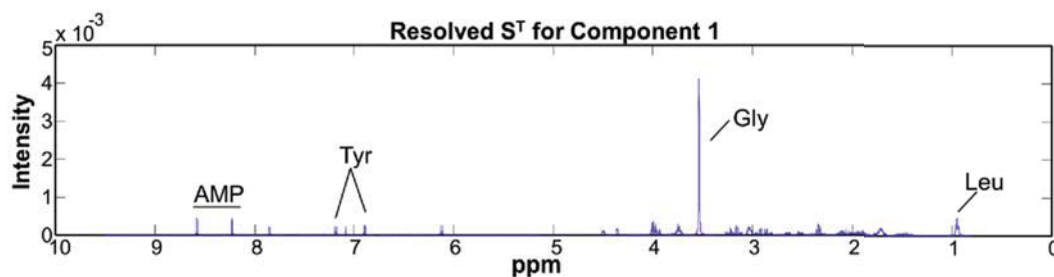


Fig. 6. 1H NMR spectra resolved by MCR-ALS for $X_{2,SNR=500}$ dataset. The purest concentrations were used as initial estimates [28] and the only constraint applied during MCR-ALS was non-negativity.

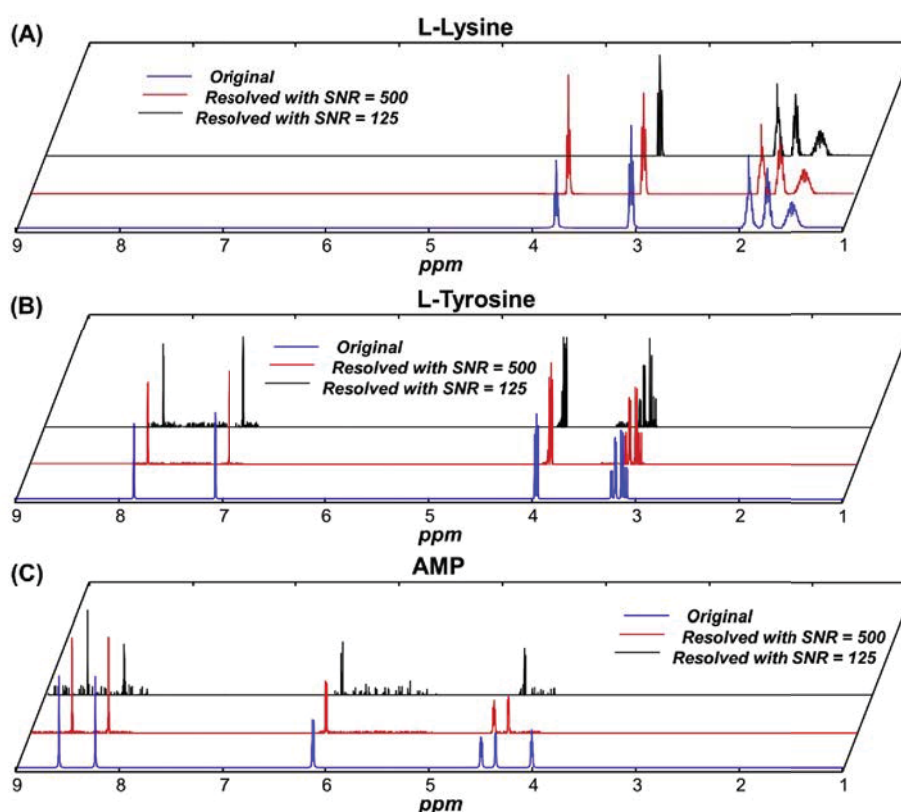


Fig. 7. ^1H NMR S^T spectra resolved by the DTC-MCR-ALS methodology in the analysis of $X_{2,\text{SNR}=500}$ (red) and $X_{2,\text{SNR}=125}$ (black) datasets. A. Resolved ^1H NMR spectra of L-Lysine. B. Resolved ^1H NMR spectra of L-Tyrosine. C. Resolved ^1H NMR spectra of AMP. In blue color are represented the original ^1H NMR spectra: $s_{\text{L-Lysine}}^0$ (A), $s_{\text{L-Tyrosine}}^0$ (B) and s_{AMP}^0 (C). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

As previously shown in section 3.1, quality of the results obtained in the MCR-ALS resolution of the two simulated datasets X_2 can be evaluated from the recoveries of the spectra and concentration profiles used for their generation (Fig. 5). In this figure, DTC-MCR-ALS allowed a better resolution of the concentration profiles than that of the spectral profiles (Fig. 5). In conclusion, DTC-MCR-ALS was proven to be a reliable strategy for achieving the ^1H NMR spectral resolution of complex samples.

3.3. DTC-MCR-ALS analysis of the experimental dataset (X_3)

Finally, DTC-MCR-ALS was also applied to the more interesting (and challenging) experimental dataset X_3 . Prior to this analysis, ^1H NMR spectra were aligned using the icoshift [34] algorithm. This alignment step results crucial, since resonances that were not perfectly aligned prior to analysis would be poorly resolved or even assigned to different metabolites. Next, the spectra were divided in 56 regions. Limits for these regions were selected using the criteria stated in the section 2.3.2. More regions were chosen for the analysis of X_3 than for X_2 , since the corresponding ^1H NMR spectra in X_3 had more resonances than in X_2 . MCR-ALS was applied to each region, and a total number of 161 metabolic features were resolved. The decision tree strategy pointed out the presence of 86 potential metabolites. From these 86 spectra, 11 grouped spectral features were discarded because they had only noise features. After resolving the dataset using the 75 remaining components, the total explained data variance was 86.9% and the lack-of-fit was 36.1%.

From multiplicities and chemical shifts of the resolved signals, some of the metabolites were already identified using NMR databases such as HMDB [9] or YMDB [35]. At least 39 out of the 75 resolved components were assigned to chemical compounds

(Table 1), including yeast metabolites (amino acids, organic acids, sugars and nucleotides, among others), the DSS NMR standard and the water signal. Relevant spectroscopic data for these molecules are given in the eighth column of Table 1, whereas resolved spectral profiles are given in the Supplementary material of Appendix D.

In the eighth column of Table 1, resonances recovered by MCR-ALS analysis are highlighted in bold. As it is shown from this list, the first recovered metabolites, apart from giving higher correlation coefficients, also had a higher number of resonances per metabolite.

It is worth mentioning that the number of correlation coefficients is linked to the number of regions where signals of a given metabolite were identified, and this number should not be directly related to the number of resonances. For instance, for the reconstruction of the ^1H NMR spectrum of D-glucose, 5 correlation coefficients were used (sixth column in Table 1), allowing the grouping of 6 features, from 6 different spectral regions. Since some of the regions contained more than one proton resonance, the number of involved regions was smaller than the total number of proton resonances.

On the other side, if a metabolite was strongly overlapped, a relatively low number of resonances were recovered. For instance, for L-threonine only one metabolite feature (1.32 ppm, d, $J = 6.7$ Hz) was resolved. As observed in the sixth column of Table 1, some of the metabolites have no correlation coefficients assigned, meaning that in this case only one feature was used to identify these metabolites. This occurs for partially masked metabolites, and for metabolites that have only one resonance (e.g. formic acid). However, despite this possible limitation, the detected resonances were characteristic enough for their identification and their corresponding quantification. Additional metabolites (or metabolite

Table 1
List of metabolites identified with DTC-MCR-ALS.

# ^a	# ^b C.	Metabolite	r ² ^c	α ^d	tree corr. ^e	First excl. corr. ^f	¹ H NMR signals ^g	HMDB number
1	1	D-Glucose	0.941	16.747	0.998, 0.997, 0.990, 0.980, 0.948	0.690	3.23 ppm (dd, J = (9.2 Hz, 7.9 Hz), 0.67 H); 3.36–3.50 ppm (m, 2 H); 3.52 ppm (dd, J = (10.3, 4.2 Hz), 0.66 Hz); 3.66–3.84 ppm (m, 2 H); 3.88 ppm (dd, J = (12.4, 1.8 Hz), 0.67 H); 4.63 ppm^h (d, 7.9 Hz, 0.67 H); 5.22 ppm (d, J = 3.7 Hz, 0.33 H)	HMDB00122
2	2	Glycerol	0.979	4.093	0.994, 0.947	0.709	3.54 ppm (dd, J = (11.8, 6.5 Hz), 2 H); 3.64 ppm (dd, J = (11.7, 4.3 Hz), 2 H); 3.78 ppm (tt, J = (6.5, 4.4 Hz), 1 H)	HMDB00131
3	3	Water (shoulder)	—	—	0.994, 0.990, 0.989, 0.993, 0.992, 0.983	0.957	5.15–5.95 ppm	
4	4	Trichalose	0.855	13.560	0.976, 0.855, 0.838, 0.763	0.282	3.44 ppm (t, J = 9.3 Hz, 2 H); 3.64 ppm (dd, J = (9.8, 3.9 Hz), 2 H); 3.73–3.78 ppm (m, 2 H); 3.79–3.88 ppm (m, 6 H); 5.18 ppm (d, J = 3.8 Hz, 2 H)	HMDB00975
5	5	DSS	0.942	9.131	0.975, 0.974,	0.937	1.76 ppm (tt, J = (10.2, 6.5 Hz), 2 H); 2.88–2.94 ppm (m, 2 H)	
6	6	L-Methionine	0.937	16.265	0.972	0.739	2.05–2.24 ppm (m, 2 H); 2.12 ppm (s, 3 H); 2.63 ppm (t, J = 7.6 Hz, 2 H); 3.85 ppm (dd, J = (7.1, 5.4), 1 H)	HMDB00696
7	7	Unknown-1	0.967	8.963	NA ⁱ	0.967	8.37 ppm (s, 1 H)	
8	8	N ⁶ -methyladenosine	0.923	13.325	NA ⁱ	0.936	2.96 ppm (s, 3 H); 6.11 ppm (d, J = 4.4 Hz, 1 H); 8.28 ppm (s, 1 H); 8.29 ppm (s, 1 H)	
9	9	Ureidosuccinic acid	0.919	26.948	0.967, 0.967, 0.965	0.874	2.44 ppm (dd, J = (15.4, 9.6 Hz), 1 H); 2.658 ppm (dd, J = (15.4, 3.4 Hz), 1 H); 4.214 ppm (dd, J = (9.7, 3.8 Hz), 1 H)	HMDB00828
10	10	Acetic acid	0.824	15.347	0.956, 0.742, 0.734	0.800	1.90 ppm (s, 3 H)	HMDB00042
11	11	L-Tyrosine	0.970	6.432	0.950	0.777	3.04 ppm (dd, J = (14.7, 7.8 Hz), 1 H); 3.19 ppm (dd, J = (14.7, 5.1 Hz), 1 H); 3.93 ppm (dd, J = (7.7, 5.2 Hz), 1 H); 6.89 ppm (d, J = 8.4 Hz, 2 H); 7.18 ppm (d, J = 8.4 Hz, 2 H)	HMDB00158
12	12	2-Isopropylmalic acid	0.993	6.522	NA ⁱ	0.945	0.84 ppm (d, J = 6.9 Hz, 3 H); 0.89 ppm (d, J = 6.9 Hz, 3 H); 1.83–1.90 ppm (m, 1 H); 2.55 ppm (d, J = 16 Hz, 1 H); 2.67 ppm (d, J = 16 Hz, 1 H)	HMDB00402
13	13	S-3-Hydroxyisobutyric acid	0.911	17.557	NA ⁱ	0.774	1.06 ppm (d, J = 6.9 Hz, 3 H); 2.60–2.69 (m, 1 H); 3.62–3.74 (m, 2 H)	HMDB00023
14	15	NAD ⁺	0.465	18.882	0.931	0.758	4.16–4.29 ppm (m, 3 H); 4.32–4.39 ppm (m, 2 H); 4.42 ppm^h (dd, J = 5.1, 2.0 Hz), 1 H); 4.44–4.52 ppm^h (m, 2 H); 4.52–4.55 ppm^h (m, 1 H); 6.03 ppm (d, J = 5.9 Hz, 1 H); 6.08 ppm (d, J = 5.3 Hz, 1 H); 8.16 ppm (s, 1 H); 8.17 ppm (d, J = 6.1 Hz); 8.15–8.21 ppm (mm, 2 Hz); 8.42 ppm (s, 1 H); 8.82 ppm (d, J = 7.6 Hz, 1 H); 9.13 ppm (d, J = 7.6 Hz, 1 H); 9.33 ppm (s, 1 H)	HMDB00902
15	16	L-Valine	0.982	3.625	0.913, 0.910	0.648	0.98 ppm (d, J = 7.1 Hz, 3 H); 1.03 ppm (d, J = 7.1 Hz, 3 H); 2.26 ppm (mm, 1 H); 3.60 ppm (d, J = 4.3 Hz, 1 H)	HMDB00883
16	19	L-Ornithine	0.834	13.305	0.889	0.749	1.65–1.78 ppm (mm, 1 H); 1.78–1.87 ppm (mm, 1 H); 1.88–1.98 ppm (mm, 2 H); 3.04 ppm (t, J = 7.5 Hz, 2 H); 3.77 ppm (t, J = 6.0 Hz, 1 H)	HMDB00214
17	20	L-Aspartic acid	0.802	11.444	0.888	0.607	2.67 ppm (dd, J = (17.4, 8.7 Hz), 1 H); 2.80 ppm (dd, J = (17.4, 3.8 Hz), 1 H); 3.89 ppm (dd, J = (8.7, 3.8 Hz), 1 H)	HMDB00191
18	21	AMP	0.851	11.034	0.887, 0.715	0.809	4.01 ppm (dd, J = (4.5, 3.4 Hz), 1 H); 4.33–4.39 ppm (m, 1 H); 4.50 ppm^h (dd, J = 5.0, 3.5 Hz), 1 H); 6.13 ppm (d, J = 5.6 Hz, 1 H); 8.26 ppm (s, 1 H); 8.59 ppm (s, 1 H)	HMDB00045
19	22	Orotic acid	0.854	19.125	NA ⁱ	0.869	6.18 ppm (s, 1 H)	HMDB00226
20	24	Citric acid	0.416	29.106	0.859	0.546	2.52 ppm (d, J = 15.8 Hz, 2 H); 2.65 ppm (d, J = 15.6 Hz, 2 H)	HMDB00094
21	25	L-Glutamine	0.797	14.886	0.854	0.785	2.08–2.18 ppm (m, 2.5 H); 2.44 (td, J = 7.3, 3.6 Hz, 1.5 H); 3.77 ppm (t, J = 6.2 Hz, 1 H)	HMDB00641
22	26	Glutathione	0.584	20.972	0.845, 0.817	0.778	2.16 ppm (q, J = 7.6 Hz, 2 H); 2.45–2.65 ppm (m, 2 H); 2.85–3.00 ppm (m, 2 H); 3.77 ppm (m, 3 H); 4.56 ppm^h (dd, J = (7.0 Hz, 5.2 Hz), 1 H)	HMDB00125
23	29	L-Arginine	0.959	5.140	0.817	0.697	1.58–1.79 ppm (m, 2 H); 1.80–2.00 ppm (m, 2 H); 3.23 ppm (t, J = 6.9 Hz, 2 H); 3.76 ppm (t, J = 6.1 Hz, 1 H)	HMDB00517
24	30	L-Histidine	0.797	15.165	0.811, 0.783, 0.599, 0.733, 0.701, 0.676	0.281	3.12 ppm (dd, J = (15.5, 7.9 Hz), 1 H); 3.23 ppm (dd, J = (15.5, 5.0 Hz), 1 H); 3.97 ppm (d, J = (7.7, 5.0 Hz), 1 H); 7.06 ppm (s, 1 H); 7.81 ppm (s, 1 H)	HMDB00177
25	37	EIGP	0.731	23.778	0.757	0.581	3.77 ppm (nd, 1H); 3.99 ppm (nd, 1 H); 4.78 ppm^h (nd, 1 H); 7.18 ppm (s, 1 H); 7.84 ppm (s, 1 H)	HMDB12208
26	39	GPC	0.945	9.372	NA ⁱ	0.743	3.22 ppm (s, 9 H); 3.56–3.71 ppm (m, 4 H); 3.83–4.00 ppm (m, 3 H); 4.28–4.36 ppm (mm, 2 H)	HMDB0086
27	40	Unknown-2	0.934	11.317	NA ⁱ	0.664	8.03 ppm (s, 1 H)	
28	41	Methyl donor	0.835	13.018	NA ⁱ	0.725	2.10 ppm (s, 3 H)	HMDB02303
29	45	Fatty acid singlet	0.967	7.421	NA ⁱ	0.701	1.24 ppm (s, 4 H)	HMDB00638
30	55	3-Methyl-2-oxovaleric acid	0.915	12.410	NA ⁱ	0.629	0.90 ppm (t, J = 7.4 Hz, 3 H); 1.09 ppm (d, J = 6.6 Hz, 3 H); 1.41–1.52 ppm (m, 1H); 1.64–1.76 ppm (m, 1 H); 2.87–3.00 ppm (m, 1 H)	HMDB00491
31	56	L-Threonine	0.909	4.742	NA ⁱ	0.551	1.32 ppm (d, J = 6.7 Hz, 3 H); 3.57 ppm (d, J = 4.9 Hz, 1 H); 4.19–4.28 ppm (m, 1 H)	HMDB00167
32	57	L-Phenylalanine	0.822	10.791	NA ⁱ	0.559	3.07–3.31 ppm (m, 2 H); 3.98 ppm (dd, J = (7.8, 5.31 Hz), 1 H); 7.29–7.45 ppm (m, 5 H)	HMDB00159
33	58	Uridine	0.328	29.372	NA ⁱ	0.588	3.72–3.83 ppm (m, 2 H); 3.86–3.93 ppm (m, 1 H); 4.09–4.16 ppm (m, 1 H); 4.22 ppm (dd, J = (5.4, 5.4 Hz), 1 H); 4.34 ppm (dd, J = (4.9, 4.9 Hz), 1 H); 5.89 ppm (d, J = 8.4 Hz, 1 H); 5.90 ppm (d, J = 4.6 Hz, 1 H); 7.86 ppm (d, J = 8.1 Hz, 1 H)	HMDB00296
34	60	L-Leucine	0.893	19.192	NA ⁱ	0.573	0.95 ppm (d, J = 5.1 Hz, 6 H); 1.61–1.78 ppm (m, 3 H); 3.60–3.75 ppm (m, 1 H)	HMDB00687
35	61	Formic acid	0.767	15.831	NA ⁱ	0.547	8.44 ppm (s, 1 H)	HMDB00142
36	62	Uracil	0.720	21.368	NA ⁱ	0.534	5.79 ppm (d, J = 7.8 Hz, 1 H); 7.53 ppm (d, J = 7.8 Hz, 1 H)	HMDB00300
37	65	L-Asparagine	0.533	19.651	NA ⁱ	0.489	2.82 ppmⁱ (d, J = 7.6 Hz, 0.33 H); 2.86 ppmⁱ (d, J = 7.4 Hz, 0.67 H); 2.94 ppm (dd, J = (16.9, 4.4 Hz), 1 H); 3.99 ppm (dd, J = (7.6, 4.3 Hz), 1 H)	HMDB00168
38	70	L-Glutamic acid	0.684	9.214	NA ⁱ	0.398	1.99–2.10 ppm (m, 2 H); 2.31–2.37 ppm (m, 2 H); 3.75 ppm (dd, J = (7.2 Hz, 4.7 Hz), 1 H)	HMDB00148
39	72	L-Alanine	0.986	5.678	NA ⁱ	0.073	1.47 ppm (d, J = 7.1 Hz, 3 H); 3.76 ppm (q, J = 7.2 Hz, 1 H)	HMDB00161

^a Number of characterized metabolites.^b Number of the MCR-ALS component.

- ^c Squared regression coefficient between metabolite concentrations obtained with BATMAN methodology and with DTC-MCR-ALS methodology.
^d Similarity angle between metabolite concentrations obtained with BATMAN methodology and with DTC-MCR-ALS methodology.
^e Correlation coefficients obtained during the decision-tree strategy.
^f Highest correlation coefficient between any of the not-grouped metabolic features and any of the grouped ones.
^g ¹H NMR data. Proton resonances in bold were recovered using the DTC-MCR-ALS methodology.
^h Region that comprised the resonance was not included in the analysis.
ⁱ Not applicable: only one metabolic feature was used to identify and quantify the metabolite.
^j Both *d* from the *dd* were enclosed in different windows.

features) that were not characteristic enough to perform metabolite assignment are shown in Supplementary material Appendix D.

In a previous work [25], some yeast metabolites (N⁶-methyladenosine, ureidosuccinic acid, orotic acid, D-erythro-imidazole-glycerol-phosphate (EIGP), and 3 singlets at 2.10 ppm, 8.03 ppm and 8.37 ppm) identified also in the present work were classified as markers of starvation stressed conditions. Since these metabolites are not at high concentrations at normal conditions in yeast cells, they could have been overlooked in previous targeted metabolomics studies. Since DTC-MCR-ALS could also resolve the spectra of these metabolites without needing additional acquisition of complementary experiments (¹H/¹H COSY, ¹H/¹³C HSQC, ¹H/¹³C HMBC, etc.), the proposed methodology can be also extended to general metabolomic studies for biomarker screening.

Relative metabolite concentrations for a given sample and metabolite can be straightforwardly calculated as the quotient between the corresponding concentration value generated in the DTC-MCR-ALS analysis, c_n^d , and the number of protons of the resolved resonances. Correlation coefficients (fourth column of Table 1) and similarity angles (fifth column of Table 1) between concentration values obtained either via conventional methods (for instance, in this case, by signal deconvolution using the Batman R-package [36]) or via the proposed untargeted methodology, confirmed the agreement between the two approaches and the reliability of the concentration values obtained for most of the molecules (Table 1). Regression lines comparing the two methods for every metabolite can be found in Supplementary material Appendix E. Regression lines for D-glucose, glycerol and L-methionine are shown in Fig. 8. As shown in this table, correlation coefficients between metabolic features change significantly depending on whether they belong to the same metabolite or not. In the sixth column of Table 1, correlations between intramolecular resonance signals selected with the decision-based tree approach are given. In the seventh column of Table 1, the higher intermolecular correlations involving a particular metabolite feature for the given metabolite are also given. For example, for trehalose metabolite, correlations were 0.76–0.97 between intramolecular signals, and lower than 0.28 for intermolecular correlations. Apart

from the fact that concentration estimates are obtained directly, the DTC-MCR-ALS method has additional advantages. For instance, water shoulder was conveniently resolved, which implies that proton integrals from resonances located on that region were not distorted by the water contribution. This represents an important advantage in comparison to other peak integration methods that cannot handle this type of background signal overlapping. This also represents an important advantage for the analysis of samples containing broad signals in the background (due to the presence of macromolecules) if the variance associated to these broad signals is large enough not to be considered as background contribution. In metabolomics studies, NMR spectra are usually normalized before analysis, either to a scalar value computed statistically (to a constant sum or better using Standard Normal Variate [2], Probabilistic Quotient Normalization [37], or Histogram Matching [38] normalization methods) or using empirical data (i.e. dry weight, creatinine signal). Since the NMR standard should be present at the same concentration in all pre-processed spectra, variations in concentration of this standard can be directly linked to instrumental variations and they should be included in the normalization calculation. However, when some resonance signals appear correlated to the NMR standard, there is evidence that the origin of these given signals are not from the studied sample, but due to impurities from the sample preparation or related to an instrumental artifact (i.e. electronic noise). Therefore, in these cases, the DTC-MCR-ALS approach results to be also a very suitable method for identifying such undesired NMR artifacts.

4. Conclusions

In this work, we have proposed a decision tree-based multivariate curve resolution (DTC-MCR-ALS) approach for the analysis of complex metabolite mixtures by ¹H NMR. This approach is based on the grouping of correlated metabolite features, and on the implementation of equality constraints without prior knowledge of sample composition or availability of ¹H NMR metabolite pure spectra. A comprehensive study about the potential use of the proposed DTC-MCR-ALS method for the analysis of ¹H NMR

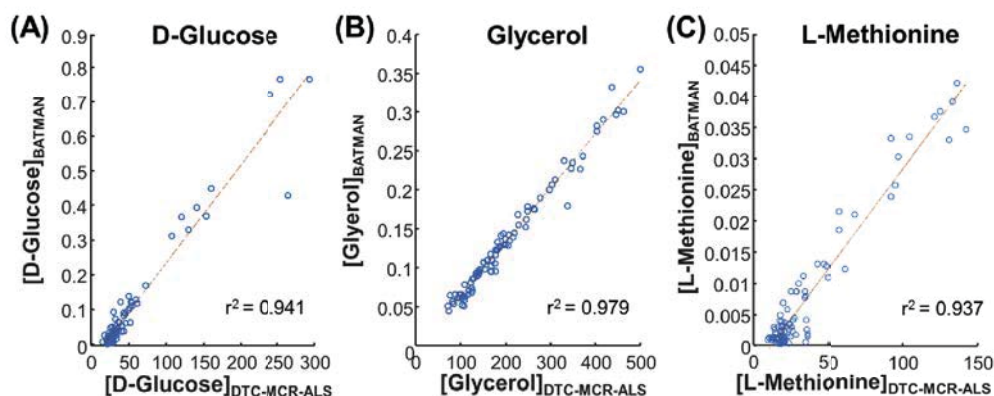


Fig. 8. Linear regression results between metabolite concentrations obtained with BATMAN and DTC-MCR-ALS methodologies. Metabolites: (A) D-Glucose, (B) glycerol and (C) L-methionine.

metabolomics datasets arrived to the following conclusions:

- 1) The performance of MCR-ALS analysis of ^1H NMR metabolomics datasets depended on the particular dataset under study, especially on the inter-correlated variability of the different metabolite concentrations, on their ^1H NMR signal (spectra) overlapping and on the signal to noise ratio, SNR, of the considered dataset.
- 2) Proper metabolite spectra resolution in complex samples (datasets) was achieved using the proposed DTC-MCR-ALS methodology when spectrum equality constraints were applied and proper selection of initial spectra estimates of the proposed metabolites was performed.
- 3) The proposed strategy has been validated using three ^1H NMR datasets: two simulated and one experimental datasets. Concentration values obtained by the proposed DTC-MCR-ALS strategy were similar to those obtained by conventional peak integration targeted methods of previously known metabolites.
- 4) DTC-MCR-ALS is a robust resolution and integration method that can also remove prevalent solvent signals (i.e. water signal in aqueous samples), and allow the detection of impurity signals from the sample preparation or even instrumental artifacts.

As a general conclusion, the proposed DTC-MCR-ALS allowed the untargeted analysis of metabolic ^1H NMR datasets. Since ^1H NMR spectra of individual metabolites and their concentration profiles can be recovered with this strategy, DTC-MCR-ALS has potential applications in both quantitative and qualitative metabolomics studies. These results presented a new framework for the analysis of complex ^1H NMR datasets, which can be automatized and be beneficial for metabolomics studies, but also for synthetic chemistry studies and in quality control processes monitored with ^1H NMR spectroscopy.

Acknowledgements

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013)/ERC Grant Agreement n. 320737 and the Spanish Ministry of Economy and Competitiveness (CTQ2015-66254-C2-1-P).

Supplementary data

Supplementary data related to this article can be found at <http://dx.doi.org/10.1016/j.aca.2017.02.010>.

References

- [1] J.K. Nicholson, J.C. Lindon, *Systems biology: metabonomics*, *Nature* 455 (7216) (2008) 1054–1056.
- [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 (6) (2015) 1612–1625, <http://dx.doi.org/10.1007/s11306-015-0812-9>.
- [3] M.B. Pisano, P. Scano, A. Murgia, S. Cosentino, P. Caboni, Metabolomics and microbiological profile of Italian mozzarella cheese produced with buffalo and cow milk, *Food Chem.* 192 (2016) 618–624, <http://dx.doi.org/10.1016/j.foodchem.2015.07.061>.
- [4] 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, *Anal. Bioanal. Chem.* (2015) 1–13, <http://dx.doi.org/10.1007/s00216-015-9042-2>.
- [5] E. Ortiz-Villanueva, J. Jaumot, F. Benavente, B. Piña, V. Sanz-Nebot, R. Tauler, Combination of CE-MS and advanced chemometric methods for high-throughput metabolic profiling, *Electrophoresis* 36 (18) (2015) 2324–2335, <http://dx.doi.org/10.1002/elps.201500027>.
- [6] 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, <http://dx.doi.org/10.1016/j.trac.2016.07.004>.
- [7] R. Tautenhahn, G.J. Patti, D. Rinehart, G. Siuzdak, XCMS online: a web-based platform to process untargeted metabolomic data, *Anal. Chem.* 84 (11) (2012) 5035–5039, <http://dx.doi.org/10.1021/ja300698c>.
- [8] M. Farrés, B. Piña, R. Tauler, Chemometric evaluation of *Saccharomyces cerevisiae* metabolic profiles using LC–MS, *Metabolomics* 11 (1) (2015) 210–224, <http://dx.doi.org/10.1007/s11306-014-0689-z>.
- [9] D.S. Wishart, T. Jewison, A.C. Guo, M. Wilson, C. Knox, Y. Liu, et al., HMDB 3.0—the human metabolome database in 2013, *Nucleic Acids Res.* 41 (D1) (2013) D801, <http://dx.doi.org/10.1093/nar/gks1065>.
- [10] Q. Cui, I.A. Lewis, A.D. Hegeman, M.E. Anderson, J. Li, C.F. Schulte, et al., Metabolite identification via the madison metabolomics consortium database, *Nat. Biotech.* 26 (2) (2008) 162–164, http://www.nature.com/nbt/journal/v26/n2/supinfo/nbt0208-162_S1.html.
- [11] S. Ravanbakhsh, P. Liu, T.C. Bjordahl, R. Mandal, J.R. Grant, M. Wilson, et al., Accurate, fully-automated NMR spectral profiling for metabolomics, *PLoS ONE* 10 (5) (2015) e0124219, <http://dx.doi.org/10.1371/journal.pone.0124219>.
- [12] A.M. Weljie, J. Newton, P. Mercier, E. Carlson, C.M. Slupsky, Targeted profiling: quantitative analysis of ^1H NMR metabolomics data, *Anal. Chem.* 78 (13) (2006) 4430–4442, <http://dx.doi.org/10.1021/ac060209g>.
- [13] W.Y. Kang, S.H. Kim, Y.K. Chae, Stress adaptation of *Saccharomyces cerevisiae* as monitored via metabolites using two-dimensional NMR spectroscopy, *FEMS Yeast Res.* 12 (5) (2012) 608–616.
- [14] T. Jézéquel, C. Deborde, M. Maucourt, V. Zhendre, A. Moing, P. Giraudeau, Absolute quantification of metabolites in tomato fruit extracts by fast 2D NMR, *Metabolomics* 11 (5) (2015) 1231–1242, <http://dx.doi.org/10.1007/s11306-015-0780-0>.
- [15] O. Cloarec, M.-E. Dumas, A. Craig, R.H. Barton, J. Trygg, J. Hudson, et al., Statistical Total Correlation Spectroscopy: an exploratory approach for latent biomarker identification from metabolic ^1H NMR data sets, *Anal. Chem.* 77 (5) (2005) 1282–1289, <http://dx.doi.org/10.1021/ac048630x>.
- [16] A. Hyvärinen, E. Oja, Independent component analysis: algorithms and applications, *Neural Netw.* 13 (4–5) (2000) 411–430, [http://dx.doi.org/10.1016/S0893-6080\(00\)00026-5](http://dx.doi.org/10.1016/S0893-6080(00)00026-5).
- [17] R. Tauler, B. Kowalski, S. Fleming, Multivariate curve resolution applied to spectral data from multiple runs of an industrial process, *Anal. Chem.* 65 (15) (1993) 2040–2047, <http://dx.doi.org/10.1021/ac00063a019>.
- [18] Y.B. Monakhova, A.M. Tsikin, T. Kuballa, D.W. Lachenmeier, S.P. Mushtakova, Independent component analysis (ICA) algorithms for improved spectral deconvolution of overlapped signals in ^1H NMR analysis: application to foods and related products, *Magnetic Reson. Chem.* 52 (5) (2014) 231–240, <http://dx.doi.org/10.1002/mrc.4059>.
- [19] H. Motegi, Y. Tsuboi, A. Saga, T. Kagami, M. Inoue, H. Toki, et al., Identification of reliable components in multivariate curve resolution-alternating least squares (MCR-ALS): a data-driven approach across metabolic processes, *Sci. Rep.* 5 (2015) 15710, <http://dx.doi.org/10.1038/srep15710>, <http://www.nature.com/articles/srep15710#supplementary-information>.
- [20] C.D. Eads, C.M. Furnish, I. Noda, K.D. Juhlin, D.A. Cooper, S.W. Morrall, Molecular factor analysis applied to collections of NMR spectra, *Anal. Chem.* 76 (7) (2004) 1982–1990, <http://dx.doi.org/10.1021/ac035301g>.
- [21] T.K. Karakach, R. Knight, E.M. Lenz, M.R. Viant, J.A. Walter, Analysis of time course ^1H NMR metabolomics data by multivariate curve resolution, *Magn. Reson. Chem.* 47 (S1) (2009) S105, <http://dx.doi.org/10.1002/mrc.2535>.
- [22] I. Montoliu, Fo-PJ. Martin, S. Collino, S. Rezzi, S. Kochhar, Multivariate modeling strategy for intercompartmental analysis of tissue and plasma ^1H NMR spectroscopy, *J. Proteome Res.* 8 (5) (2009) 2397–2406.
- [23] J. Hao, M. Liebecke, W. Astle, M. De Iorio, J.G. Bundy, T.M.D. Ebbels, Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN, *Nat. Protoc.* 9 (6) (2014) 1416–1427, <http://dx.doi.org/10.1038/nprot.2014.090>.
- [24] H.J. Muncey, R. Jones, M. De Iorio, T.M. Ebbels, Metasimulo: simulation of realistic NMR metabolic profiles, *BMC Bioinforma.* 11 (1) (2010) 1–11, <http://dx.doi.org/10.1186/1471-2105-11-496>.
- [25] F. Puig-Castellví, I. Alfonso, B. Piña, R. Tauler, ^1H NMR metabolomic study of auxotrophic starvation in yeast using multivariate curve resolution-alternating least squares for pathway analysis, *Sci. Rep.* 6 (2016) 30982, <http://dx.doi.org/10.1038/srep30982>.
- [26] J. Jaumot, A. de Juan, R. Tauler, G.U.I. MCR-ALS, 2.0: new features and applications, *Chemom. Intell. Lab. Syst. Syst.* 140 (2015) 1–12, <http://dx.doi.org/10.1016/j.chemolab.2014.10.003>.
- [27] W.H. Press, *Numerical Recipes 3rd Edition: the Art of Scientific Computing*, Cambridge University Press, 2007.
- [28] W. Windig, D.A. Stephenson, Self-modeling mixture analysis of second-derivative near-infrared spectral data using the SIMPLSMA approach, *Anal. Chem.* 64 (22) (1992) 2735–2742, <http://dx.doi.org/10.1021/ac00046a015>.
- [29] H. Abdollahi, R. Tauler, Uniqueness and rotation ambiguities in multivariate curve resolution methods, *Chemom. Intell. Laboratory Syst.* 108 (2) (2011) 100–111, <http://dx.doi.org/10.1016/j.chemolab.2011.05.009>.
- [30] R. Tauler, A. Smilde, B. Kowalski, Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution, *J. Chemom.* 9 (1) (1995) 31–58, <http://dx.doi.org/10.1002/jcem.1180090105>.
- [31] C.Y. Airiau, H. Shen, R.G. Brereton, Principal component analysis in liquid chromatography proton nuclear magnetic resonance: differentiation of three regio-isomers, *Anal. Chim. Acta* 447 (1–2) (2001) 199–210, [http://dx.doi.org/10.1016/S0003-2670\(01\)01233-8](http://dx.doi.org/10.1016/S0003-2670(01)01233-8).

- [32] H. Shen, C.Y. Airiau, R.G. Brereton, Resolution of LC/1H NMR data applied to a three-component mixture of polyaromatic hydrocarbons, *J. Chemom.* 16 (4) (2002) 165–175, <http://dx.doi.org/10.1002/cem.699>.
- [33] Q. Xu, J.R. Sachs, T.-C. Wang, W.H. Schaefer, Quantification and identification of components in solution mixtures from 1D proton NMR spectra using singular value decomposition, *Anal. Chem.* 78 (20) (2006) 7175–7185, <http://dx.doi.org/10.1021/ac0606857>.
- [34] F. Savorani, G. Tomasi, S.B. Engelsen, icoshift: a versatile tool for the rapid alignment of 1D NMR spectra, *J. Magn. Reson* 202 (2) (2010) 190–202, <http://dx.doi.org/10.1016/j.jmr.2009.11.012>.
- [35] T. Jewison, C. Knox, V. Neveu, Y. Djoumbou, A.C. Guo, J. Lee, et al., YMDB: the yeast metabolome database, *Nucleic Acids Res.* 40 (D1) (2012) D815, <http://dx.doi.org/10.1093/nar/gkr916>, D20.
- [36] J. Hao, W. Astle, M. De Iorio, T.M.D. Ebbels, BATMAN—an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model, *Bioinformatics* 28 (15) (2012) 2088–2090, <http://dx.doi.org/10.1093/bioinformatics/bts308>.
- [37] F. Dieterle, A. Ross, G. Schlotterbeck, H. Senn, Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics, *Anal. Chem.* 78 (13) (2006) 4281–4290, <http://dx.doi.org/10.1021/ac051632c>.
- [38] R.J.O. Torgrip, K.M. Åberg, E. Alm, I. Schuppe-Koistinen, J. Lindberg, A note on normalization of biofluid 1D 1H-NMR data, *Metabolomics* 4 (2) (2008) 114–121, <http://dx.doi.org/10.1007/s11306-007-0102-2>.

SUPPLEMENTARY MATERIAL FOR SCIENTIFIC ARTICLE IV

Untargeted Assignment and Automatic Integration of ^1H NMR metabolomic datasets using a Multivariate Curve Resolution Approach.

Authors: Puig-Castellví F., Alfonso I., Tauler R.

Citation reference: *Anal. Chim. Acta* (2017), 964: 55-66.

DOI: 10.1016/j.aca.2017.02.010

Appendix 1. Quantitation of yeast metabolic extracts and generation of simulated datasets.

1. YEAST GROWTH

S. cerevisiae BY4741 (MATa; *his3Δ1*; *leu2Δ0*; *met15Δ0*; *ura3Δ0*) cells were pre-cultured in YPD medium on an orbital shaker (150 rpm) at 30°C overnight. Fresh YPD medium was inoculated with the pre-culture to an optical density at 600 nm (OD₆₀₀) of 0.1 and divided into volumes of 50 ml, which were grown either 15 °C, 30 °C, 37 °C or 40 °C (six replicates each, shaking at 150 rpm). After reaching an OD₆₀₀ of 0.6-0.8, cultures were arrested on ice. Cell harvesting of a 45-ml fraction for every sample was performed by centrifugation at 4000 g for 5 min and discarding the supernatant. Cells were washed afterwards with 100 mM sodium phosphate buffer (pH 7.0). The resulting pellets were stored at -80 °C and lyophilized.

2. METABOLITE EXTRACTION

Metabolites were extracted by following the protocol published in a previous work[1].

3. ¹H NMR EXPERIMENT

¹H NMR spectra were recorded in a 400 MHz Varian spectrometer, using a spectrometer frequency of 400.14 MHz with a OneNMR Probe and a ProTune System (Agilent). Spectral size range covered from -2 to 12 ppm, consisting in 65,538 data points. The number of scans was 512 and the relaxation delay was 5 seconds.

4. DATA PREPROCESSING

¹H NMR spectra were preprocessed with MestreNova v.9.0 (Mestrelab Research, Spain). NMR Spectra preprocessing consisted in an exponential apodization of 0.5 Hz, a manual phasing and a baseline correction with Bernstein polynomial of 3rd order. After adjusting the reference to DSS (4,4-dimethyl-4-silapentane-1-sulfonic acid), 4.66 - 5.16 ppm, 3.30 - 3.37 ppm and 7.63 -7.70 ppm regions were removed. Data points with chemical shift higher than 9.40 ppm or lower than 0.75 ppm were also removed. The final NMR dataset consisted on a data matrix of 24 spectra (rows) having 37,476 ppm values (columns) each one. This data matrix was stored in an ASCII file format.

^1H NMR spectra were normalized using Probabilistic Quotient Normalization (PQN)[2] to correct possible sample size effects. In this case, the reference spectrum used was the median spectrum obtained from the ^1H NMR spectra of yeast samples cultured at standard conditions. The quotient values were calculated using the median quotient values obtained from dividing the intensity values of the region of 0.8 – 3.8 ppm of every ^1H NMR spectrum by the reference spectrum.

5. METABOLITE QUANTITATION

L-Glutamate, L-lysine, citric acid, L-asparagine, glycine, L-histidine, L-leucine, L-ornithine, adenosine monophosphate and L-tyrosine presence in the ^1H NMR of the yeast extracts was confirmed against a home-made ^1H NMR spectra library and also using the Yeast Metabolome Data Base library[3](YMDB).

Proton resonances of these metabolites were further integrated using BATMAN R-package[4].

Relative concentration values obtained with this methodology are presented in **TableS1** below:

Table S1. Mean and standard deviation relative concentration values for the quantified metabolites.

	15°C		30°C		37°C		40°C	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Leu	2.66E-03	2.22E-04	2.32E-03	2.47E-04	2.41E-03	3.83E-04	1.88E-03	3.57E-04
Lys	1.12E-02	1.06E-03	2.50E-02	1.30E-03	2.84E-02	1.98E-03	2.76E-02	1.11E-03
Glu	3.18E-02	3.92E-03	3.92E-02	1.73E-03	4.51E-02	2.87E-03	5.25E-02	3.88E-03
Cit	9.51E-03	1.67E-03	3.94E-03	9.74E-04	4.28E-03	6.57E-04	4.80E-03	1.45E-03
Asn	4.54E-03	1.08E-03	5.22E-03	3.42E-04	5.30E-03	1.77E-03	4.43E-03	9.36E-04
Orn	3.81E-03	5.72E-04	1.71E-03	5.20E-04	1.25E-03	3.92E-04	1.27E-03	7.08E-04
Gly	2.36E-03	7.27E-04	3.92E-03	4.28E-04	6.86E-03	1.45E-03	2.82E-03	9.11E-04
Tyr	5.10E-04	1.55E-04	6.13E-04	1.06E-04	8.30E-04	1.12E-04	6.39E-04	8.85E-05
His	1.49E-03	1.87E-04	3.32E-03	1.51E-04	4.06E-03	8.15E-04	3.43E-03	3.51E-04
AMP	1.01E-03	3.47E-04	1.33E-03	1.06E-04	1.88E-03	2.86E-04	1.46E-03	3.99E-04

6. GENERATION OF SIMULATED DATASETS

Fitted metabolite ^1H NMR spectra were generated automatically during BATMAN analysis and stored in a *.txt* file in the output folder. For more information about the output data, see[5].

Each fitted ^1H NMR spectra was imported to Matlab R2014b (The Mathworks Inc. Natick, MA, USA) as an intensity vector with as many variables as spectral data points used during BATMAN analysis. Next, interpolation was applied in each vector using as a reference vector a common ppm vector of 10,801 chemical shifts.

Continuing from the previous step, in every interpolated ^1H NMR spectra, \mathbf{s}_j , the following operation was performed:

$$\mathbf{s}_j^0 = \mathbf{s}_j \frac{nH_j}{\sum_{p=1}^{p=length(ppm)} S_{j,p}}$$

where nH_j is the proton number of the corresponding j metabolite. With this operation, the inherent correspondence between proton intensity and metabolite concentration is established for the studied metabolites.

Each simulated mixture spectra, \mathbf{x}_i , was modelled according to the equation below:

$$\mathbf{x}_i = \sum_{j=1}^J c_{i,j}^0 \mathbf{s}_j^0 + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma)$$

Thus, \mathbf{x}_i is calculated as the sum of all metabolic contributions, defined by the product of each original individual ^1H NMR metabolite profile, \mathbf{s}_j^0 , and a concentration value, $c_{i,j}^0$ for the given i sample and j metabolite.

For \mathbf{X}_1 dataset, with randomly distributed concentrations, concentrations were defined using the following Matlab code:

```
C_X1 = rand(60,10);
```

For \mathbf{X}_2 dataset, concentrations were defined with the following Matlab code:

```
C_X2_15 = zeros(15,10);
```

```

C_X2_30 = zeros(15,10);

C_X2_37 = zeros(15,10);

C_X2_40 = zeros(15,10);

for i = 1:10

C_X2_15(:,i) = (2*SD15(i)) .* rand(15,1) + (mean15(i)-SD15(i));

C_X2_30(:,i) = (2*SD30(i)) .* rand(15,1) + (mean30(i)-SD30(i));

C_X2_37(:,i) = (2*SD37(i)) .* rand(15,1) + (mean37(i)-SD37(i));

C_X2_40(:,i) = (2*SD40(i)) .* rand(15,1) + (mean40(i)-SD40(i));

end

C_X2 = [C_X2_15;C_X2_30;C_X2_37;C_X2_40];

```

mean15, mean30, mean37 and mean40 represent the mean concentration values of the 10 metabolites at the four evaluated temperatures (first, third, fifth and seventh column in **TableS1** respectively), whereas SD15, SD30, SD37 and SD40 contain the corresponding standard deviation associated to these mean concentrations (second, fourth, sixth and eighth columns in **TableS1**).

Therefore,

$$X1 = C_X1 * SJ_O;$$

$$X2 = C_X2 * SJ_O;$$

Finally, noise was added in each ^1H NMR spectra of the given datasets as:

$$\sigma = \max(\max(X1)) / \text{SNR};$$

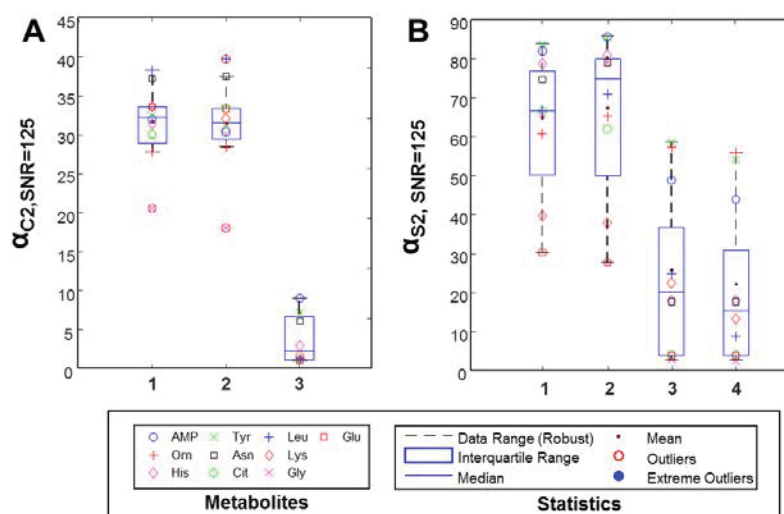
$$X1_WITH_NOISE = X1 + \text{normrnd}(0, \sigma, 60, 10801);$$

$$\sigma = \max(\max(X2)) / \text{SNR};$$

$$X2_WITH_NOISE = X2 + \text{normrnd}(0, \sigma, 60, 10801);$$

SNR is a scalar value pre-defined by the user. In the analyses, this value was either 125 or 500.

Appendix 2. Comparison of MCR-ALS results depending on initial estimates and on spectrum equality constraints in $\mathbf{X}_{2, \text{SNR}=125}$ dataset.



FigS1. Comparison of MCR-ALS results depending on initial estimates and on spectrum equality constraints in $\mathbf{X}_{2, \text{SNR}=125}$ dataset. (A). Similarity angle between the original concentrations, \mathbf{c}_j^0 , and the MCR-ALS resolved ones, \mathbf{c}_n^r . (B). Similarity angle between the metabolic ^1H NMR known spectra (\mathbf{s}_j^0) and the MCR-ALS resolved ones (\mathbf{s}_n^r). Method used: **1**, MCR-ALS with purest \mathbf{C} [6]; **2**, MCR-ALS with purest \mathbf{S}^T [6]; **3**, DTC-MCR-ALS; **4**, DTC-MCR-ALS, but only variables of correlated ^1H NMR regions were used to calculate similarity angles. Numerical values shown in boxplots are given in more detail in **Appendix 3** of Supplementary material.

Appendix 3. Similarity angles and regression coefficients between the original and resolved values (^1H NMR spectra or concentrations) for \mathbf{X}_2 dataset.

Table S1. Similarity angles and regression coefficients between the original concentrations, \mathbf{c}_j^0 , and the resolved ones, \mathbf{c}_n^r , in $\mathbf{X}_{2,\text{SNR}=500}$ dataset.

Metabolites	\mathbf{s}_j^0 ¹		C ²		DTC-MCR-ALS ³	
	angle	r2	angle	r2	angle	r2
AMP	11.7602	0.9790	12.2595	0.9772	1.1664	0.9998
L-Ornithine	12.4757	0.9764	13.3453	0.9730	9.2542	0.9870
L-Histidine	7.1619	0.9922	6.1465	0.9943	1.6496	0.9996
L-Tyrosine	11.5833	0.9796	13.1713	0.9737	1.6045	0.9996
L-Asparagine	15.8378	0.9620	19.5931	0.9421	4.4169	0.9970
Citric acid	10.2642	0.9840	18.6836	0.9473	0.1887	1.0000
L-Leucine	16.8116	0.9573	23.4082	0.9177	1.0656	0.9998
L-Lysine	2.0237	0.9994	1.2023	0.9998	2.3767	0.9991
Glycine	7.8188	0.9907	8.4528	0.9891	0.2657	1.0000
L-Glutamate	0.8205	0.9999	9.7183	0.9856	0.0308	1.0000

¹ \mathbf{s}_j^0 : original templates were used as initial estimates; ²C: purest concentration profiles[6] were used as initial estimates; ³DTC-MCR-ALS: DTC-MCR-ALS approach was used.

Table S2. Similarity angles and regression coefficients between the original concentrations, \mathbf{c}_j^0 , and the resolved ones, \mathbf{c}_n^r , in $\mathbf{X}_{2,\text{SNR}=125}$ dataset.

Metabolites	\mathbf{s}_j^0 ¹		C ²		DTC-MCR-ALS ³	
	angle	r2	angle	r2	angle	r2
AMP	31.0530	0.8567	31.8807	0.8491	8.9760	0.9878
L-Ornithine	27.0596	0.8905	27.8064	0.8845	40.4335	0.7612
L-Histidine	32.5487	0.8429	31.3591	0.8539	2.9949	0.9986
L-Tyrosine	32.1437	0.8467	32.4889	0.8435	7.1318	0.9923
L-Asparagine	37.1181	0.7974	37.2075	0.7965	6.1571	0.9942
Citric acid	30.0881	0.8653	29.9772	0.8662	1.1605	0.9998
L-Leucine	38.9595	0.7776	38.2634	0.7852	1.2771	0.9998
L-Lysine	30.8968	0.8581	33.6897	0.8321	1.6071	0.9996
Glycine	37.8475	0.7896	20.5646	0.9363	0.9873	0.9999
L-Glutamate	27.2930	0.8887	33.4938	0.8339	1.0593	0.9998

¹ \mathbf{s}_j^0 : original templates were used as initial estimates; ²C: purest concentration profiles[6] were used as initial estimates; ³DTC-MCR-ALS: DTC-MCR-ALS approach was used.

Table S3. Similarity angles and regression coefficients between the metabolic ^1H NMR templates (\mathbf{s}_j^0) and the resolved ones (\mathbf{s}_n^r) in $\mathbf{X}_{2,\text{SNR}=500}$ dataset.

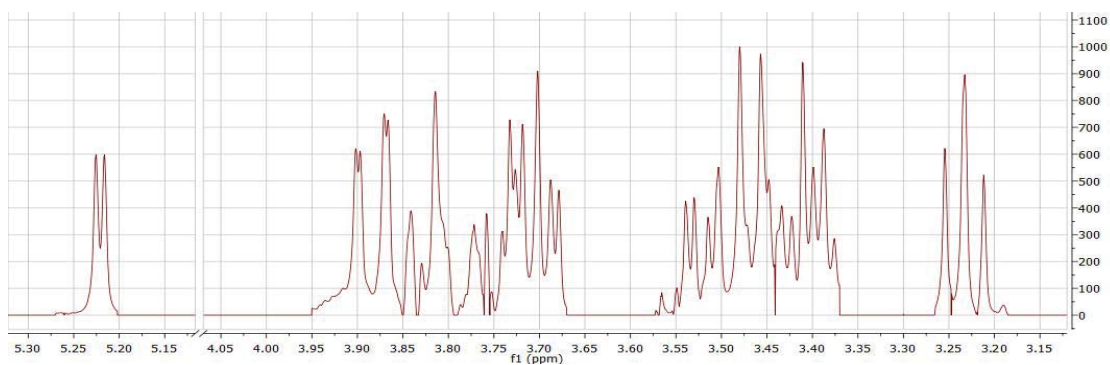
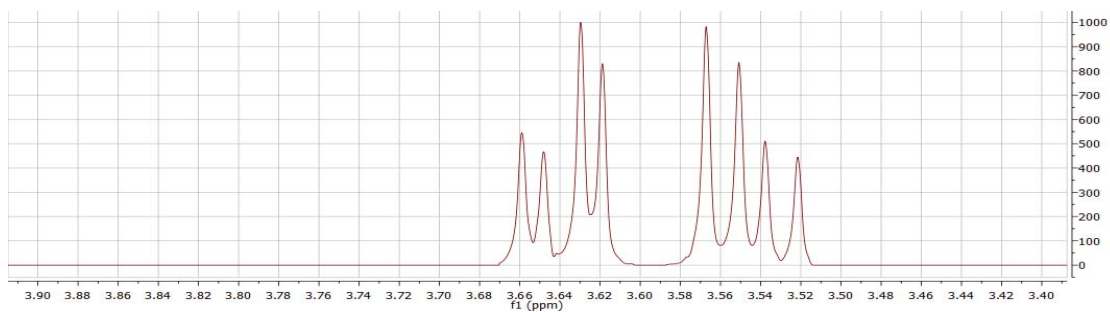
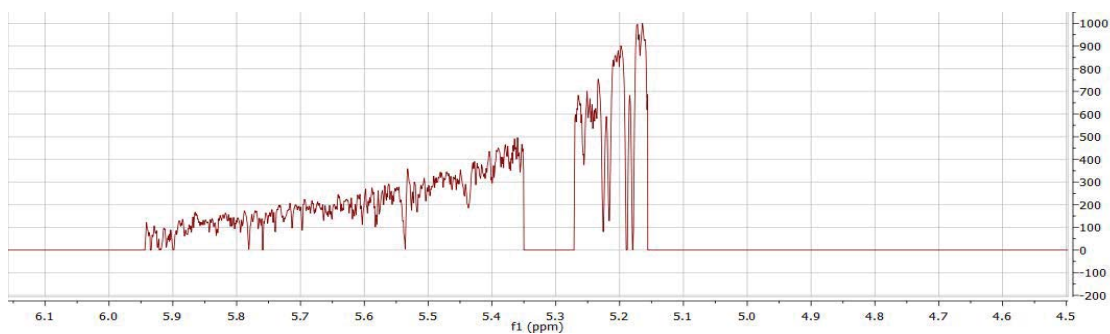
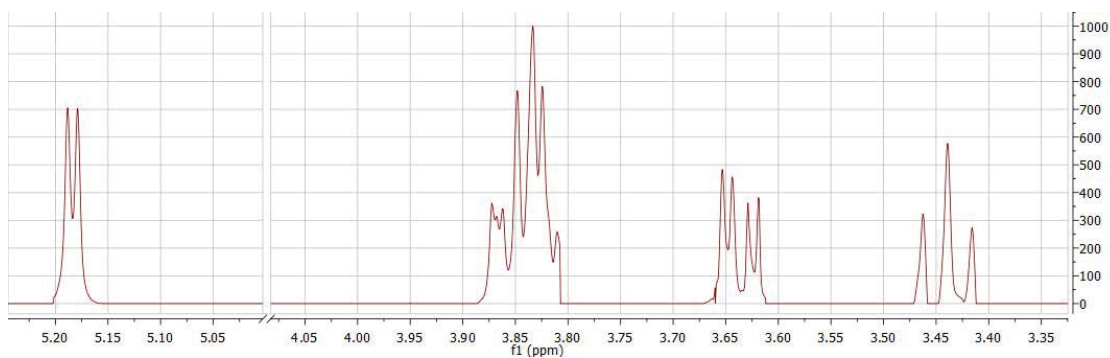
<i>Metabolites</i>	\mathbf{s}_j^0 ¹		\mathbf{C} ²		DTC(1) ³		DTC(2) ⁴	
	angle	r2	angle	r2	angle	r2	angle	r2
AMP	68.2880	0.3699	65.0799	0.4214	21.6799	0.9293	8.3275	0.9895
L-Ornithine	40.8688	0.7562	33.7142	0.8318	31.7879	0.8500	18.9927	0.9456
L-Histidine	64.4692	0.4310	62.2442	0.4657	8.0514	0.9901	8.0320	0.9902
L-Tyrosine	74.2773	0.2710	81.0864	0.1549	22.3988	0.9246	11.3189	0.9806
L-Asparagine	41.4279	0.7498	44.5624	0.7125	5.7772	0.9949	5.6583	0.9951
Citric acid	18.0526	0.9508	54.2470	0.5843	2.1578	0.9993	2.1543	0.9993
L-Leucine	25.8526	0.8999	33.6729	0.8322	21.3812	0.9312	5.1645	0.9959
L-Lysine	8.4882	0.9890	47.3520	0.6775	11.6249	0.9795	11.6115	0.9795
Glycine	23.3036	0.9184	23.2452	0.9188	1.2732	0.9998	1.1863	0.9998
L-Glutamate	3.1108	0.9985	7.9510	0.9904	24.9409	0.9067	0.8084	0.9999

$^1\mathbf{s}_j^0$: original templates were used as initial estimates; $^2\mathbf{C}$: purest concentration profiles[6] were used as initial estimates; $^3\text{DTC}(1)$: DTC-MCR-ALS approach was used, and similarity angles were calculated using variables containing information of all the ^1H NMR spectral domain; $^4\text{DTC}(2)$: DTC-MCR-ALS approach was used, but only variables from correlated ^1H NMR regions were used to calculate the similarity angles.

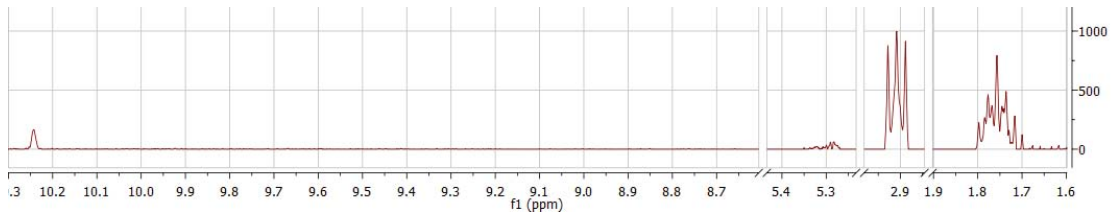
Table S4. Similarity angles and correlation coefficients between the metabolic ^1H NMR templates (\mathbf{s}_j^0) and the resolved ones (\mathbf{s}_n^r) in $\mathbf{X}_{2,\text{SNR}=125}$ dataset.

<i>Metabolites</i>	\mathbf{s}_j^0 ¹		\mathbf{C} ²		DTC(1) ³		DTC(2) ⁴	
	angle	r2	angle	r2	angle	r2	angle	r2
AMP	84.4442	0.0968	81.9920	0.1393	48.8373	0.6582	44.0606	0.7186
L-Ornithine	70.5299	0.3333	60.6205	0.4906	57.3915	0.5389	55.9924	0.5593
L-Histidine	80.7321	0.1611	78.8948	0.1926	17.8513	0.9519	17.8428	0.9519
L-Tyrosine	84.1449	0.1020	83.8984	0.1063	58.5574	0.5216	54.1759	0.5853
L-Asparagine	80.1943	0.1703	74.7239	0.2635	17.7051	0.9526	17.6678	0.9528
Citric acid	60.3737	0.4943	66.5556	0.3979	3.9579	0.9976	3.9561	0.9976
L-Leucine	69.3761	0.3522	66.6607	0.3962	24.8072	0.9077	8.6753	0.9886
L-Lysine	39.5769	0.7708	39.7726	0.7686	22.3114	0.9251	13.1440	0.9738
Glycine	60.7699	0.4883	65.8529	0.4091	2.6447	0.9989	2.6040	0.9990
L-Glutamate	24.4477	0.9103	30.2976	0.8634	3.7958	0.9978	3.7925	0.9978

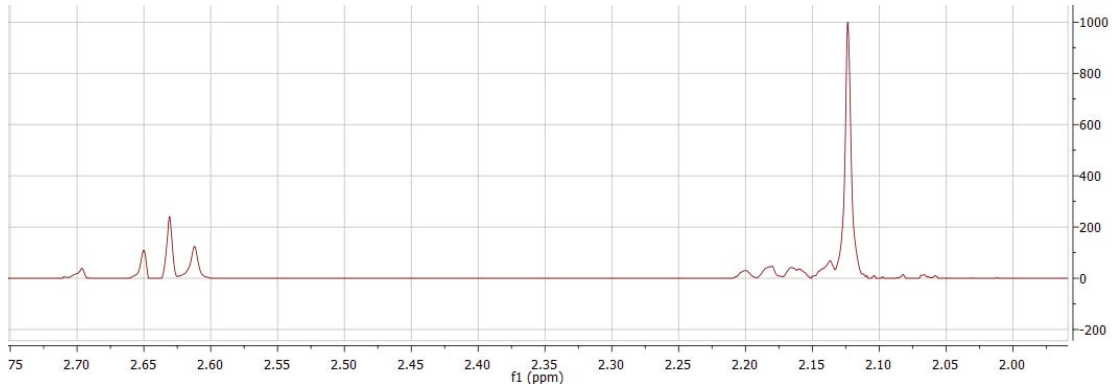
$^1\mathbf{s}_j^0$: original templates were used as initial estimates; $^2\mathbf{C}$: purest concentration profiles[6] were used as initial estimates; $^3\text{DTC}(1)$: DTC-MCR-ALS approach was used, and similarity angles were calculated using variables containing information of all the ^1H NMR spectral domain; $^4\text{DTC}(2)$: DTC-MCR-ALS approach was used, but only variables from correlated ^1H NMR regions were used to calculate the similarity angles.

Appendix 4. ^1H NMR spectral profiles for the 75 resolved components.**Component 1:** D-Glucose**Component 2:** Glycerol**Component 3:** Water shoulder**Component 4:** Trehalose

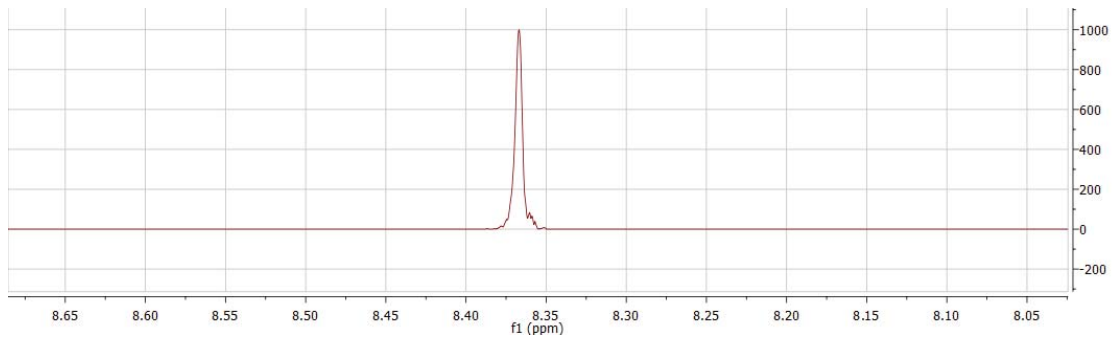
Component 5: DSS



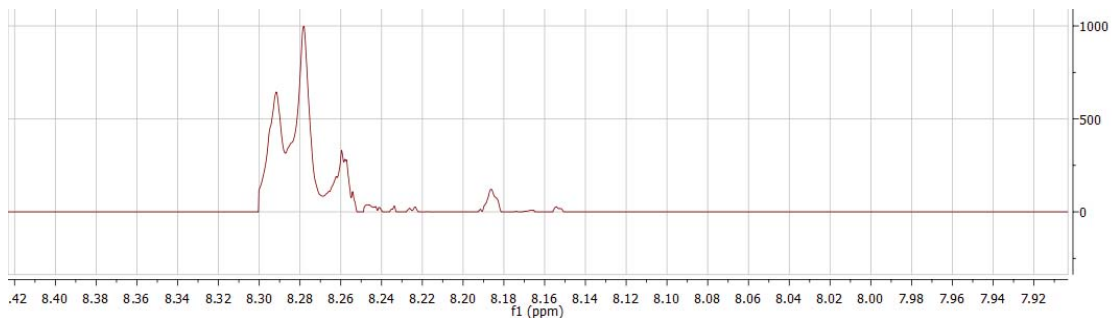
Component 6: L-Methionine



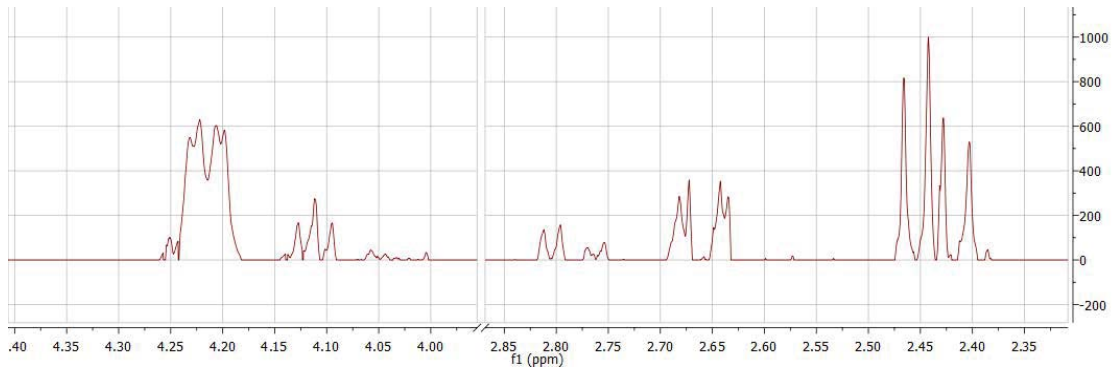
Component 7: 8.37 ppm (s)



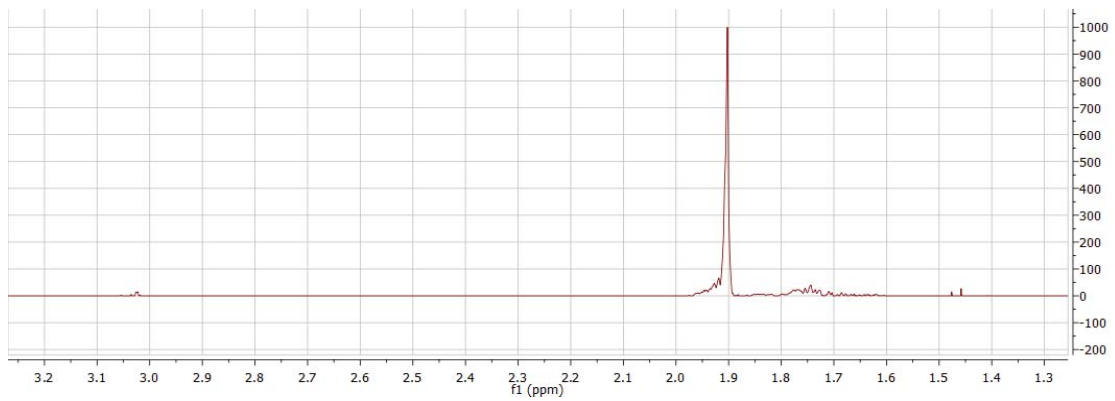
Component 8: N⁶-methyladenosine



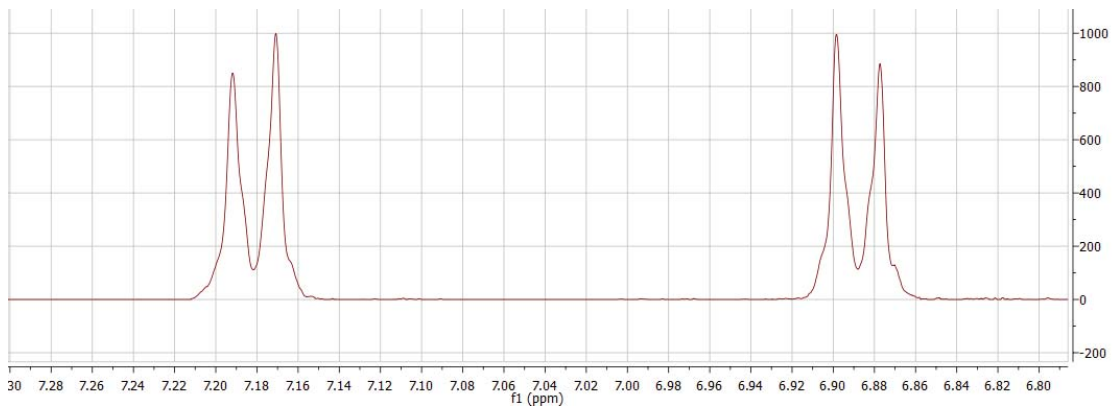
Component 9: Ureidosuccinic acid



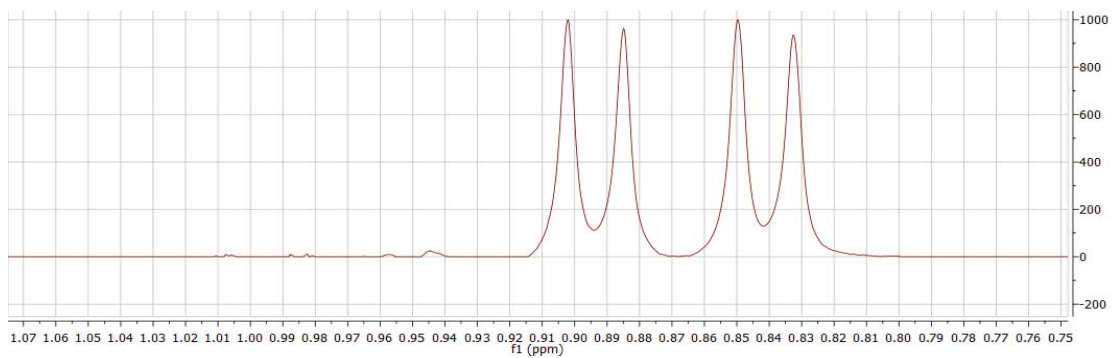
Component 10: Acetic acid



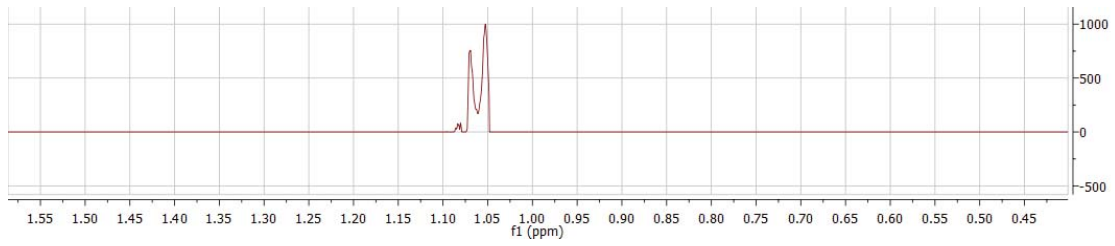
Component 11: L-Tyrosine



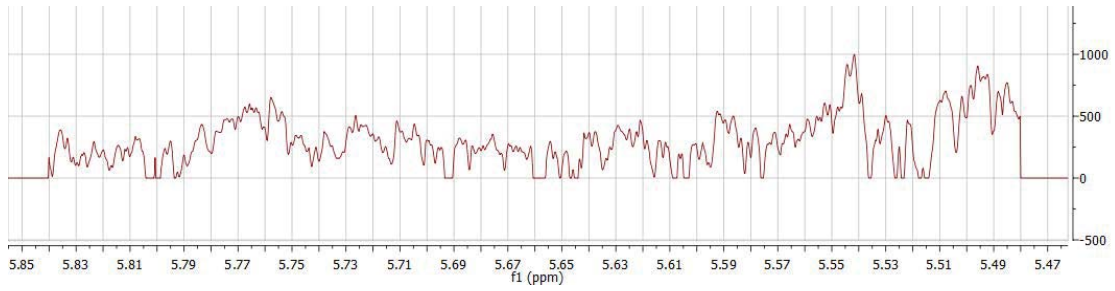
Component 12: 2-Isopropylmalic acid



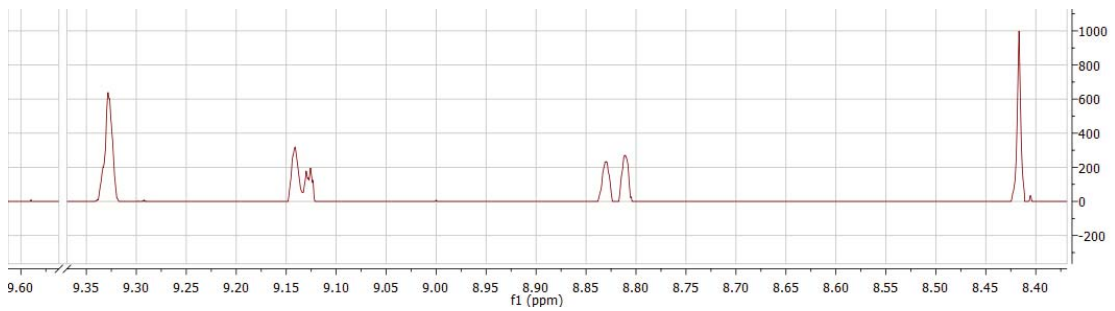
Component 13: S-3-Hydroxyisobutyric acid



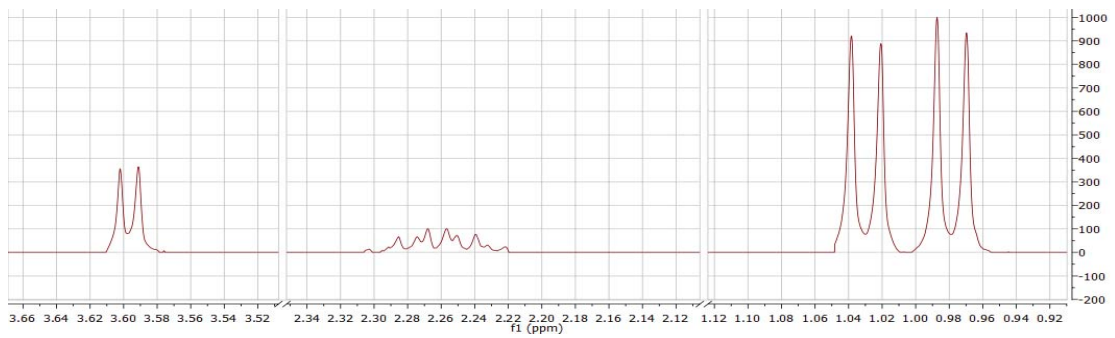
Component 14: Residual contribution of water



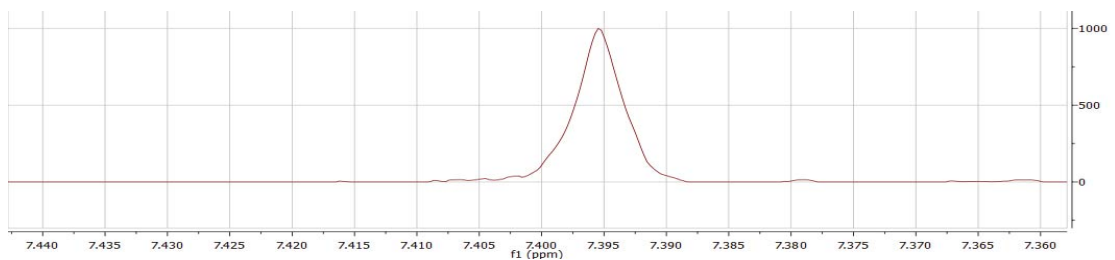
Component 15: NAD⁺



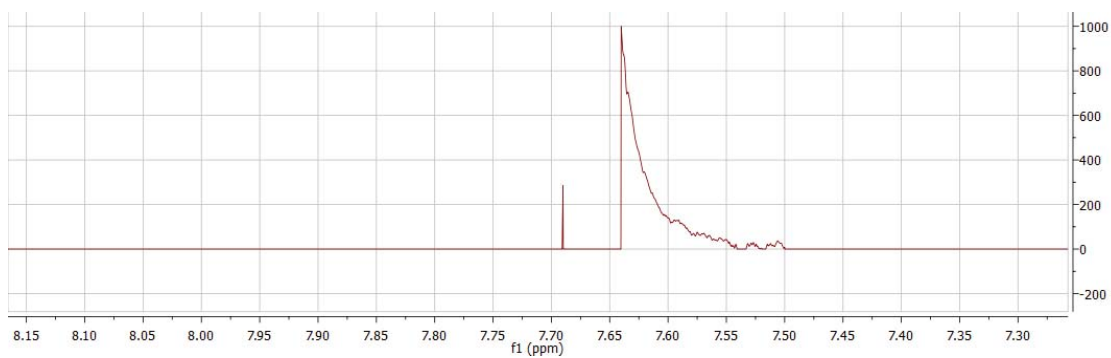
Component 16: L-Valine



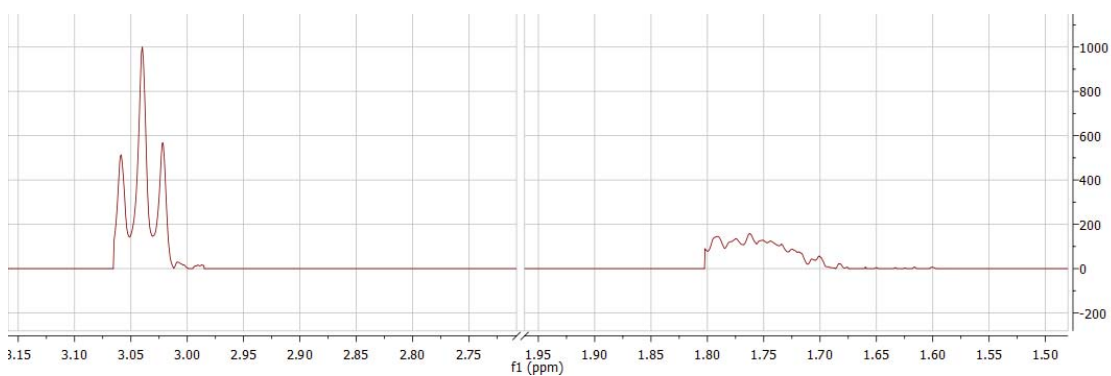
Component 17: 7.39 ppm (s)



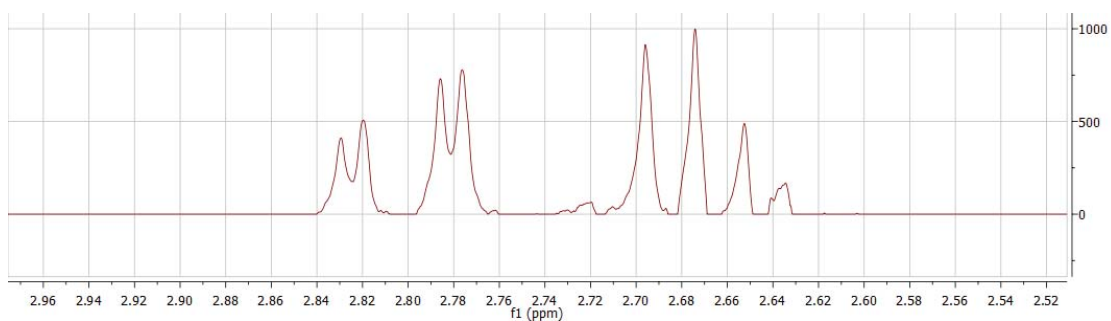
Component 18: Right shoulder of 7.66 ppm (s)



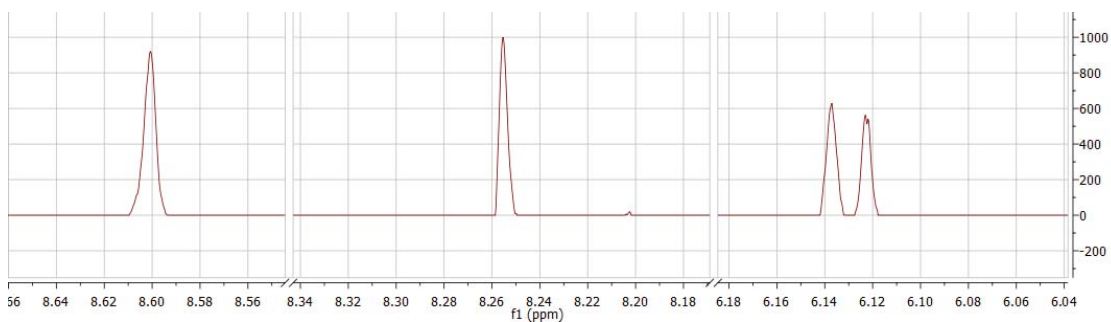
Component 19: L-Ornithine



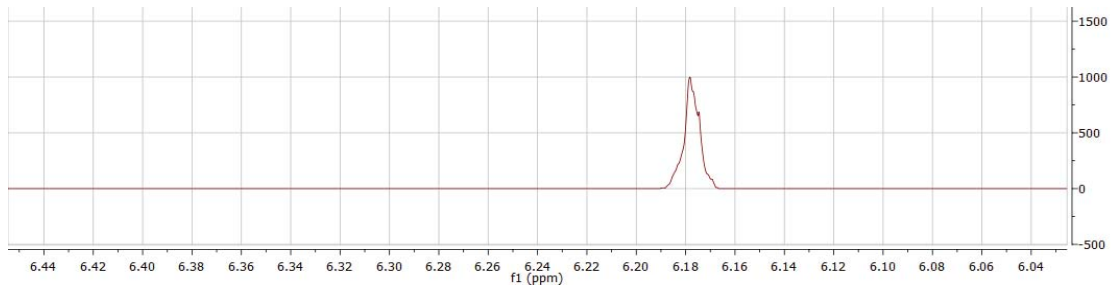
Component 20: L-Aspartic acid



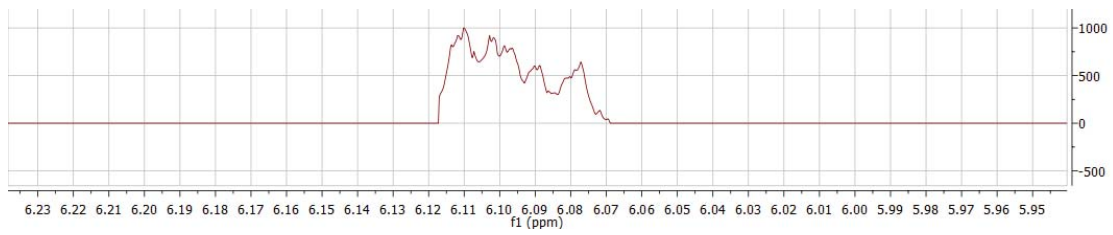
Component 21: AMP



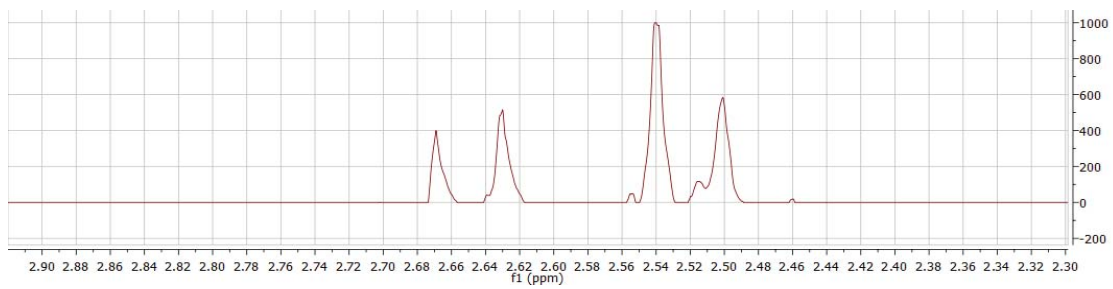
Component 22: Orotic acid



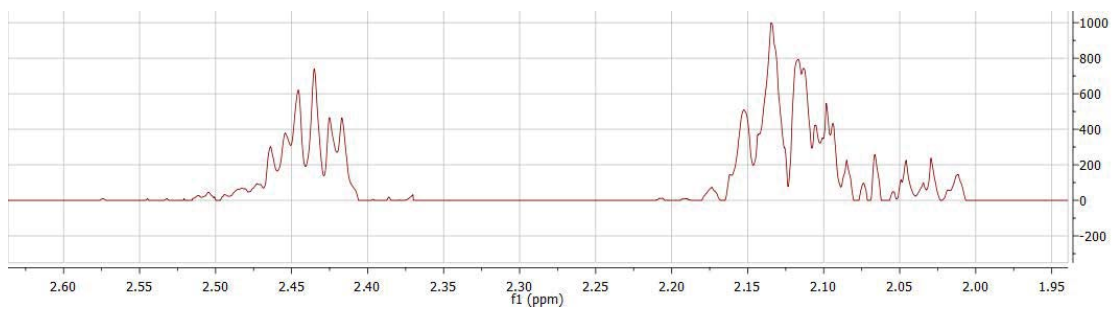
Component 23: 6.09 ppm (m)



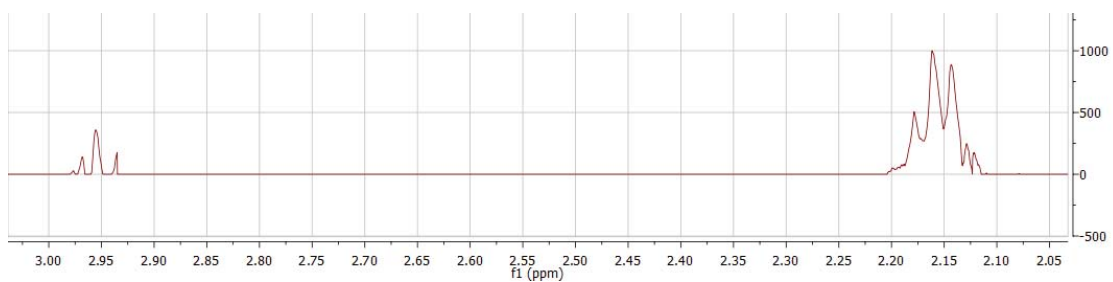
Component 24: Citric acid

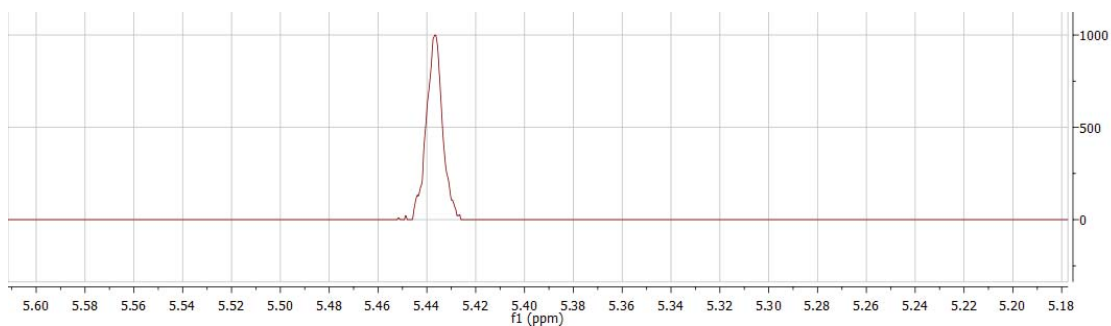
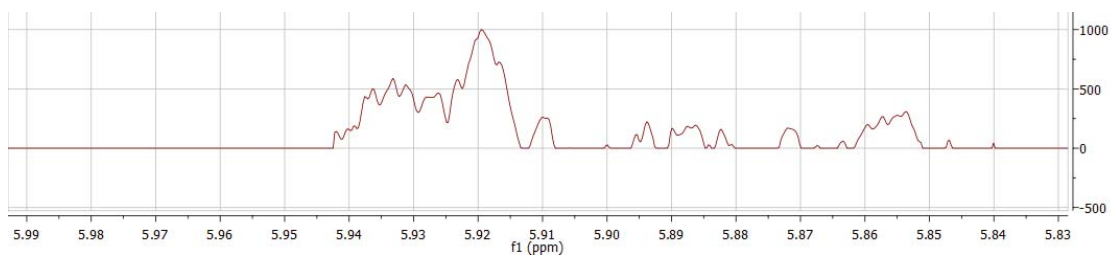
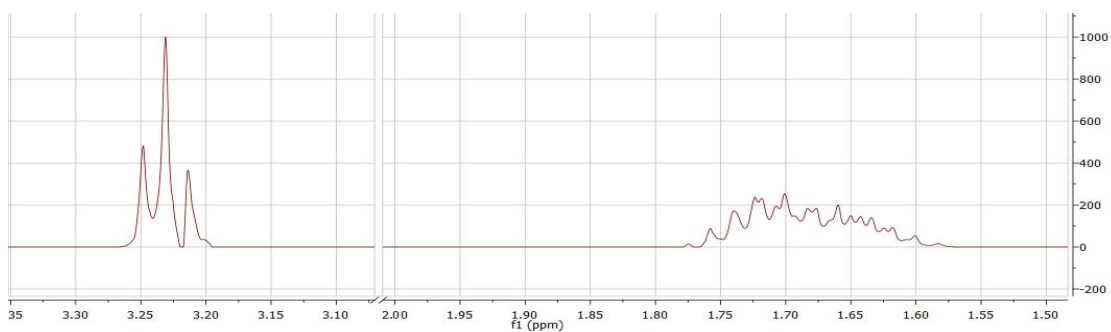
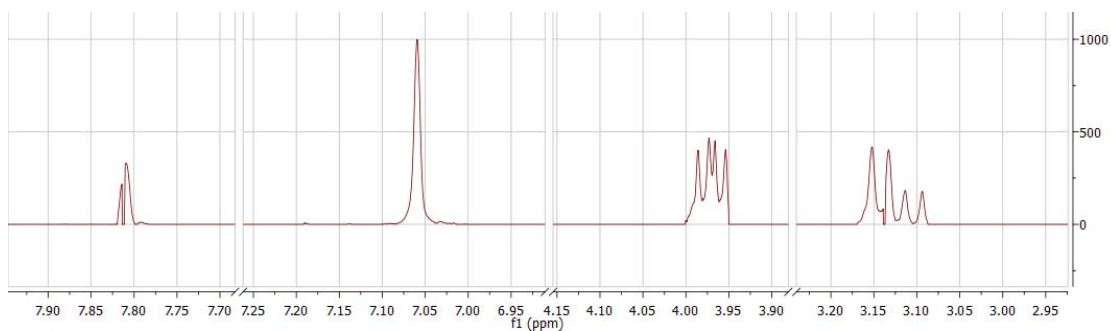


Component 25: L-Glutamine

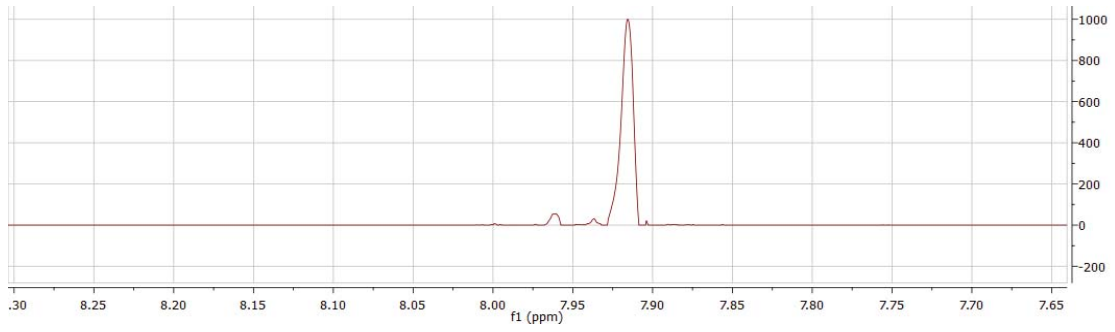


Component 26: Glutathione

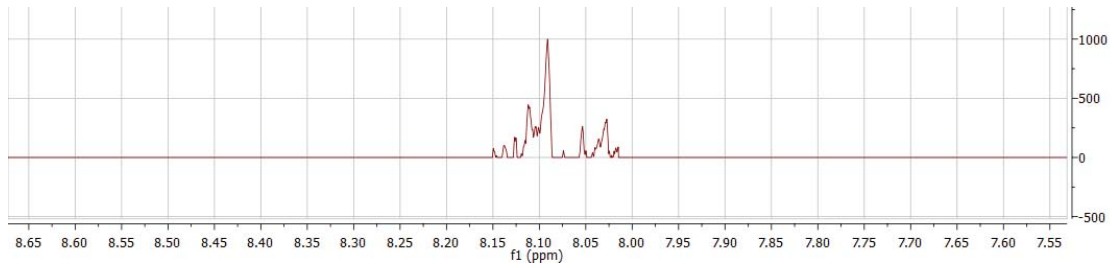


Component 27: 5.44 (s)**Component 28: Noise****Component 29: L-Arginine****Component 30: L-Histidine**

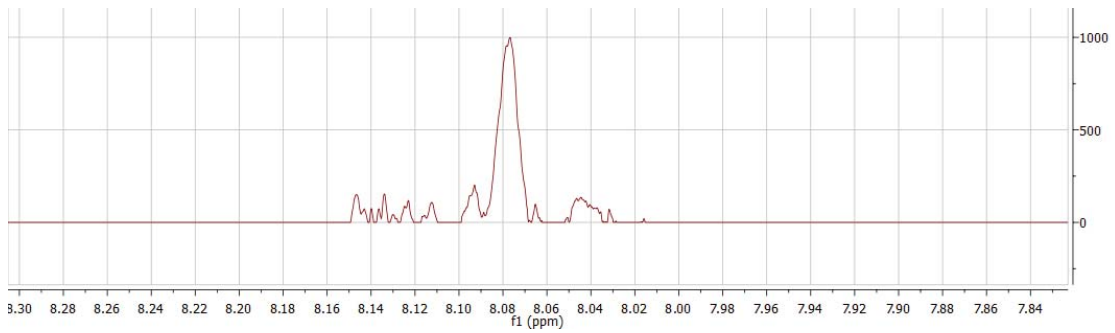
Component 31: 7.91 ppm (s)



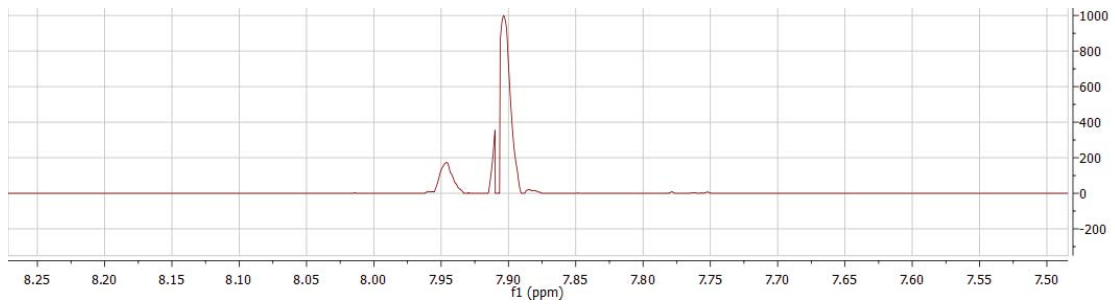
Component 32: 8.10 ppm (m)



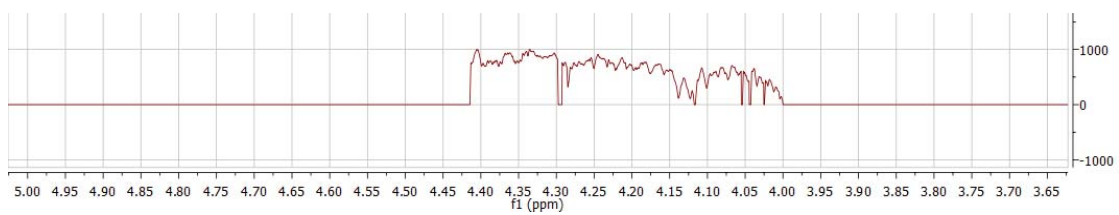
Component 33: 8.08 ppm (s)



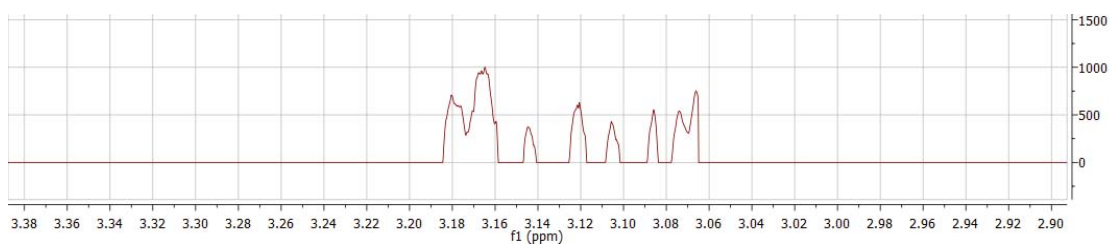
Component 34: 7.90 ppm (s)



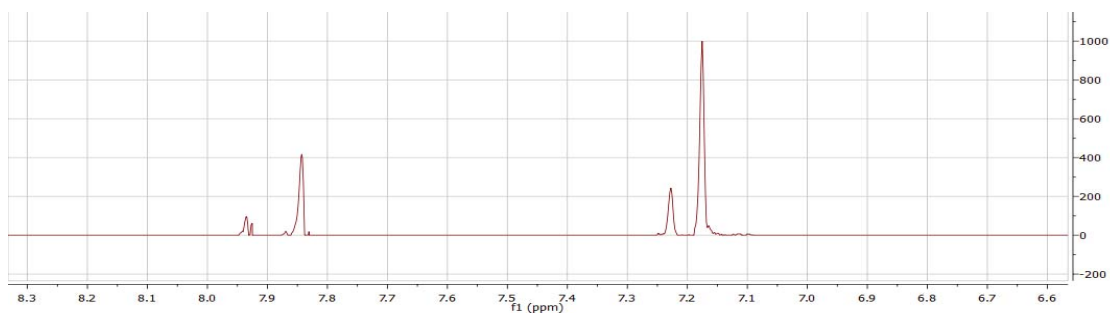
Component 35: Noise



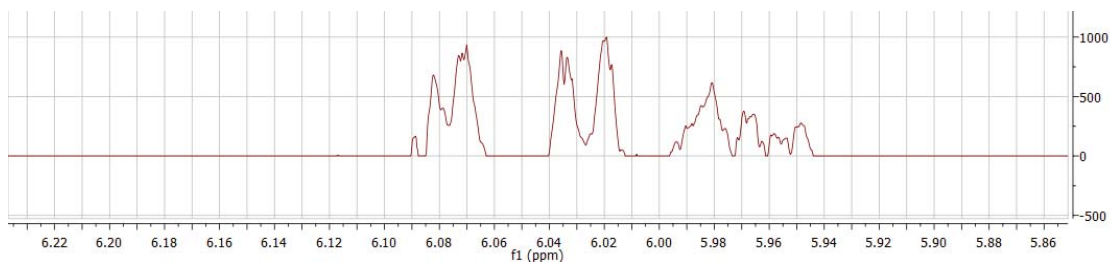
Component 36: Noise



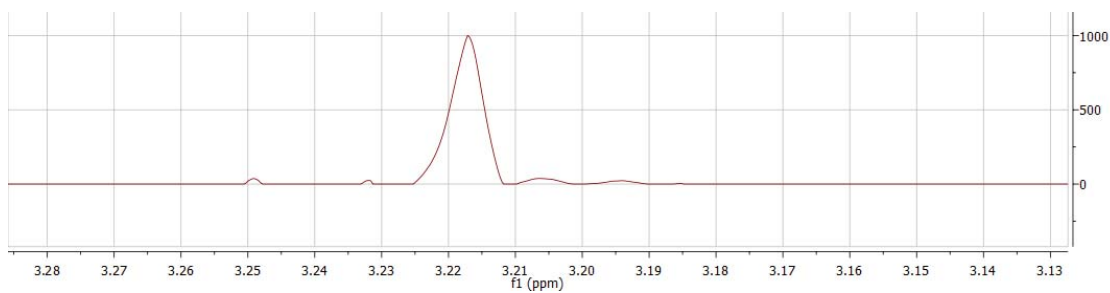
Component 37: EIGP



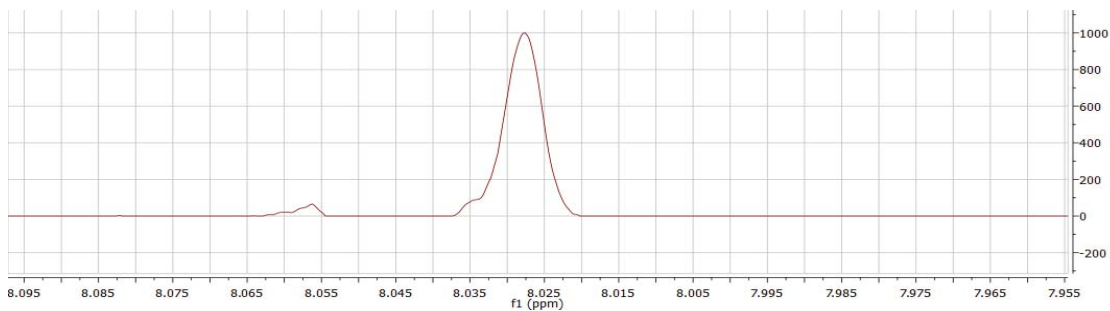
Component 38: 6.03 ppm (*d*, $J=6.1$ Hz), 6.08 ppm (*d*, $J=4.1$ Hz), 5.96 ppm (*m*)



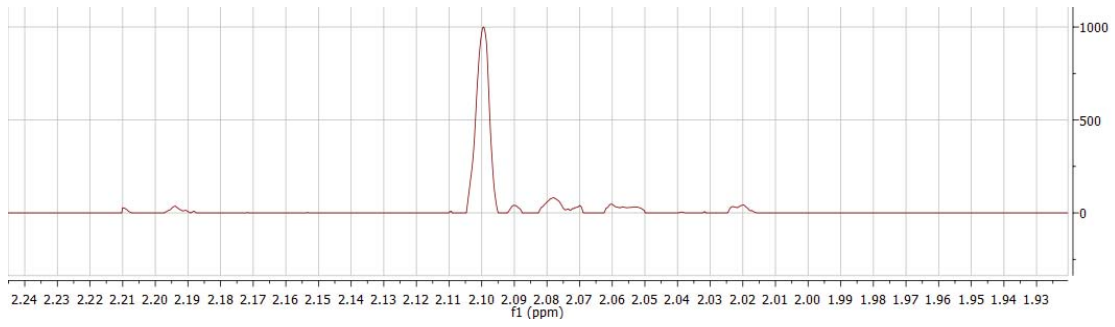
Component 39: GPC



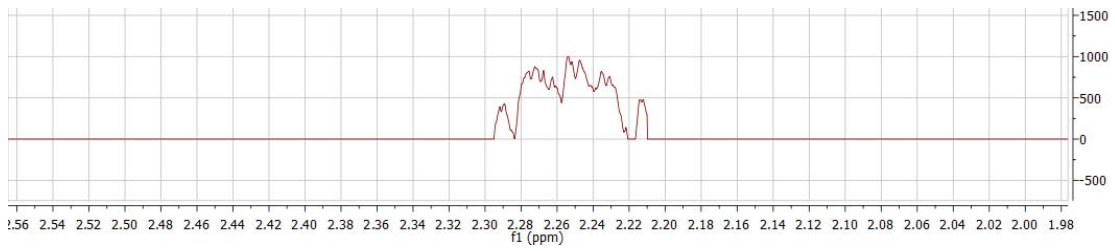
Component 40: 8.03 ppm (*s*)



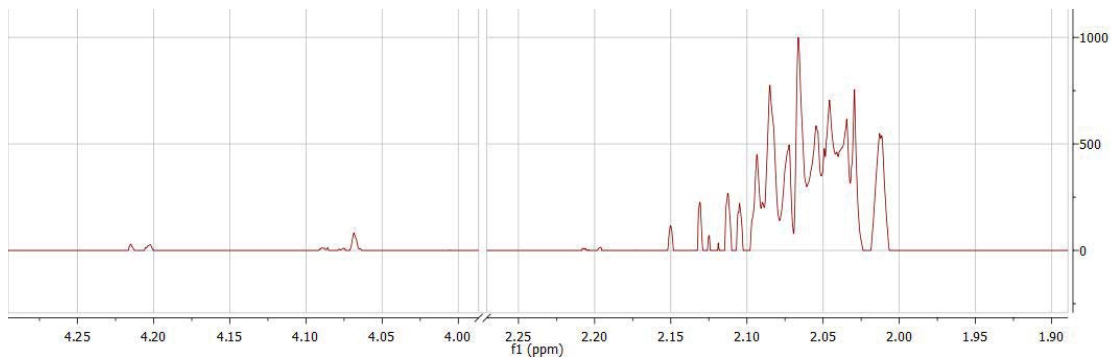
Component 41: 2.10 ppm (s)



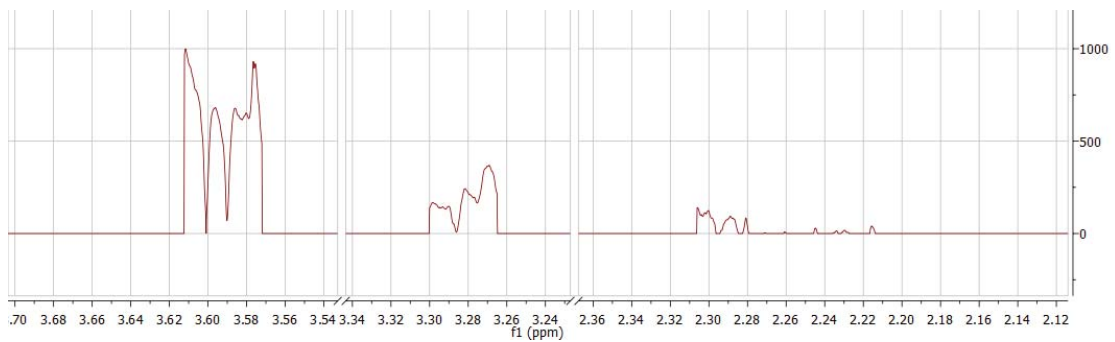
Component 42: 2.26 ppm (m)

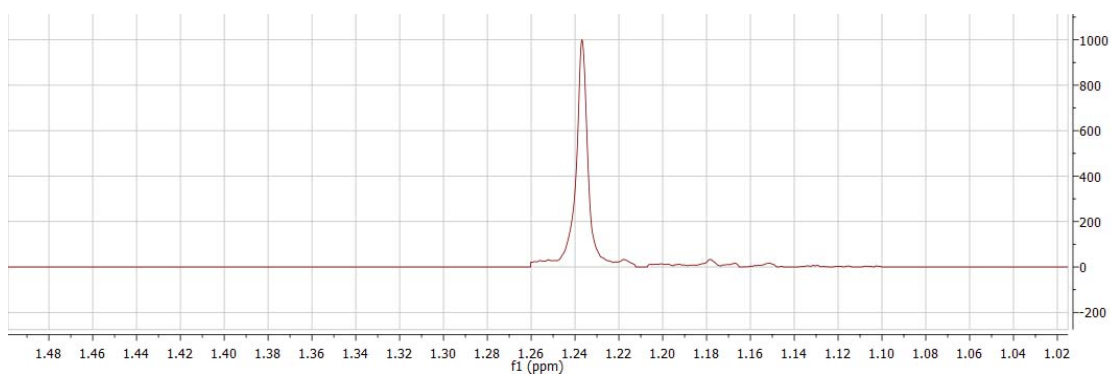
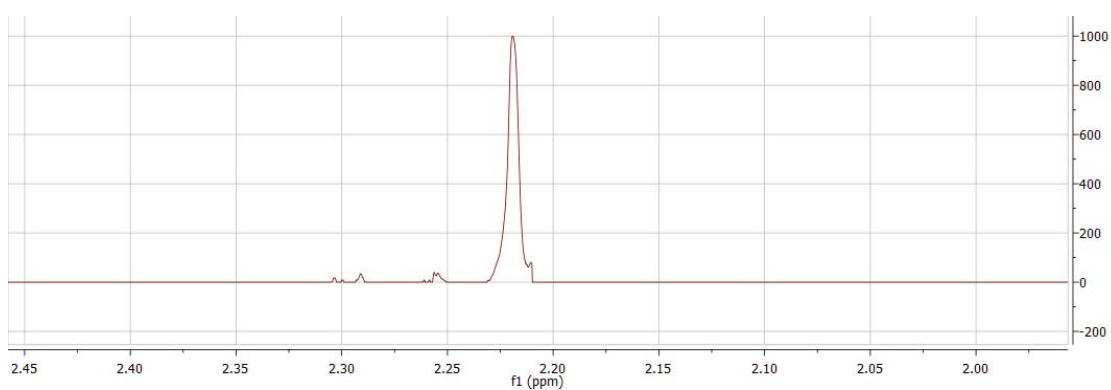
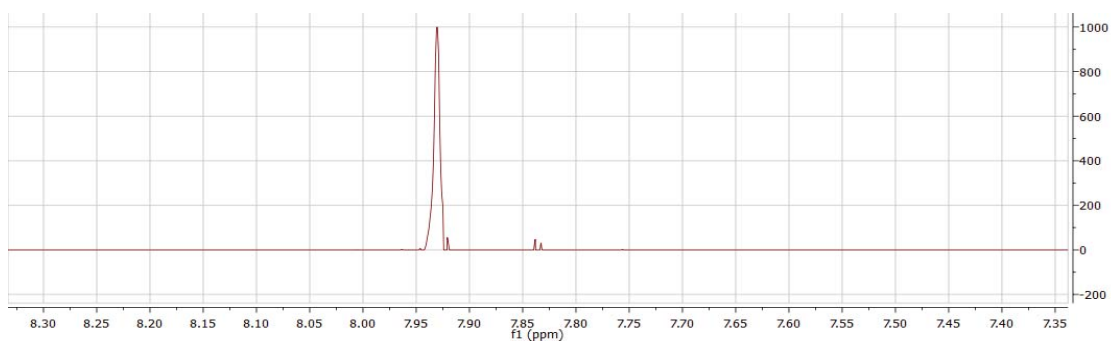
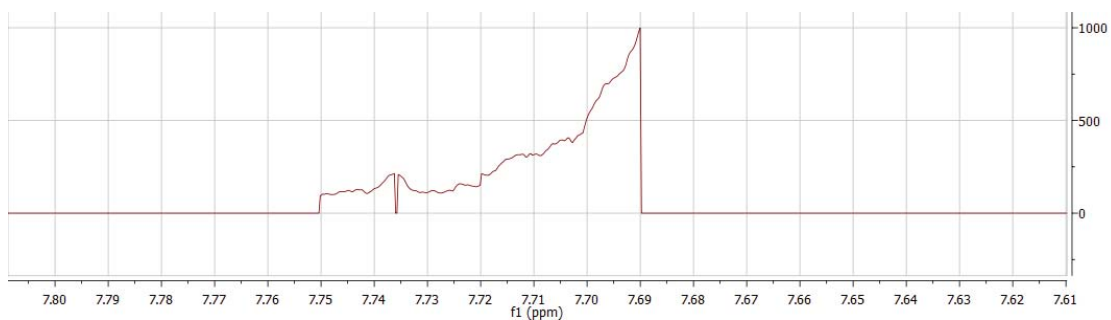


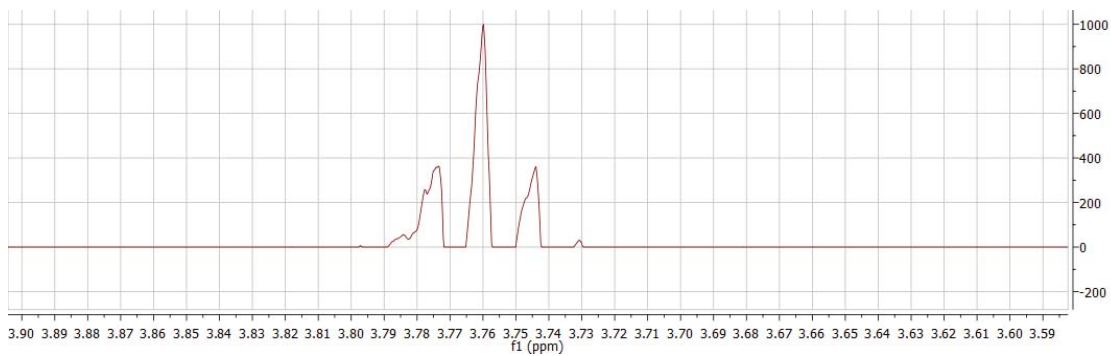
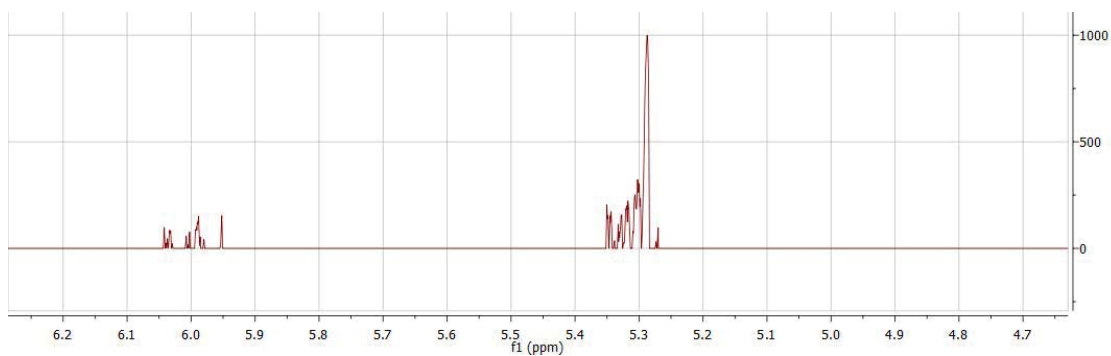
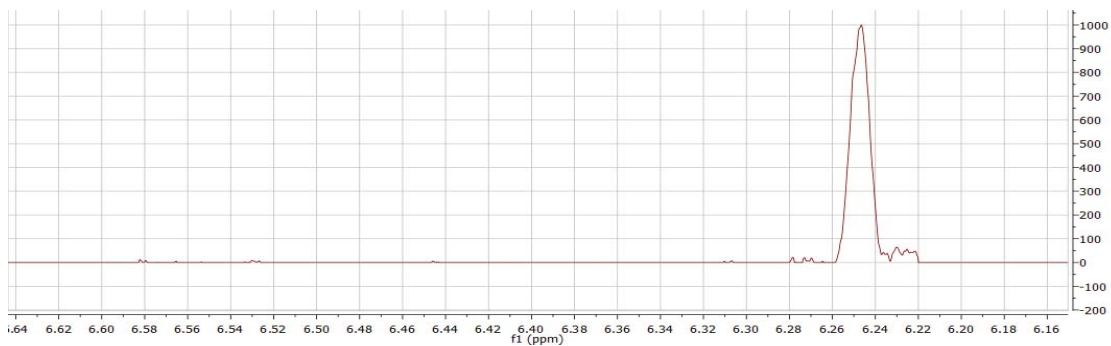
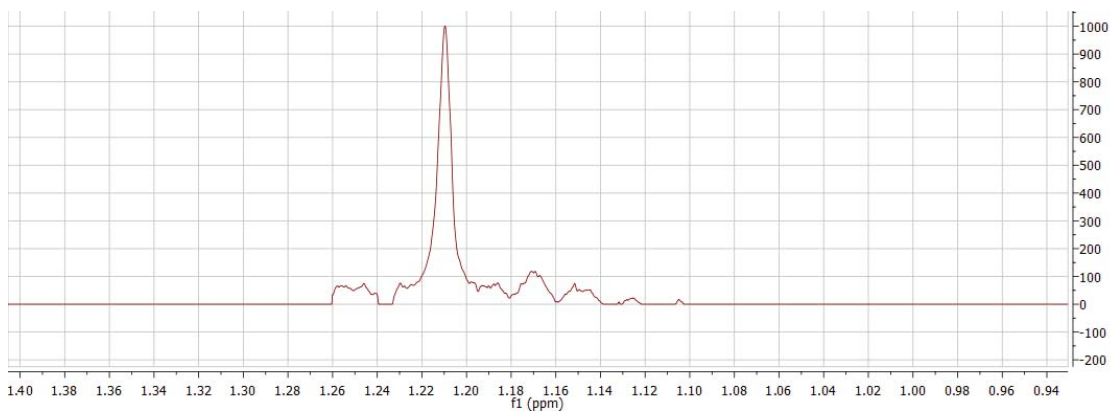
Component 43: 2.06 ppm (m)

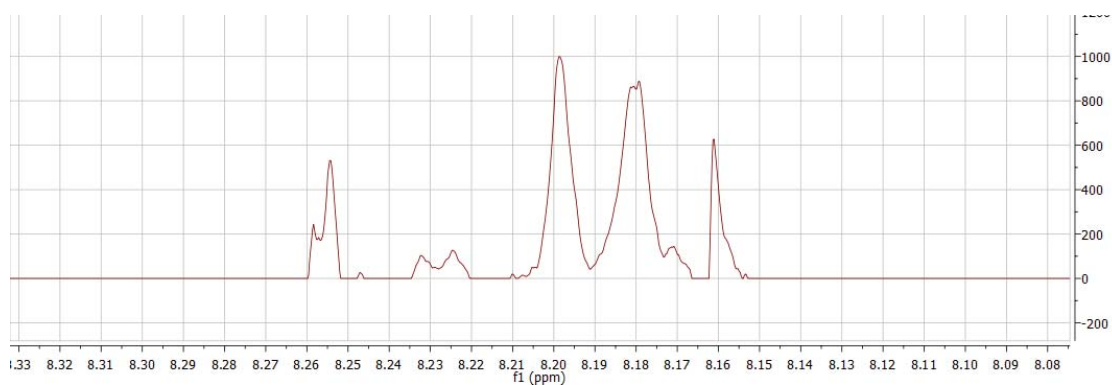
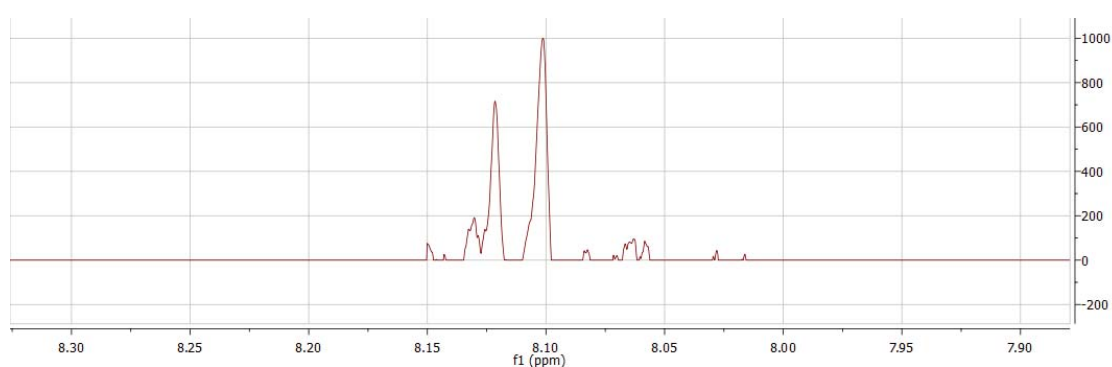
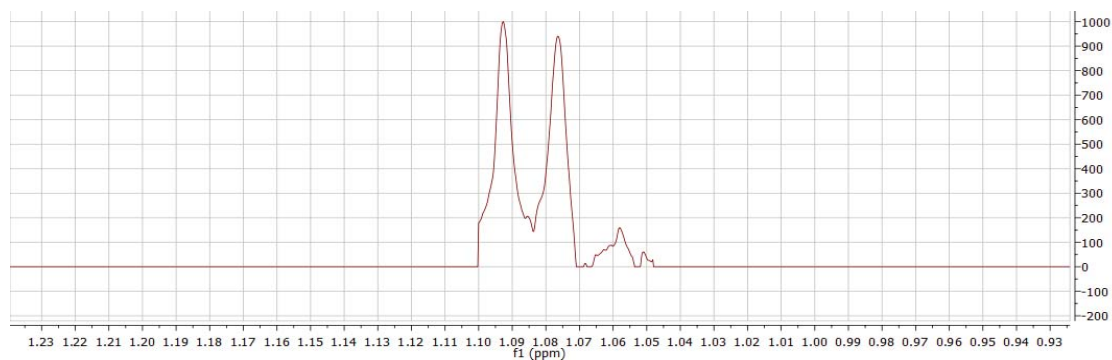
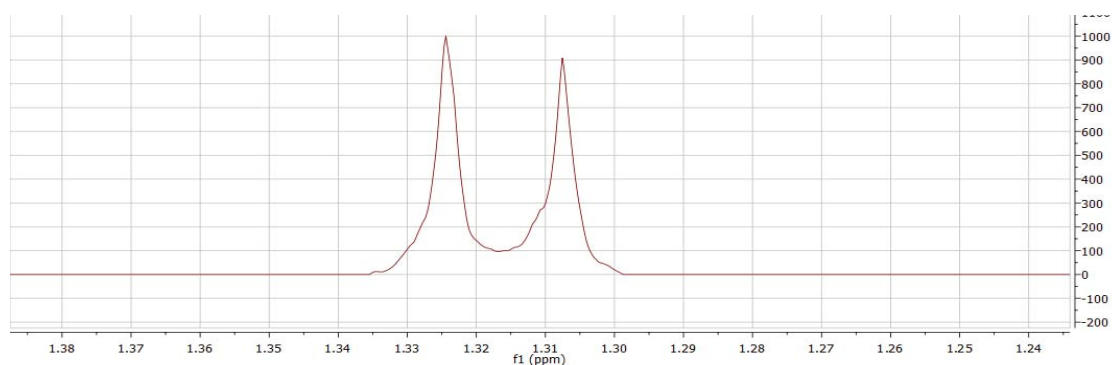


Component 44: Noise

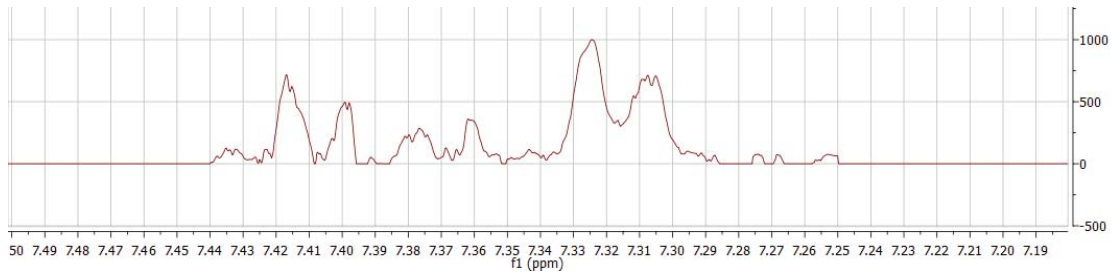


Component 45: Fatty acid singletComponent 46: 2.22 ppm (s)Component 47: 7.93 ppm (s)Component 48: Left shoulder of 7.66 ppm (s)

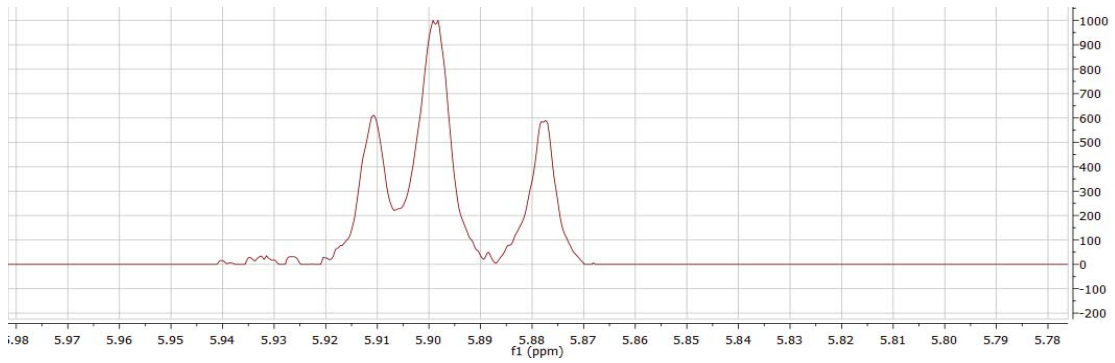
Component 49: 3.77 ppm (t, J=5.4 Hz)Component 50: NoiseComponent 51: 6.25 ppm (s)Component 52: 1.21 ppm (s)

Component 53: 8.18 ppm (m)Component 54: 8.11 ppm (d, $J=8.1$ Hz)Component 55: 3-Methyl-2-oxovaleric acidComponent 56: L-Threonine

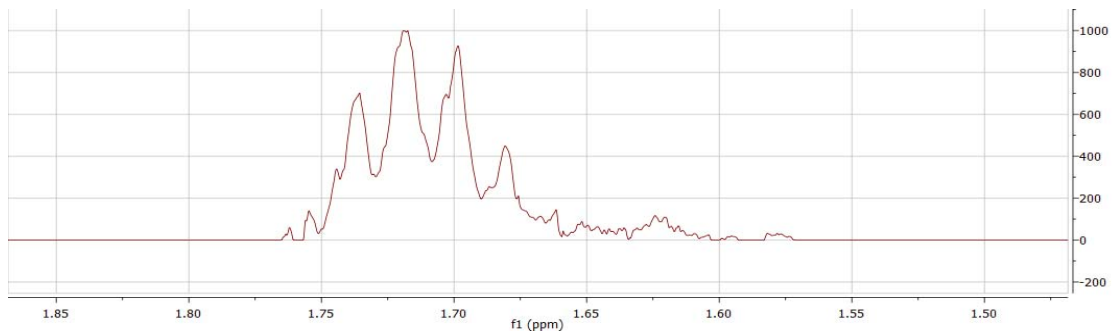
Component 57: L-Phenylalanine



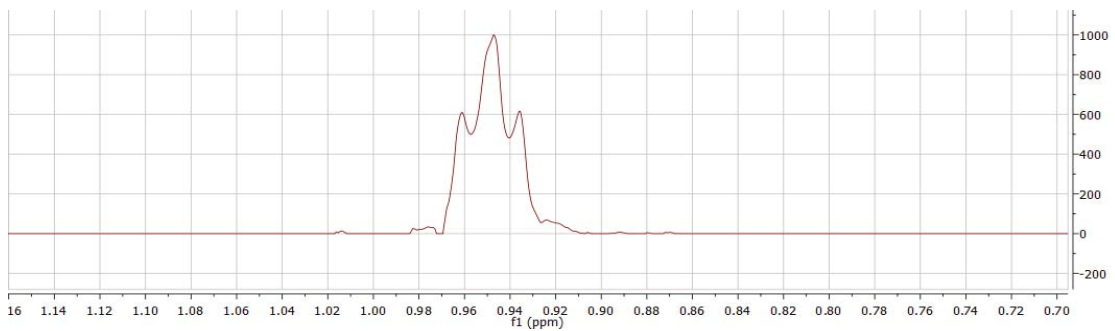
Component 58: Uridine

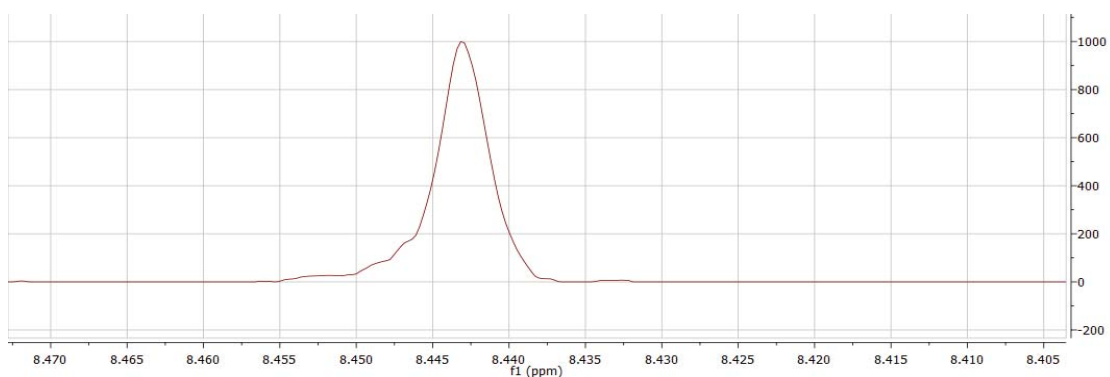
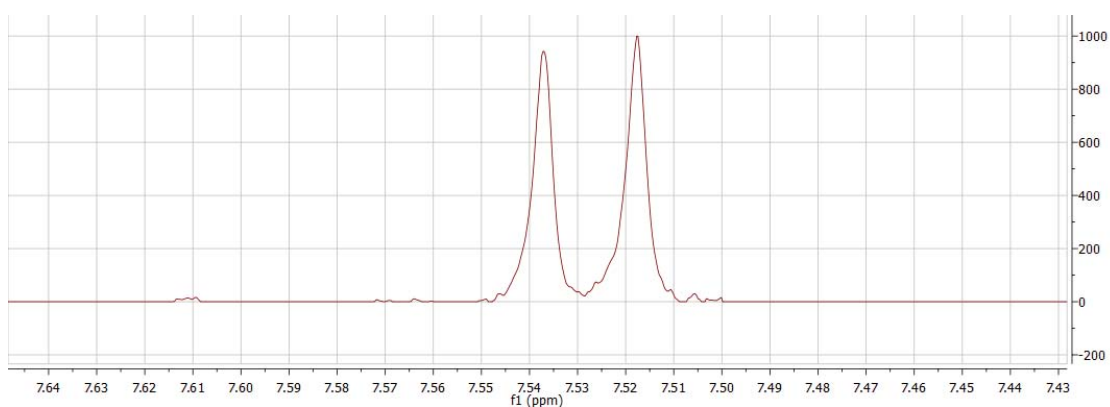
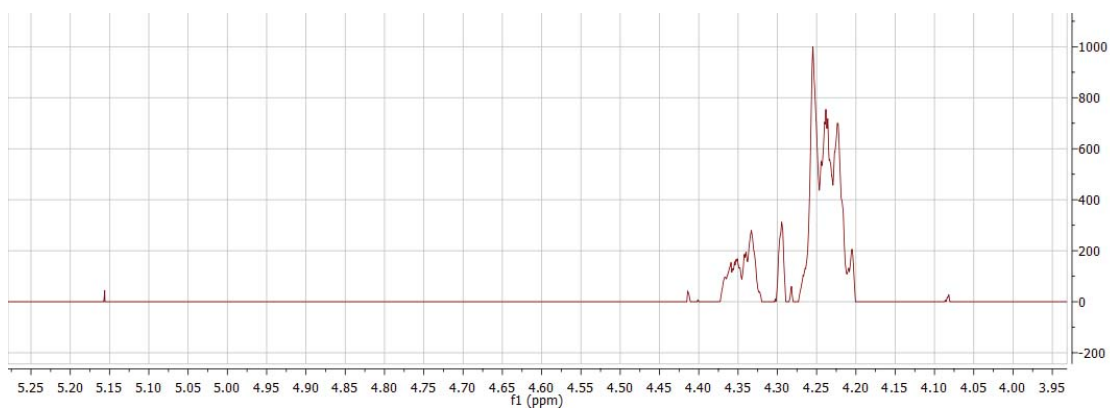
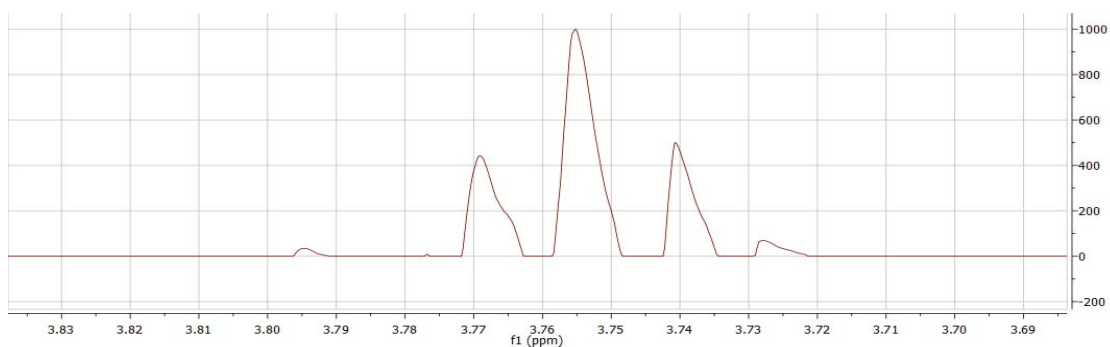


Component 59: 1.72 ppm (m)

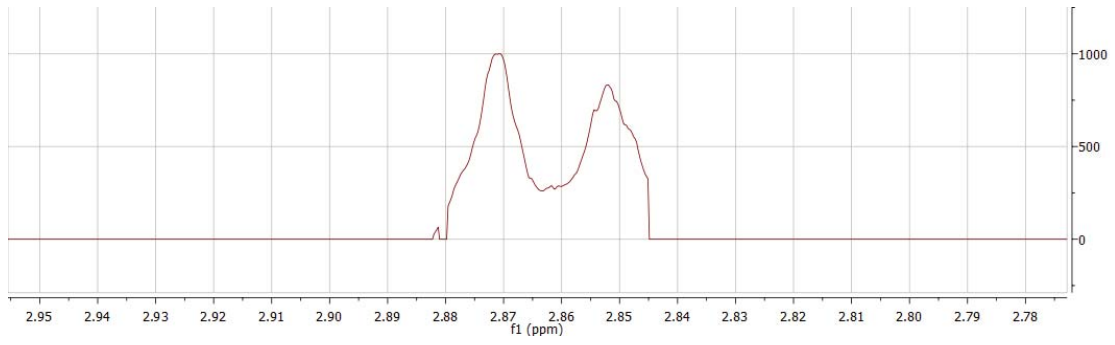


Component 60: L-Leucine

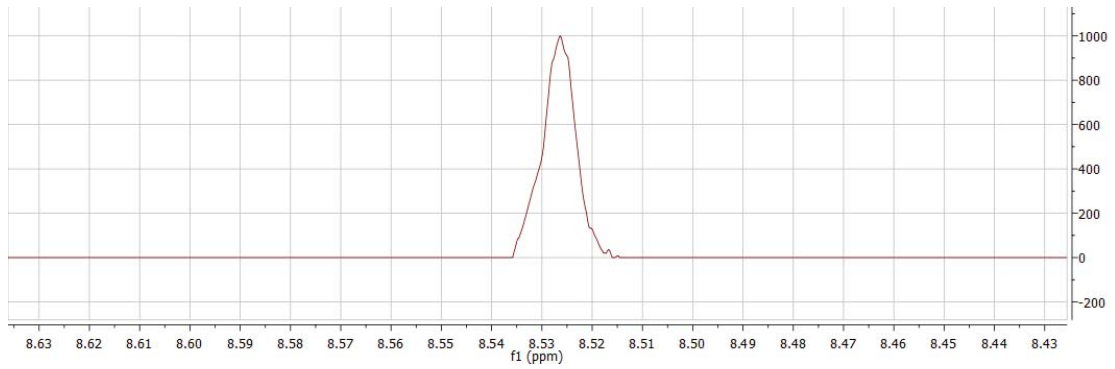


Component 61: Formic acidComponent 62: UracilComponent 63: NoiseComponent 64: 3.75 ppm (*t*, $J=5.7$ Hz)

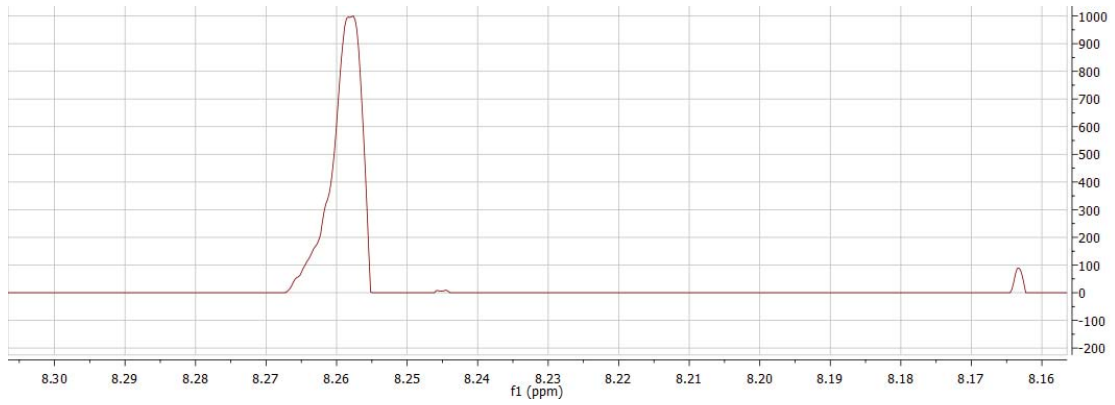
Component 65: L-Asparagine



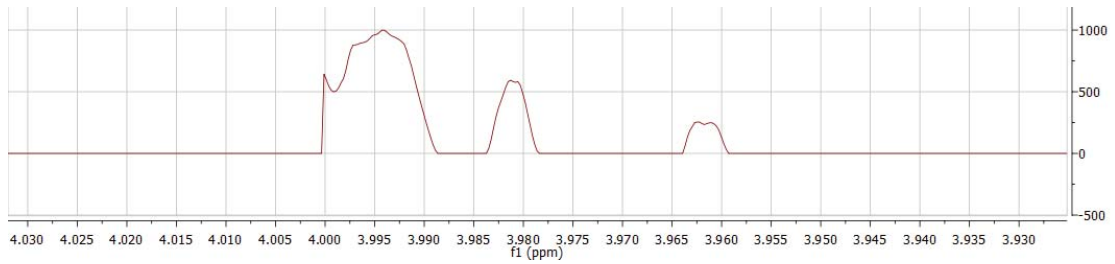
Component 66: 8.53 ppm (s)



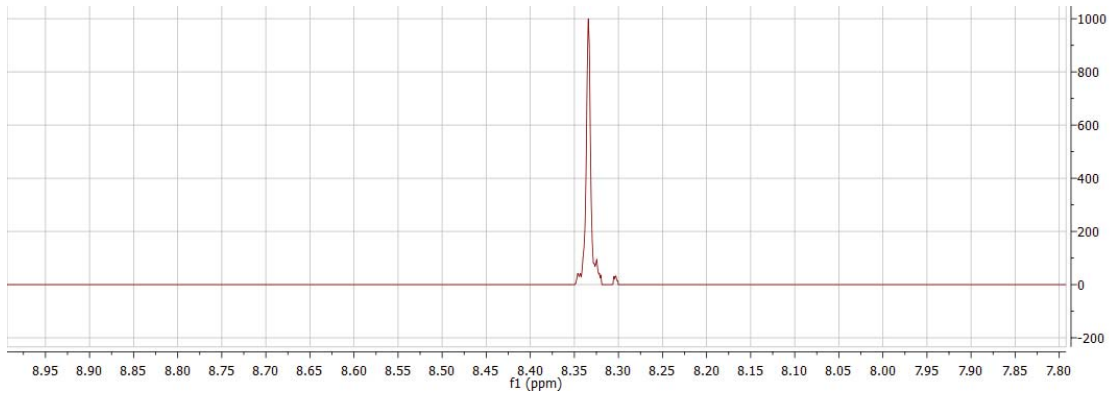
Component 67: 8.26 ppm (s)



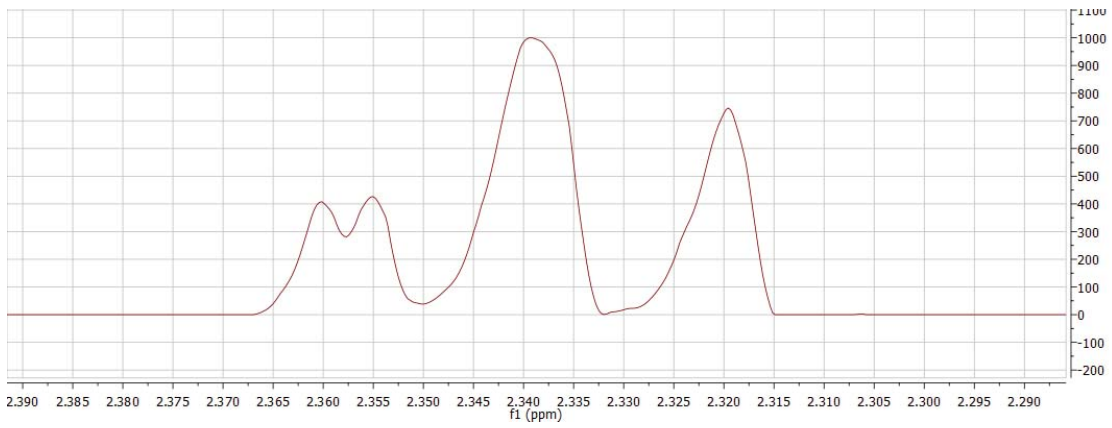
Component 68: Noise



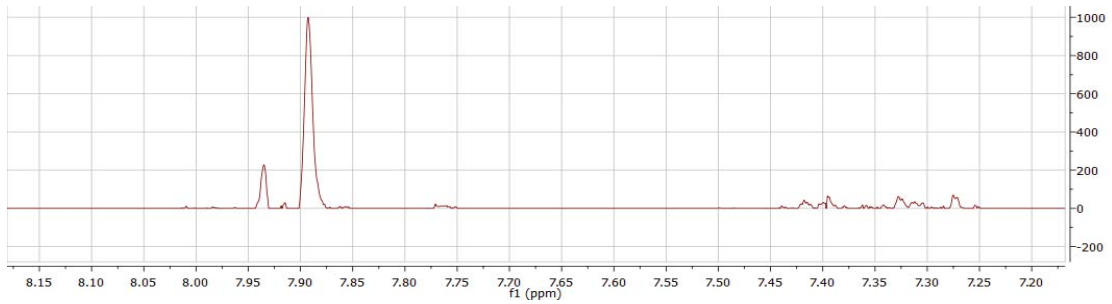
Component 69: 8.33 ppm (s)



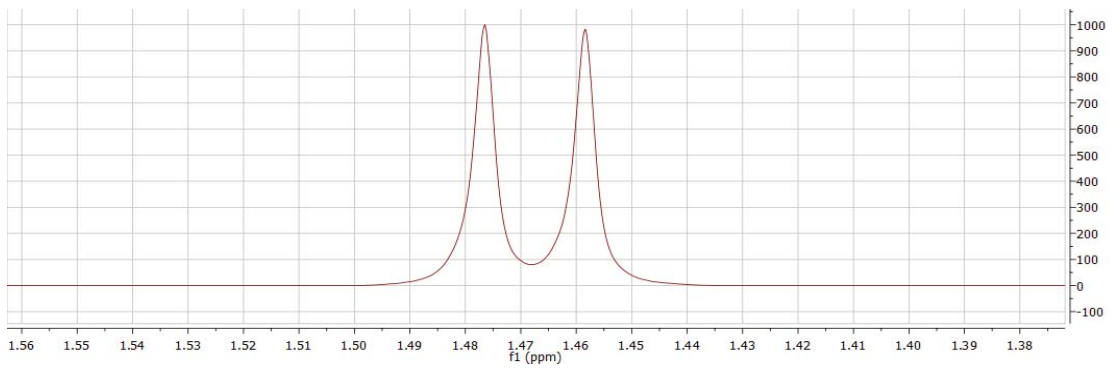
Component 70: L-Glutamic acid



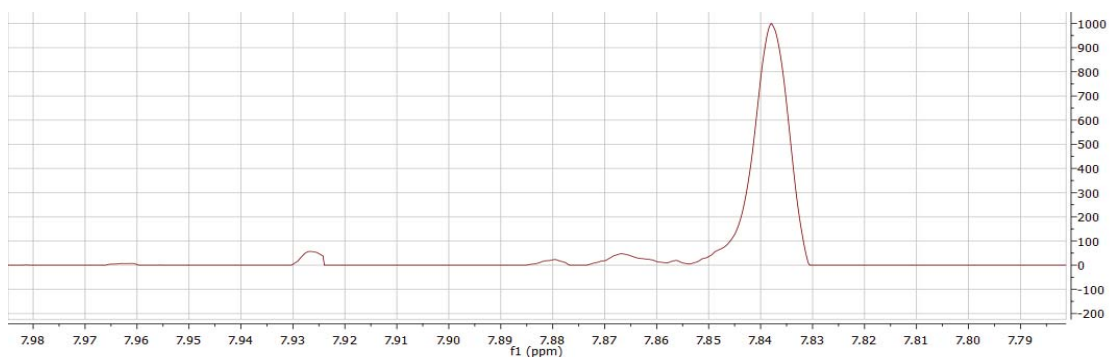
Component 71: 7.89 ppm (s)



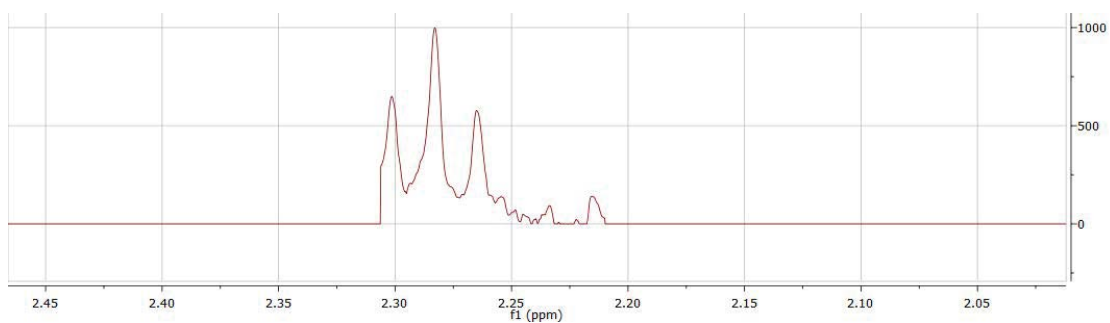
Component 72: L-Alanine



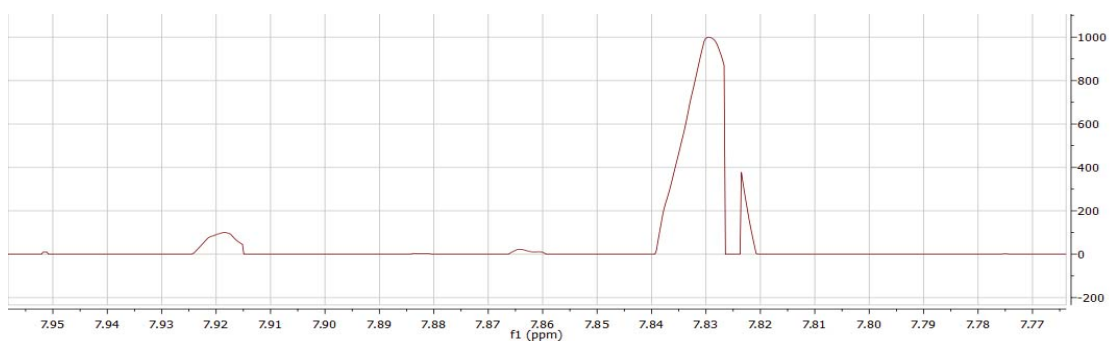
Component 73: 7.84 ppm (s)

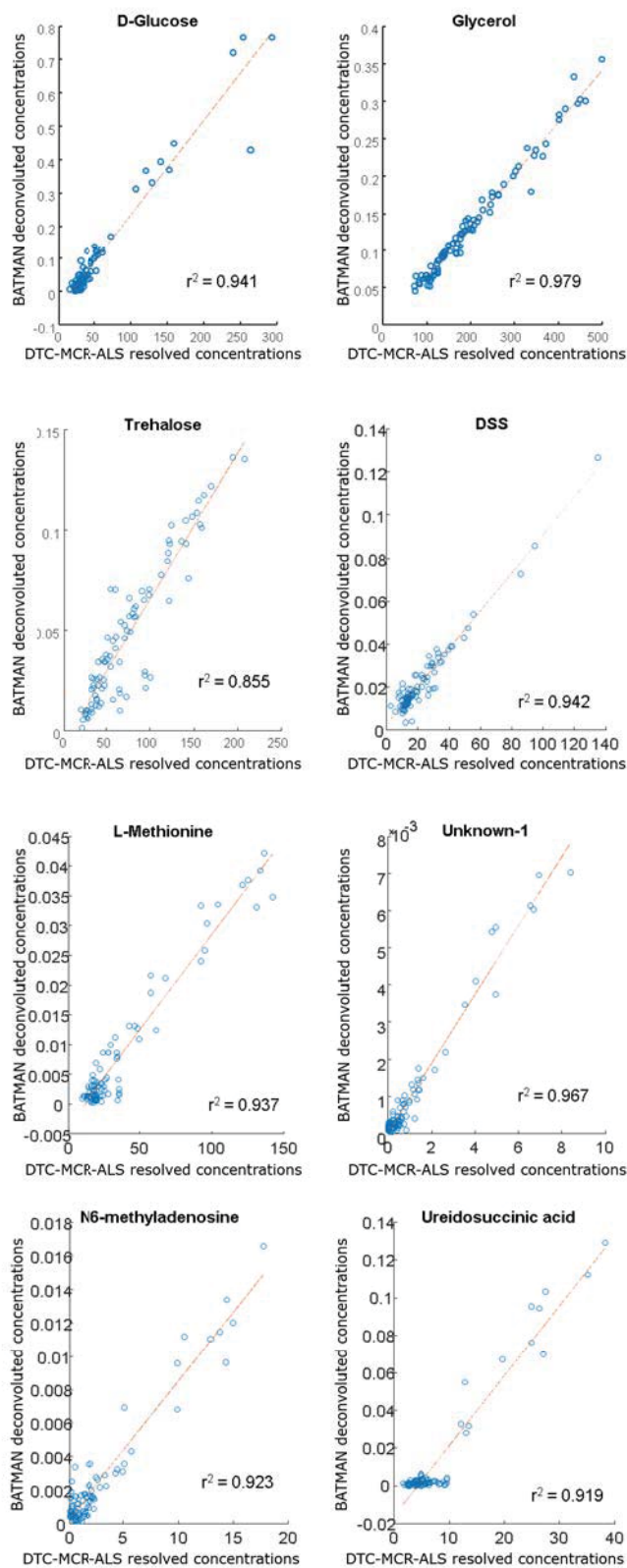


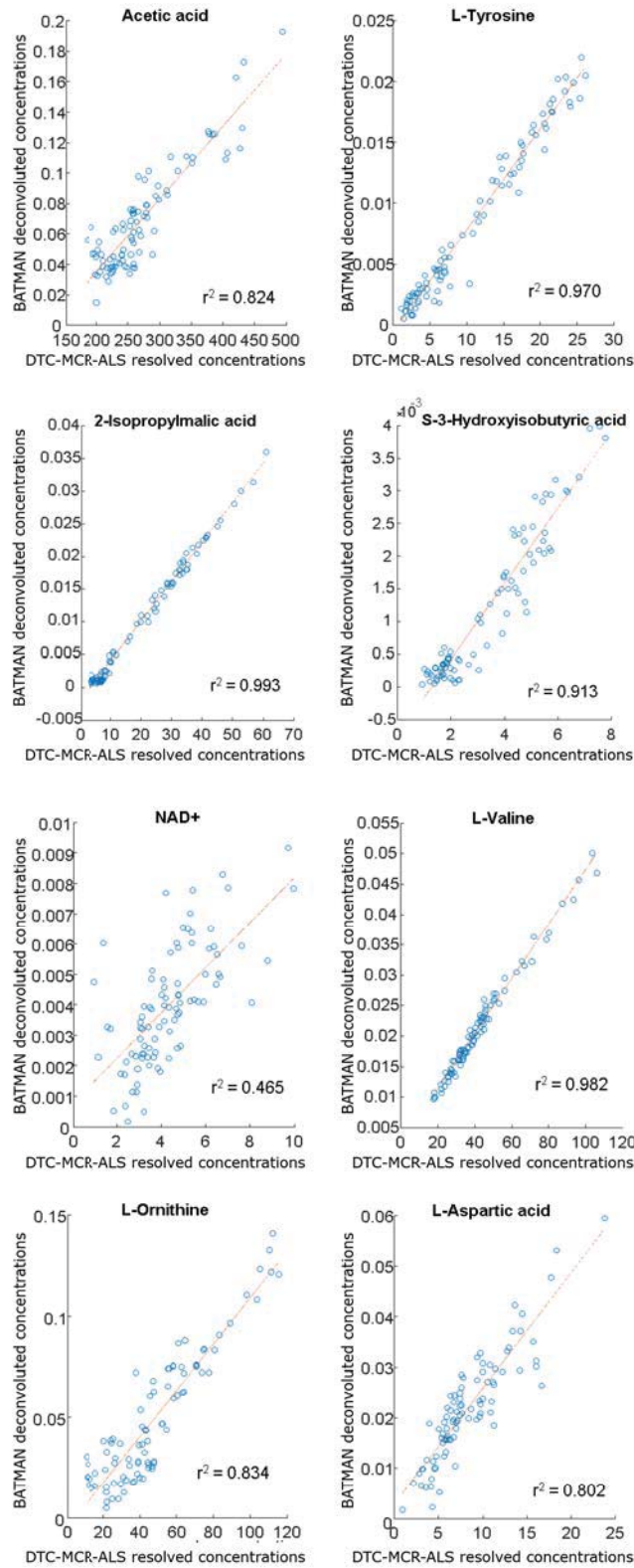
Component 74: 2.28 ppm (t, J=7.3 Hz)

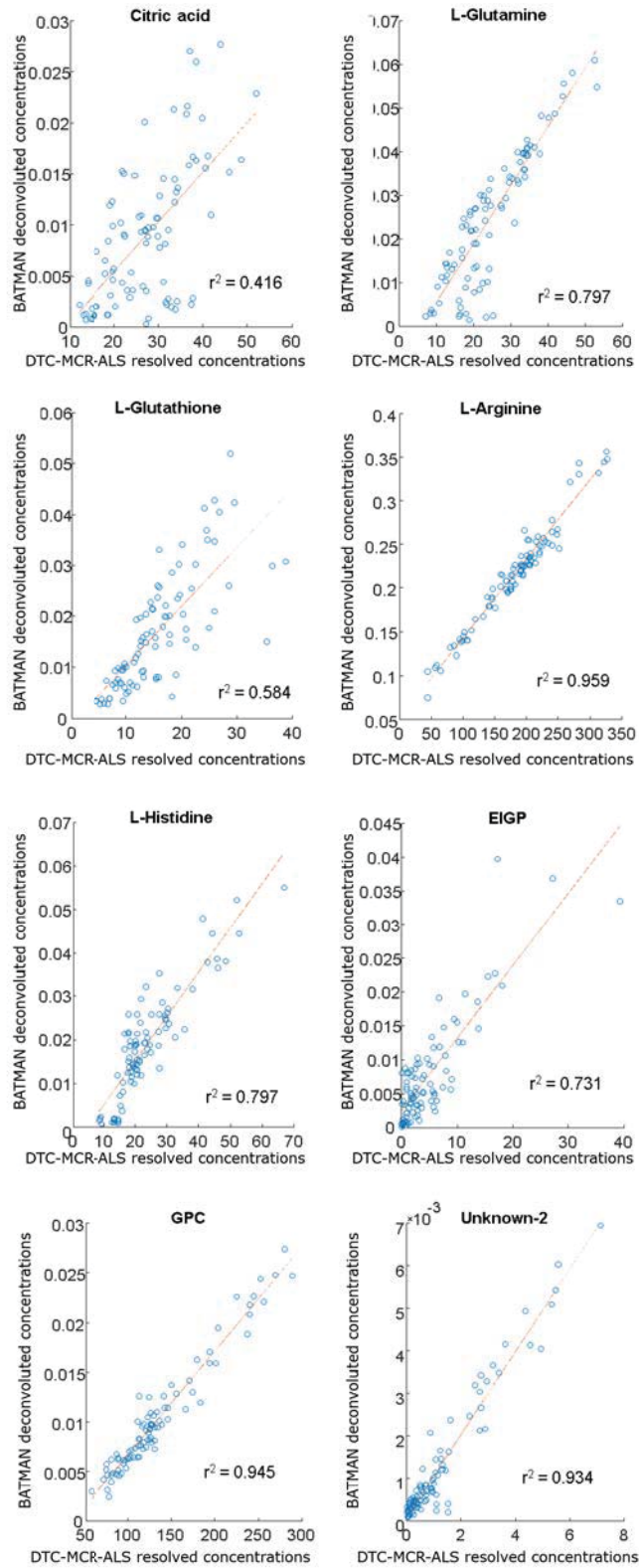


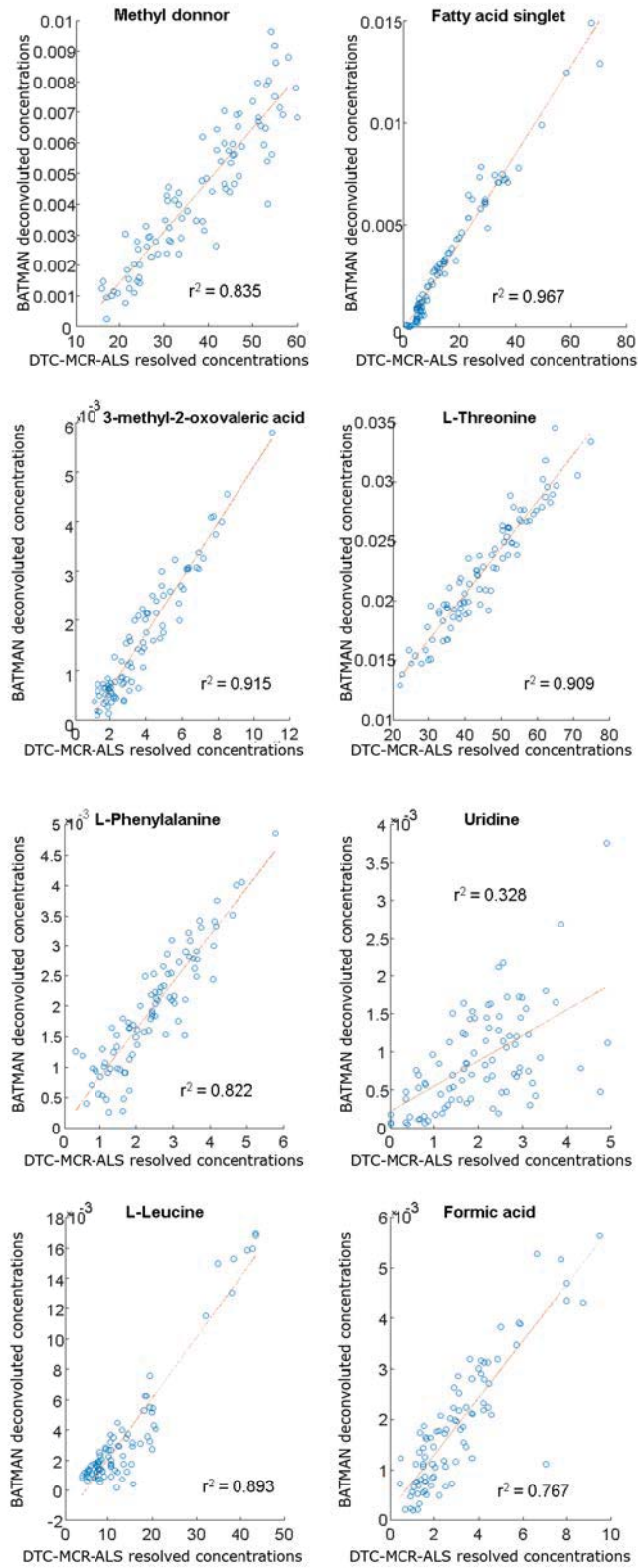
Component 75: 7.83 ppm (s)

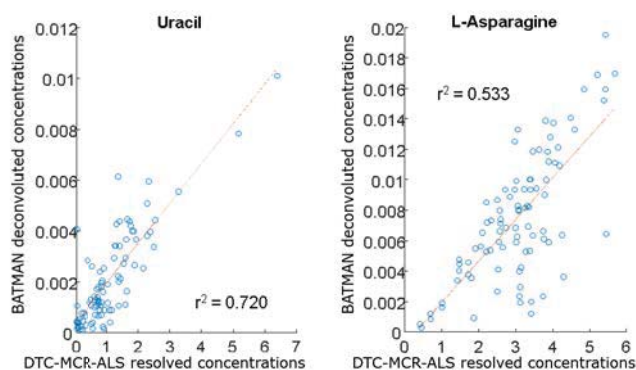


Appendix 5. Linear regressions between metabolite concentrations obtained with BATMAN methodology and with DTC-MCR-ALS methodology.









REFERENCES

1. Puig-Castellví F, Alfonso I, Piña B, Tauler R (2015) A quantitative ^1H NMR approach for evaluating the metabolic response of *Saccharomyces cerevisiae* to mild heat stress. *Metabolomics* 11: 1612-1625.
2. Dieterle F, Ross A, Schlotterbeck G, Senn H (2006) Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in ^1H NMR Metabonomics. *Anal Chem* 78: 4281-4290.
3. Jewison T, Knox C, Neveu V, Djoumbou Y, Guo AC, *et al.* (2012) YMDB: the Yeast Metabolome Database. *Nucleic Acids Res* 40: D815-D820.
4. Hao J, Astle W, De Iorio M, Ebbels TMD (2012) BATMAN—an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics* 28: 2088-2090.
5. Hao J, Liebeke M, Astle W, De Iorio M, Bundy JG, *et al.* (2014) Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN. *Nat Protoc* 9: 1416-1427.
6. Windig W, Stephenson DA (1992) Self-modeling mixture analysis of second-derivative near-infrared spectral data using the SIMPLISMA approach. *Analytical Chemistry* 64: 2735-2742.

2.2 SCIENTIFIC ARTICLE V

Compression of multidimensional NMR spectra allows a faster and more accurate analysis of complex samples.

Authors: Puig-Castellví F., Pérez Y., Piña B., Tauler R., Alfonso I.

Citation reference: *Chem. Comm.* (2018), 54:3090-3093.

DOI: 10.1039/C7CC09891J

ChemComm

Chemical Communications

rsc.li/chemcomm



ISSN 1359-7345



COMMUNICATION

Ignacio Alfonso *et al.*

Compression of multidimensional NMR spectra allows a faster and more accurate analysis of complex samples



Cite this: *Chem. Commun.*, 2018, 54, 3090

Received 27th December 2017,
Accepted 29th January 2018

DOI: 10.1039/c7cc09891j

rsc.li/chemcomm

Compression of multidimensional NMR spectra allows a faster and more accurate analysis of complex samples†

Francesc Puig-Castellví,^{ib}^a Yolanda Pérez,^{ib}^b Benjamín Piña,^{id}^a Romà Tauler^{id}^a and Ignacio Alfonso^{ib}^{*c}

We propose an approach to efficiently compress and denoise multidimensional NMR spectral data, improving their corresponding storage, handling, and analysis. This method has been tested with 2D homonuclear, 2D and 3D heteronuclear, and 2D phase-sensitive NMR spectral data and shown to be especially powerful for 2D NMR metabolomics studies.

The processing, storage, and handling of a large amount of data are becoming challenging tasks in different areas of science, such as chemistry,¹ biology,² medicine,³ physics,⁴ environmental science,⁵ and metabolomics.⁶ In this regard, the characterization of the concentration levels of metabolites in biological samples is usually done by Mass Spectrometry (MS)⁷ and Nuclear Magnetic Resonance (NMR) spectroscopy. Consequently, handling metabolomics data has become a critical aspect, though hampered by the large amount of information to be stored and efficiently transferred. Several options have been proposed to improve data storage of metabolomics datasets, such as the databases HMDB⁸ and MetaboLights,⁹ and also to facilitate data sharing by the use of metabolomics-oriented vendor-independent formats.⁶

High-resolution multidimensional NMR spectra are relatively large (10–500 MB) and contain a considerable amount of noisy data (> 90.0–99.8%), making its compression strongly recommendable. Some 2D NMR data compression approaches already exist (*i.e.* ROI,¹⁰ wavelet¹¹), but they retain the meaningless noise. Other approaches replace noise with zero values, implying that the number of variables is (inefficiently)

conserved,¹² consequently producing unnecessarily very sparse data sets.

¹H–¹³C HSQC is among the most common type of 2D spectra. However, even for the more complex mixtures, only 0.01–2% of the data points in ¹H–¹³C HSQC is linked to real resonances (Fig. S1, ESI†). All this noise contribution in the ¹H–¹³C HSQC spectra is especially troublesome for metabolomics purposes. First, in all organisms, metabolites are present in a wide range of concentrations, some of their respective resonances being close to the background noise. Therefore, some signals could be hidden behind the noise. Second, the contribution of noise in these spectra may mask the detection of the biological patterns when chemometric methods, such as Principal Component Analysis (PCA),¹³ a common tool in NMR metabolomics, are used. PCA produces a linear combination of the original spectral variables into a reduced number of new orthogonal variables or components, which contain most of the original data variance. Since experimental noise is often high, some of the new components may have embedded a significant noise contribution, distorting the metabolomics analysis and possibly leading to misinterpretation of the results.

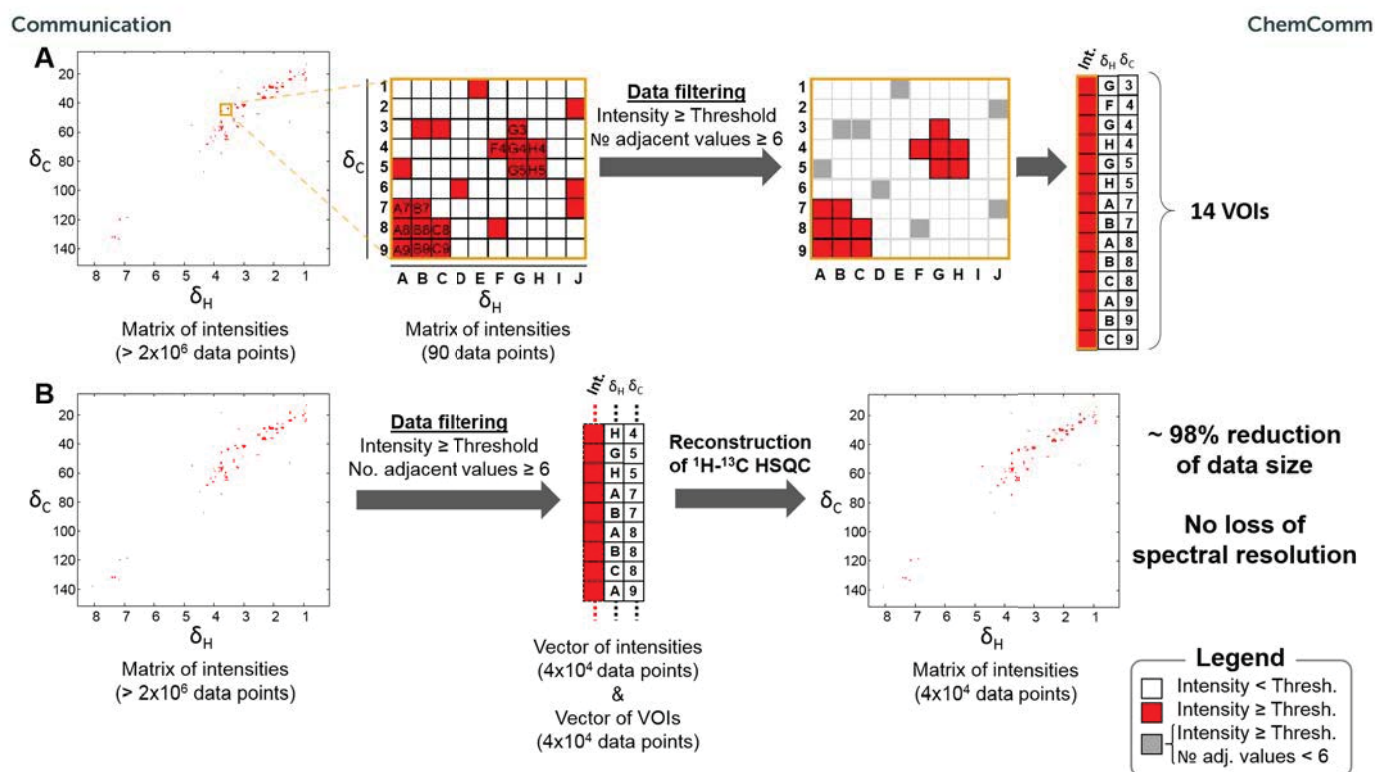
To overcome this difficulty, we propose in this work the VOI (Variable-Of-Interest) strategy. VOI detects those variables with chemical meaning and discards those that only describe noise by applying two criteria. First, variables having signals below a fixed threshold are directly discarded. Second, from all the remaining variables, only the ones giving peak shapes are kept. To give a peak shape signal, a minimal number of connected (clustered) variables of interest (or *minvoi*) must be found. For a 2D NMR spectrum, the selected data are stored as a 3-row data matrix: one of the rows contains all the VOIs, and the other two contain the associated chemical shifts in the two dimensions, *f*₁ and *f*₂. Since these *f*₁ and *f*₂ ppm values associated with every selected intensity data point are maintained, the 2D NMR spectrum can be easily reconstructed from this VOI matrix. Scheme 1 summarizes how the VOI strategy would work in a ¹H–¹³C HSQC NMR example. Due to the apparent simplicity of the VOI approach, it can be extended to 3D NMR spectra

^a Department of Environmental Chemistry, Institute of Environmental Assessment and Water Research (IDAEA-CSIC), Jordi Girona 18-26, 08034 Barcelona, Spain

^b NMR Facility, Institute of Advanced Chemistry of Catalonia (IQAC-CSIC), Jordi Girona 18-26, 08034 Barcelona, Spain

^c Department of Biological Chemistry and Molecular Modelling, Institute of Advanced Chemistry of Catalonia (IQAC-CSIC), Jordi Girona 18-26, 08034 Barcelona, Spain. E-mail: ignacio.alfonso@iqac.csic.es

† Electronic supplementary information (ESI) available: Methods and additional figures. See DOI: 10.1039/c7cc09891j



Scheme 1 Summary of the steps of the proposed VOI strategy. (A) VOI of a highlighted region of a ¹H-¹³C HSQC NMR spectrum. (B) VOI approach of the whole ¹H-¹³C HSQC spectrum and its reconstruction from the VOI matrix.

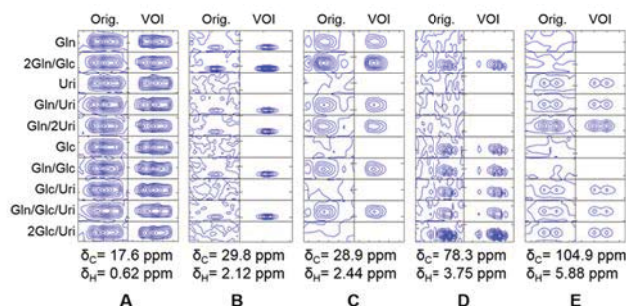


Fig. 1 Highlighted 2D NMR regions before and after application of the VOI-compressing algorithm. All samples contain 0.2 mM DSS. Selected peaks are from DSS (A), L-glutamine (B and C), D-glucose (D) and uridine (E).

(as shown in the ESI[†]) or even with higher dimensionality. The VOI-filtering approach has been validated in the analysis of different examples of single 2D NMR spectra, including homonuclear, heteronuclear, and phase-sensitive NMR experiments, and also of an example of a single 3D NMR HNC0 spectrum of a natively unfolded protein region (ESI[†]). For all the tested cases, after optimization of the *minvoi* and threshold parameters, all resonances were well conserved in the processed

spectra, and the number of variables was reduced to a number of values between 2.2% and 14.7% (see Section S7 for parameters' optimization and Sections S14–S16, ESI[†] for the results) of the total number of measured values.

The VOI approach can also be applied for the simultaneous analysis of multiple 2D NMR spectra, as long as a third criterion is applied to cope with the fact that some key resonances may not be present in some samples. In these cases, the VOIs selected in at least one sample are always retained (regardless they do not pass the selection criteria in any other sample). The resulting dataset will contain as many rows as VOIs in all analysed samples, plus the two rows containing the associated chemical shifts in *f1* and *f2*.

A set of ten ¹H-¹³C HSQC NMR spectra from binary, ternary and quaternary mixtures of 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS) containing uridine, D-glucose, and L-glutamine at different concentrations (see Table S1 in the ESI[†] for details) was used for testing the implementation and validation of the VOI proposed procedure.

When the VOI approach was applied to this dataset, the number of selected VOIs was between 324 (0.015% of the total number of data-points) and 844 (0.040%), depending on the

Table 1 VOI data compression results in the analysis of the investigated experimental datasets

Dataset	Number spectra	¹ H- ¹³ C HSQC dimensions	Total data points	<i>minvoi</i>	Threshold	VOIs	Data reduction ^a	PCA running time ^b
Synthetic mixture	10	1024 × 2048	2 097 152	20	9000	1352	146 MB → 74 KB	4.152 s → 0.006 s
Yeast extracts	32	1024 × 2048	2 097 152	24	6000	30 904	467 MB → 5.35 MB	17.729 s → 0.300 s

^a Original data intensities were stored in the ASCII.txt format. VOI results were stored also in the ASCII.txt format. ^b Intel workstation with 2.40 GHz, 128 GB RAM, and 6 cores; for more details see the results in the ESI.

considered sample. The number of clusters (or peaks) ranged between 9 and 21, and the median of the number of data-values per cluster was 32. When the ten VOI-compressed spectra were fused into one single dataset, the total number of VOIs increased up to 1352 (0.064%). Thus, with more analyzed samples, more meaningful VOIs representative of sample composition in the dataset will be detected. This increase can be explained because not all of the peaks were exactly defined in all the samples by the same VOIs. In terms of computer storage requirements, as much as 1970-fold of file size reduction was achieved, and the analysis by PCA was 692 times faster.

Since VOI datasets stored both δ_C and δ_H , VOI-processed spectra can be easily transformed to their original dimensions of 1024×2048 values. To proceed with the data conversion, all the previously discarded data-points were filled with zeros.

Some selected regions from the ten ^1H - ^{13}C HSQC NMR are highlighted in Fig. 1. In the odd columns, the regions from the original spectra are shown. In the even columns, the regions from the reconstructed VOI-processed spectra are presented. The figure clearly illustrates how VOI processing allows a much better visual identification of the peaks, as they appear now free from the surrounding noise. Since reconstructed VOI-processed spectra were filtered from noise, peak integration becomes much easier by directly summing the intensity values from all data-values that define a given peak. In our dataset, peak integrals from VOI-processed data resulted between 3% and 29% lower than the equivalent peak integrals of the non-processed spectra. These dissimilarities reflect the large amount of noise in the original (non-VOI) experimental spectra. The tiniest peaks showed larger differences, since noise contributions were more significant on them.

PCA^{13b} was applied to the experimental datasets to estimate the number of independent variance sources. When PCA was applied to a single VOI-processed 2D spectrum, the number of variance sources coincided with the number of resonances. In contrast, if PCA was applied to the original raw ^1H - ^{13}C HSQC spectrum, the number of variance sources grew up to *ca.* 500 (Fig. S3, ESI[†]). Thus, all these extra variable sources found in the original 2D spectra were describing mostly noise. Thus, PCA directly applied to the raw original data is affected by the noise, while PCA on the VOI-processed data will be much less affected by noise and gives more reliable results.

In order to confirm the efficiency of the VOI method, this methodology was additionally validated using a more realistic dataset, consisting of 32 ^1H - ^{13}C HSQC spectra from different metabolic yeast extracts. To make the variable selection more challenging, the threshold level was reduced from 9000 to 6000 (Table 1), which is considered to be within the experimental noise level (Fig. S4, ESI[†]). To minimize the number of false-positive VOI values, *minvoi* was increased up to 24 (Table 1). As a result, only 30 904 (1.47%) VOIs were finally selected from 174 clusters with a median of 44 data-points per cluster. VOI compression reduced the file size 88 times, and PCA analysis resulted 59 times faster (Table 1).

Interestingly, when the same dataset was compressed using the same threshold, 35.4% of the variables were

retained (Fig. S4C, ESI[†]). If the *minvoi* filter is not used, single high noise and signal artefacts can be selected when an increasing number of analysed spectra are considered.

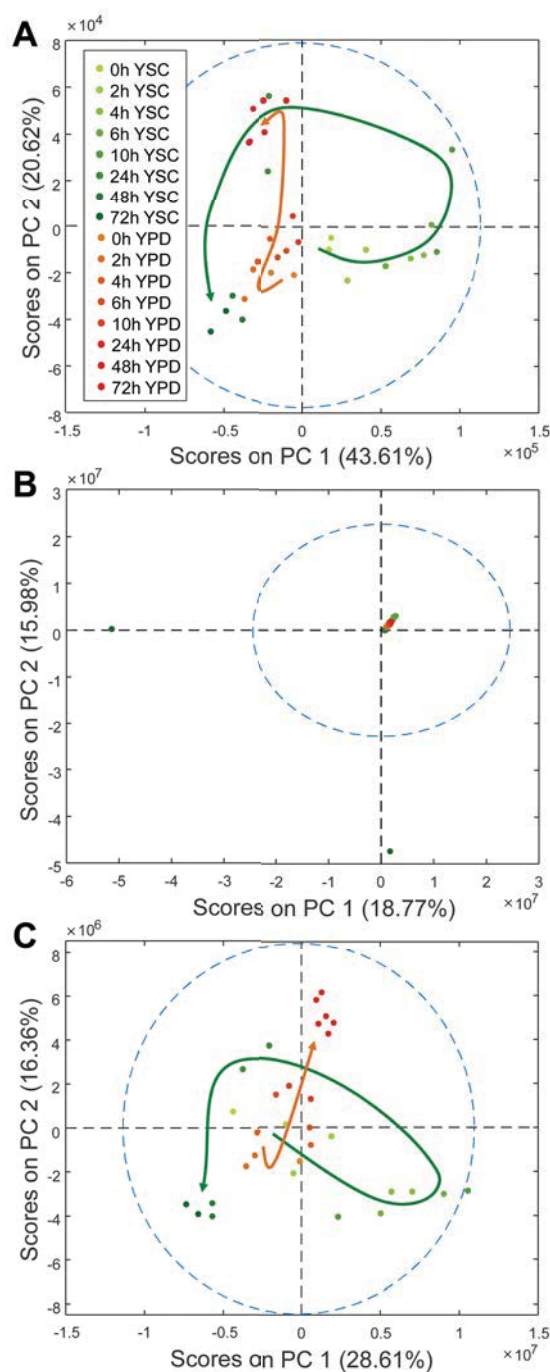


Fig. 2 PCA score plots of yeast extracts in the analysis of (A) ^1H NMR data, (B) ^1H - ^{13}C HSQC NMR data or (C) VOI-processed ^1H - ^{13}C HSQC data. Yeast (represented by red and green dot symbols) were cultured in two different growth media, Yeast Synthetic Complete (YSC, green samples) and Yeast Peptone Dextrose (YPD, red samples) media for 72 hours. The intensity of the dot symbol colours indicates time evolution as given in the panel of top (A). Arrows show the evolution of the two groups of samples over time. The difference in score distribution between B with A and C is due to the presence of noise in B. For more details on the culturing conditions protocol, see Section S2 in the ESI[†].

Moreover, if the signal thresholds were increased to compensate this effect, some of the VOIs from real resonances were lost. Thus, we concluded that the application of the two filtering criteria is required to ensure a significant data compression with proper variable selection.

Finally, we compared the scores obtained from PCA decomposition (see the Methods section S9, ESI[†]) of the original ¹H-¹³C HSQCs and of the VOI-processed ¹H-¹³C HSQC NMR spectra. Application of PCA to the latter (Fig. 2C) or to the ¹H NMR dataset (Fig. 2A) resulted in a similar sample distribution in the score plots, and therefore, the two data types can be used indistinctively for sample monitoring. However, when PCA was applied to the original ¹H-¹³C HSQC metabolomics raw dataset, the distribution pattern was different, and some of the samples were considered as outliers (Fig. 2B), highlighting again the need for removing noise before carrying out a multivariate analysis. In addition, since ¹H-¹³C HSQC NMR data contain a lower number of overlapping peaks than ¹H NMR data, the identification of the metabolites responsible for the observed variations in the PCA analysis resulted to be much simpler for the (filtered) 2D NMR dataset.

In summary, we demonstrate that noise in multidimensional NMR spectra can be selectively removed without losing spectral resolution using the VOI approach presented here. After noise removal with the proposed VOI strategy, a single multidimensional NMR spectrum dataset can be compressed up to 2000 times, making its analysis substantially faster. Moreover, peak integration also becomes a straightforward process with the achieved noise filtering properties of the proposed VOI method. The proposed VOI approach strongly facilitates the handling of multidimensional NMR data and it can be used with other types of spectroscopic datasets, such as those obtained in 2D Raman spectroscopy. Additionally, it has the potential for speeding-up the acquisition of high-throughput multidimensional NMR when combined with Non-Uniform Sampling (NUS).¹⁴

Research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement No. 320737.

Conflicts of interest

There are no conflicts to declare.

References

- 1 I. V. Tetko, O. Engkvist, U. Koch, J.-L. Reymond and H. Chen, *Mol. Inf.*, 2016, **35**, 615–621.
- 2 V. Marx, *Nature*, 2013, **498**, 255.
- 3 C. H. Lee and H.-J. Yoon, *Kidney Res. Clin. Pract.*, 2017, **36**, 3–11.
- 4 A. Klimentov, M. Grigorieva, A. Kiryanov and A. Zarochentsev, *J. Instrum.*, 2017, **12**, C06044.
- 5 Y. Liu, M. Qiu, C. Liu and Z. Guo, *Pers. Ubiquit. Comput.*, 2017, **21**, 55–65.
- 6 P. Rocca-Serra, R. M. Salek, M. Arita, E. Correa, S. Dayalan, A. Gonzalez-Beltran, T. Ebbels, R. Goodacre, J. Hastings, K. Haug, A. Koulman, M. Nikolski, M. Oresic, S. A. Sansone, D. Schober, J. Smith, C. Steinbeck, M. R. Viant and S. Neumann, *Metabolomics*, 2016, **12**, 14.
- 7 C. H. Johnson and F. J. Gonzalez, *J. Cell. Physiol.*, 2012, **227**, 2975–2981.
- 8 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. Bjorn Dahl, R. Perez-Pineiro, R. Eisner, F. Allen, V. Neveu, R. Greiner and A. Scalbert, *Nucleic Acids Res.*, 2013, **41**, D801–D807.
- 9 N. S. Kale, K. Haug, P. Conesa, K. Jayseelan, P. Moreno, P. Rocca-Serra, V. C. Nainala, R. A. Spicer, M. Williams, X. Li, R. M. Salek, J. L. Griffin and C. Steinbeck, *Curr. Protoc. Bioinformatics*, 2016, **53**, 14.13.11.
- 10 I. A. Lewis, S. C. Schommer and J. L. Markley, *Magn. Reson. Chem.*, 2009, **47**, S123–S126.
- 11 J. C. Cobas, P. G. Tahoces, M. Martin-Pastor, M. Penedo and F. J. Sardina, *J. Magn. Reson.*, 2004, **168**, 288–295.
- 12 Z. Zolnai, S. Macura and J. L. Markley, *J. Magn. Reson.*, 1988, **80**, 60–70.
- 13 (a) R. Bro and A. K. Smilde, *Anal. Methods*, 2014, **6**, 2812–2831; (b) I. Jolliffe, *Wiley StatsRef: Statistics Reference Online*, John Wiley & Sons, Ltd, 2014.
- 14 M. Mobli and J. C. Hoch, *Prog. Nucl. Magn. Reson. Spectrosc.*, 2014, **83**, 21–41.

SUPPLEMENTARY MATERIAL FOR SCIENTIFIC ARTICLE V

Compression of multidimensional NMR spectra allows a faster and more accurate analysis of complex samples.

Authors: Puig-Castellví F., Pérez Y., Piña B., Tauler R., Alfonso I.

Citation reference: *Chem. Comm.* (2018), 54:3090-3093.

DOI: 10.1039/C7CC09891J

Supplementary Experimental Procedures

1. Preparation of NMR samples of synthetic mixtures

NMR samples were prepared in 600 μ L of 0.25 mM sodium phosphate deuterated buffer (pH 7.0) with 200 mM DSS. Mixtures were prepared following the concentrations indicated in the table below:

Table S1. Metabolite concentrations used.

Sample name	DSS (in mM)	D-glucose (in mM)	Uridine (in mM)	L-glutamine (in mM)
Gln	0.25	0	0	1.6
2Gln/Glc	0.25	0.35	0	3.2
Uri	0.25	0	0.24	0
Gln/Uri	0.25	0	0.24	1.6
Gln/2Uri	0.25	0	0.48	1.6
Glc	0.25	0.35	0	0
Gln/Glc	0.25	0.35	0	1.6
Glc/Uri	0.25	0.35	0.24	0
Gln/Glc/Uri	0.25	0.35	0.24	1.6
2Glc/Uri	0.25	0.7	0.24	0

2. Preparation of NMR samples of yeast extracts

Yeast Growth. *S. cerevisiae* S288C cells were pre-cultured in YPD (1 % yeast extract, 1 % peptone, 2 % glucose) medium on an orbital shaker (150 rpm) at 30 °C overnight. All following cultures were cultured with these shaking and temperature conditions. 2 L of YNB Synthetic Complete medium (YSC, 1.7 g/L Yeast Nitrogen Base without amino acids and sulfate (Difco), 5 g/L (NH₄)₂SO₄) were inoculated with 200 μ L of the pre-culture sample and left at the same temperature and shaking conditions until the culture reached an absorbance at 600 nm (A_{600}) of approximately 0.8 - 1. Pellets from these resulting cultures were collected by centrifuging the cultures, but not washed, at 2000 rpm for 3 min and 4 °C. Pellets were used right after for inoculating Erlenmeyer's containing either YSC medium or YPD medium.

Sample collection. 100 ml aliquots of every culture were collected six times during three days (0h, 2h, 4h, 6h, 10h, 24h, 48h and 72h). Samples were arrested with a cold shock in ice and cell were harvested by centrifugation at 4000 g for 3 min, discarding the supernatant. Cells were washed twice in 100 mM Na₂HPO₄ pH 7.0 followed by a centrifugation at 4700 g for 3 min. Resulting pellets were stored at -80 °C and lyophilized.

Metabolite extraction. Metabolites were extracted by following the protocol published in a previous work. 1800 μL of a solution of methanol-chloroform 1:2 (4 °C) were added to the pellet, followed by a vigorous vortexing. A cold shock was then applied to the pellets for 5 times using the following procedure: the pellets were submerged in liquid nitrogen for 1 minute and consequently thawing in ice for 2 minutes. 400 μL of water were added to create the biphasic system. After homogenization by vortexing, a 3 min centrifugation at 16,500 rpm and 4 °C was carried out. The aqueous phase (upper part) was collected. This process was repeated and samples were freeze-dried afterwards.

NMR sample preparation. Aqueous metabolites samples were dissolved in 650 μL of deuterated phosphate buffer (25 mM Na_2DPO_4 , pH 7.0) in D_2O with 0.2 mM DSS as internal standard. Samples were centrifuged at 9,168 g for 5 min and the supernatant was collected and introduced into 5 mm NMR tubes.

3. Preparation of NMR samples of complex molecules

Cyclosporin A (Sigma-Aldrich) sample was prepared by dissolving 12mg in 0.55 mL deuterated benzene. 0.4 mM ^{15}N labeled Ubiquitin sample (Sigma-Aldrich) was prepared by dissolving the protein in 50 mM phosphate buffer (pH 6.2, 90% $\text{H}_2\text{O}/\text{D}_2\text{O}$). Peptide AcCNPFDLEC (Genscript HK Limited) was dissolved in $\text{DMSO}-d_6$ (1.5 mg, 2.5 mM). Lyophilized recombinant double labeled ^{13}C - ^{15}N natively unfolded protein fragment was dissolved in 400 μL of 20 mM acetate buffer with 50 mM NaCl (pH 5, 90% $\text{H}_2\text{O}/\text{D}_2\text{O}$) and introduced in a 5 mm Shigemi NMR tube.

4. NMR acquisition parameters

All NMR spectra were acquired at 298 K on a 500 MHz AvanceIII HD NMR spectrometer equipped with a TCI cryoprobe from Bruker. All pulse sequences used are from Bruker TopSpn3.5pl6.

Synthetic mixture samples set. 1D NOESY spectra were recorded using the *noesygprr1d* pulse sequence and the following parameters: 256 scans, 4 seconds of relaxation delay, spectral width of 10 kHz and an acquired spectral size of 32k (final spectral size of 64k). The 90° pulse width (between 8.5 and 10.5 μs) and presaturation power for water suppression were measured for every sample before experiment acquisition. ^1H - ^{13}C HSQC NMR spectra were recorded using the *hsqcetgprrsisp2.2.be* pulse sequence and the following parameters: 12 scans, 3 seconds of relaxation delay, spectral width of 12.7 kHz and an acquired spectral size of 548 data points in f1 dimension and of 1,536 data points in f2 dimension. The spectra were phase and baseline corrected and referenced to the DSS reference peak. After zero-filling, final spectral size in the ^1H - ^{13}C HSQC NMR spectra were 1,024 data points in f1 (^{13}C) dimension and 2,048 data points in f2 (^1H) dimension.

Yeast extracts samples set. 1D NOESY spectra were recorded using the *noesygprr1d* pulse sequence and the following parameters: 256 scans, 4 seconds of relaxation delay, spectral width of 10 kHz and an acquired spectral size of 32k (final spectral size of 64k). The 90° pulse width

(between 8.5 and 10.5 μ s) and presaturation power for water suppression were measured for every sample before experiment acquisition. 2D ^1H - ^{13}C HSQC NMR spectra were recorded using the *hsqcetgpprsisp2.2.be* pulse sequence and the following parameters: 12 scans, 3 seconds of relaxation delay, spectral width of 20.7 kHz in f1 and 7.9 kHz in f2, and an acquired spectral size of 548 data points in f1 dimension and of 1,536 data points in f2 dimension. The spectra were phase and baseline corrected and referenced to the DSS reference peak. After zero-filling, final spectral size in the ^1H - ^{13}C HSQC NMR spectra were 1,024 data points in f1 (^{13}C) dimension and 2,048 data points in f2 (^1H) dimension.

Cyclosporin A sample. 2D ^1H - ^1H TOCSY NMR spectrum was recorded using the *mlevphpp* pulse sequence and the following parameters: 8 scans, 1 seconds of relaxation delay, pulse width of 7.5 μ s, spectral width of 6 kHz in both dimensions, and an acquired spectral size of 128 data points in f1 (^1H) dimension and of 1,024 data points in f2 (^1H) dimension, resulting in a final spectral size of 1,024 data points per dimension.

Ubiquitin sample. 2D ^1H - ^{15}N HSQC NMR spectrum was recorded using the *hsqcetf3gpsi* pulse sequence and the following parameters: 2 scans, 1 second of relaxation delay, pulse width of 8 μ s, spectral width of 8 kHz for ^1H channel and 1.7 kHz for ^{15}N channel, and an acquired spectral size of 64 data points in f1 dimension and of 1,024 data points in f2 dimension. After zero-filling, final spectral size in the ^1H - ^{15}N HSQC NMR spectra were 256 data points in f1 dimension and 2,048 data points in f2 dimension.

Peptide sample 2D ^1H - ^1H TOCSY NMR and 2D ^1H - ^1H ROESY NMR spectra were recorded using the *mlevtgp* and *roesyphpp.2* pulse sequences, respectively. For both pulse sequences, the following parameters were used: 8 scans, 2 seconds of relaxation delay, pulse width of 8.0 s, spectral width of 6.5 kHz in both dimensions, and an acquired spectral size of 256 data points in f1 dimension and of 1,024 data points in f2 dimension, resulting in a final spectral size of 1,024 data points per dimension.

Natively unfolded protein sample: 3D HNCO spectrum was acquired using *hncogp3d* pulse sequence and the following acquisition parameters: 8 scans, 1 second of relaxation delay, pulse width of 8 μ s, spectral width of 8.1/1.8/2.0 kHz for $^1\text{H}/^{15}\text{N}/^{13}\text{C}$ channels, and acquired spectral size of 2048/72/128 for $^1\text{H}/^{15}\text{N}/^{13}\text{C}$ dimensions.

5. VOI filtering function

The VOI filtering function (*voi2D.m*), and their equivalent for phase-sensitive 2D NMR spectra (*voi2Df.m*) and for 3D NMR spectra (*voi3D.m*) were implemented in Matlab programming language, and they can be downloaded from <https://github.com/f-puig/VOI>.

voi2D (and *voi2Df*) can filter a ^1H - ^{13}C HSQC NMR matrix of 1,024 x 2,048 data-points in less than 0.5 seconds (time measured in an Intel workstation with 2.40 GHz, 128 GB RAM and 6 cores). *voi3D* can filter a 3D HNCO NMR spectrum of 1024 x 256 x 256 data-points in less than 150 seconds (time measured in the same Intel workstation).

Four different outputs are generated from the application of *voi2D* on a 2D NMR spectrum:

1. *VOImatrix*: 3-row data matrix containing the vector of filtered intensities and the two vectors containing the two measured δ .
2. *filtered_NMR*: 2D NMR filtered data matrix of equal in size than the input 2D NMR matrix, but with zero values on those positions considered to be noise.
3. *indexes*: list of positions in the input 2D NMR matrix that contain the filtered intensities.
4. *peak_arrays*: lists of filtered positions for every cluster.

VOImatrix has the compressed 2D NMR matrix. *filtered_NMR* has the reconstructed 2D NMR spectrum and it is very convenient for representing contour plots. The *indexes* vector is used to create the matrix of 2 or more VOI-processed spectra (see section 6 below). Finally, *peak_arrays* is used to determine the number of clusters per spectrum and the number of points per cluster.

6. Combining 2 or more VOI-processed spectra.

To combine two or more VOI-processed spectra, we need first to ensure that the spectra have the same dimensions and that the ppm1 and ppm2 measured values are the same. Otherwise, the spectra will need to be interpolated first using one of the spectra of the dataset as a reference. VOI algorithm is first applied separately for every 2D NMR spectrum. Next, all the lists of VOI positions (*indexes*) are combined into one long matrix, excluding all repeated instances, with the list of common and uncommon VOIs.

To generate the matrix of VOI-processed spectra, an empty matrix with the same number of rows as samples and with the same number of columns as VOIs is created. Then, the first row is filled with the vector of selected intensities from the first spectra, and this process is repeated for the remaining spectra. Finally, all negative values are converted to 0.

7. Setting-up the *threshold* and *minvoi* parameters.

To define the *threshold* level, the most practical option is by checking the signal intensities of the 2D NMR spectrum using only one dimension (either ppm1 or ppm2). This means that, if the intensity values are represented on the first dimension (ppm1), we will have as many plotted lines as measured ppm2 values. Examples of the used 2D NMR datasets are given below. From this representation, it is easy to establish an intensity *threshold* value higher than the observed noise. The selection of the threshold for two 2D NMR spectra is shown in detail in **Fig1A** and **Fig1B**. This threshold value can be also proposed from NMR regions where no meaningful resonances (only noise) are present.

By fixing the threshold level to the maximum intensity value detected in these NMR regions where only noise is present, filtered variables would only be representative of peak resonances. To avoid losing signal information related to the base of the signal resonances close to noise, the threshold value has to be decreased. For instance, in **Fig1C**, the maximum noise intensity measured was around 12,000, and the chosen threshold was decreased to 6,000. Most of the noise values comprised between 6,000 and 12,000 were also filtered after application of the *minvoi* (minimum number of adjacent points that define a peak) parameter. For all datasets tested in this work, fixing the threshold level to the half of the maximum noise value gave satisfactory filtering results.

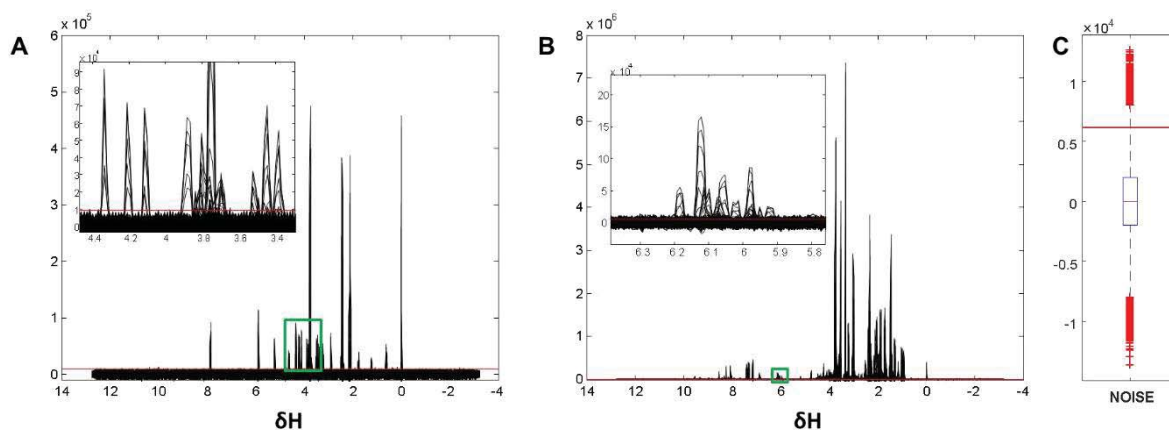


Figure S1. 2D NMR spectra plotted using only ppm1 (δ_H) dimension. A) Representative spectrum from dataset 1 (synthetic mixture). B) Representative spectrum from dataset 2 (yeast extract). C) Noise intensity values in dataset 2. Red horizontal line shows the applied threshold.

For phase-sensitive spectra, two thresholds were used: one for peaks in phase (positive peaks) and another one for peaks in antiphase (negative peaks). Since noise values did not depend of the phase (noise is always centered to zero, see **Fig S1C**), the threshold level for both phases, in absolute values, was the same. Therefore, the threshold level was estimated in the positive phase and changed the sign for the negative antiphase. To estimate the threshold level in the positive phase, the same procedure as for a phase-insensitive NMR spectrum described above was used.

To define the minimum number of adjacent points that define a NMR peak (*minvoi*), 2D NMR spectra were investigated using the typical NMR MestreNova (Mestrelab Research S.L.) or Topspin (Bruker BioSpin GmbH) platforms. First, the smallest true peak was selected. In these two popular NMR platforms, such operation is rather fast, efficient and easy to use. Afterwards, under the MATLAB environment, the selected smallest peak was visualized, and the number of variables (or pixels) that define this peak are counted. The obtained value is the maximum recommended value for the *minvoi* parameter, which gives a satisfactory NMR signal filtering by adjusting this *minvoi* parameter by a factor between 0.7 and 1.

8. NMR preprocessing

NMR spectra have been automatically referenced, phased and baseline corrected using TopSpin (Bruker, Germany) routines.

NMR preprocessing of ^1H NMR datasets. ^1H NMR Bruker files were imported to MestreNova v.11.0 (Mestrelab Research), and an exponential apodization of 0.2 Hz was applied on each one of them. In MestreNova v.11.0 environment, spectra were converted into ASCII format and imported to Matlab R2016a (The Mathworks Inc. Natick, MA, USA). In Matlab, data was first normalized using Probabilistic Quotient Normalization (PQN) ^[1] using an in-house function, followed by a mean-centering using the PLS toolbox 8.2.0 (Eigenvector Research Inc., Wenatchee, WA, USA). Regions of water (4.41 - 5.16 ppm), methanol (3.30 - 3.37 ppm),

chloroform (7.64 -7.69 ppm) and DSS (< 0.7 ppm) were removed. Data points which chemical shifts were higher than 9.7 ppm were also removed.

NMR preprocessing of ^1H - ^{13}C HSQC NMR datasets. ^1H - ^{13}C HSQC Bruker files were directly imported to Matlab R2016a using BBIO Toolbox Matlab scripts kindly provided by Bruker BioSpin GmbH, producing one (ppm1 x ppm2) matrix per spectrum. Every data matrix was then unfolded into a vector, and all vectors were merged into one matrix with as many rows as samples, and as many columns as measured ppm values. Then, data matrix was mean-centered. Before Principal Component Analysis (PCA)^[2] of the yeast extract metabolomics dataset, the data were normalized using the sample factors from PQN^[1] of the 1D NOESY dataset, and the same proton regions that were removed in the equivalent 1D NOESY spectra were excluded for the analysis. PQN is a convenient tool to normalize NMR spectra from time-course experiments, in which the total amount of metabolites increases over time because the studied organism is growing during the course of the experiment^[3], i.e. to correct for sample size effects. Quotients used in PQN normalization are estimated from comparing the intensities relative to significant resonances of every spectrum to a reference spectrum^[1].

NMR preprocessing of VOI datasets. Regions of water ($\delta_{\text{H}} = 4.41 - 5.16$ ppm), methanol ($\delta_{\text{H}} = 3.30 - 3.37$ ppm), chloroform ($\delta_{\text{H}} = 7.64 - 7.69$ ppm) and DSS ($\delta_{\text{H}} < 0.7$ ppm) were removed. Resulting VOI datasets were mean-centered.

9. Principal Component Analysis

In this study, we have applied PCA to single ^1H - ^{13}C HSQC NMR spectra and to datasets containing several ^1H - ^{13}C HSQC NMR spectra. PCA performs an orthogonal decomposition of the analyzed spectral data sets in matrix form under the constraints of maximum variance and normalization. See refs^[2] for more details about the PCA method.

In the first scenario, $\mathbf{X}_{(\text{ppm1}, \text{ppm2})}$ has as many rows as ppm variables in the f1 dimension (ppm1), and as many columns as ppm variables in the f2 dimension (ppm2). On the other hand, in the second scenario, $\mathbf{X}_{(m, \text{ppm1} \times \text{ppm2})}$ has as many rows as investigated NMR spectra or samples, and as many columns as the total number of ppm variables in both f1 and f2 dimensions.

With PCA, data compression is performed by selecting only the principal components associated with the largest singular values^[2c, 4] which will give information about the systematic variation of the data and do not describe the experimental noise, which are usually not associated to the components with the largest singular values.

In PCA, to decide the number of principal components to be considered, the singular values associated to the investigated data matrix are plotted and their sizes compared (see **Fig S3** in the Supplementary Results section for examples). It is assumed that singular values related with the relevant information of the dataset are larger than those related with random noise whose magnitude decreases slowly. Thus, the number of the largest singular values indicates the possible number of systematic variance sources (see **Fig S3**).

A more detailed explanation of the PCA method can be found elsewhere^[2a, 2b].

In PCA decomposition, each new orthogonal variable explains a percentage of the variance of the initial dataset. Thus, with the analysis of the explained variance associated to every orthogonal variable, the complexity of the data can be investigated. For instance, when most of the variance (apart from noise) of a dataset containing 100 samples and 500 variables is explained by only two components, it means that most of the variance present in these 500 variables can be described by a linear combination of these two factors or components (frequently called principal components). In most of the cases, natural phenomena are driven by a limited number of physical independent sources of systematic variance apart from random experimental noise variance sources, which can be discarded.

In this study, PCA was applied directly under MATLAB R2016a environment.

Principal Component Analysis of a single ^1H - ^{13}C HSQC NMR spectrum. For all the tested cases in this study, \mathbf{X} is the original ^1H - ^{13}C HSQC NMR spectra, containing 1,024 ppm values in f1 dimension ($m=1,024$), and 2,048 ppm values in f2 dimension ($n=2,048$).

Principal Component Analysis of multiple ^1H - ^{13}C HSQC NMR datasets. PCA was applied directly to each mean-centered unfolded data set (see preprocessing of ^1H - ^{13}C HSQC NMR datasets and NMR preprocessing of VOI datasets).

Supplementary Results and Discussion

10. Example of a representative ^1H - ^{13}C HSQC NMR spectrum from a metabolic yeast extract

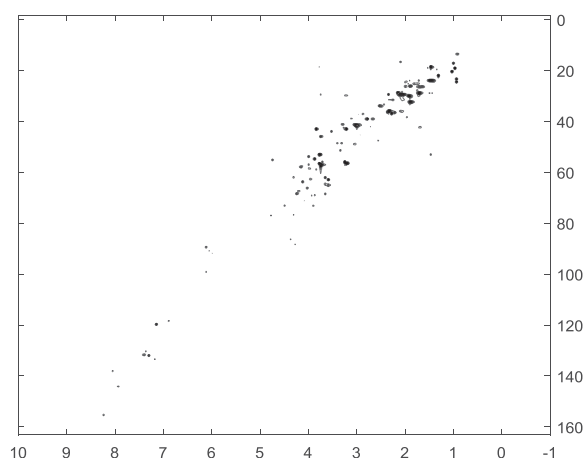


Figure S2. ^1H - ^{13}C HSQC of a yeast metabolic extract.

11. Principal Component Analysis of a single ^1H - ^{13}C HSQC NMR spectrum

When PCA is applied to a single 2D NMR spectral matrix, we should expect that the number of components with singular values^[2c, 4] different to zero (in absence of noise) will be close to the number of detected resonances, since the only differences in intensity between the different rows should come from these resonances.

However, due to the unavoidable presence of experimental noise, as seen in **FigS3** and **TableS2**, the number of components (different to zero) in every acquired HSQC was very high (close to 500, blue lines). When the VOI-compression noise filtering algorithm was applied to the same 2D NMR spectra, the number of components decreased approximately to the number of detected resonances (red lines in **FigS3**). Some of the differences between the number of detected resonances can be explained because some of the resonances appear in the same carbon chemical shift (i.e. short-range and long-range couplings of L-glutamine and some carbons from the glucose ring).

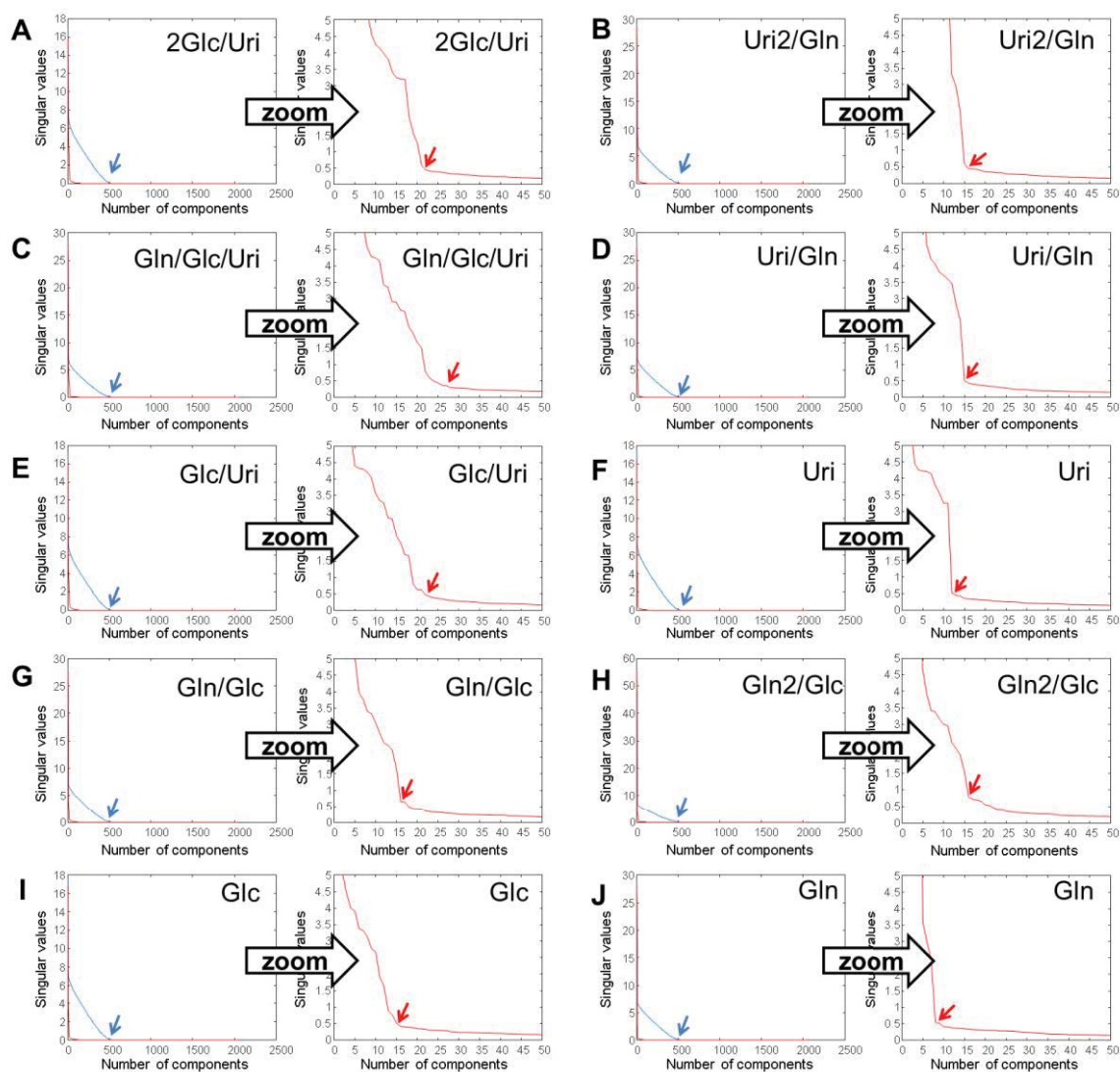


Figure S3. Plot of singular values associated to the original and VOI-processed ^1H - ^{13}C spectra. Blue lines and arrows give the singular values of the original 2D NMR spectra, while the red lines and arrows give the singular values of the VOI-processed data.

Table S2. Number of resonances and components found in the original and VOI-processed ^1H - ^{13}C spectra.

Sample name	Number of resonances*	Number of components (original spectra)	Number of components (VOI-processed spectra)
Gln	10	~500	8
2Gln/Glc	20	~500	15
Uri	12	~500	12
Gln/Uri	18	~500	16
Gln/2Uri	18	~500	16
Glc	14	~500	15
Gln/Glc	20	~500	16
Glc/Uri	22	~500	22
Gln/Glc/Uri	28	~500	28
2Glc/Uri	22	~500	22

*4 resonances were assigned to DSS, 6 to L-glutamine, 10 to D-glucose, and 8 to uridine.

12. Principal Component Analysis of multiple ^1H - ^{13}C HSQC NMR datasets

To reduce time and computational demands, only the first m ($m=10$ for dataset 1, $m=32$ for dataset 2) components were calculated.

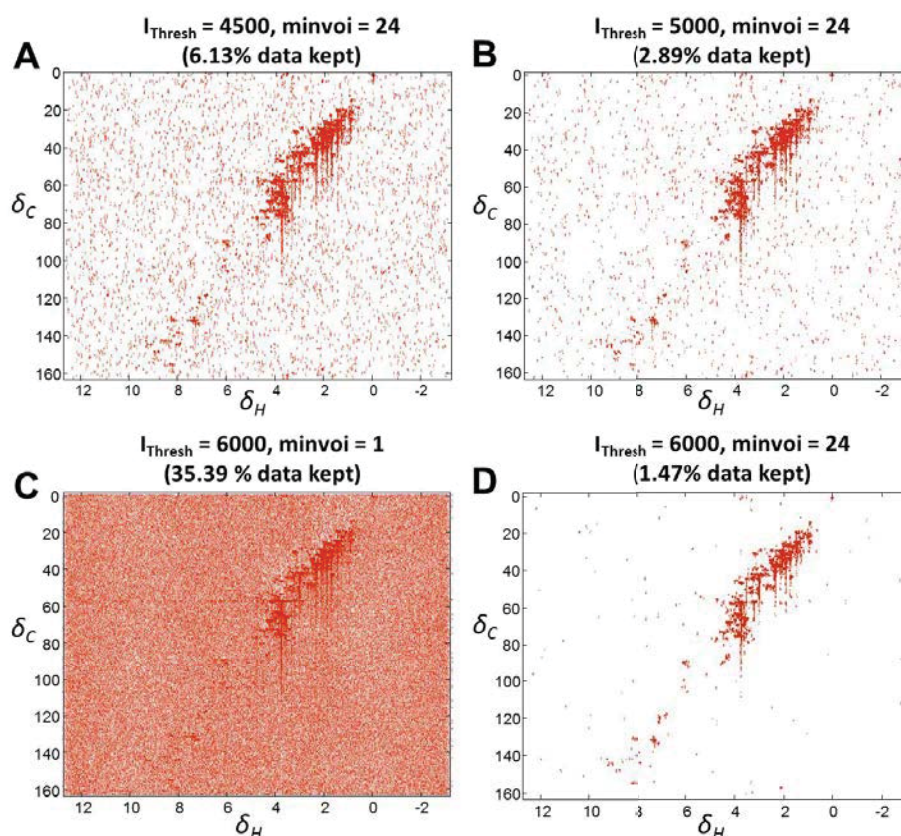
13. Effect on *threshold* and *minvoi* parameters

Figure S4. Selected VOIs (in red) for different applied threshold and *minvoi* levels.

When the *threshold* value was increased, a lower amount of random noise was included in the data. This is easily seen when figures A, B and D and compared.

On the other hand, not using the criterion of the minimum number of adjacent points that define a peak (*minvoi*) but maintaining the same threshold level (**Figure S4C**) results in a lower selective power. This is also explained because variables are selected when in at least one 2D NMR spectrum are higher than the *threshold*. Thus, due to the random distribution of noise, if threshold is at the same level as noise (as it is in many practical situations), the larger number of spectra analyzed, the larger number of noisy variables will be included in the data set. The only way to minimize this problem without using the *minvoi* criterion is by increasing the signal threshold value, although then this may result in the loss of some variables which are smaller than the fixed threshold level. Therefore, the simultaneous optimization of these two parameters should be better performed simultaneously.

14. VOI processing applied on 2D ^1H - ^1H TOCSY experiments

Example 1: Cyclosporine sample (threshold = 100,000, minvoi = 10)

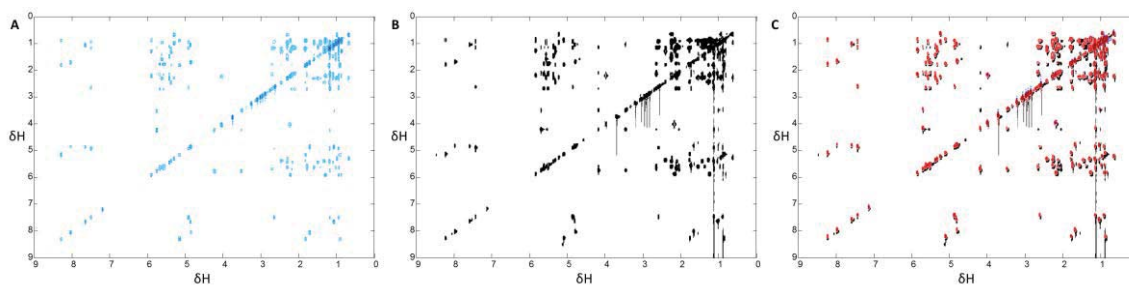


Figure S5. A) Contour plot of the reconstructed 2D ^1H - ^1H TOCSY of cyclosporine sample. B) Selected VOIs (in black). C) NMR spectrum in Fig S5B overlapped with the original NMR spectrum (contour plot obtained in MestreNova NMR suite).

After application of the VOI algorithm, only 22,998 variables were selected (**Fig S5B**), which are the 2.2% of the total set of variables.

In **Fig S5A**, the contour plot of the selected variables is shown. More intense peaks are colored with deep blue, whereas less intense peaks are colored with light blue. From comparing **FigS5A** and **FigS5B**, it is observed that the use of contour plots can be misleading for peak identification, as the smallest peaks may not even be plotted if not enough contour curves are used. In **FigS5C**, NMR peaks found in the reconstructed VOI-processed NMR spectrum coincide with the ones detected in the original NMR spectrum.

Example 2: AcCNPFDLEC sample (threshold = 14,000, minvoi = 40)

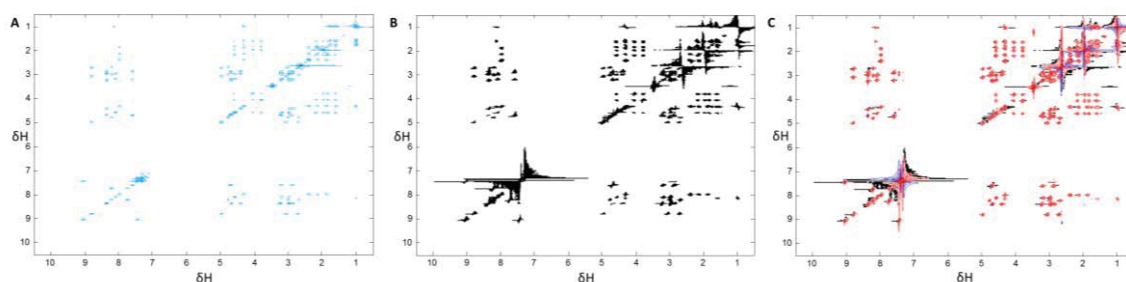


Figure S6. A) Contour plot of the reconstructed 2D ^1H - ^1H TOCSY of AcCNPFDLEC sample. B) Selected VOIs (in black). C) NMR spectrum in Fig S6B overlapped with the original NMR spectrum (contour plot obtained in MestreNova NMR suite).

After application of VOI algorithm, only 29,833 variables were selected (**Fig S6B**), which corresponds to 2.8% of the total set of variables.

15. VOI processing applied on 2D ^1H - ^{15}N HSQC experiments

Example 3: Ubiquitin sample (threshold = 3,000, minvoi = 40)

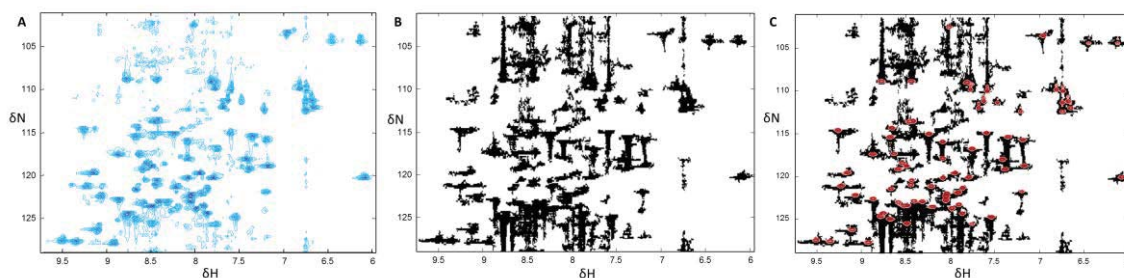


Figure S7. A) Contour plot of the reconstructed 2D ^1H - ^{15}N HSQC of ubiquitin sample. B) Selected VOIs (in black). C) Overlapped NMR spectrum of Fig S7B with the original NMR spectra (contour plot obtained in MestreNova NMR suite).

After application of VOI algorithm, only 18,660 variables were selected (**Fig S7B**), which corresponds to 14.6 % of the total set of variables. In this example, the number of selected variables was higher than in the previous examples because most peaks present tails along f1 that were also selected. With this example, it is proven that even for not very well resolved NMR spectra in both dimensions (the HSQC used here has 213 data points in f1 and 602 data points in f2) the analysis can be performed properly by the proposed VOI approach.

16. VOI processing applied on 2D ^1H - ^1H ROESY experiments

Example 4: AcCNPFDLEC sample

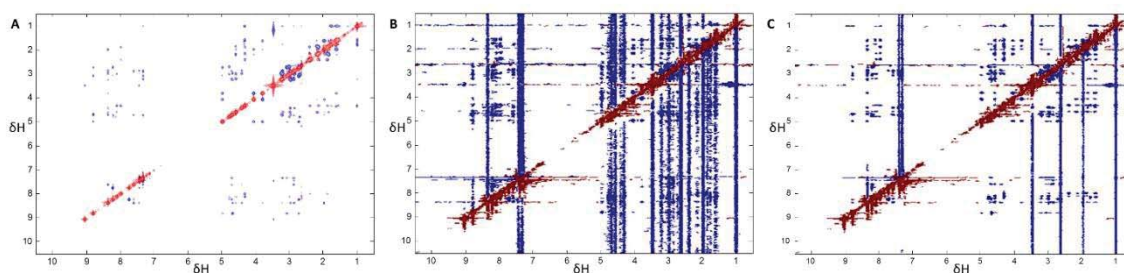


Figure S8. A) Contour plot of the original 2D ^1H - ^1H ROESY of AcCNPFDLEC sample (in MestreNova NMR suite). B) Contour plot of the selected VOIs in the analysis of the 2D ^1H - ^1H ROESY of AcCNPFDLEC sample (in MATLAB with a threshold_positive = 1,400; threshold_negative = -1,400; minvoi = 25). C) Contour plot of the selected VOIs in the analysis of the 2D ^1H - ^1H ROESY of AcCNPFDLEC sample (in MATLAB threshold_positive = 1,400; threshold_negative = -4,000; minvoi = 25). Blue color is associated to negative intensity values, and red color is associated to positive intensity values.

The VOI-processing strategy can be also applied to the phase-sensitive 2D NMR spectra such as in 2D ^1H - ^1H ROESY experiments. As stated before in section 7, to deal with positive and negative peaks, two threshold levels were used.

In the example 4 (**Figure S8**), a 2D ^1H - ^1H ROESY experiment was processed using the strategy based on VOIs. When the two threshold levels were 1,400 and -1,400, corresponding approximately to the 50% of the maximum and minimum noise levels, respectively, 161,396 variables were selected (15.4% of the total set of variables). In **FigS8B**, it is observed that most of the selected variables did not correspond to noise, but to structured negative bands found mostly along f_1 . These structures were consequence of the NMR pulse sequence used. In order to remove most of these meaningless data values with systematic (not random) information, negative threshold can be set up a bit lower. When this parameter was fixed at -4,000, the number of selected variables decreased down to 72,790, which corresponds to the 7.9% of the total set of original variables.

17. VOI processing applied on a 3D HNC0 experiment

Example 5: Protein sample (threshold = 150,000, $\text{minvoi} = 40$)

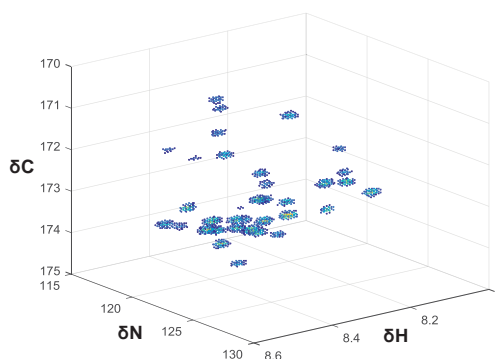


Figure S9. 3D plot of a highlighted region of the filtered HNC0 NMR spectrum, where intensity is represented by a color scale.

The VOI-processing strategy can be also applied to 3D NMR spectra such as in 3D HNC0 NMR experiments.

In the example 5 (**Figure S9**), a 3D HNC0 NMR experiment was processed using the strategy based on VOIs. The input 3D NMR spectra consisted on a cubic dataset with dimensions of 1,024 x 256 x 256 (after Topspin processing of 2048/72/128 acquired TD points, applying zero-filling and strip transform with STSI=1024), giving a total of 67,108,864 variables that occupy 256 MB (file 3rrr). These dimensions correspond to 1,024 δ_{H} values (from 4.77 to 12.89 ppm), 256 δ_{N} values (from 100.09 to 136.09 ppm) and 256 δ_{C} values (from 165 to 181 ppm), respectively.

In the 2D VOI-processing strategy, variables are searched on the 8 different positions contained in the X-Y plane (upper-left, up, upper-right, left, right, lower-left, low, lower-right). For the VOI-processing strategy extended to 3D NMR data, 26 positions are considered instead (8 positions

for the same X-Y plane than the investigated variable (with $z=0$), and 9 positions for the X-Y planes with $z=+1$ and $z=-1$).

When the threshold level was set to 150,000 and the *minvoi* was set to 40, 125,678 variables were selected (0.19% of the total set of variables). These variables were grouped in 61 clusters (individual or overlapped 3D resonances). In **Figure S9**, a highlighted region of the reconstructed 3D NMR spectrum, with 89 δ_H values, 61 δ_N values and 101 δ_C values, is shown.

In **Figure S10**, five ^1H - ^{13}C slices at different ^{15}N chemical shift from **Figure S9** are given.

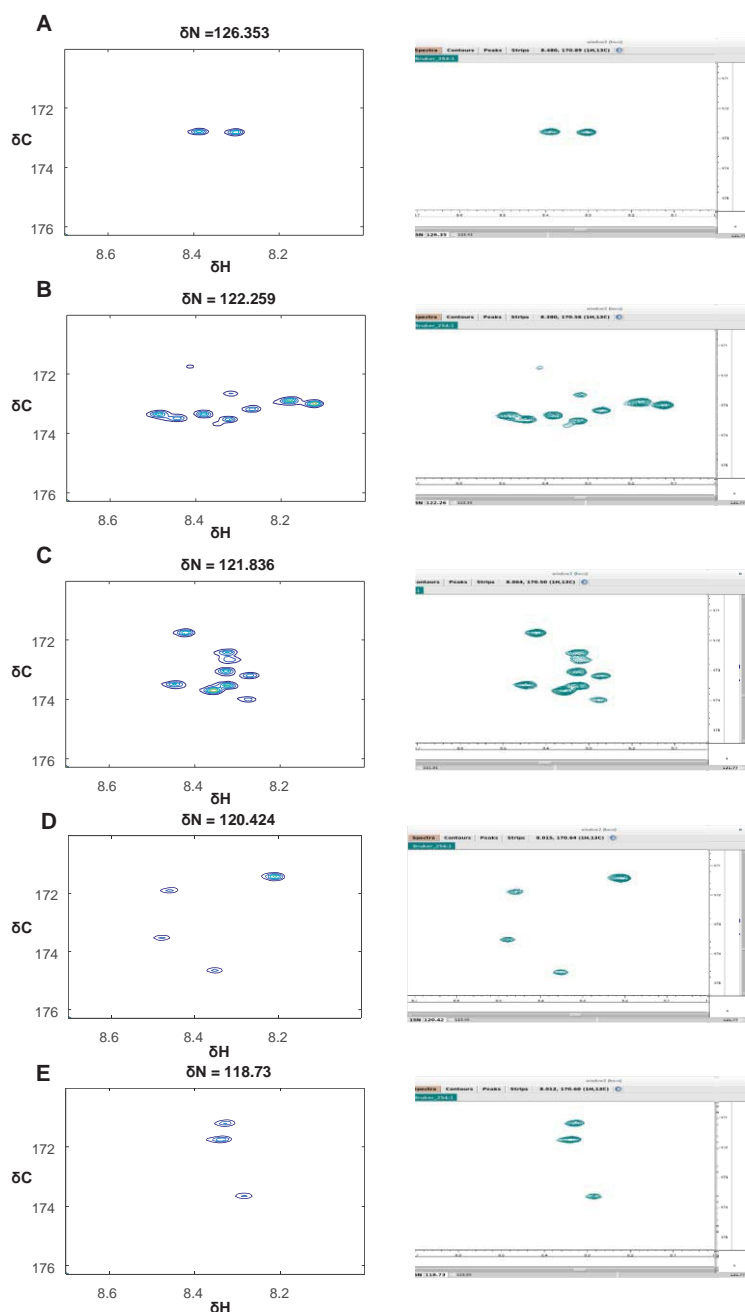


Figure S10. Selected ^1H - ^{13}C slices at different ^{15}N chemical shifts (A-E) from the 3D region highlighted in **Figure S9**. The left spectra were the reconstructed spectra after VOI filtering, while the right ones are the original raw data viewed in CcpNmr software.

References

- [1] F. Dieterle, A. Ross, G. Schlotterbeck, H. Senn, *Anal. Chem.* **2006**, *78*, 4281-4290.
- [2] al. Jolliffe, in *Wiley StatsRef: Statistics Reference Online*, John Wiley & Sons, Ltd, **2014**; bR. Bro, A. K. Smilde, *Anal. Methods* **2014**, *6*, 2812-2831; cH. Abdi, L. J. Williams, *Wiley Interdisciplinary Reviews: Computational Statistics* **2010**, *2*, 433-459.
- [3] F. Puig-Castellví, I. Alfonso, B. Piña, R. Tauler, *Scientific Reports* **2016**, *6*, 30982.
- [4] H. Abdi, in *Encyclopedia of measurement and statistics* (Ed.: N. J. Salkind), SAGE Publications, **2007**, pp. 907-912.

2.3 SCIENTIFIC ARTICLE VI

Comparative analysis of ^1H NMR and ^1H - ^{13}C HSQC NMR metabolomics to understand the effects of medium composition in yeast growth.

Authors: Puig-Castellví F., Pérez Y., Piña B., Tauler R., Alfonso I.

Citation reference: Submitted

DOI: Submitted

Comparative analysis of ^1H NMR and ^1H - ^{13}C HSQC NMR metabolomics to understand the effects of medium composition in yeast growth

Francesc Puig-Castellví^a, Yolanda Pérez^b, Benjamín Piña^a, Romà Tauler^a, Ignacio Alfonso^{c*}

^a Department of Environmental Chemistry, Institute of Environmental Assessment and Water Research (IDAEA-CSIC), Jordi Girona 18-26, 08034 Barcelona, Spain; ^b NMR Facility, Institute of Advanced Chemistry of Catalonia (IQAC-CSIC), Jordi Girona 18-26, 08034 Barcelona, Spain; ^c Department of Biological Chemistry and Molecular Modelling, Institute of Advanced Chemistry of Catalonia (IQAC-CSIC), Jordi Girona 18-26, 08034 Barcelona, Spain

ABSTRACT: In nuclear magnetic resonance (NMR) metabolomics, most of the studies have been focused on the analysis of one-dimensional proton (1D ^1H) NMR, while the analysis of other nuclei, such as ^{13}C , or other NMR experiments are still underrepresented. The preference of 1D ^1H NMR metabolomics lies on the fact that it has good sensitivity in a short acquisition time, albeit it lacks spectral resolution since it presents a high overlapping degree. In this study, the growth metabolism of yeast (*Saccharomyces cerevisiae*) has been analyzed by 1D ^1H NMR and by two-dimensional (2D) ^1H - ^{13}C HSQC (Heteronuclear Single Quantum Coherence) NMR spectroscopy, leading to the detection of more than 50 metabolites with both analytical approaches. These two analyses allow for a better understanding of the strengths and intrinsic limitations of the two types of NMR approaches. The two datasets (1D and 2D NMR) were investigated with PCA, ASCA, PLS-DA and PLSR chemometric methods, and similar results were obtained regardless of the data type used. However, data analysis time for the 2D NMR dataset was substantially reduced when compared to the data analysis of the corresponding ^1H NMR dataset because, for the 2D NMR data, signal overlapping was not a major problem and deconvolution was not required. The comparative study described in this work can be useful for the future design of metabolomics workflows, to assist in the selection of the most convenient NMR platform and to guide in the posterior data analysis of biomarker selection.

INTRODUCTION

Metabolomics¹ is the research field focused on the characterization of metabolites in cell extracts, tissues and living organisms for disease diagnosis,² biomarker discovery,³ and phenotyping,⁴ among others.⁵⁻⁷

Metabolites are commonly detected using Nuclear Magnetic Resonance spectroscopy (NMR) or mass spectrometry (MS) hyphenated to a chromatographic technique.⁸ NMR is a non-destructive technique, with minimal sample preparation, and it can provide inherently quantitative measurements. On the other hand, MS is much more sensitive and therefore a larger number of metabolites (in the order of hundreds or even thousands) can be detected, although metabolite quantitation requires the use of metabolite standards and construction of calibration curves.

Different active nuclei can be measured by NMR (such as ^1H and ^{13}C), expanding the possibilities of the technique for structural assignment.^{9,10} Regarding that, the most common NMR experiments correspond to one-dimensional (1D) spectra where a given nucleus is directly observed and analyzed. Additionally, two-dimensional (2D) experiments allow obtaining information about connectivity between nuclei, as a very powerful structural assignment tool. For instance, in a conventional ^1H - ^{13}C HSQC (Heteronuclear Single Quantum Coherence) 2D NMR spectrum, the observed resonances reveal the ^1H and the ^{13}C nuclei connected through a direct C-H covalent bond.¹¹

Even though an increasing number of NMR pulse sequences exists nowadays, most of the NMR metabolomics studies are based on the analysis of ^1H NMR datasets,¹²⁻¹⁴ since they can be acquired relatively fast with a good sensitivity and resolution, while the analysis of datasets based on other NMR pulse sequences, such as 2D homonuclear or heteronuclear, is still not so frequent.¹⁵⁻¹⁹ Nevertheless, recent strategies have emerged for increasing sensitivity in solution NMR,²⁰ some of them specially designed for high-throughput 2D experiments.²¹

Despite few new 2D pulse sequences are quantitative²²⁻²⁴ (cross-peak intensity proportional to the concentration of magnetically equivalent nuclei), routine 2D NMR experiments are not quantitative for several reasons.²⁵ First of all, 2D cross-peaks intensities depend on metabolite relaxation times and acquisition relaxation parameters, since relaxation delay is minimized in 2D experiments to reduce acquisition time for each t1 increment. Moreover, pulses excitation profiles and diversity in the actual values of J -couplings also produce differences in the intensity of the observed signals. We must bear in mind that the evolution time used for the experiments is optimized for an average $^1J_{\text{CH}}$ of 145 Hz, which is a compromise between the real values that can be observed for the different C-H groupings within molecule and between different molecules. For these routine 2D NMR spectra, absolute concentrations can still be obtained in a less straightforward approach using calibration curves.²⁶

Due to these limitations, 2D NMR spectra are used in metabolomics studies mostly for structure elucidation with selected samples^{27,28}, while 2D NMR metabolomics studies are less common^{29,30}. Nevertheless, 2D NMR metabolomics still has several advantages over 1D ¹H NMR metabolomics, as they allow for a better structural analysis, and resonance overlapping is reduced due to the existence of the second dimension.

The choice between 1D ¹H NMR and any 2D NMR when designing metabolomics studies not only affects the number and characteristics of the detected resonances, but also the posterior processing workflow. From a mathematical point of view, a single 1D NMR spectrum is a vector of intensities, while a 2D NMR spectrum is a matrix. Due to the higher simplicity of the former, most commercial or open-access tools for deconvoluting NMR spectra only apply to 1D data (i.e. Chenomx (Chenomx Inc., Alberta, Canada), AMIX (Bruker, Billerica, MA, USA), Batman³¹), while 2D NMR deconvolution methods are almost inexistent.³² In addition, chemometric analysis of 2D NMR spectra by Principal Component Analysis (PCA) or Partial Least Squares – Discriminant Analysis (PLS-DA) is not as straightforward as for 1D NMR spectral datasets. In these methods, each sample must correspond to one vector. Accordingly, 2D NMR spectra data matrices can be also analyzed in the conventional way if they are unfolded to data vectors.³³

Few 2D NMR metabolomics studies have been published until now, which used two main analytical strategies. In most of them, a careful resonance assignment was initially performed, and resonances were individually enclosed in Regions-Of-Interest (or ROI) segments³⁴ that were bucketed afterwards. Then, these sets of buckets were analyzed with PCA.^{15-17,30,35,36} Moreover, in a reduced number of papers, samples were investigated using PCA or PLS-DA directly on the vectorized form of the 2D spectral datasets.^{29,33,37-39}

Despite being less common, the chemometric analysis of all data points from the NMR spectra might provide more comprehensive results than the chemometric analysis of the buckets, since bucketing implies an important loss of spectral resolution (e.g., for ¹H NMR spectra, buckets a typically constructed with a 0.04 ppm width).

Comparative analyses between 1D and 2D NMR have been previously performed on different sets of metabolomics data, but the potential of the full data was unexploited, as the 2D NMR spectra were only analyzed either after bucketing^{36,40-42} or by just univariate analysis⁴³, while any further exploration of the data (assignment of detected resonances, resonance integration, and the chemometric analysis of the resonance integrals) were left out

In this work, we have performed an exhaustive comparative study between ¹H NMR and ¹H-¹³C HSQC NMR analyses (combining both untargeted and targeted approaches) of metabolomics samples from *Saccharomyces cerevisiae* (yeast) extracts. Specifically, yeast was grown in two different liquid media and their metabolism was characterized at 8 different time-points of a 3-day period. The two media used, YPD (Yeast Peptone Dextrose) and YSC (Yeast nitrogen base Synthetic Complete), are broadly used in yeast lab routines, and results from this analysis should be of interest for improving lab methodologies involving yeast.

Finally, in addition to the evaluation of the biological results, we have also explored and discussed the similarities and differences on the analysis workflow and on the results that were ob-

tained from either 1D or 2D NMR metabolomics analysis, giving an insight of the particular benefits and weaknesses of the two approaches.

EXPERIMENTAL SECTION

Yeast Growth. *S. cerevisiae* S288C cells were pre-cultured in YPD (1 % yeast extract, 1 % peptone, 2 % glucose) medium on an orbital shaker (150 rpm) at 30 °C overnight. All following cultures were cultured with these shaking and temperature conditions. 2 L of YNB Synthetic Complete medium (YSC, 1.7 g/L Yeast Nitrogen Base without amino acids and sulphate (Difco), 5 g/L (NH₄)₂SO₄) were inoculated with 200 µl of the pre-culture sample and left at the same temperature and shaking conditions until the culture reached an absorbance at 600 nm (*A*₆₀₀) of approximately 0.8 - 1. Pellets from these resulting cultures were collected by centrifuging the cultures, but not washed, at 613 g for 3 min and 4 °C. Pellets were used right after for inoculating erlenmeyers containing either YSC medium or YPD medium.

Sample collection. 100 ml aliquots of every culture were collected eight times during three days (0h, 2h, 4h, 6h, 10h, 24h, 48h, and 72h). For every culture and time-point, four replicates were collected. Samples were arrested with a cold shock in ice and cell were harvested by centrifugation at 4000 g for 3 min, discarding the supernatant. Cells were washed twice in Na₂HPO₄ 100 mM pH 7.0 followed by a centrifugation at 4700 g for 3 min. Pellets were stored at -80 °C and lyophilized.

Metabolite extraction. Metabolites were extracted by following the protocol published in a previous work.¹²

NMR sample preparation. Aqueous samples were dissolved in 650 µl of deuterated phosphate buffer (Na₂DPO₄ 25 mM, pH 7.0) in D₂O with DSS 0.2 mM as internal standard. Samples were centrifuged at 9,000 g for 5 min and the supernatant was collected and introduced into the NMR tube.

NMR spectroscopy. All NMR data were acquired using a Bruker Avance-IIIHD 500 MHz spectrometer equipped with a z-axis pulsed field gradient triple resonance (¹H, ¹³C, ¹⁵N) TCI cryoprobe.

¹H NMR experiments. 1D NOESY spectra were recorded using the *noesygppld* pulse sequence, using 256 scans, 4 seconds of relaxation delay, spectral width of 10 kHz and an acquired spectral size of 32k (final spectral size of 65 k after zero-filling). In total, every 1D NOESY experiment lasts for 30 minutes. The 90° pulse width (between 8.5 and 10.5 µs) and presaturation power for water suppression were measured for every sample before experiment acquisition.

¹H-¹³C HSQC NMR experiments. 2D ¹H-¹³C HSQC NMR spectra were recorded using a pulse sequence with water presaturation, sensitivity improvement and shaped adiabatic pulses for all 180-degree pulses on f2 channel (*hsqcetgprrisp2.2*) and the following parameters: 12 scans, 3 seconds of relaxation delay, spectral width of 20.7 kHz in f1 and 7.9 kHz in f2, and an acquired spectral size of 548 data points in f1 dimension and of 1,536 data points in f2 dimension. After zero-filling, final spectral size in the ¹H-¹³C HSQC NMR spectra were 1,024 data points in ¹³C dimension and 2,048 data points in f2 ¹H dimension. The total acquisition for each ¹H-¹³C HSQC NMR experiment was 6 h.

Preprocessing of NMR spectra. NMR spectra have been automatically referenced, phased and baseline corrected using

TopSpin (Bruker BioSpin GmbH, Billerica, MA, USA) routines. In addition, an exponential apodization of 0.2 Hz with MestreNova v.11.0 (Mestrelab Research, Spain) was applied to the set of ^1H NMR spectra.

The ^1H NMR spectra were stored in ASCII file format and imported to Matlab R2014a (The Mathworks Inc., Natick, MA, USA) as a data matrix of 64 rows (samples) and 65,598 columns (ppm values). Then, regions of 4.41 - 5.16 ppm (water), 3.30 - 3.37 ppm (methanol), 7.64 - 7.69 ppm (chloroform), below 0.7 ppm (DSS), and above 9.7 ppm (empty) were excluded from the analysis. With this step, the number of variables was reduced to 26,633. Next, minor resonance misalignments were corrected with *icoshift*⁴⁴. Finally, in order to eliminate sample size effects, ^1H NMR spectra were normalized using the Probabilistic Quotient Normalization (PQN)^{45,46} method.

On the other hand, the set of ^1H - ^{13}C HSQC NMR spectra were imported to Matlab R2014a using the Matlab scripts from the BBIO Toolbox Matlab scripts provided by Bruker BioSpin GmbH. For every 2D spectrum, a matrix with 1,024 rows (δ_{C}) and 2,048 columns (δ_{H}) was generated. Next, for every spectrum, noise variables were discarded using the Variable-Of-Interest (VOI) strategy published recently.⁴⁷ In VOI strategy, variables relative to noise are discarded, whereas variables with chemical meaning are kept. Selected variables needed to accomplish two criteria: first, they needed to be higher than a given threshold; and second, these intensity values should be agglomerated forming a cluster (resonance) of a minimum size. In our study, we set the intensity threshold at 6,000, and the minimum number of clustered variables was set to 24. A detailed description regarding the application of the VOI strategy can be consulted in the *Supplementary methods*. With VOI, the total number of selected variables was 65,881. Finally, the set of VOI-filtered ^1H - ^{13}C HSQC NMR spectra were normalized using the same PQN quotient values obtained in the PQN normalization of the ^1H NMR data.

Metabolite identification. Metabolite assignment was performed by a detailed targeted metabolite profiling analysis of the ^1H NMR or ^1H - ^{13}C HSQC NMR signals using the Yeast Metabolome Data Base library,⁴⁸ the Biological Magnetic Resonance Data Bank,⁴⁹ and the NMR spectral library BBIREFCODE from AMIX software (Bruker Inc.).

Integration of ^1H NMR resonances. Relative metabolite quantifications of the ^1H NMR spectral matrix were performed using BATMAN R-package³¹.

Integration of ^1H - ^{13}C HSQC NMR resonances. Relative metabolite quantifications were performed by a first segmentation of the denoised ^1H - ^{13}C HSQC NMR spectra, where each segment only includes a single resonance (for each metabolite, the most intense one or the one with less interfering overlap), followed by the sum of all the intensity values contained in each segment. For more information, see⁴⁷.

Principal Component Analysis (PCA). PCA was applied to the ^1H NMR and ^1H - ^{13}C HSQC NMR spectral matrices. Before PCA analysis, each data matrix was mean-centered.

Analysis of resonance integral datasets. The two datasets of resonance integrals were analyzed with ANOVA – Simultaneous Component Analysis (ASCA) and Partial Least Squares (PLS) chemometric methods. ASCA is a multivariate method that combines the power of ANOVA to separate variance sources with the advantages of Simultaneous Component Anal-

ysis (SCA) to the modeling of the individual separate effect matrices.^{50,51} In this work, ASCA was used to evaluate whether each of the individual factors (culture medium and time) produces a significant effect on yeast metabolism, and to evaluate whether the two factors interact (each medium produces a different metabolic response over time). Before ASCA analysis, these two matrices were scaled by dividing them with the standard deviation of the corresponding metabolite signal in the yeast culture control group (time 0h),⁵² and the effect of each factor was evaluated using a permutation test with 10,000 permutations.

PLS⁵³ is a regression method that allows correlating a relatively set of \mathbf{y} variables to a large set of \mathbf{X} -variables. As a result, a reduced number of new linear combination of the independent original \mathbf{X} -values (called Latent Variables, LV) is obtained that correlates optimally with the variation in \mathbf{y} . When the \mathbf{y} variables contain numerical information, PLS is referred as PLS Regression (PLSR). When the \mathbf{y} variables are categorical for discriminant purposes, PLS is referred as PLS-DA (Discriminant Analysis). In this work, the most influent \mathbf{X} variables in the model were calculated using their Variable Importance on Projection (VIP) scores⁵⁴. \mathbf{X} -variables associated with VIP scores greater than one are considered to be relevant on the PLS model.⁵⁵

In this study, 3 different type of PLS models per dataset have been performed: one discriminant PLS-DA model to distinguish between samples cultured in the two culture media (32 samples per class); and one regression PLSR model for each of the two sets of samples cultured in the same medium (YSC- or YPD-cultured samples). For the PLS-DAs, the \mathbf{y} vector distinguishes between the two media (YSC= 0 or YPD= 1), whereas for the PLSR, the \mathbf{y} vector has the sample collection times.

PCA, ASCA, PLSR and PLS-DA were performed using PLS toolbox 7.8.0 (Eigenvector Research Inc., Wenatchee, WA, USA). For PCA and PLS analyses, Cross-Validation with Venetian Blinds was used.

RESULTS AND DISCUSSION

Metabolite identification

In general, resonance assignment in NMR metabolomics is troublesome for 1D ^1H NMR and 2D ^1H - ^{13}C HSQC datasets. In the case of 1D ^1H NMR data, proton chemical shifts, multiplicity, and coupling constants can be measured. However, in some cases, the multiplicity pattern cannot be recognized because some of the resonances are masked by other neighboring intense signals due to the large signal overlapping. In addition, since some moieties are common for various metabolites (i.e. the trimethylammonium moiety in choline-containing metabolites, detected as a singlet at $\delta_{\text{H}} = \sim 3.2$ ppm), the full characterization of a single isolated resonance may not be sufficient to confirm a metabolite. A stack plot of representative ^1H NMR spectra of the studied samples is given in **Figure 1A**.

In the case of ^1H - ^{13}C HSQC data, resonances corresponding to direct C-H bonds ($^1J_{\text{CH}}$) are detected as cross-peaks between the two nuclei dimensions. Despite some resonances may overlap, the existence of an additional dimension attenuates this overlapping when compared with the corresponding ^1H NMR data. For instance, one of the few examples of signal overlapping is found at $\delta_{\text{H}} = 3.742 \pm 0.024$ ppm and $\delta_{\text{C}} = 57.16 \pm 0.46$ ppm. This cluster includes resonances relative to the $\text{C}_{\alpha}\text{-H}$ of L-lysine, L-ornithine, L-glutamine, L-glutamate, and L-arginine.

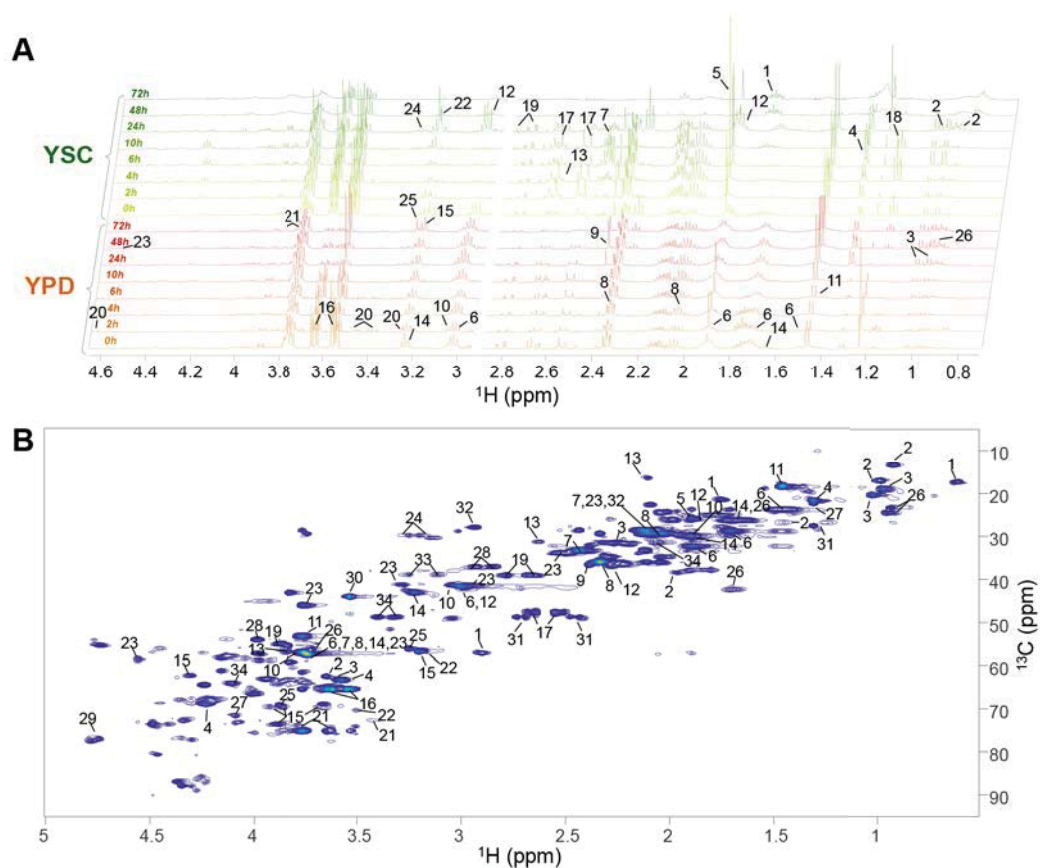


Figure 1. Representative NMR spectra. **A.** Stack plot of ^1H NMR samples (region showed $\delta_{\text{H}} = 0.7\text{--}4.7$ ppm). **B.** Contour plot of a $^1\text{H}\text{--}^{13}\text{C}$ HSQC NMR spectrum (region shown: $\delta_{\text{H}} = 0.7\text{--}5$ ppm, $\delta_{\text{C}} = 0\text{--}100$ ppm). Listed metabolites: **1** DSS, **2** L-isoleucine, **3** L-valine, **4** L-threonine, **5** acetic acid, **6** L-lysine, **7** L-glutamine, **8** L-glutamate, **9** succinic acid, **10** L-ornithine, **11** L-alanine, **12** GABA, **13** L-methionine, **14** L-arginine, **15** GPC, **16** glycerol, **17** citric acid, **18** ethanol, **19** L-aspartate, **20** glucose, **21** trehalose, **22** choline, **23** GSSG, **24** L-histidine, **25** betaine, **26** L-leucine, **27** L-lactic acid, **28** L-asparagine, **29** AMP, **30** glycine, **31** citramalic acid, **32** GSH, **33** L-phenylalanine, **34** L-proline.

In addition, for some abundant metabolites, resonances of indirect C-H bonds ($^3J_{\text{CH}}$ or longer) are detected, allowing for a better metabolite identification. $^3J_{\text{CH}}$ couplings were observed for glycerol, L-alanine, L-leucine, L-isoleucine, L-lysine, and L-valine. In **Figure 1B**, the contour plot of a single $^1\text{H}\text{--}^{13}\text{C}$ HSQC NMR is shown.

From the analysis of the ^1H NMR dataset, 54 metabolites were identified, comprising mostly amino acids, nucleotides, sugars and organic acids. On the other hand, when the $^1\text{H}\text{--}^{13}\text{C}$ HSQC dataset was analyzed, the corresponding cross-peaks for 55 metabolites were assigned, of which 50 were also detected by 1D ^1H NMR. Overall, a total of 59 different metabolites were detected.

Metabolites that were detected in the ^1H NMR but not in $^1\text{H}\text{--}^{13}\text{C}$ HSQC NMR were (S)-2-isopropylmalate, fumaric acid, oxalacetic acid, and a choline derivative. Metabolites detected in $^1\text{H}\text{--}^{13}\text{C}$ HSQC but not satisfactorily confirmed in ^1H NMR were citramalic acid, malic acid, pyroglutamic acid, β -alanine, and cysteineglutathione disulfide (CYSSG). Metabolites that were only detected by ^1H NMR corresponded to those that were found below the detection limits by $^1\text{H}\text{--}^{13}\text{C}$ HSQC. On the other hand, resonances from metabolites that were only detected by $^1\text{H}\text{--}^{13}\text{C}$ HSQC NMR were masked by other resonances in ^1H NMR. For CYSSG, it is worth mentioning that its characteristic signal ($\delta_{\text{H}} = 4.75$ ppm and $\delta_{\text{C}} = 55.3$ ppm) was masked by the

water peak (and partly suppressed by the 1D NOESY pulse sequence). The final list of assigned resonances for both datasets is given in **Table S1**.

Metabolite quantitation

Here, we have investigated the robustness of 2D NMR in providing reliable relative concentration (or fold-change) measurements by performing calibration curves between integrals from 1D data with integrals from 2D data. Integrals from ^1H NMR data were obtained with Batman R-package^{31,56}, whereas 2D integrals were obtained using our previously proposed VOI strategy¹⁷.

After performing the integration for the NMR signals of the two datasets, we noticed that time spent on integration of 2D NMR data is drastically reduced when compared to the same analysis in the 1D NMR data integration. This is caused because deconvolution is a necessary preliminary step in the integration analysis of 1D NMR data but not for 2D NMR data integration (see **methods** section).

Then, for every metabolite, the vector of 1D integrals was regressed with the corresponding vector of 2D integrals and the correlation coefficient between them was calculated. To avoid obtaining falsely low correlation coefficients, samples with integrated 2D cross-peak resonances below the detection limit were excluded in the regression. A table containing the correlation coefficients for every detected metabolite is provided in **Table S2**.

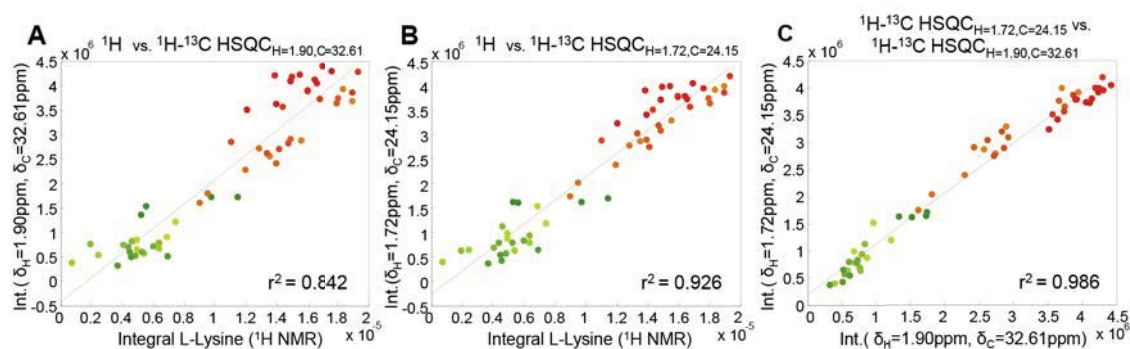


Figure 2. Linear regressions using resonance integrals from L-lysine. A–B) Regression between the ^1H NMR Batman integrals and the ^1H - ^{13}C HSQC NMR integrals from A) $\delta_{\text{H}}=1.72$ ppm and $\delta_{\text{C}}=29.1$ ppm or from B) $\delta_{\text{H}}=1.90$ ppm and $\delta_{\text{C}}=32.6$ ppm. C) Regressions between the ^1H - ^{13}C HSQC NMR integrals (cross-peak at $\delta_{\text{H}}=1.90$ ppm and $\delta_{\text{C}}=32.6$ ppm) and the ^1H - ^{13}C HSQC NMR integrals (cross-peak at $\delta_{\text{H}}=1.72$ ppm and $\delta_{\text{C}}=29.1$ ppm). Each dot represents the same analyzed sample, green and red colors denote for the liquid medium used (YSC, green; YPD, red), and sample collection time is indicated by color darkness (early time-points in lighter colors, and late time-points in darker colors).

Results from these analyses suggest that different methods may produce different outcomes. This is not an intrinsic problem of the NMR data type, but of its quality. For concentrated metabolites, a good regression was obtained, indicating that the two methods can provide equally satisfactory semi-quantitative measurements. However, for 1D NMR data, these measurements can be undermined if the peak integrals were obtained from an overlapped region. Even though a Lorentzian model can be used to deconvolute the integral from the analyzed overlapped region, results can differ from the true ones if the model captured wrongly intensities from other peaks.

This has been observed, for instance, for L-lysine (Figure 2). L-lysine concentration was obtained from the ^1H NMR data by deconvolution of the multiplet peak at 1.47 ppm and of the triplet peak at 3.01 ppm, which were simpler to deconvolute than the other L-lysine resonances. On the other hand, in the 2D ^1H - ^{13}C HSQC NMR dataset, two integrals related to L-lysine were obtained from integration of the cross-peaks found at $\delta_{\text{H}}=1.90$ ppm and $\delta_{\text{C}}=32.6$ ppm, and at $\delta_{\text{H}}=1.72$ ppm and $\delta_{\text{C}}=29.1$ ppm, both perfectly isolated from the rest of the signals. When we compared the two sets of integrals, the regression using the 2D NMR data presented a convincing match ($r^2=0.986$, Figure 2C), whereas it was considerably worse for 2D NMR data when it was compared with the 1D data ($r^2=0.842$ and $r^2=0.926$, Figure 2A and Figure 2B, respectively). This implies that both 1D and 2D data can be useful for semi-quantitation, and that the deconvolution may be a source of error and caution should be taken. In any case, regardless of the type of NMR data used, we consider that it is always preferable to calculate integrals from non-overlapped resonances than to use deconvolution tools to integrate overlapped resonances.

From our data, the metabolites that can be much better integrated in the 2D ^1H - ^{13}C HSQC NMR data are citric acid, oxidized glutathione, L-asparagine, L-aspartic acid and L-serine. For all of them, all their characteristic resonances lay in complex (and crowded) areas in the ^1H NMR spectrum.

Finally, due to a lower sensitivity in the HSQC, some metabolites were not detected in all samples. In our study, 4 metabolites were only detected in the ^1H NMR samples, and some others (i.e. L-tryptophan, thiamine, nicotinamide mononucleotide) were not detected for most of the ^1H - ^{13}C HSQC spectra although they were measurable in the corresponding ^1H NMR

spectra. Thus, in these cases, it is preferable to use integral values from deconvoluting ^1H NMR spectra.

Explorative untargeted analysis with PCA

PCA was used on the two NMR spectral (1D or 2D) datasets to assess, as an untargeted approach, whether the spectral differences produced a significant impact on the observed results.

Variables from both assigned and unassigned resonances were considered in the two PCA analyses, but noisy variables were excluded before analysis. Before PCA analysis of spectral data containing an abundant number of noisy variables, such as for NMR data, so it is recommended to discard those noisy variables which could compromise data analysis and interpretation.^{33,39,47} For the ^1H NMR dataset, variables from empty regions were removed, whereas for the ^1H - ^{13}C HSQC dataset, noisy variables were discarded with the VOI strategy (see **method section** for more information about variable filtering⁴⁷).

First two principal components explained 62.95% of the data variance for the 1D NMR dataset (Figure 3A), and 55.08% for the 2D NMR dataset (Figure 3C). A similar score plots distribution was observed for the two PCA analyses (Figure 3A and Figure 3C). A linear evolution over time of the YPD-cultured sample scores (orange-red) was observed, whereas for YSC-cultured sample scores (green), their trajectory changed after the first 24h. Positive scores on PC1 were associated with the early growth (0–24h) of YSC-cultured samples, while positive scores on PC2 were associated with the growth of YPD-cultured samples. This comparison showed that chemometric analysis of 2D NMR spectral data is as informative as chemometric analysis of 1D NMR spectral data, although not so commonly recognized in the previous literature (with very few exceptions^{29,33,37–39}). Furthermore, the computational times for the two analyses were also similar.

In addition, for these two type of analyses, PCA loading plots highlighted the same metabolites, with resonances between $\delta_{\text{H}}=1$ ppm and $\delta_{\text{H}}=5$ ppm (loadings for PC1 are given in Figure 3B (for 1D NMR) and Figure 3D (for 2D NMR)). However, since 2D NMR data has less spectral overlapping, better-defined loadings were unambiguously obtained. Thus, 2D NMR data should be preferred when unambiguous assignment of the variable loadings is required.³⁷

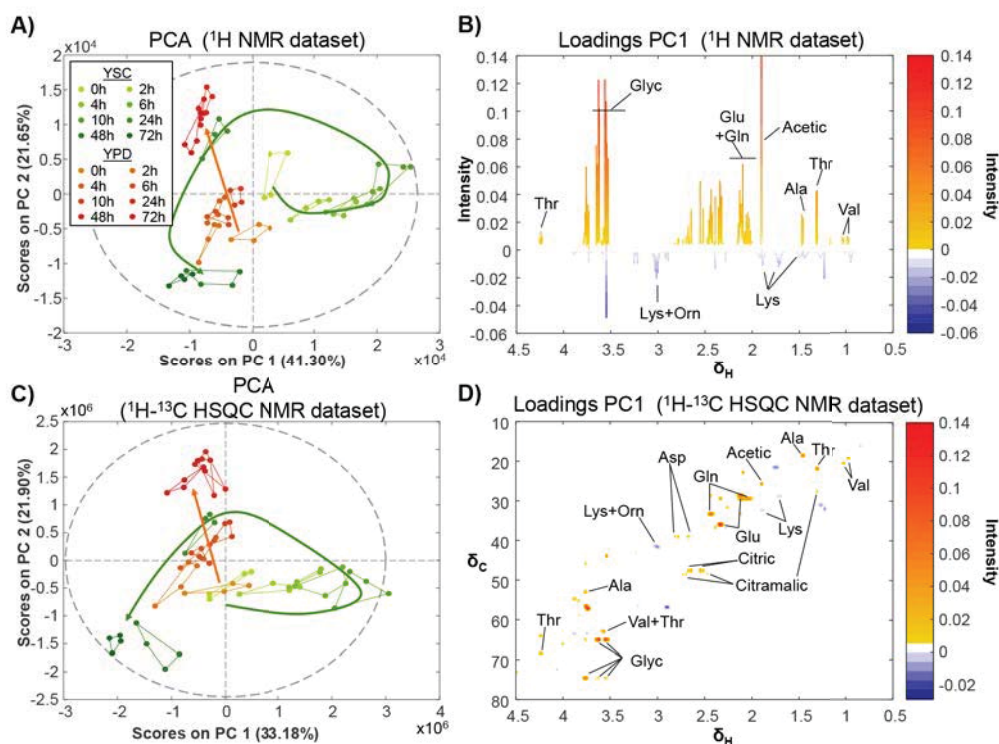


Figure 3. PCA. **A)** and **C)** PCA scores plot of the **(A)** ^1H NMR dataset and the **(C)** ^1H - ^{13}C HSQC NMR dataset. **B)** and **D)** Loadings of PC1 of the **(B)** ^1H NMR dataset and in the **(D)** ^1H - ^{13}C HSQC NMR dataset. Samples in the scores plot (**A**, **C**) are colored according to the legend in **A**, while variables in the loading plots (**B**, **D**) are colored in red for positive loadings and in blue for negative loadings. In the loadings plots (**B**, **D**), only the proton region between 0.5 ppm and 4.5 ppm is shown. Green and orange arrow denote score progression over time of the YSC- and YPD-cultured samples, respectively.

Study of the effects of medium composition and time

The two resonance integrals datasets were then analyzed by ASCA and PLS (PLSR and PLS-DA) chemometric methods. ASCA was first used to evaluate the importance of the two factors (culture medium and time) on yeast metabolism, as well as the interaction between them. The two ASCA analyses (one per NMR dataset) revealed that both factors and their interaction were significant ($p=0.0001$). These results confirmed that using a different culture medium would affect the speed of the metabolic events occurring inside yeast cells.

Effects of each of the two factors were then examined by PLS analysis of the integral matrices obtained from each of the two NMR datasets. Metabolic differences caused by the two different media are summarized in **Fig4**, whereas metabolites that better correlate with yeast growth over time in YPD and in YSC media are summarized in **Fig5** and in **FigS2**, respectively.

2 LVs were selected in the PLS-DA models of the samples cultured at the two different media (**Fig4**). 88.04% and 34.14% of the y - and X -variances were explained respectively for the 1D dataset (**Fig4A**). In the case of the 2D dataset, 81.56% and 37.66% of the y - and X -variances were explained (**Fig4C**). Interestingly, for the two type of analyses, metabolites associated to the highest VIPs ($\text{VIP} > 1$) were mostly the same (**Fig4B** and **Fig4D**). According to these two PLS-DA models, metabolites that allow optimal class separation were GABA, glycerol, L-arginine, L-lysine, and L-ornithine, among others.

In the case of YPD-cultured samples, only one LV component was needed in the PLSR model to obtain a good prediction (**Fig5A** and **Fig5C**). For the 1D NMR dataset, 96.28% of the y -

variance was explained considering the 45.08% of the X -variance. For the 2D NMR dataset, 90.57% of the y -variance was explained considering 46.96% of the X -variance (**Fig5C**). Analogously to the PLS-DA analysis of **Fig4**, for the two analyses displayed in **Fig5**, a similar list of metabolites was highlighted (**Fig5B** and **Fig5D**). Metabolites correlated to yeast growth in YPD medium were GABA, glycerophosphocholine, glycine, L-phenylalanine, and L-proline, among others.

Due to the much higher complexity of yeast metabolism in YSC medium compared to YPD medium, as also observed from PCA results shown in **Fig3**, 3 LVs components were needed to achieve a good prediction (**FigS2A** and **FigS2C**), in both cases needing more than the 97% of the X -variance to explain at most only 63% of the y -variance (see supplementary material). 27 metabolites were detected as significant in the PLSR VIPs selection method of the 1D NMR dataset and 22 metabolites in the case of the 2D NMR datasets (14 metabolites were in common in both types of analysis). Metabolites correlated to yeast growth in YSC medium were L-alanine, L-asparagine, L-tyrosine, L-valine, and succinic acid, among others.

Biological interpretation

The observed metabolomic differences between cells grown in rich (YPD) and minimal (YSC) media are likely related to the different response to starvation occurring in these two different media. Prior to entry into stationary phase, yeast cultures progress through a series of growth phases, including the exponential phase (high glucose, fermentative metabolism), the diauxic shift (low glucose, transition to respiratory metabolism), the post-diauxic phase (low nutrients, respiratory metabolism) and, finally, the stationary phase (lack of nutrients, no growth),

which typically occurs after 48-72 h of growth.⁵⁷ It is at this point where the medium of the culture determines the fate of cells. Cells grown in YPD have a more prolonged period of hypo-metabolism in stationary phase, which allows them to remain viable for several weeks. In contrast, yeast cells grown in synthetic media, like YSC, maintain a high metabolism rate in the stationary phase, which results in a considerable loss of viability in only a few days after the exhaustion of the medium⁵⁸.

Our results confirmed very similar metabolic patterns for both cultures for the first 24h of incubation, roughly coinciding with the entry into stationary phase. After this point, the two cultures diverge, the differences becoming maximal after 72 h of culture. At this point, YPD-grown cells would likely enter into a low-metabolic, resilient state (therefore maintaining the physiological levels of essential metabolites), whereas those YSC-grown ones are at their limit of viability, after having consumed all available nutrients.^{57,58}

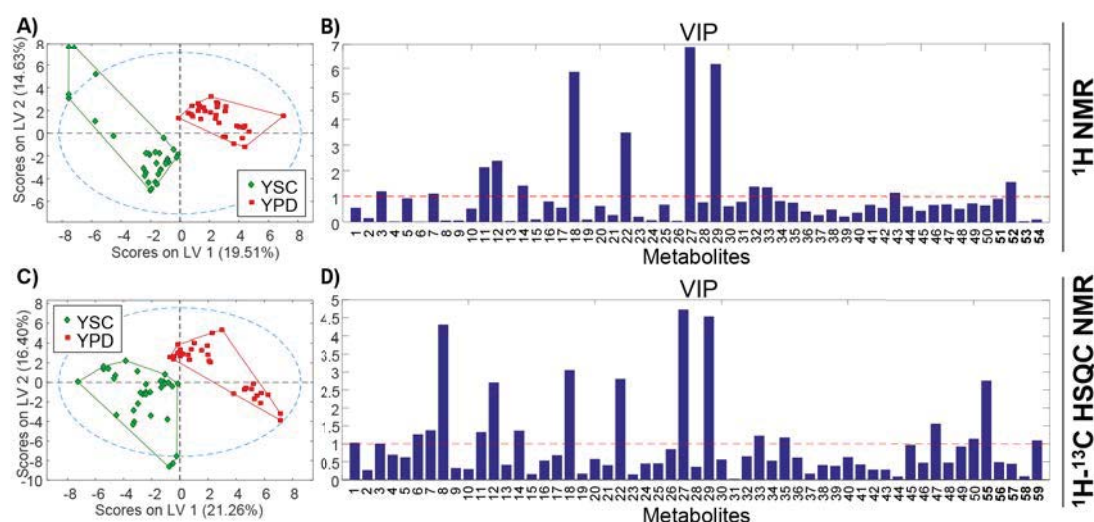


Figure 4. PLS-DA. A) and C) Scores plot for the PLS-DA analysis of the (A) ¹H NMR dataset and the (C) ¹H-¹³C HSQC NMR dataset. B) and D) VIP values for the variables in the (B) ¹H NMR dataset and in the (D) ¹H-¹³C HSQC NMR dataset. Metabolite list: 1 acetic acid, 2 adenosine, 3 AMP, 4 ATP, 5 betaine, 6 choline, 7 citraconic acid, 8 citric acid, 9 CMP, 10 ethanol, 11 GABA, 12 glycerol, 13 glycerophosphocholine, 14 glycine, 15 GMP, 16 GSSG, 17 L-alanine, 18 L-arginine, 19 L-asparagine, 20 L-aspartic acid, 21 L-glutamic acid, 22 L-glutamine, 23 L-histidine, 24 L-isoleucine, 25 L-lactic acid, 26 L-leucine, 27 L-lysine, 28 L-methionine, 29 L-ornithine, 30 L-phenylalanine, 31 L-proline, 32 L-serine, 33 L-threonine, 34 L-tryptophan, 35 L-tyrosine, 36 L-valine, 37 NAD⁺, 38 NADP, 39 NMN, 40 succinic acid, 41 taurine, 42 thiamine, 43 thiaminePP, 44 trehalose, 45 UDP-Glc, 46 UDP-Nac-glc-NH₂, 47 uracil, 48 uridine, 49 α-glucose, 50 β-glucose, 51 2-(S)-isopropylmalic acid, 52 choline-derivative, 53 fumaric acid, 54 oxalacetic acid, 55 citramalic acid, 56 CYSSG, 57 malic acid, 58 pyroglutamic acid, 59 β-alanine.

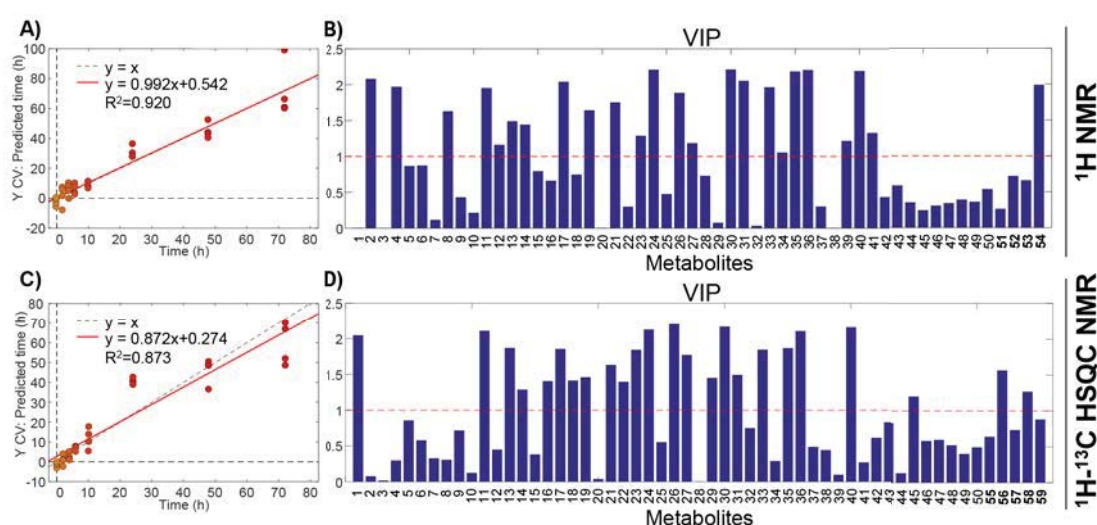


Figure 5. PLSR of YPD-cultured samples. A) and C) Regression between experimental and predicted collection times for the (A) ¹H NMR dataset and the (C) ¹H-¹³C HSQC NMR dataset. B) and D) VIP values for the variables in the (B) ¹H NMR dataset and in the (D) ¹H-¹³C HSQC NMR dataset. Metabolite names for the numbers in B) and in D) are listed in the caption of Figure 4.

CONCLUSIONS

The metabolism of *Saccharomyces cerevisiae* cultured under unrestricted conditions has been characterized indistinctively by 1D ^1H NMR and 2D ^1H - ^{13}C HSQC NMR spectroscopy. The intrinsic differences in signal resolution and sensitivity inherent to each of the NMR pulse sequences used caused that the lists of metabolites detected by the two analyses were slightly different. These intrinsic differences also affected the estimation of the resonance integrals. For instance, resonance integrals from low concentrated metabolites were poorly estimated from the 2D NMR data, while deconvolution approaches may not perfectly integrate overlapped resonances, such as the ones in ^1H NMR spectra of metabolomics samples.

Untargeted analysis of 2D NMR data has been confirmed to be a reliable strategy to study yeast metabolism because the information present in the second dimension allows for a better signal resonance assignment, which results on an improvement in the understanding of the biological studied system (yeast in this case). In addition, for low overlapped ^1H - ^{13}C HSQC NMR spectra, resonance integration after application of the VOI denoising strategy is much faster than the corresponding integration of ^1H NMR spectra needing their preliminary deconvolution. PCA, ASCA and PLS analyses revealed to be especially useful to analyze both 1D and 2D NMR spectra, leading to similar results. Using both approaches, metabolic events occurring during yeast growth were confirmed to be highly influenced by the culture medium composition.

ASSOCIATED CONTENT

Supporting Information

Table S1. Resonance assignment for ^1H NMR and ^1H - ^{13}C HSQC NMR data (PDF).

Table S2. Correlation coefficients between integrals from the ^1H NMR analysis and the ^1H - ^{13}C HSQC NMR analysis (PDF).

Figure S1. PLSR of YSC-cultured samples (PDF).

The Supporting Information is available free of charge on the ACS Publications website.

AUTHOR INFORMATION

Corresponding Author

* Tel: +34-934006100

E-mail address: ignacio.alfonso@iqac.csic.es

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

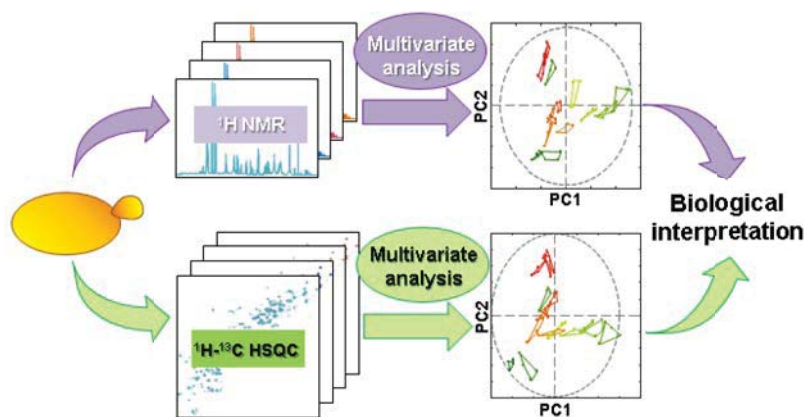
The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 320737. The 500-MHz spectrometer was purchased in part through a Research Infrastructure MINECO-FEDER fund (Grant CSIC13-4E-2076).

REFERENCES

(1) Johnson, C. H.; Ivanisevic, J.; Siuzdak, G. *Nat Rev Mol Cell Biol* **2016**, *17*, 451-459.

- (2) Kang, J.; Zhu, L.; Lu, J.; Zhang, X. *Journal of Neuroimmunology* **2015**, *279*, 25-32.
- (3) Diaz, S.; Pinto, J.; Graca, G.; Duarte, I.; Barros, A.; Galhano, E.; Pita, C.; Almeida Mdo, C.; Goodfellow, B.; Carreira, I.; Gil, A. *J Proteome Res* **2011**, *10*, 3732 - 3742.
- (4) Dalmau, N.; Jaumot, J.; Tauler, R.; Bedia, C. *Molecular BioSystems* **2015**, *11*, 3397-3406.
- (5) Putri, S. P.; Nakayama, Y.; Matsuda, F.; Uchikata, T.; Kobayashi, S.; Matsubara, A.; Fukusaki, E. *Journal of Bioscience and Bioengineering* **2013**, *115*, 579-589.
- (6) Wishart, D. S. *Nature Reviews Drug Discovery* **2016**, *15*, 473.
- (7) Wishart, D. S. *Trends in Food Science & Technology* **2008**, *19*, 482-493.
- (8) Dunn, W. B.; Bailey, N. J. C.; Johnson, H. E. *Analyst* **2005**, *130*, 606-625.
- (9) Batta, G.; Kövér, K.; Szántay, C. *Methods for Structure Elucidation by High-Resolution NMR: Applications to Organic Molecules of Moderate Molecular Weight*; Elsevier Science, 1997.
- (10) Williams, A.; Martin, G.; Rovnyak, D. *Modern NMR Approaches to the Structure Elucidation of Natural Products: Volume 2: Data Acquisition and Applications to Compound Classes*; Royal Society of Chemistry, 2016.
- (11) Castañar, L.; Parella, T. In *Annual Reports on NMR Spectroscopy*, Webb, G. A., Ed.; Academic Press, 2015, pp 163-232.
- (12) Puig-Castellví, F.; Alfonso, I.; Piña, B.; Tauler, R. *Metabolomics* **2015**, *11*, 1612-1625.
- (13) Kruger, N. J.; Troncoso-Ponce, M. A.; Ratcliffe, R. G. *Nat. Protocols* **2008**, *3*, 1001-1012.
- (14) Cuperlovic-Culf, M.; Cormier, K.; Touaibia, M.; Reyjal, J.; Robichaud, S.; Belbraouet, M.; Turcotte, S. *International Journal of Cancer* **2016**, *138*, 2439-2449.
- (15) Kang, W. Y.; Kim, S. H.; Chae, Y. K. *FEMS Yeast Res.* **2012**, *12*, 608-616.
- (16) Chae, Y. K.; Kim, S. H.; Nam, Y.-K. *Chemistry & Biodiversity* **2013**, *10*, 1816-1827.
- (17) Chae, Y. K.; Kim, S. H. *Bulletin of the Korean Chemical Society* **2016**, *37*, 1612-1617.
- (18) Chylla, R. A.; Van Acker, R.; Kim, H.; Azapira, A.; Mukerjee, P.; Markley, J. L.; Storme, V.; Boerjan, W.; Ralph, J. *Biotechnology for Biofuels* **2013**, *6*, 45.
- (19) Deyrup, S. T.; Eckman, L. E.; McCarthy, P. H.; Smedley, S. R.; Meinwald, J.; Schroeder, F. C. *Proceedings of the National Academy of Sciences* **2011**, *108*, 9753-9758.
- (20) Lee, J. H.; Okuno, Y.; Cavagnero, S. *Journal of magnetic resonance (San Diego, Calif. : 1997)* **2014**, *241*, 18-31.
- (21) Hansen, A. L.; Li, D.; Wang, C.; Brüscheiler, R. *Angewandte Chemie International Edition* **2017**, *56*, 8149-8152.
- (22) Hu, K. F.; Westler, W. M.; Markley, J. L. *J Am Chem Soc* **2011**, *133*.
- (23) Martineau, E.; Tea, I.; Akoka, S.; Giraudeau, P. *NMR Biomed* **2012**, *25*, 985-992.
- (24) Mauve, C.; Khelifi, S.; Gilard, F.; Mouille, G.; Farjon, J. *Chemical Communications* **2016**, *52*, 6142-6145.
- (25) Koskela, H. In *Annual Reports on NMR Spectroscopy*; Academic Press, 2009, pp 1-31.
- (26) Lewis, I. A.; Schommer, S. C.; Hodis, B.; Robb, K. A.; Tonelli, M.; Westler, W. M.; Sussman, M. R.; Markley, J. L. *Analytical Chemistry* **2007**, *79*, 9385-9390.
- (27) Marchev, A.; Yordanova, Z.; Alipieva, K.; Zahmanov, G.; Rusinova-Videva, S.; Kapchina-Toteva, V.; Simova, S.; Popova, M.; Georgiev, M. I. *Biotechnology Letters* **2016**, *38*, 1621-1629.
- (28) Sobolev, A. P.; Mannina, L.; Proietti, N.; Carradori, S.; Daglia, M.; Giusti, A. M.; Antiochia, R.; Capitani, D. *Molecules (Basel, Switzerland)* **2015**, *20*, 4088-4108.
- (29) Izrayelit, Y.; Robinette, S. L.; Bose, N.; von Reuss, S. H.; Schroeder, F. C. *ACS Chemical Biology* **2013**, *8*, 314-319.
- (30) Chae, Y. K.; Kim, S. H.; Markley, J. L. *PLOS ONE* **2017**, *12*, e0177233.
- (31) Hao, J.; Astle, W.; De Iorio, M.; Ebbels, T. M. D. *Bioinformatics* **2012**, *28*, 2088-2090.

- (32) Chylla, R. A.; Hu, K.; Ellinger, J. J.; Markley, J. L. *Anal Chem* **2011**, *83*.
- (33) Hedenström, M.; Wiklund, S.; Sundberg, B.; Edlund, U. *Chemometrics and Intelligent Laboratory Systems* **2008**, *92*, 110-117.
- (34) Lewis, I. A.; Schommer, S. C.; Markley, J. L. *Magn Reson Chem* **2009**, *47*.
- (35) Kang, C.-M.; Seong Hyeon, J.; Ra Kim, S.; Kyeong Lee, E.; Jin Yun, H.; Young Kim, S.; Kee Chae, Y. *Chemistry & Biodiversity* **2015**, *12*, 1696-1705.
- (36) Guerrini, M.; Rudd, T. R.; Mauri, L.; Macchi, E.; Fareed, J.; Yates, E. A.; Naggi, A.; Torri, G. *Analytical Chemistry* **2015**, *87*, 8275-8283.
- (37) Arbogast, L. W.; Delaglio, F.; Schiel, J. E.; Marino, J. P. *Analytical Chemistry* **2017**, *89*, 11839-11845.
- (38) Robinette, S. L.; Ajredini, R.; Rasheed, H.; Zeinomar, A.; Schroeder, F. C.; Dossey, A. T.; Edison, A. S. *Analytical Chemistry* **2011**, *83*, 1649-1657.
- (39) Sharma, R.; Gogna, N.; Singh, H.; Dorai, K. *RSC Advances* **2017**, *7*, 29860-29870.
- (40) Nadal-Desbarats, L.; Aidoud, N.; Emond, P.; Blasco, H.; Filipiak, I.; Sarda, P.; Bonnet-Brilhault, F.; Mavel, S.; Andres, C. R. *Analyst* **2014**, *139*, 3460-3468.
- (41) Guennec, A. L.; Giraudeau, P.; Caldarelli, S. *Analytical Chemistry* **2014**, *86*, 5946-5954.
- (42) Van, Q. N.; Issaq, H. J.; Jiang, Q.; Li, Q.; Muschik, G. M.; Waybright, T. J.; Lou, H.; Dean, M.; Uitto, J.; Veenstra, T. D. *Journal of Proteome Research* **2008**, *7*, 630-639.
- (43) Allen, P. J.; Wise, D.; Greenway, T.; Khoo, L.; Griffin, M. J.; Jablonsky, M. *Metabolomics* **2015**, *11*, 1131-1143.
- (44) Savorani, F.; Tomasi, G.; Engelsen, S. B. *J. Magn. Reson.* **2010**, *202*, 190-202.
- (45) Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H. *Anal. Chem.* **2006**, *78*, 4281-4290.
- (46) Puig-Castellví, F.; Alfonso, I.; Piña, B.; Tauler, R. *Scient Reports* **2016**, *6*, 30982.
- (47) Puig-Castellví, F.; Pérez, Y.; Piña, B.; Tauler, R.; Alfonso *Chemical Communications* **2018**.
- (48) Jewison, T.; Knox, C.; Neveu, V.; Djoumbou, Y.; Guo, A. C.; I. J.; Liu, P.; Mandal, R.; Krishnamurthy, R.; Sinelnikov, I.; Wilson, . Wishart, D. S. *Nucleic Acids Res.* **2012**, *40*, D815-D820.
- (49) Ulrich, E. L.; Akutsu, H.; Doreleijers, J. F.; Harano, Y.; Ioanni Y. E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z.; Nakatani, E.; Schulte, C. F.; Tolmie, D. E.; Kent Wenger, R.; Yao, H.; Markl J. L. *Nucleic Acids Research* **2008**, *36*, D402-D408.
- (50) Smilde, A. K.; Jansen, J. J.; Hoefsloot, H. C. J.; Lamers, R.-J. N.; van der Greef, J.; Timmerman, M. E. *Bioinformatics* **2005**, *3043-3048*.
- (51) Jansen, J. J.; Hoefsloot, H. C. J.; van der Greef, J.; Timmerman M. E.; Westerhuis, J. A.; Smilde, A. K. *J. Chemometr.* **2005**, *19*, 4 481.
- (52) Timmerman, M. E.; Hoefsloot, H. C. J.; Smilde, A. K.; Ceuleme E. *Metabolomics* **2015**, *11*, 1265-1276.
- (53) Geladi, P.; Kowalski, B. R. *Anal. Chim. Acta* **1986**, *185*, 1-17.
- (54) Wold, S.; Sjöström, M.; Eriksson, L. *Chemometrics Intell. L. Syst.* **2001**, *58*, 109-130.
- (55) Chong, I.-G.; Jun, C.-H. *Chemometrics Intell. Lab. Syst.* **2005**, *103-112*.
- (56) Hao, J.; Liebeke, M.; Astle, W.; De Iorio, M.; Bundy, J. Ebbels, T. M. D. *Nat. Protoc.* **2014**, *9*, 1416-1427.
- (57) Wernerwashburne, M.; Braun, E.; Johnston, G. C.; Singer, R. *Microbiol. Rev.* **1993**, *57*, 383-401.
- (58) Chen, Q. H.; Ding, Q. X.; Keller, R. N. *Biogerontology* **2005** 1-13.



SUPPLEMENTARY MATERIAL FOR SCIENTIFIC ARTICLE VI

Comparative analysis of ^1H NMR and ^1H - ^{13}C HSQC NMR metabolomics to understand the effects of medium composition in yeast growth.

Authors: Puig-Castellví F., Pérez Y., Piña B., Tauler R., Alfonso I.

Citation reference: Submitted

DOI: Submitted

SUPPLEMENTARY METHODS

1. Variable of Interest (VOI) methodology

The VOI filtering function (*voi2D.m*) was implemented in Matlab programming language and it can be downloaded from <https://github.com/f-puig/VOI>.

To define the *threshold* level, the most practical option is by checking the signal intensities of the 2D NMR spectrum projected on one of the two dimensions (either ppm1 or ppm2). From this representation, it is easy to establish an intensity *threshold* value higher than the observed noise. The chosen threshold value should be located below the upper-limit of the noise level (see example in **Figure S1**).

It is also possible to establish the *threshold* level after the analysis of the distribution of the intensities from noise variables. This can be performed by selecting a NMR region where no resonances were detected and building the corresponding box plot. For the tested 2D NMR spectra, satisfactory results were obtained by fixing the threshold level to the half of the maximum noise value gave satisfactory filtering results.

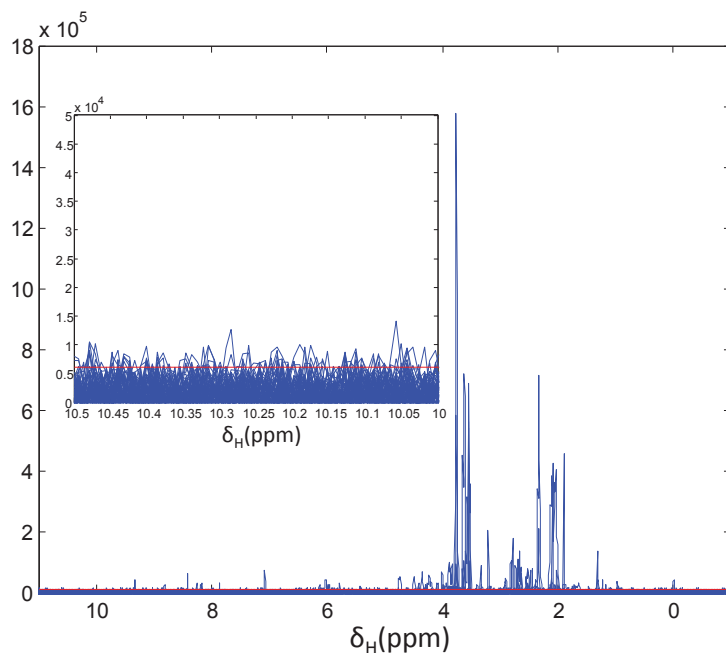


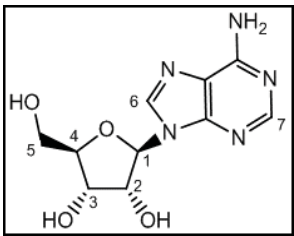
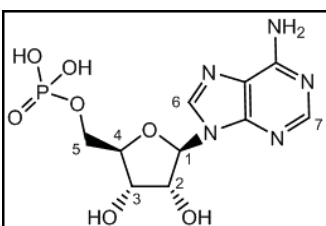
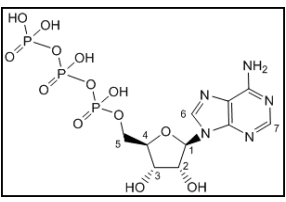
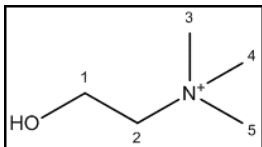
Figure S1. 2D NMR spectra plotted using only ppm1 (δ_H) dimension. The chosen threshold is indicated with a red line.

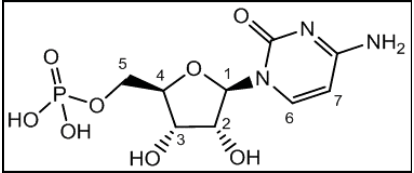
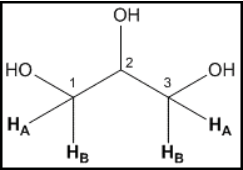
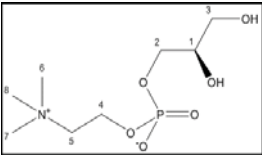
On the other hand, to define the minimum number of adjacent points that define a NMR peak (*minvoi*), 2D NMR spectra were opened under MestreNova software (Mestrelab Research S.L.), and the smallest detectable peak was selected. Afterwards, under the MATLAB environment, this peak was selected, and the number of variables (or pixels) that define this peak were counted. The optimal *minvoi* parameter corresponds to this count, adjusted by a factor between 0.7 and 1. This factor is applied to prevent the filtering of peaks not visualized in MestreNova because of their small size.

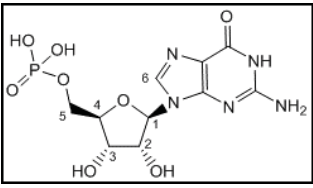
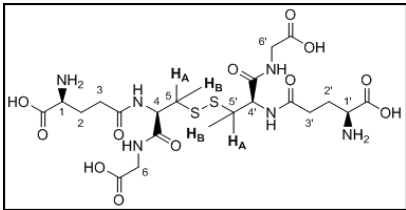
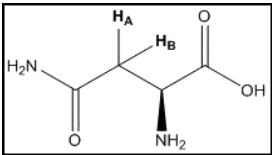
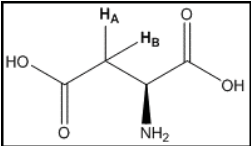
A more exhaustive description of this method can be found in the original manuscript (DOI: 10.1039/C7CC09891J).

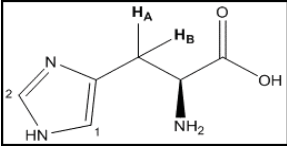
SUPPLEMENTARY TABLES

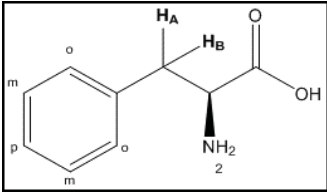
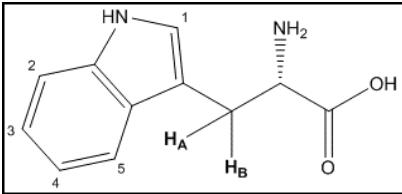
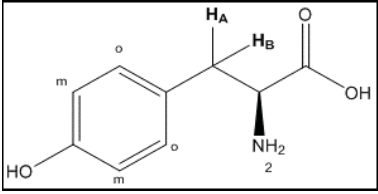
Table S1. Resonance assignment for ^1H NMR and ^1H - ^{13}C HSQC NMR data.

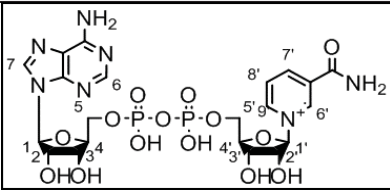
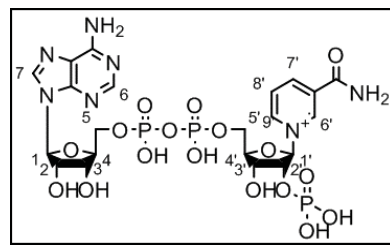
#	Metabolite	Group	δ_{H} (ppm)	δ_{C} (ppm)	Proton multiplicity	J_{HH} (Hz)
1	Acetic acid (HMDB00042)	CH ₃	1.903	26.00	s	
2	Adenosine (HMDB00050) 	1 - CH	6.058	91.02	d	6.0
		2 - CH	4.73	76.67	na. in ^1H NMR	
		3 - CH	4.42	73.47	na. in ^1H NMR	
		4 - CH	4.28	88.59	na. in ^1H NMR	
		5 - CH ₂	3.91	64.38	na. in ^1H NMR	
			3.85	64.38	na. in ^1H NMR	
		6 - CH	8.328	143.39	s	
7 - CH	8.160	155.55	s			
3	AMP (HMDB00045) 	1 - CH	6.094	89.18	d	6.1
		2 - CH	4.77	77.27	na. in ^1H NMR	
		3 - CH	4.51	73.28	na. in ^1H NMR	
		4 - CH	4.37	87.14	na. in ^1H NMR	
		5 - CH ₂	4.03	66.20	na. in ^1H NMR	
		6 - CH	8.564	142.8	s	
		7 - CH	8.160	155.55	s	
4	ATP (HMDB00538) 	1 - CH	6.153	89.25	na. in ^1H NMR	
		2 - CH	4.42	86.6	na. in ^1H NMR	
		3 - CH	4.82	76.96	na. in ^1H NMR	
		4 - CH	4.65	72.96	na. in ^1H NMR	
		5 - CH ₂	4.31	67.78	na. in ^1H NMR	
			4.25	67.71	na. in ^1H NMR	
		6 - CH	8.535	142.63	s	
7 - CH	8.240	155.00	s			
5	Betaine (HMDB00043)	CH ₂	3.887	68.98	s	
		3 x CH ₃	3.253	56.09	s	
6	Choline (HMDB00097) 	1 - CH ₂	4.049	58.26	na. in ^1H NMR	
		2 - CH ₂	3.507	70.14	na. in ^1H NMR	
		3 - CH ₃ , 4 - CH ₃ , 5 - CH ₃	3.190	56.61	s	
7	Citraconic acid (HMDB000634)	CH ₃	1.910	23.22	na. in ^1H NMR	
		CH	5.503	123.00	d	
8	Citric acid (HMDB00094)	CH _A	2.520	48.03	d	15.4
			2.551	48.00		

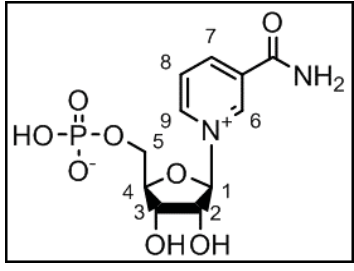
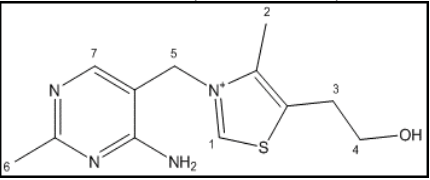
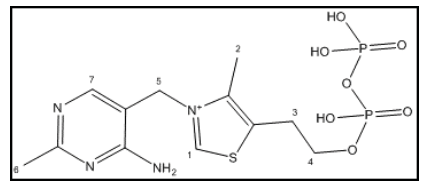
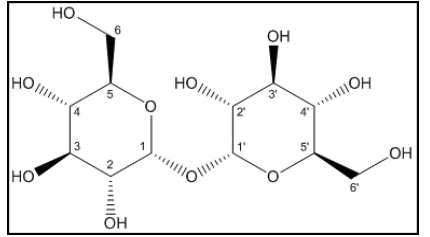
#	Metabolite	Group	δ_H (ppm)	δ_C (ppm)	Proton multiplicity	J_{HH} (Hz)
		CH _B	2.642	47.98	d	15.4
			2.674	48.04		
9	CMP (HMDB00095) 	1 - CH	5.980	91.02	m	
		2 - CH	4.33	72.38	na. in ¹ H NMR	
		3 - CH	4.33	77.14	na. in ¹ H NMR	
		4 - CH	4.23	86.05	na. in ¹ H NMR	
		5 - CH ₂	4.05	65.63	na. in ¹ H NMR	
			3.97	65.59	na. in ¹ H NMR	
		6 - CH	8.080	144.71	d	8.1
7 - CH	6.110	99.31	d	8.00		
10	DSS	α CH ₂	2.908	56.97	m	
		β CH ₂	1.754	21.72	m	
		γ CH ₂	0.625	17.59	m	
		δ CH ₃ , δ' CH ₃ , δ'' CH ₃	0.000	0.00	s	
11	Ethanol (HMDB00108)	CH ₃	1.170	19.5	t	7.1
		CH ₂	3.640	60.13	q	7.1
12	GABA (HMDB00112)	α CH ₂	2.284	37.11	t	7.4
		β CH ₂	1.892	26.34	m	7.4
		γ CH ₂	3.002	41.98	t	7.6
13	Glycerol (HMDB00131) 	1 - CH _B , 3 - CH _B	3.549	65.32	dd	11.7, 6.5
		1 - CH _A , 3 - CH _A	3.635	65.18	dd	11.7, 4.4
		2 - CH	3.767	74.85	tt	6.5, 4.4
14	Glycerophosphocholine (HMDB00086) 	1 - CH	3.902	73.35	na. in ¹ H NMR	
		2 - CH	3.863	69.23	na. in ¹ H NMR	
			3.939	69.23	na. in ¹ H NMR	
		3 - CH	3.662	64.74	na. in ¹ H NMR	
		4 - CH ₂	4.312	62.19	na. in ¹ H NMR	
		5 - CH ₂	3.666	68.73	na. in ¹ H NMR	
6 - CH ₃ , 7 - CH ₃ , 8 - CH ₃	3.214	56.69	s			
15	Glycine (HMDB00123)	α CH ₂	3.547	44.15	s	

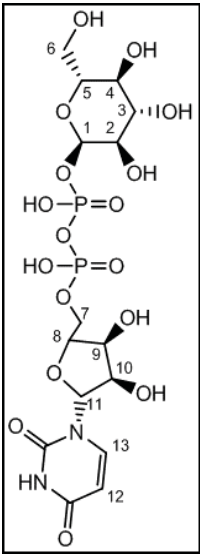
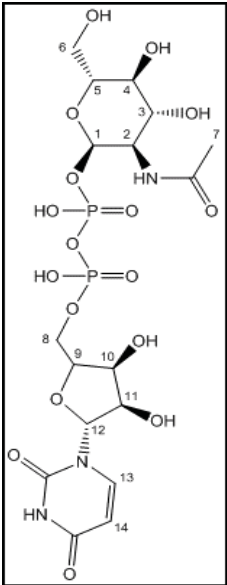

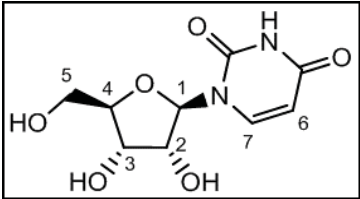
#	Metabolite	Group	δ_H (ppm)	δ_C (ppm)	Proton multiplicity	J_{HH} (Hz)
16	GMP (HMDB01397) 	1 - CH	5.914	89.27	d	6.4
		2 - CH	4.74	76.79	na. in 1H NMR	
		3 - CH	4.48	73.42	na. in 1H NMR	
		4 - CH	4.329	86.16	na. in 1H NMR	
		5 - CH ₂	3.98	66.13	na. in 1H NMR	
		6 - CH	8.192	140.59	s	
17	GSSG (bmse000170) 	1 - CH, 1' - CH	3.77	56.83	na. in 1H NMR	
		2 - CH ₂ , 2' - CH ₂	2.147	28.93	m	
		3 - CH ₂ , 3' - CH ₂	2.519	34.11	m	
		4 - CH, 4' - CH	4.563	58.34	dd	7.0, 5.2
		5 - CH _A , 5' - CH _A	2.970	41.76	m	
		5 - CH _B , 5' - CH _B	3.28	41.40	na. in 1H NMR	
			3.31	41.41	na. in 1H NMR	
6 - CH ₂ , 6' - CH ₂	3.76	46.15	na. in 1H NMR			
18	L-alanine (HMDB00161)	α CH	3.766	53.21	q	7.2
		β CH ₃	1.470	18.85	d	7.1
19	L-arginine (HMDB00517)	α CH	3.747	57.26	t	6.1
		β CH ₂	3.229	43.19	t	6.9
		γ CH ₂	1.653	26.57	m	
			1.709	26.62	m	
δ CH ₂	1.900	30.95	m			
20	L-asparagine (HMDB00168) 	α CH	3.994	53.99	dd	7.7, 4.3
		CH _A	2.833	37.11	dd	16.9, 7.7
			2.873	37.14		
		CH _B	2.927	37.15	dd	16.9, 4.3
2.961	37.15					
21	L-aspartic acid (HMDB00191) 	α CH	3.884	54.94	dd	8.8, 3.8
		β CH _A	2.822	39.25	dd	17.4, 3.8
			2.785	39.22		
		β CH _B	2.684	39.21	dd	17.4, 8.8
			2.667	39.24		
			2.654	39.27		
2.630	39.30					
22	L-glutamic acid (HMDB00148)	α CH	3.75	57.24	dd	7.2, 4.7
		β CH ₂	2.084	29.68	m	
		γ CH ₂	2.340	36.19	m	

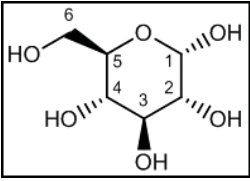
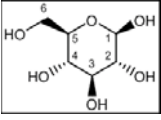
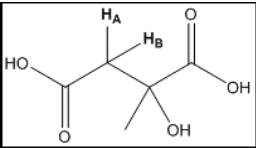
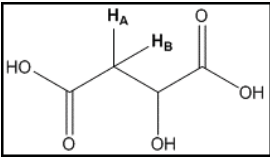
#	Metabolite	Group	δ_H (ppm)	δ_C (ppm)	Proton multiplicity	J_{HH} (Hz)	
23	L-glutamine (HMDB00641)	α CH	3.75	57.36	m	7.5, 3.7	
		β CH ₂	2.122	28.95	m		
		γ CH ₂	2.439	33.56	td		
24	L-histidine (HMDB00177) 	α CH	3.994	57.11	na. in ¹ H NMR		
		CH _A	3.205	30.03	na. in ¹ H NMR		
		CH _B	3.261	30.01	na. in ¹ H NMR		
			3.292	30.01	na. in ¹ H NMR		
		1 - CH	7.068-7.151	119.83	s		
		2 - CH	7.829-8.060	138.20	s		
25	L-isoleucine (HMDB00172)	α CH	3.661	62.27	d	4.0	
		β CH	1.966	38.60	m		
		γ CH ₃	0.999	17.39	d	7.4	
		δ CH ₂	1.461	27.18	m		
		ϵ CH ₃	0.927	13.70	t	7.4	
26	L-lactic acid (HMDB00190)	CH ₃	1.313	23.05	d	7.0	
		CH	4.103	71.33	q	6.9	
27	L-leucine (HMDB00687)	α CH	3.720	56.15	m		
		β CH ₂	1.698	42.50	m		
		γ CH	1.706	26.64	m		
		δ CH ₃	0.943	23.62	d		6.1
		δ' CH ₃	0.953	24.73	d		6.1
28	L-lysine (HMDB00182)	α CH	3.747	57.27	t	6.1	
		β CH ₂	1.894	32.59	m		
		γ CH ₂	1.460	24.14	m		
		δ CH ₂	1.716	29.10	m		
		ϵ CH ₂	3.014	41.82	m		
29	L-methionine (HMDB00696)	α CH	3.851	56.48	m		
		β CH ₂	2.17	nd.	na. in ¹ H NMR		
		δ CH ₃	2.120	16.62	s		
		γ CH ₂	2.633	31.50	t		7.5
30	L-ornithine (HMDB00214)	α CH	3.769	56.83	m		
		δ CH ₂	3.043	41.57	t		7.6
		β CH ₂	1.913	30.24	m		
		γ CH ₂	1.787	25.48	m		

#	Metabolite	Group	δ_H (ppm)	δ_C (ppm)	Proton multiplicity	J_{HH} (Hz)
31	L-phenylalanine (HMDB00159) 	α CH	3.985	58.72	na. in 1H NMR	
		CH _A	3.098	39.12	na. in 1H NMR	
			3.139	39.09		
		CH _B	3.26	39.05	na. in 1H NMR	
			3.288	39.05		
		2 x $^{\circ}CH$	7.318	132.04	m	
2 x mCH	7.413	131.83	m			
pCH	7.361	130.38	m			
32	L-proline (HMDB00162)	α CH	4.119	63.93	dd	8.6, 6.4
		β CH ₂	2.066	31.69	na. in 1H NMR	
			2.333	31.69	na. in 1H NMR	
		γ CH ₂	1.990	26.47	na. in 1H NMR	
		δ CH ₂	3.335	48.78	na. in 1H NMR	
3.400	48.78		na. in 1H NMR			
33	L-serine (HMDB00167)	α CH	3.833	59.09	dd	5.6, 3.8
		β CH ₂	3.963	62.88	m	
34	L-threonine (HMDB00167)	α CH	3.575	63.15	d	4.9
		β CH	4.243	68.62	m	
		γ CH ₃	1.319	22.16	d	6.6
35	L-tryptophan (HMDB00929) 	α CH	4.048	58.26	na. in 1H NMR	
		CH _A	3.46	nd.	na. in 1H NMR	
		CH _B	3.29	nd.	na. in 1H NMR	
		1 - CH	7.307	127.76	s	
		2 - CH	7.513	114.63	d	8.2
		3 - CH	7.711	121.12	d	8.0
		4 - CH	7.182	122.09	m	
5 - CH	7.265	124.78	m			
36	L-tyrosine (HMDB00158) 	α CH	3.925	58.76	na. in 1H NMR	
		CH _A	3.057	38.13	na. in 1H NMR	
		CH _B	3.202	38.16	na. in 1H NMR	
			3.172	38.19	na. in 1H NMR	
		2 x mCH	6.884	118.54	d	8.4
2 x $^{\circ}CH$	7.179	133.46	d	8.4		
37	L-valine (HMDB00883)	α CH	3.599	63.08	d	4.3
		β CH	2.258	31.78	m	
		γ CH ₃	0.981	19.36	d	7.1
		γ' CH ₃	1.037	20.66	d	7.1
38	NAD ⁺ (HMDB00092)	1 - CH	6.026	89.36	d	5.9
		2 - CH	4.37	86.46	na. in 1H NMR	

#	Metabolite	Group	δ_H (ppm)	δ_C (ppm)	Proton multiplicity	J_{HH} (Hz)
		3 - CH	4.51	73.03	na. in 1H NMR	
		4 - CH	4.76	76.69	na. in 1H NMR	
		5 - CH ₂	4.24	68.05	na. in 1H NMR	
			4.21	68.05	na. in 1H NMR	
		6 - CH	8.134	155.44	s	
		7 - CH	8.411	142.50	s	
		1' - CH	6.075	102.75	d	5.3
		2' - CH	4.49	80.34	na. in 1H NMR	
		3' - CH	4.43	73.36	na. in 1H NMR	
		4' - CH	4.55	89.63	na. in 1H NMR	
		5' - CH ₂	4.36	67.55	na. in 1H NMR	
			4.24	67.55	na. in 1H NMR	
		6' - CH	9.322	142.64	s	
		7' - CH	8.816	148.44	d	8.1
		8' - CH	8.183	131.33	d	6.1
9' - CH	9.14	145.00	d	5.3		
39	NADP (HMDB00237)	1 - CH	6.093	89.19	d	5.9
	2 - CH	4.37	85.82	na. in 1H NMR		
	3 - CH	4.61	72.71	na. in 1H NMR		
	4 - CH	4.97	78.83	na. in 1H NMR		
	5 - CH ₂	4.28	68.05	na. in 1H NMR		
		4.19	68.05	na. in 1H NMR		
	6 - CH	8.090	155.20	s		
	7 - CH	8.390	142.87	s		
	1' - CH	6.030	102.68	m		
	2' - CH	4.45	80.33	na. in 1H NMR		
	3' - CH	4.41	73.47	na. in 1H NMR		
	4' - CH	4.49	89.69	na. in 1H NMR		
	5' - CH ₂	4.32	67.65	na. in 1H NMR		
		4.21	67.54	na. in 1H NMR		
	6' - CH	9.286	142.64	s		
	7' - CH	8.800	148.34	m		
8' - CH	8.170	131.30	m			
9' - CH	9.100	145.04	d	5.3		

#	Metabolite	Group	δ_H (ppm)	δ_C (ppm)	Proton multiplicity	J_{HH} (Hz)
40	NMN (HMDB00229) 	1 - CH	6.191	103.01	d	5.8
		2 - CH	4.64	80.44	na. in 1H NMR	
		3 - CH	4.61	90.93	na. in 1H NMR	na. in 1H NMR
		4 - CH	4.47	74.03	na. in 1H NMR	
		5 - CH ₂	4.2	65.74	na. in 1H NMR	
			4.03	65.74	na. in 1H NMR	
		6 - CH	9.566	142.46	s	8.1
		7 - CH	8.994	149.01	d	
		8 - CH	8.309	131.22	m	
		9 - CH	9.326	145.81	d	5.9
41	Succinic acid (HMDB00254)	2 x CH ₂	2.395	36.81	s	
42	Taurine (HMDB00251)	α CH	3.233	50.06	t	6.1
		β CH	3.429	38.96	t	6.1
43	Thiamine (HMDB00235) 	1 - CH	9.443	nd.	s	
		2 - CH ₃	2.545	13.77	na. in 1H NMR	
		3 - CH ₂	3.164	31.99	na. in 1H NMR	
		4 - CH ₂	3.88	63.45	na. in 1H NMR	
		5 - CH ₂	5.433	53.69	s	
		6 - CH ₃	2.474	26.72	na. in 1H NMR	
		7 - CH	8.029	159.82	s	
44	Thiamine-PP (HMDB01372) 	1 - CH	9.429	nd.	s	
		2 - CH ₃	2.575	13.89	na. in 1H NMR	
		3 - CH ₂	3.164	31.99	na. in 1H NMR	
		4 - CH ₂	3.341	51.64	na. in 1H NMR	
		5 - CH ₂	5.413	53.81	s	
		6 - CH ₃	2.480	26.67	na. in 1H NMR	
		7 - CH	8.029	159.82	s	
45	Trehalose (HMDB00975) 	1 - CH, 1' - CH	5.184	95.94	d	3.8
		2 - CH, 2' - CH	3.635	73.76	m	
		3 - CH, 3' - CH	3.835	75.03	m	9.3
		4 - CH, 4' - CH	3.439	72.41	t	
		5 - CH, 5' - CH	3.768	74.78	m	
		6 - CH _A , 6' - CH _A	3.839	63.11	m	
		6 - CH _B , 6' - CH _B	3.762	63.16	m	

#	Metabolite	Group	δ_H (ppm)	δ_C (ppm)	Proton multiplicity	J_{HH} (Hz)
46	UDP-Glc (HMDB00286) 	1 - CH	5.590	98.32	dd	7.2, 3.5
		2 - CH	3.53	74.3	na. in 1H NMR	
		3 - CH	3.76	75.6	na. in 1H NMR	
		4 - CH	3.46	71.92	na. in 1H NMR	
		5 - CH	3.88	75.4	na. in 1H NMR	
		6 - CH2	3.78	63.13	na. in 1H NMR	
			3.86	63.13	na. in 1H NMR	
		7 - CH2	4.19	67.6	na. in 1H NMR	
			4.23	67.6	na. in 1H NMR	
		8 - CH	4.27	85.9	na. in 1H NMR	
		9 - CH	4.36	72.4	na. in 1H NMR	
		10 - CH	4.36	76.6	na. in 1H NMR	
		11 - CH	5.976	91.19	m	
12 - CH	5.975	105.4	m			
13 - CH	7.940	144.27	d	8.1		
47	UDP-Nac-glc-NH2 (HMDB00290) 	1 - CH	5.504	97.18	dd	7.2, 3.1
		2 - CH	3.53	74.3	na. in 1H NMR	
		3 - CH	3.76	75.6	na. in 1H NMR	
		4 - CH	3.46	71.92	na. in 1H NMR	
		5 - CH	3.88	75.4	na. in 1H NMR	
		6 - CH2	3.86	63.13	na. in 1H NMR	
		7 - CH3	2.102	23.03	na. in 1H NMR	
		8 - CH2	4.19	67.6	na. in 1H NMR	
			4.23	67.6	na. in 1H NMR	
		9 - CH	4.27	85.9	na. in 1H NMR	
		10 - CH	4.36	72.4	na. in 1H NMR	
		11 - CH	4.36	76.6	na. in 1H NMR	
		12 - CH	5.976	91.19	m	
		13 - CH	5.975	105.4	m	
14 - CH	7.94	144.27	d	8.1		
48	Uracil (HMDB00300) 	1 - CH	5.788	103.76	d	7.8
		2 - CH	7.52	146.15	d	7.7
49	Uridine (HMDB00296) 	1 - CH	5.904	92.08	d	4.6
		2 - CH	4.343	76.31	na. in 1H NMR	
		3 - CH	4.221	72.19	na. in 1H NMR	
		4 - CH	nd.	nd.	na. in 1H NMR	
		5 - CH	3.893	63.61	na. in 1H NMR	
		6 - CH	5.896	105.02	d	8.2

#	Metabolite	Group	δ_H (ppm)	δ_C (ppm)	Proton multiplicity	J_{HH} (Hz)	
		7 - CH ₂	7.870	144.57	d	8.5	
50	α -Glucose (HMDB00122) 	1 - CH	5.219	94.82	d	3.8	
		2 - CH	3.531	74.33	m		
		3 - CH	3.701	75.51	m		
		4 - CH	3.389	72.33	m		
		5 - CH	3.825	74.14	m		
		6 -CH ₂	3.815	63.22	m		
51	β -Glucose (HMDB00122) 	1 - CH	4.636	98.63	d	8.0	
		2 - CH	3.230	76.85	dd		7.8, 9.2
		3 - CH, 5 - CH	3.455	78.59	m		
		4 - CH	3.389	72.33	m		
		6 -CH ₂	3.725	63.48	m		
			3.894	63.33	m		
52	2-Isopropylmalate (HMDB00402)	CH ₃	0.843	nd.	d	6.9	
		CH ₃	0.896	nd.	d	6.9	
53	Fumaric acid (HMDB00134)	2 x CH	6.500	nd.	s		
54	Oxalacetic acid (HMDB00223)	CH ₂	2.375	nd.	s		
55	Choline-derivate	3 x CH ₃	3.208	nd.	s		
56	Citramalic acid (HMDB00426) 	CH _A	2.736	48.79	na. in ¹ H NMR		
			2.703	48.89			
		CH _B	2.467	48.68		na. in ¹ H NMR	
			2.439	48.77			
		CH ₃	1.317	27.95		na. in ¹ H NMR	
57	CYSSG (HMDB00656)	1 - CH, 1' - CH	3.77	56.83	na. in ¹ H NMR		
		2 - CH ₂ , 2' - CH ₂	2.147	28.93	m		
		3 - CH ₂ , 3' - CH ₂	2.519	34.11	m		
		4 - CH, 4' - CH	4.754	55.33	na. in ¹ H NMR		
		5 - CH _A , 5' - CH _A	2.97	41.76	m		
			3.28	41.40	na. in ¹ H NMR		
		5 - CH _B , 5' - CH _B	3.31	41.41	na. in ¹ H NMR		
			3.761	46.15	na. in ¹ H NMR		
58	L-malic acid (HMDB00156) 	α CH	4.275	73.11	na. in ¹ H NMR		
		β CH _A	2.643	45.28	na. in ¹ H NMR		
			2.675	45.30	na. in ¹ H NMR		
		β CH _B	2.326	45.41	na. in ¹ H NMR		
			2.38	45.32	na. in ¹ H NMR		

#	Metabolite	Group	δ_{H} (ppm)	δ_{C} (ppm)	Proton multiplicity	J_{HH} (Hz)
59	Pyroglutamic acid (HMDB00267)	α CH	4.164	60.96	na. in ^1H NMR	
		β CH ₂	2.491	27.99	na. in ^1H NMR	
			2.041	27.99	na. in ^1H NMR	
		γ CH ₂	2.387	32.32	na. in ^1H NMR	
60	β -alanine (HMDB00056)	α CH ₂	2.541	36.24	na. in ^1H NMR	
		β CH ₂	3.165	39.32	na. in ^1H NMR	

nd.: Not detected

na. in ^1H NMR: not assigned in ^1H NMR

Table S2. Correlation coefficients between integrals from the ^1H NMR analysis and the ^1H - ^{13}C HSQC NMR analysis.

Metabolite name	^1H NMR	^1H - ^{13}C HSQC NMR	Correlation
2-isopropylmalate	x		-
Acetic acid	x	x	0.696
Adenosine	x	x	*
AMP	x	x	0.603
ATP	x	x	*
b-alanine		x	-
Betaine	x	x	0.945
Choline	x	x	*
Citraconic acid	x	x	0.852
Citramalic acid		x	-
Citric acid	x	x	0.075
CMP	x	x	*
Cys-GSH		x	-
Ethanol	x	x	0.995
Fumaric acid	x		-
GABA	x	x	0.914
Glycerol	x	x	0.940
Glycerophosphocholine	x	x	0.784
Glycine	x	x	0.956
GMP	x	x	*
GSSG	x	x	0.305
L-Alanine	x	x	0.970
L-arginine	x	x	0.546
L-Asparagine	x	x	0.015
L-Aspartic acid	x	x	0.490
L-Glutamic acid	x	x	0.842
L-Glutamine	x	x	0.991
L-Histidine	x	x	0.158
L-Isoleucine	x	x	0.965
L-Lactic acid	x	x	0.741
L-Leucine	x	x	0.847
L-Lysine	x	x	0.926
L-Methionine	x	x	0.776
L-Ornithine	x	x	0.652
L-Phenylalanine	x	x	0.939
L-Proline	x	x	0.586
L-Serine	x	x	0.108
L-Threonine	x	x	0.888
L-Tryptophan	x		-
L-Tyrosine	x	x	0.861

Metabolite name	¹ H NMR	¹ H- ¹³ C HSQC NMR	Correlation
L-Valine	x	x	0.991
Malic acid		x	-
NAD	x	x	0.910
NADP	x	x	*
NMN	x	x	0.081
Oxalacetic acid	x		-
Pyroglutamic acid		x	-
Succinic acid	x	x	0.919
Taurine	x	x	0.428
Thiamine	x	x	*
ThiaminePP	x	x	*
Trehalose	x	x	0.611
UDP-Glc	x	x	*
UDP-Nac-glc-NH ₂	x	x	*
Uracil	x	x	*
Uridine	x	x	*
α-Glucose	x	x	*
β-Glucose	x	x	*

* Detected in the ¹H-¹³C HSQC NMR only in a few samples

SUPPLEMENTARY FIGURES

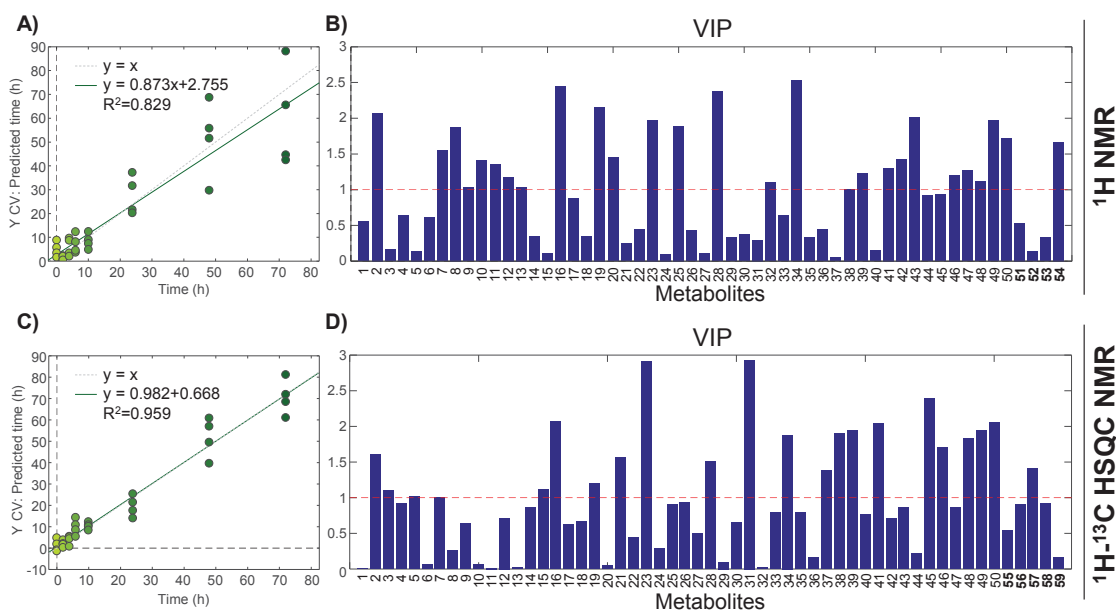


Figure S2. PLSR of YSC-cultured samples. A) and C) Regression between experimental and predicted collection times for the (A) ^1H NMR dataset and the (C) ^1H - ^{13}C HSQC NMR dataset. B) and D) VIP values for the variables in the (B) ^1H NMR dataset and in the (D) ^1H - ^{13}C HSQC NMR dataset. Listed metabolites: 1 acetic acid, 2 adenosine, 3 AMP, 4 ATP, 5 betaine, 6 choline, 7 citraconic acid, 8 citric acid, 9 CMP, 10 ethanol, 11 GABA, 12 glycerol, 13 glycerophosphocholine, 14 glycine, 15 GMP, 16 GSSG, 17 L-alanine, 18 L-arginine, 19 L-asparagine, 20 L-aspartic acid, 21 L-glutamic acid, 22 L-glutamine, 23 L-histidine, 24 L-isoleucine, 25 L-lactic acid, 26 L-leucine, 27 L-lysine, 28 L-methionine, 29 L-ornithine, 30 L-phenylalanine, 31 L-proline, 32 L-serine, 33 L-threonine, 34 L-tryptophan, 35 L-tyrosine, 36 L-valine, 37 NAD^+ , 38 NADP, 39 NMN, 40 succinic acid, 41 taurine, 42 thiamine, 43 thiaminePP, 44 trehalose, 45 UDP-Glc, 46 UDP-Nac-glc-NH₂, 47 uracil, 48 uridine, 49 α -glucose, 50 β -glucose, 51 2-(S)-isopropylmalic acid, 52 choline-derivative, 53 fumaric acid, 54 oxalacetic acid, 55 citramalic acid, 56 CYSSG, 57 malic acid, 58 pyroglutamic acid, 59 β -alanine.

For the PLSR of **Figure S2A-B** (using resonance integrals from 1D data), 97.19% of the y-variance was explained by the 56.87% of the X-variance. For the PLSR of **Figure S2C-D** (using resonance integrals from 2D data), 98.58% of the y-variance was explained by 62.10% of the X-variance.

3 DISCUSSION OF THE RESULTS

3.1 MCR-ALS AND NMR DATA

MCR-ALS has been shown to be a powerful chemometric approach to resolve the NMR spectral profiles (fingerprints) of the chemical compounds from their mixtures. However, the quality of this performance depends on (i) the type of analyzed NMR data, and (ii) how the concentrations of the metabolites change in the analyzed dataset.

As commented in the introduction section of this Chapter, with chemometrics, 1D NMR spectra of the constituent compounds can be resolved from the 2D NMR spectra of their mixtures. However, this is only true for a few types of 2D NMR spectra. In particular, the analyzed 2D NMR spectra must show, for the measured compounds, cross-peak correlations among all intramolecular resonances in at least one of the two dimensions. Examples of 2D NMR spectra that fulfill this condition are ^1H - ^1H TOCSY NMR (**Fig. 4.4**) and DOSY NMR (**Fig. 4.5B**) spectra. Conversely, when MCR-ALS is applied to some other type of 2D NMR spectra, it will resolve every resonance in a separate component.

MCR-ALS can also be applied to different type of data arrays of ^1H NMR spectra, such as PGSE NMR (**Fig. 4.5A**) or ^1H NMR spectra of mixture samples. However, these two type of NMR spectra are not directly comparable.

PGSE NMR spectra give a data matrix for every single sample, and every row from each of these data matrices corresponds to a 1D NMR spectrum from the sample using a different magnetic field gradient strength during the spectra acquisition. In the PGSE NMR spectra, the detected resonances decay with increasing magnetic field gradient strengths at different rates that depend on the diffusion properties of the corresponding compounds (**Fig. 4.5A**).

On the other hand, ^1H NMR spectra give a data vector for every sample and, to arrange a data matrix of ^1H NMR spectra, two or more samples are appended column-wisely (**Fig 2.27**, see page 52). Thus, in this data matrix, the differences among rows came from the differences in sample composition among samples.

To resolve by MCR-ALS a chemical mixture measured using any of these two NMR pulse sequences, each one of the chemical compounds must give an independent evolution along the acquired ^1H NMR spectra. This means that, in the MCR-ALS analysis of PGSE NMR datasets, the ^1H NMR of the different chemical species will only be satisfactorily resolved if they have different diffusion properties. On the other hand, for the direct MCR-ALS analysis of arrays of ^1H NMR spectra, a good resolution can only be achieved if the concentrations of the observed metabolites are changing independently.

The analysis of PGSE NMR spectral data arrays results much easier than the equivalent analysis of the ^1H NMR spectra of mixtures arranged in a data matrix. This occurs because, in the PGSE NMR arrays, resonances do not present variations in chemical shifts among the acquired spectra, since all 1D NMR spectra are obtained from the same sample.

On the other hand, the analysis of data arrays of ^1H NMR spectra can be more problematic. In order to resolve the pure ^1H NMR spectra of the metabolites from their mixture, the experimental workflow analysis has to be applied consistently for all the samples. Moreover, if samples are not prepared using similar protocols, a resonance may appear at different chemical shifts in the different samples and this makes more difficult their analysis. Subtle variations of environmental variables (*e.g.*, ionic strength, pH) may also produce a change in the structure of the metabolites and, as a result, their resonances appear shifted. Having said this, in metabolomics studies, NMR samples are prepared following robust protocols in order to minimize the effect of these environmental variables in the chemical shifts.

In situations where shifted resonances are present, extra MCR-ALS (bilinear) components will be needed to represent all the shifting variants of the same ^1H NMR spectra of a pure metabolite [358]. Most of the times, this drawback can be circumvented by applying peak alignment algorithms [102], although the performance of these algorithms is dependent on how misaligned are the resonances and on the degree of overlapping of these resonances.

It is also necessary to use the same NMR pulse sequence with the same acquisition and processing parameters for all the samples. Otherwise, for the same metabolic abundances, different intensity values will be obtained in the different recorded spectra.

For instance, for the acquisition of a typical ^1H NMR spectrum, the same relaxation delay (RD, which corresponds to the time left for all nuclei to align with the magnetic field) should be used. Moreover, for this acquisition parameter, it is important to use an RD that allows the full relaxation of all nuclei. Otherwise, each nucleus will show a different correspondence between the real concentration and the measured intensity, and the analysis could not be considered as inherently quantitative anymore. As stated in [section 3.5.5 of Chapter 2](#), for ^1H NMR metabolomics experiments, an RD of at least 5 seconds is recommended. This difference in intensity response associated with different RD values is highlighted in [Figure 4.6](#).

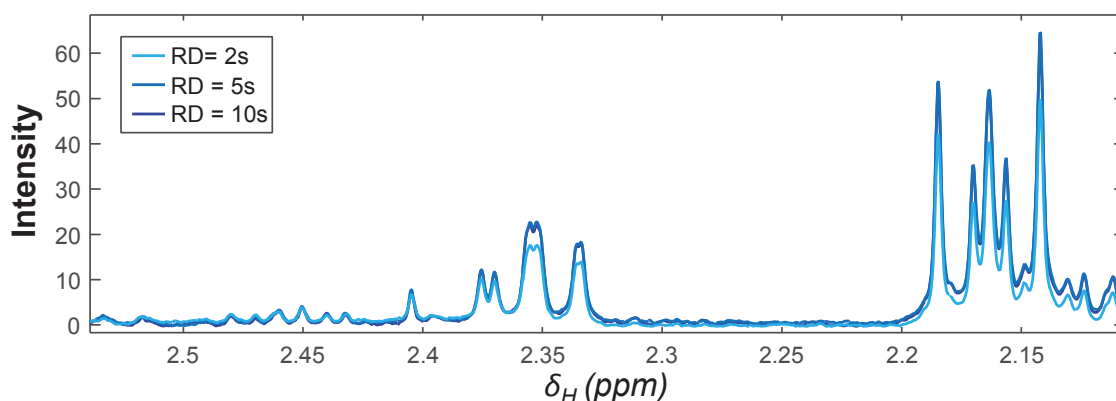


Figure 4.6. A ^1H NMR spectra of a yeast extract acquired using different relaxation delays (RD). A RD of 2 seconds does not allow the full relaxation of all proton nuclei.

3.2 MCR-ALS AS A BIOMARKER DETECTION TOOL

MCR-ALS has been successfully used to extract the concentrations and spectra of pure compounds from UV [359], NIR [360], and HPLC-MS [206], among other analytical techniques.

For (^1H) NMR data, MCR-ALS has shown promising results in the study of acid-base equilibria [358]. This good performance is achieved when the concentration of the different chemical species varies significantly along all the screened samples and special measures are taken for the lability of the proton resonances [131] when pH is changed.

For other datasets with low variance of metabolite concentrations, the best results were obtained after application of appropriate spectral window constraints [356,357] (Fig. 2.42, page 72). With this type of constraints, MCR rotation ambiguities are reduced and the resolution is improved since the number of possible metabolites per spectral window is limited.

In Scientific Article IV, we have investigated the resolution of a complex NMR metabolomics system using MCR-ALS combined with spectral window constraints. In order to define these constraints, a set of predefined windows is first selected, and MCR-ALS is then applied to every one of these windows. The resolved \mathbf{S}^T profiles (eq. 2.15) in these MCR-ALS analyses should be assignable to pure metabolite resonances. Nevertheless, in some cases, the resolved \mathbf{S}^T profiles contained mixed contributions from different metabolites, because resonances from different metabolites could not be properly resolved into separate components. To circumvent this situation, the considered spectral window should be divided into two or more smaller spectral windows. This process is represented in Figure 4.7.

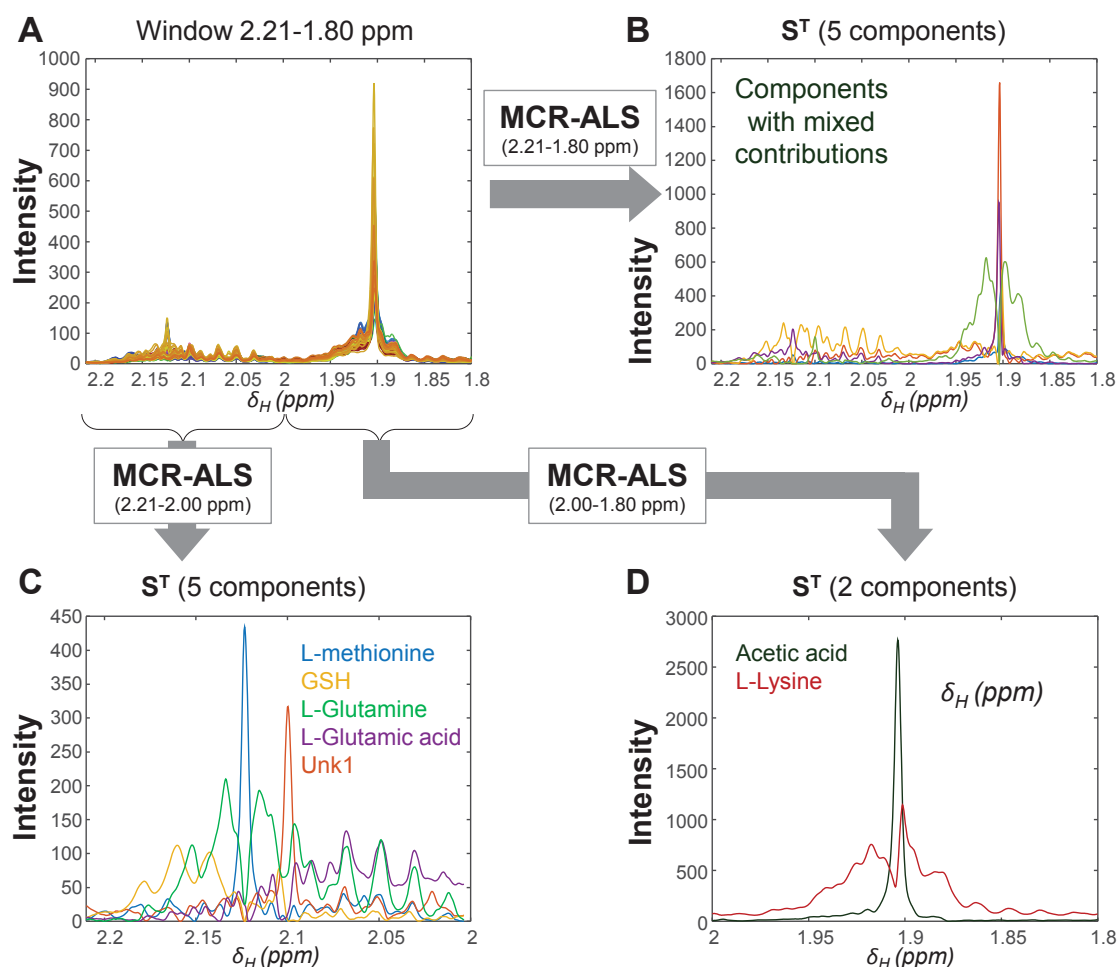


Figure 4.7. MCR-ALS of one spectral window ($\delta_H = 2.21\text{--}1.80$ ppm). **A)** ^1H NMR window from a metabolomics dataset consisting of 90 samples [243]. **B)** Resolved \mathbf{s}_k do not correspond to the expected resonances. The increase of the number of components does not improve MCR-ALS resolution. **C-D)** Splitting the window into two smaller windows ($\delta_H = 2.21\text{--}2.00$ ppm, and $\delta_H = 2.00\text{--}1.80$ ppm) allows a more comprehensive MCR-ALS resolution. Resolved \mathbf{s}_k were then tentatively assigned to metabolites.

This process is very time-demanding since, for the analysis of all the ^1H NMR spectral domain, dozens of windows may be needed. In addition, for each window, MCR-ALS must be tested with a different number of components in order to identify the optimal number.

For the studied metabolomics dataset in [243], 56 spectral windows were needed to separate all the resonances. Moreover, in case the metabolic variance is lower, even more spectral windows could be needed. For example, in another metabolomics study (analyzing zebrafish extracts, not published yet) presenting a similar amount of detected resonances, the number of windows needed was 98 because the changes of the concentration of the metabolites among samples were lower.

Since the application of MCR-ALS to different NMR spectral windows can provide good estimates of the pure resonances (\mathbf{S}^T matrix in [eq. 2.15](#)) and of the pure concentrations (\mathbf{C} matrix in [eq. 2.15](#)), this chemometric method has been applied as a resonances integration tool [361]. However, in [362], it was already stated that resonances non-perfectly aligned (for instance, using *icoshift* correction [102]) could not be integrated via this method. For shifted resonances that present no or low overlapping, resonance integral estimates can be also obtained through the calculation of the second derivative [362]. For highly overlapped and shifted resonances, resonance integral estimates can be estimated using deconvolution approaches or using NMR pulse sequences that separate resonances from the confounding metabolites.

In the metabolomics dataset examined in Scientific Article IV, the spectral data were correctly aligned, and therefore, this dataset could be properly decomposed as a set of spectral features (\mathbf{s}_k in Scientific Article IV) and their corresponding set of concentrations for each sample (\mathbf{c}_k).

From the set of \mathbf{c}_k vectors (concentration profiles), it is possible to construct a row-wise augmented data matrix, with as many rows as analyzed samples and as many columns as resolved k features. Since the \mathbf{c}_k vectors relative to resonances from the same metabolite will be highly correlated, it is possible to investigate this correspondence by the application of a correlation-based approach. For instance, if the augmented matrix of \mathbf{c}_k vectors is analyzed with hierarchical clustering, the \mathbf{c}_k vectors clustered together are likely to be from the same metabolite. [Figure 4.8](#) shows that 6 features ($k= 57, 61, 63, 70, 73,$ and 90) have their \mathbf{c}_k vectors clustered in [Figure 4.8A](#) and [4.8B](#). When the \mathbf{s}_k profiles from these 6 k features are combined in the same spectral profile, the reconstructed ^1H NMR spectrum is confirmed to correspond to glucose ([Fig. 4.8C](#)).

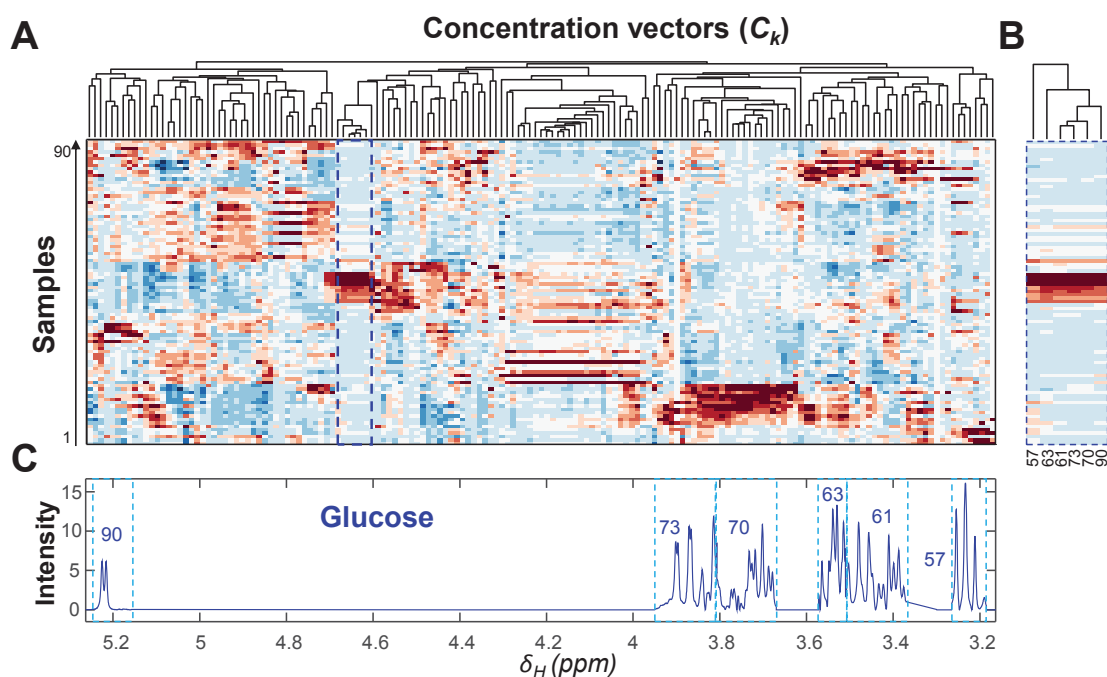


Figure 4.8. **A)** Heat-map representation with hierarchical clustering of the augmented matrix of \mathbf{c}_k vectors. Clustering of 6 \mathbf{c}_k vectors is highlighted with a dashed blue line. **B)** Zoom-in of the region highlighted in **A**. Clustered \mathbf{c}_k vectors are 57, 61, 63, 70, 73, and 90. **C)** The combination of \mathbf{s}_k vectors with $k=57, 61, 63, 70, 73,$ and 90 allow the reconstruction of the ^1H NMR spectrum of glucose.

Whether two or more features are from the same metabolite can be sometimes difficult to determine from the hierarchical clustering. In order to assist in this decision process, we have proposed the method called Decision Tree of Correlations (or DTC). This method consists in the calculation of the pair-wise correlation among all \mathbf{c}_k vectors, and the sequential clustering of these \mathbf{c}_k vectors from the more correlated ones to the least correlated ones. This operation is performed repeatedly until the next \mathbf{c}_k vector to be chosen is not linked to the same metabolite as the already clustered \mathbf{c}_k vectors. This can be assessed by comparison of the corresponding \mathbf{s}_k spectral features. For example, if the next \mathbf{c}_k vector to be chosen is associated to a \mathbf{s}_k descriptive of noise, or descriptive of a resonance with intensities in a different order or magnitude than the resonances of the previously grouped \mathbf{s}_k spectral features, then this considered feature is not included in the list of the previously clustered features. At this point, a new DTC cluster analysis is initiated with the remaining features, and this process is repeated until all features are grouped. A more profound explanation of this method is given in [section 2.3.2](#) and in [Figure 3](#) from Scientific Article IV.

This approach is different than the standard workflow used in NMR metabolomics. Typically, in NMR metabolomics, the resonance assignment is first carried out, and only the assigned resonances are integrated. This is commonly referred as targeted analysis. However,

when the MCR-ALS method is used, it is possible to first detect which resonances are important for describing the biological system (which resonances show variations in intensity among the screened conditions), and it is then possible to perform their assignment. Thus, this approach should be considered a non-targeted approach.

Other non-targeted analyses for NMR data can be proposed, such as the application of bilinear decomposition methods (*e.g.*, PCA, PLS) to ^1H NMR datasets, but they do not result as effective as the MCR-ALS approach. In the analysis by PCA or PLS of a ^1H NMR dataset, data is usually mean-centered or Pareto scaled with the aim of reducing noise contribution, but these two data pretreatments cause that not only noise but also that smaller resonances are underrepresented in the chemometric analysis.

On the other hand, in the MCR-ALS preprocessing analysis proposed in this Thesis, the relevance of every \mathbf{s}_k spectral feature or resonance is studied using their associated \mathbf{c}_k vectors resolved, which already contain the resonance integral values. Since these data have the noise filtered, they can be properly auto-scaled, and small resonances will become as important as large resonances in the PCA or PLS model. Therefore, this approach allows for a more comprehensive characterization of the studied biological system than the other direct non-targeted approaches in NMR metabolomics.

This proposed MCR-ALS approach shows the best performance when the metabolic variance is high. Therefore, it results a very powerful chemometric tool for the identification of possible metabolite biomarkers. Furthermore, despite being time-demanding, it is much faster than integrating one by one each of the detected resonances (for a ^1H NMR spectrum, which can be in the order of hundreds to thousands) for each sample since, in the MCR-ALS approach, the integral values for all samples and metabolites can be obtained in one single step.

3.3 MCR-ALS AS A RESONANCES INTEGRATION TOOL

As introduced in the previous section, MCR-ALS can be used to resolve resonance integrals since, for every feature, a vector of concentrations will be obtained. For instance, in **Figure 4.8**, 6 different (although very similar) concentration profiles were obtained for glucose metabolite.

In order to unify these concentration values, we performed a simultaneous MCR-ALS analysis of the whole ^1H NMR dataset (**Fig. 4.9**). To bypass the problem of the spectral ambiguities, spectral window constraints were imposed. These spectral window constraints

were designed using the knowledge gained during the preliminary screening of the ^1H NMR data.

In the previous analysis, resonances from each metabolite were resolved in different spectral windows. Thus, for every spectral window, every component (one metabolite) should be resolved only in this spectral window.

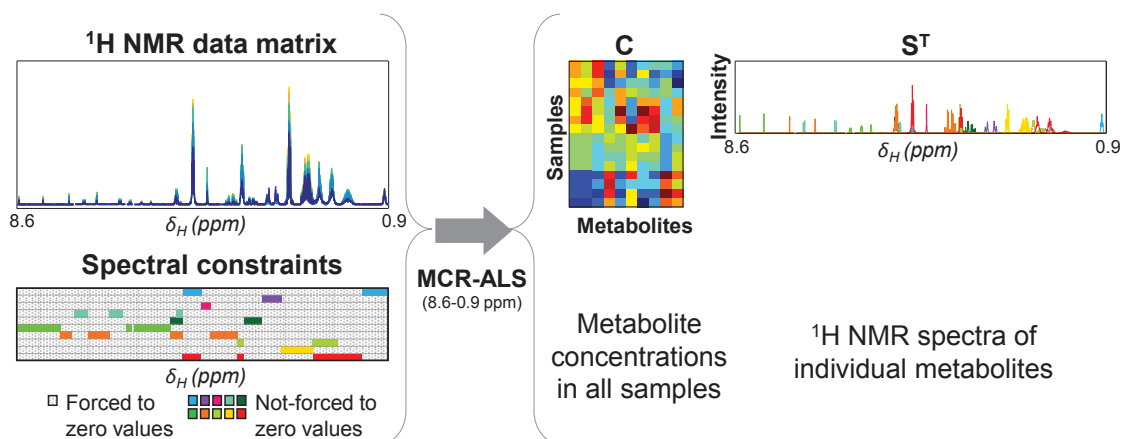


Figure 4.9. Simultaneous MCR-ALS analysis of the whole ^1H NMR dataset using spectral windows.

After the MCR-ALS analysis, not only one concentration value is obtained per sample and metabolite, but also all the s_k features from the same component will be joined together.

In Scientific Article IV, we tested this approach using 75 components. From the resolved \mathbf{S}^T , we were able to assign 39 metabolites. However, it is worth mentioning that not all the spectral ambiguities were completely removed, and that the ^1H NMR spectra for some of the metabolites were better resolved than for others.

On the other hand, the relative concentrations of these components, from the concentration profiles **C** matrix, were better recovered than the corresponding ^1H NMR spectra, in the \mathbf{S}^T matrix. This is because, as in more traditional NMR integration methods, the concentration of a given metabolite can be obtained from the analysis of just one resonance. Thus, even if the ^1H NMR spectrum of a metabolite is only partially resolved, if some of the resolved resonances coincide with the real ones in the true ^1H NMR spectrum of this metabolite, the corresponding concentrations associated to this metabolite will be close to the real ones. This can be easily seen in **Table 1** from the Scientific Article IV [243]. For example, in this table, one of the three resonances from glycerol could not be properly resolved by the MCR-ALS analysis of the whole ^1H NMR dataset. Nevertheless, since the other two resonances were perfectly resolved, the concentrations obtained with this approach coincided well with the correct ones ($r^2=0.994$).

3.4 NOISE INFLUENCES THE RANK OF 2D NMR DATA

Theoretically, any 2D NMR dataset could also be analyzed using the previously proposed DTC-MCR-ALS methodology. As a preliminary step, the set of 2D spectral windows should be first defined. Then, for every sample, the data from every window should be vectorized, and a column-wise augmented matrix (**Fig. 2.27**, see page 52) is built combining all the vectors from the same windows. Finally, these augmented matrices are analyzed by MCR-ALS.

As said above, one of the main reasons for using the DTC-MCR-ALS method is to overcome the problem of resonance overlapping observed in ^1H NMR data, very frequently encountered for instance in metabolomics studies. However, in 2D NMR data, resonance overlapping is not so strong as in ^1H NMR data.

During the analyses of 2D and 3D NMR datasets (Scientific Article V), we observed that, although some resonances can partly overlap, these overlapped resonances are clearly distinguishable from the rest. Moreover, in the ^1H - ^{13}C HSQC NMR spectra of metabolic extracts from yeast (Scientific Article VI), for all assigned metabolites, relative integrals could be straightforwardly obtained from isolated resonances. Thus, since resonance assignment and integration from 2D NMR data is much easier, we decided not to use the previously described DTC-MCR-ALS strategy for the analysis of this type of data.

Nevertheless, despite individual resonances are better separated in 2D NMR data, we observed that the structure of the 2D NMR data is by far more complex than for 1D NMR data. This observation was already considered in Jaumot *et al.* [160], where it was confirmed that resolution of 2D NMR data requires a higher number of components than the number of chemical species present in the system, indicating that 2D NMR data do not follow well the postulated bilinear model where every component refers to a single chemical species.

In the precedent work, it was defended that the number of components needed to reconstruct a single 2D NMR spectrum coincides with the number of cross-peaks resonances [343]. To corroborate this statement, we estimated the number of components for two distinct 2D NMR spectra (**Fig. 4.10A** and **Fig. 4.10B**) by means of SVD [363]. In these analyses, the estimated number of components was only close to the number of cross-peak resonances for the second case (circa 50 components, **Fig. 4.10F**). On the other hand, for the 2D NMR spectrum of **Fig. 4.10A**, the number of components deviated largely from the expected (estimated number of components ≈ 500 , **Fig. 4.10B**). These results suggest that the number of components is greatly conditioned by the non-bilinear data structure and the lower SNR in this case (compare **Figure 4.10A** with **Figure 4.10B**).

The importance of noise in 2D NMR spectra was confirmed in a posterior analysis included in Scientific Article V (**Supplementary Material S11**) [159]. In this article, we determined that the removal of noise produces that the estimated number of components becomes closer to the number of detected resonances.

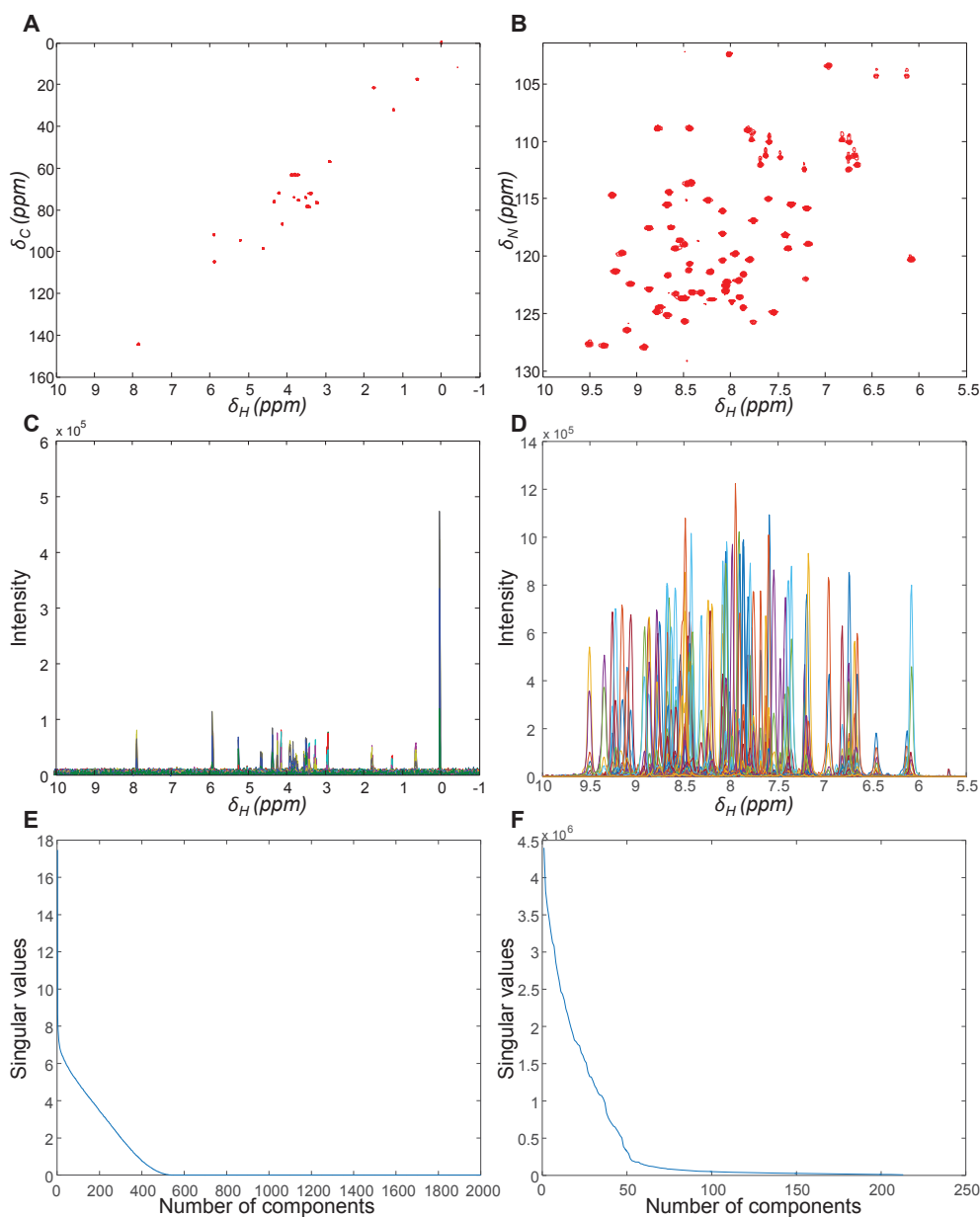


Figure 4.10. SVD analysis of two 2D NMR datasets. **A)** ^1H - ^{13}C HSQC NMR of a mixture of 3 compounds. **B)** ^1H - ^{15}N HSQC NMR of ubiquitin protein. **C)** and **D)** 1D projection of the ^1H - ^{13}C HSQC NMR from **A** and **B** in the ^1H -dimension, respectively. Different colors denote different intensities in f_i (δ_c). **E)** and **F)** Plot of the singular values associated to the ^1H - ^{13}C HSQC NMR from **A** and **B**.

Moreover, in the same article, we observed that removing the noisy variables from a 2D NMR metabolomics dataset had an important effect on the PCA analysis. This was observed

in the scores distribution from the denoised 2D NMR dataset (**Fig2C** in Scientific Article V [159]), which were very similar to the scores distribution obtained from applying PCA analysis on a 1D NMR dataset of the same samples (**Fig2A** in Scientific Article V [159]). On the contrary, results from the PCA analysis on the raw 2D NMR dataset gave a different scores distribution (**Fig2B** in Scientific Article V [159]). In agreement with this observation, in [364,365], variables descriptive of noise from 2D NMR datasets were removed prior the chemometrics data analysis to improve results interpretation.

In previous works [364,365], variables with intensities below a given threshold were considered to be noise and excluded. For 2D NMR spectra containing resonances close to the detection limit, as in 2D NMR spectra from metabolomics studies, to establish the appropriate intensity threshold results very challenging.

After manual evaluation of different 2D NMR spectra, we estimated that every resonance is commonly defined by at least 10 data values. Having this into account, we decided to implement a variable filtering algorithm that uses two parameters, the threshold level and a minimal number of contiguous values that constitute a resonance. This second parameter was implemented as the minimal number of adjacent variables with intensities higher than the threshold level.

This approach has been called as the VOI (Variables of Interest) strategy. When the threshold level is lowered, the most intense noise values will surpass this threshold, but since they will not be found clustered but randomly distributed over all the spectra, they will be still filtered. This strategy has been described in more detail in Scientific Article V [159].

3.4.1 Signal sparseness in 2D NMR spectra

Application of VOI on 2D NMR datasets revealed that 2D NMR spectra are sparse and only a few number of variables are representative of meaningful resonances. For example, for the investigated ^1H - ^{13}C HSQC NMR metabolomics dataset in the Scientific Article V and VI, with more than 60 detected metabolites (one sample is shown in **Figure 4.3**), only 2% of the spectral data was really related with their resonances. Moreover, this sparsity is even more accentuated for 2D NMR data of simpler mixtures, and even more for spectral data of higher dimensionality. For instance, for the 3D HNC0 spectrum of an unfolded protein, only 0.19% of the variables were representative of meaningful resonances (Scientific Article V, **Supplementary Material S17**).

Therefore, biased results obtained in the PCA analysis of raw 2D NMR datasets (when compared to the 1D NMR data, mentioned in the previous section) were caused mostly because ~98% of the variables from the dataset were only descriptive of noise.

3.5 VOI APPROACH IS A ROBUST 2D INTEGRATION METHOD

In [section 3.5.1](#) of [Chapter 2](#), it was mentioned that non-overlapped 1D NMR resonances can be integrated by summing the intensities of all the data-points that define these resonances.

For 2D NMR spectral data, most NMR integration tools use a similar approach. In this approach, each resonance is encapsulated by a user-defined region, and the intensities for all the variables enclosed in this region are summed.

In rNMR [34], the user-defined region is rectangular-shaped, while in MestReNova (MestreLab, Inc.), the shape of the selected region is an ellipse. However, resonances are not rectangular nor perfect ellipses. This means that the measured resonance integrals will be always higher than the real ones because some variables representative of noise will be included inside these user-defined regions and summed together with the relevant ones.

In [Figure 4.11](#), after noise removal using the VOI approach, the limits and shapes of the resonances are much better appreciated. In [Figure 4.11B](#), VOI-filtered resonances are not perfect ellipses, as in the original 2D NMR spectrum ([Figure 4.11A](#)).

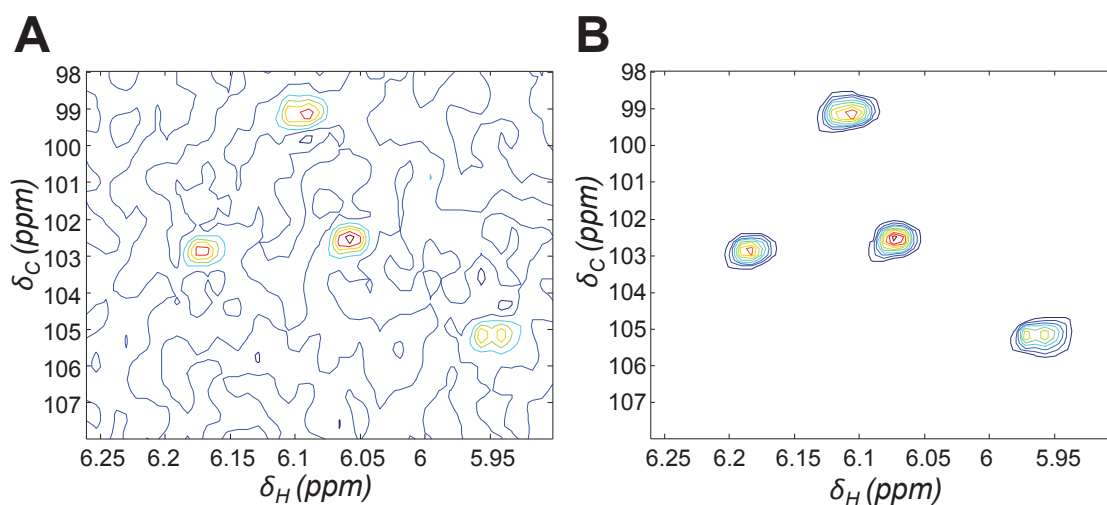


Figure 4.11. Filtering of noise in 2D ^1H - ^{13}C HSQC NMR data. **A)** Original data. **B)** VOI-filtered data from **A**.

When resonances are integrated using VOI approach, only the variables inside each peak are summed. Because of this, resonances can be integrated regardless of their shape. In [Figure 4.12](#), variables from the same cluster are colored with the same color, showing that the clustered variables have different sizes and shapes.

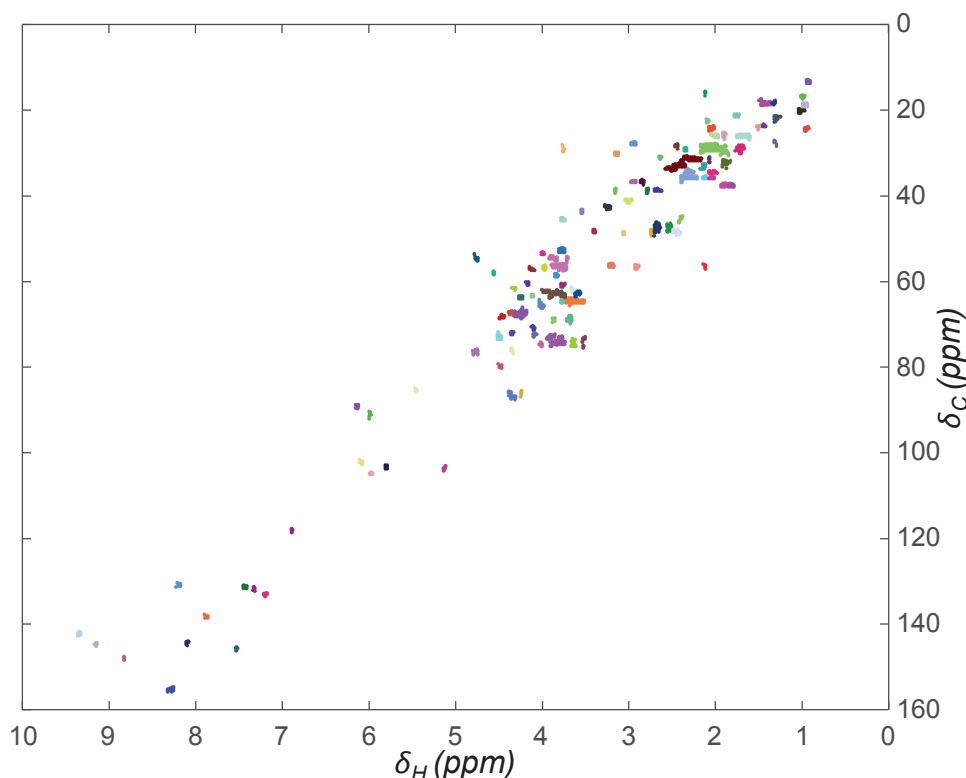


Figure 4.12. ^1H - ^{13}C HSQC NMR of a yeast extract. Each set of clustered variables has been colored with one color, denoting that clusters of variables are detected regardless of their peak shape.

Differences in the shape of each colored cluster in **Figure 4.12** are due to different *spin-spin* coupling constants and also because some of the clusters include two or more overlapped resonances. In other systems, additional differences in shape could be caused by other reasons, such as incomplete phasing.

Thus, with VOI strategy, only the cluster of variables representative of an isolated resonance is required to obtain a reliable resonance integral value.

These integrals can be calculated inside the MATLAB[®] computer and visualization environment or, if preferred, using other NMR suites capable of importing the VOI-processed data. For instance, in **Figure 4.13**, the integration of VOI-processed 2D NMR spectra by MestReNova software is shown.

In this figure, when original data are directly analyzed using ellipses with different sizes (green ellipses in **Figure 4.13A** and **4.13B**), different resonance integral values are obtained. On the contrary, when original 2D NMR spectral data are VOI-filtered, the use of ellipses with different sizes (green ellipses in **Fig. 4.14C** and **4.14D**) did not affect the estimation of the resonance integrals, since variables descriptive of noise were replaced by zero values after application of VOI.

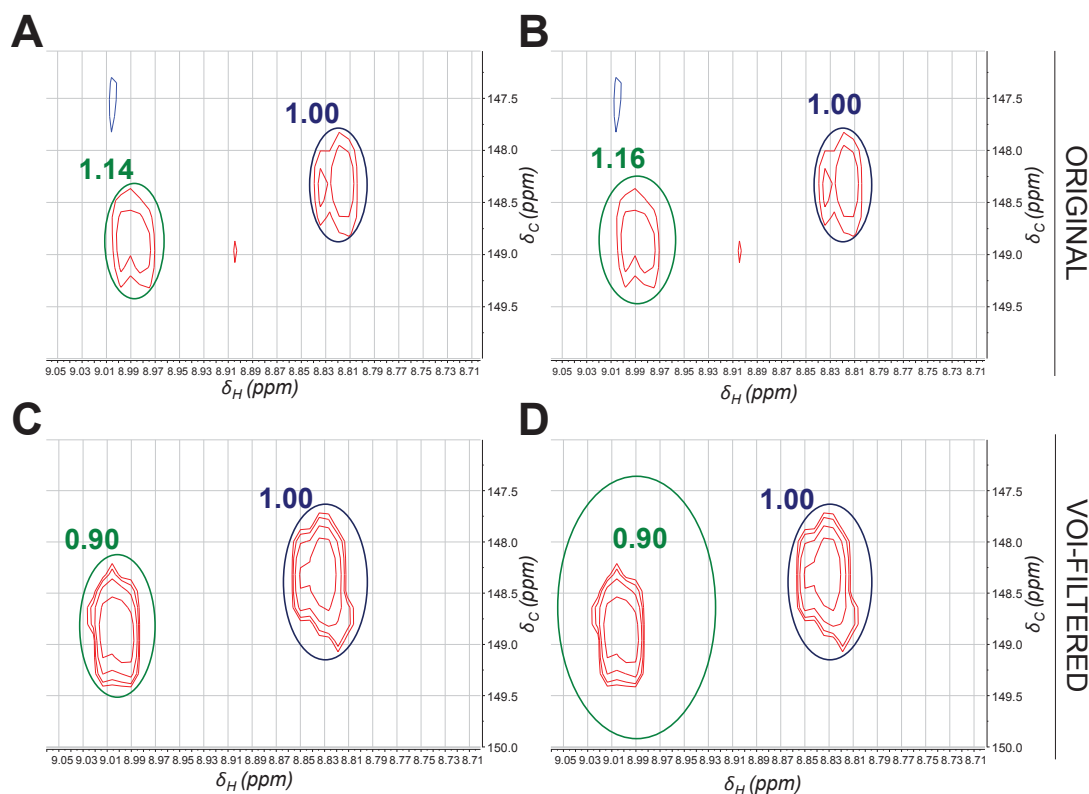


Figure 4.13. Integration of a 2D NMR spectrum in MestReNova. **A-B)** Experimental data. **C-D)** The same data after application of VOI. The resonance circled by a green ellipse in **A** and **C** was integrated again with a much broader ellipse in **B** and **D**, respectively. The resonance circled in blue acts as the reference integral.

3.6 ^1H NMR AND ^1H - ^{13}C HSQC NMR METABOLOMICS: DOES THE DIMENSIONALITY MATTER?

Regardless ^1H NMR or ^1H - ^{13}C HSQC NMR spectral data were used, similar results were obtained when they were investigated by chemometric methods such as PCA. Thus, not only similar scores distribution were obtained (as mentioned in [Section 3.4](#)), but the corresponding loadings highlighted the same metabolites ([Fig3](#) in Scientific Article VI), leading to the same biological interpretations. Nonetheless, since 2D NMR spectral data provides more information relative to compounds structure and it presents less signal overlapping, the highlighted resonances were much more straightforwardly assigned.

Other aspects that differ between acquiring ^1H NMR spectral data and 2D NMR data are the sensitivity, acquisition time and quantitative power.

Sensitivity depends on several factors, such as the gyromagnetic ratios (γ), the relative abundance of the measured nuclei, and the number of scans, among others. For naturally enriched samples, measuring carbon nuclei is less sensitive than proton nuclei (1.1%

abundance of ^{13}C , γ_{C} is approximately one-fourth of the γ_{H}). Besides, for 2D NMR data, a less number of scans is used than in 1D NMR to avoid acquiring the signals during very long time periods. As a consequence, 2D ^1H - ^{13}C HSQC NMR is substantially more insensitive than ^1H NMR. Thus, it was observed that one of the main things that have hindered the development of 2D ^1H - ^{13}C HSQC NMR metabolomics (if compared to ^1H NMR metabolomics) is the considerable long acquisition times needed, which produces a substantial increase of expenses. In Scientific Article V and VI, more than 7 hours were used to acquire each 2D ^1H - ^{13}C HSQC NMR spectrum of a metabolites yeast extract, while 30 minutes were spent to acquire each ^1H NMR spectrum.

However, these acquisition times can be drastically reduced if only data-points representative of the relevant resonances were acquired, which can be obtained for instance using the Absolute Minimal Sampling (AMS) [366] approach. In AMS, only data relative to a reduced list of frequencies is acquired. Thus, to reduce significantly acquisition time, the frequencies to be measured would be those previously detected with VOI approach in a representative sample [89].

Moreover, since more than a 10-fold time reductions can be achieved by AMS [89,366], a larger number of scans can be considered, improving the quality of the acquired 2D NMR spectrum.

Not all 2D NMR pulse sequences produce inherently quantitative results [163], although quantitative information can be reached if calibration curves are used [367]. Nevertheless, in metabolomics, most of the data interpretation is based on semi-quantitative data and metabolite fold-changes and, therefore, inherently quantitative data is not commonly required. In Scientific Article VI, we demonstrated that the investigation of an integral dataset from either 1D NMR or from 2D NMR metabolomics data by ASCA and PLS led to similar interpretation of the results.

In this article, we compared the two sets of integrals, and we determined that discrepancies between them were caused by two reasons. First, because the lowest concentrated metabolites were hardly detected in the 2D NMR data. And second, because deconvolution, used for the 1D NMR dataset (because resonances were extremely overlapped) but not used for the 2D NMR dataset, might capture unwillingly intensities from other peaks or from noise.

With the recent and upcoming advances [5,368] in NMR sensitivity, high-throughput 2D NMR metabolomics can be regarded as a promising tool because the time spent in the data analysis is considerably reduced when compared to 1D NMR metabolomics. Resonances

from 2D NMR data are more easily assigned and the integration step is faster because deconvolution is not usually required in this case.

4 CONCLUSIONS

The scientific research included in this Chapter can be summarized in the following specific conclusions:

- ^1H NMR spectral data from metabolomic samples are intrinsically complex, because they are crowded with hundreds of resonances from several dozens of metabolites. Moreover, for some of these metabolites, their concentrations are varying similarly within defined ranges because they are co-regulated in order to maintain the organism alive. For these two reasons, the direct analysis of these datasets can produce unreliable results.
- MCR-ALS resolution of these datasets can be performed by the application of a windowing approach. That is, by resolving the whole ^1H NMR spectral dataset in small windows. With this approach, resonances from co-regulated metabolites found in different windows will be resolved separately because they are analyzed in different MCR-ALS analysis.
- A simultaneous MCR-ALS analysis of all the ^1H NMR spectral windows can be performed afterwards by using properly designed spectral window constraints. With these constraints, it is imposed that resonances can only be resolved in the spectral windows pre-defined by the analyst user.
- The proposed Decision Tree of Correlations (DTC) approach is confirmed to be a satisfactory strategy to build these spectral window constraints without requiring previous knowledge of samples composition.
- The MCR-ALS analysis of ^1H NMR metabolomics datasets assessed by window spectral constraints allow the resolution of the concentration and ^1H NMR spectra profiles of the pure chemical species. This resolution may not be achieved for all metabolites in the samples, depending on their concentration variance and on the degree of spectral overlapping. Concentration profiles were normally better resolved than spectra profiles and their integration usually produce good results.
- 2D NMR spectra of metabolomics samples are also intrinsically complex. However, their spectra overlapping is less prominent due to the existence of the second dimension. However, SNR for 2D NMR spectra are commonly worse than for ^1H NMR spectra.
 - 2D NMR spectra are sparse and more than 98 % of the variables (in the ^1H - ^{13}C HSQC NMR) are usually only descriptive of noise. The presence of this very large amount of noise hampers any chemometric analysis performed on the data.
 - The Variables of Interest (VOI) approach has been proposed as a method to filter noisy 2D (and 3D) NMR datasets. With this approach, only those variables

representative of meaningful resonances are kept, while variables relative to noise are discarded. Only variables that surpass a user-defined threshold and found contiguously clustered forming a resonance are selected.

- The intensities of the isolated resonances within these specific clusters can be directly summed separately.
- Regardless of the type of data used, ^1H NMR and ^1H - ^{13}C HSQC NMR gave similar results when they were analyzed with chemometric methods.
- Since resonance assignment and resonance integration are more easily performed for 2D NMR data than for 1D NMR data, the total time spent for the analysis of 2D NMR metabolomics datasets is significantly shorter.

Chapter 5

Conclusions



In this Thesis, the application of different chemometric methods to investigate the effect of environmental perturbations on the metabolism of yeast using NMR spectroscopic data is demonstrated.

The analytical and biological conclusions resulting from the work performed in this Thesis are presented below:

Analytical conclusions

1. An analytical workflow for metabolomics studies in yeast is proposed, including the sample preparation, the acquisition of the NMR spectra, the import of data, the pre-processing of data, the application of chemometric methods, the interpretation of the chemometric results, and the export of results.
2. NMR metabolomics datasets are very complex to analyze because every NMR spectrum contains hundreds of resonances from dozens of metabolites. In ^1H NMR metabolomics datasets, resonances are strongly overlapped. On the other hand, in ^1H - ^{13}C HSQC NMR datasets, data overlapping is less important, but 2D FT-NMR spectra are predominantly constituted by variables representative of noise, and the SNR is worse than for ^1H NMR spectra.
3. Metabolic variance in NMR metabolomics datasets is usually very restricted since metabolic processes are highly regulated in living systems. Because of this, the direct application of resolution methods cannot in general extract satisfactorily the NMR spectra and concentrations of the different metabolite constituents of the analysed samples. In this Thesis, we have proven that this type of datasets can be reliably resolved through the application of the MCR-ALS method on small size NMR spectral windows, and also through the application of the DTC-MCR-ALS approach developed in this Thesis, which uses spectral windows constraints. The advantage of using these two approaches lies in the fact that, even though these datasets usually contain a large number of resonances from several dozens of metabolites whose concentrations change very little, the number of detected metabolites in every window is small, which makes their analysis feasible.
4. The influence of noise in 2D NMR spectra can be removed after application of the proposed VOI approach. After noise removal, chemometric analysis of 2D NMR spectral datasets led to the same results as the analysis of the corresponding dataset of 1D NMR spectra. Moreover, removal of noise from 2D NMR spectra allows for a faster and more accurate resonances integration. This is due to: (i) noise-filtered 2D NMR spectra provide better structural information than the one-dimensional ones; (ii) resonances from 2D NMR spectra are less overlapped; and (iii) resonances integration from these NMR data are faster. These particularities cause that, in

overall, the data analysis of 2D NMR metabolomics datasets is faster than the data analysis of 1D NMR metabolomics datasets.

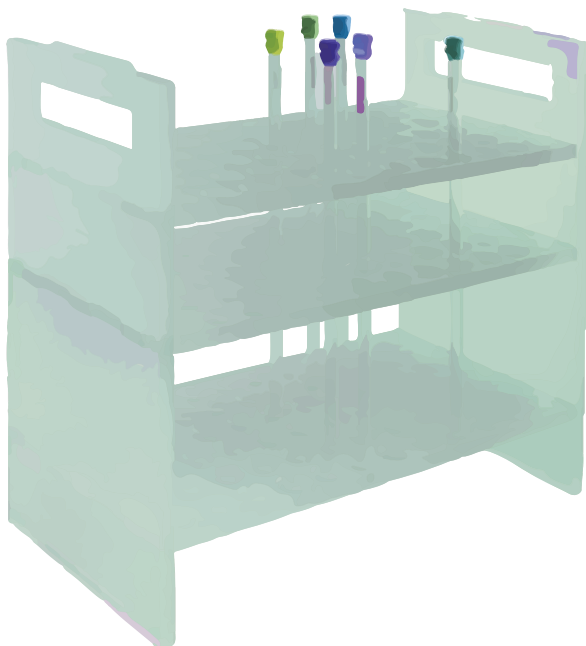
Biological conclusions

5. Application of the MCR-ALS method is shown to be an effective method for elucidating the metabolic processes occurring in biological organisms under environmental stress or any other biological perturbation. In order to extract the maximal possible biological information, it is important to use datasets as much comprehensive as possible. This includes time-course experiments and metabolomics experiments where two or more analytical platforms (*e.g.*, NMR and UHPLC-MS) are used.
6. *Saccharomyces cerevisiae* (strain BY4741) has shown to suffer an adaptation to changes in the growth temperature at a metabolic level. At low temperatures, metabolic pathways showing more activity are those required to maintain cells alive. For this reason, compounds characteristic from this biological state are metabolites from the Krebs cycle. In the lipidome, at low temperatures, phosphatidylinositols, short diacylglycerides and triacylglycerides with a low number of unsaturations are also accumulated. At optimal temperatures, amino acids and nucleotides are abundant because they are used to build cell structures and to promote growth. Regarding the lipid fraction, several phospholipid species including phosphatidylethanolamines and phosphatidylserines are found abundant. Finally, at higher temperatures, metabolites from fermentation pathways, and metabolites characteristic from stress are found at elevated concentrations. On the other hand, at higher temperatures, the lipidome accumulates some diacylglycerides, phosphatidylcholines, and long poly-unsaturated triacylglycerides.
7. In *Saccharomyces cerevisiae* (strain BY4741), the growth in drop-out media causes the complete metabolic, nutrient-specific, de-regulation. When these yeast cells grow under normal conditions, they follow the typical cell growth explained by the lag, the exponential, the diauxic shift and the stationary growth phases. On the other hand, starved cells follow different cell growth patterns. L-leucine-starved cells entered into a state of quiescence. The cell growth of uracil-starved cells was arrested, although the metabolic activity of the biosynthesis of uracil precursors was uncontrollably elevated. L-methionine-starved cells show a delay of the exponential growth phase. Finally, L-histidine-starved cells show a softly repressed yeast growth, attributed to the fact that a large number of resources were invested inefficiently to the biosynthesis of L-histidine precursors.

8. In *Saccharomyces cerevisiae* (strain S288C), using culture media containing different medium composition produces changes on yeast growth. Despite YPD (rich) and YSC (minimal) medium contain all nutrients required for the cells to grow, due to the limited variety of nutrients of YSC medium, yeast cells in YSC medium present a high metabolic activity, as most metabolites are *de novo* synthesized from glucose, ammonia, sulfate, and phosphate. On the contrary, YPD-cultured cells can directly uptake nutrients from the medium, resulting in a slightly faster cell growth. As a consequence of the major biosynthetic activity of YSC-cultured cells, a much diverse metabolome was observed under this condition than in YPD-cultured cells. At the stationary phase, cell growth is arrested due to the absence of nutrients in the medium. At this point, the two cultures confronted this situation differently. On one hand, cells grown in YPD medium enter into a hypo-metabolism state that can be prolonged for several weeks. On the other hand, cells grown in YSC medium maintain a high metabolism rate during this growth phase, resulting in a rapid loss of viability.

Chapter 6

References



1. Shiina I, Umezaki Y, Murata T, Suzuki K, Tono T. 2018. Asymmetric Total Synthesis of (+)-Coprophilin. *Synthesis* **50**: 1301-1306.
2. Dalitz F, Cudaj M, Maiwald M, Guthausen G. 2012. Process and reaction monitoring by low-field NMR spectroscopy. *Progress in Nuclear Magnetic Resonance Spectroscopy* **60**: 52-70.
3. Markwick PRL, Malliavin T, Nilges M. 2008. Structural Biology by NMR: Structure, Dynamics, and Interactions. *PLOS Computational Biology* **4**: e1000168.
4. Fallahi F, Oliver R, Mandalia SS, Jonker L. 2014. Early MRI diagnostics for suspected scaphoid fractures subsequent to initial plain radiography. *European Journal of Orthopaedic Surgery & Traumatology* **24**: 1161-1166.
5. Nagana Gowda GA, Raftery D. 2017. Recent Advances in NMR-Based Metabolomics. *Analytical Chemistry* **89**: 490-510.
6. Frisch R, Stern O. 1933. Über die magnetische Ablenkung von Wasserstoffmolekülen und das magnetische Moment des Protons. I. *Zeitschrift für Physik* **85**: 4-16.
7. Rabi II, Millman S, Kusch P, Zacharias JR. 1939. The Molecular Beam Resonance Method for Measuring Nuclear Magnetic Moments. The Magnetic Moments of ${}^6\text{Li}$, ${}^7\text{Li}$ and ${}^{19}\text{F}$. *Physical Review* **55**: 526-535.
8. Magill FN. 1991 *The Nobel Prize Winners: 1901-1944*. Salem Press.
9. Bloch F, Hansen WW, Packard M. 1946. Nuclear Induction. *Physical Review* **69**: 127-127.
10. Purcell EM, Torrey HC, Pound RV. 1946. Resonance Absorption by Nuclear Magnetic Moments in a Solid. *Physical Review* **69**: 37-38.
11. Becker ED. 1993. A Brief History of Nuclear Magnetic Resonance. *Analytical Chemistry* **65**: 295A-302A.
12. Arnold JT, Dharmatti SS, Packard ME. 1951. Chemical Effects on Nuclear Induction Signals from Organic Compounds. *The Journal of Chemical Physics* **19**: 507-507.
13. Lowe IJ, Norberg RE. 1957. Free-Induction Decays in Solids. *Physical Review* **107**: 46-61.
14. Ernst RR, Anderson WA. 1966. Application of Fourier Transform Spectroscopy to Magnetic Resonance. *Review of Scientific Instruments* **37**: 93-102.
15. Shampo MA, Kyle RA, Steensma DP. 2012. Richard Ernst—Nobel Prize for Nuclear Magnetic Resonance Spectroscopy. *Mayo Clinic Proceedings* **87**: e109-e109.
16. Bax A, Lerner L. 1986. Two-dimensional nuclear magnetic resonance spectroscopy. *Science* **232**: 960-967.
17. Bell JD, Brown JCC, Sadler PJ. 1989. NMR studies of body fluids. *NMR in Biomedicine* **2**: 246-256.
18. Nicholson JK, Wilson ID. 1989. High resolution proton magnetic resonance spectroscopy of biological fluids. *Progress in Nuclear Magnetic Resonance Spectroscopy* **21**: 449-501.
19. Palmer AG, Patel DJ. 2002. Kurt Wüthrich and NMR of Biological Macromolecules. *Structure* **10**: 1603-1604.
20. Bartels LW, Bakker CJ. 2004. Nobel Prize for physiology or medicine in 2003 awarded to the fathers of magnetic resonance imaging. *Nederlands Tijdschrift voor Geneeskunde* **148**: 117-119.

21. Nicholson JK, Lindon JC, Holmes E. 1999. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* **29**: 1181-1189.
22. Hermsen CC, Crommert JV, Fredix H, Sauerwein RW, Eling WM. 1997. Circulating tumour necrosis factor alpha is not involved in the development of cerebral malaria in *Plasmodium berghei*-infected C57Bl mice. *Parasite Immunology* **19**: 571-577.
23. Mounet F, Lemaire-Chamley M, Maucourt M, Cabasson C, Giraudel JL, *et al.* 2007. Quantitative metabolic profiles of tomato flesh and seeds during fruit development: Complementary analysis with ANN and PCA. *Metabolomics* **3**: 273-288.
24. Behrends V, Ryall B, Wang X, Bundy JG, Williams HD. 2010. Metabolic profiling of *Pseudomonas aeruginosa* demonstrates that the anti-sigma factor MucA modulates osmotic stress tolerance. *Molecular Biosystems* **6**: 562-569.
25. Liebeke M, Meyer H, Donat S, Ohlsen K, Lalk M. 2010. A metabolomic view of *Staphylococcus aureus* and Its Ser/Thr kinase and phosphatase deletion mutants: involvement in cell wall biosynthesis. *Chemistry & Biology* **17**: 820-830.
26. Nagato EG, D'eon JC, Lankadurai BP, Poirier DG, Reiner EJ, *et al.* 2013. ¹H NMR-based metabolomics investigation of *Daphnia magna* responses to sub-lethal exposure to arsenic, copper and lithium. *Chemosphere* **93**: 331-337.
27. Akhtar MT, Mushtaq MY, Verpoorte R, Richardson MK, Choi YH. 2016. Metabolic effects of cannabinoids in zebrafish (*Danio rerio*) embryos determined by ¹H NMR metabolomics. *Metabolomics* **12**: 44.
28. Markley JL, Brüschweiler R, Edison AS, Eghbalnia HR, Powers R, *et al.* 2017. The future of NMR-based metabolomics. *Current Opinion in Biotechnology* **43**: 34-40.
29. Lankadurai BP, Nagato EG, Simpson MJ. 2013. Environmental metabolomics: an emerging approach to study organism responses to environmental stressors. *Environmental Reviews* **21**: 180-205.
30. Puig-Castellví F, Alfonso I, Piña B, Tauler R. 2015. A quantitative ¹H NMR approach for evaluating the metabolic response of *Saccharomyces cerevisiae* to mild heat stress. *Metabolomics* **11**: 1612-1625.
31. Worley B, Powers R. 2014. MVAPACK: A Complete Data Handling Package for NMR Metabolomics. *ACS Chemical Biology* **9**: 1138-1144.
32. Xia J, Mandal R, Sinelnikov IV, Broadhurst D, Wishart DS. 2012. MetaboAnalyst 2.0—a comprehensive server for metabolomic data analysis. *Nucleic Acids Research* **40**: W127-W133.
33. Ravanbakhsh S, Liu P, Bjordahl TC, Mandal R, Grant JR, *et al.* 2015. Accurate, Fully-Automated NMR Spectral Profiling for Metabolomics. *PLOS ONE* **10**: e0124219.
34. Lewis IA, Schommer SC, Markley JL. 2009. rNMR: open source software for identifying and quantifying metabolites in NMR spectra. *Magnetic Resonance in Chemistry* **47**: S123-S126.
35. Hao J, Astle W, De Iorio M, Ebbels TMD. 2012. BATMAN—an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics* **28**: 2088-2090.
36. Tauler R. 1995. Multivariate curve resolution applied to second order data. *Chemometrics and Intelligent Laboratory Systems* **30**: 133-146.

37. Weljie AM, Newton J, Mercier P, Carlson E, Slupsky CM. 2006. Targeted Profiling: Quantitative Analysis of ^1H NMR Metabolomics Data. *Analytical Chemistry* **78**: 4430-4442.
38. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, *et al.* 1995. NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *Journal of Biomolecular NMR* **6**: 277-293.
39. Hao J, Astle W, de-Iorio M, Ebbels T. 2011. BATMAN--an R package for the automated quantification of metabolites from NMR spectra using a Bayesian Model. *Bioinformatics* **28**: 2088-2090.
40. R Core Team. 2013 *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
41. Jaumot J, de Juan A, Tauler R. 2015. MCR-ALS GUI 2.0: New features and applications. *Chemometrics and Intelligent Laboratory Systems* **140**: 1-12.
42. Foundation PS. Python Language Reference. <http://www.python.org>.
43. Mathworks. 2011. Global Optimization Toolbox: User's Guide (r2011b). www.mathworks.com.
44. Willcott MR. 2009. MestRe Nova. *Journal of the American Chemical Society* **131**: 13180-13180.
45. Bruker. AMIX. www.bruker.com/bruker/amix.
46. Chenomx. Chenomx. <https://www.chenomx.com/>.
47. ACD/Labs. ACD/NMR. <http://www.acdlabs.com/>.
48. Bruker. TopSpin® software. www.bruker.com/bruker/topspin.
49. Tauler R, Izquierdo-Ridorsa A, Casassas E. 1993. Simultaneous analysis of several spectroscopic titrations with self-modelling curve resolution. *Chemometrics and Intelligent Laboratory Systems* **18**: 293-300.
50. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. 2017. *cluster: Cluster Analysis Basics and Extensions*.
51. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, *et al.* 2015 *gplots: Various R Programming Tools for Plotting Data*.
52. Csardi G, Nepusz T. 2006. The igraph software package for complex network research. *InterJournal Complex Systems*: 1695.
53. Fajjes M, Mars AE, Smid EJ. 2007. Comparison of quenching and extraction methodologies for metabolome analysis of *Lactobacillus plantarum*. *Microbial Cell Factories* **6**: 27.
54. Meyer H, Weidmann H, Lalk M. 2013. Methodological approaches to help unravel the intracellular metabolome of *Bacillus subtilis*. *Microbial Cell Factories* **12**: 69.
55. Palomino-Schätzlein M, Molina-Navarro M, Tormos-Pérez M, Rodríguez-Navarro S, Pineda-Lucena A. 2013. Optimised protocols for the metabolic profiling of *S. cerevisiae* by ^1H -NMR and HRMAS spectroscopy. *Analytical and Bioanalytical Chemistry* **405**: 8431-8441.
56. Sasidharan K, Soga T, Tomita M, Murray DB. 2012. A Yeast Metabolite Extraction Protocol Optimised for Time-Series Analyses. *PLOS ONE* **7**: e44283.
57. Geier FM, Want EJ, Leroi AM, Bundy JG. 2011. Cross-Platform Comparison of *Caenorhabditis elegans* Tissue Extraction Strategies for Comprehensive Metabolome Coverage. *Analytical Chemistry* **83**: 3730-3736.

58. Kim HK, Verpoorte R. 2010. Sample preparation for plant metabolomics. *Phytochemical Analysis* **21**: 4-13.
59. Wu X, Li N, Li H, Tang H. 2014. An optimized method for NMR-based plant seed metabolomic analysis with maximized polar metabolite extraction efficiency, signal-to-noise ratio, and chemical shift consistency. *Analyst* **139**: 1769-1778.
60. Brown SAE, Simpson AJ, Simpson MJ. 2008. Evaluation of sample preparation methods for nuclear magnetic resonance metabolic profiling studies with *Eisenia fetida*. *Environmental Toxicology and Chemistry* **27**: 828-836.
61. Lin CY, Wu H, Tjeerdema RS, Viant MR. 2007. Evaluation of metabolite extraction strategies from tissue samples using NMR metabolomics. *Metabolomics* **3**: 55-67.
62. Diémé B, Lefèvre A, Nadal-Desbarats L, Galineau L, Madji Hounoum B, *et al.* 2017. Workflow methodology for rat brain metabolome exploration using NMR, LC-MS and GC-MS analytical platforms. *Journal of Pharmaceutical and Biomedical Analysis* **142**: 270-278.
63. Anwar MA, Vorkas PA, Li JV, Shalhoub J, Want EJ, *et al.* 2015. Optimization of metabolite extraction of human vein tissue for ultra performance liquid chromatography-mass spectrometry and nuclear magnetic resonance-based untargeted metabolic profiling. *Analyst* **140**: 7586-7597.
64. Emwas A-H, Roy R, McKay RT, Ryan D, Brennan L, *et al.* 2016. Recommendations and Standardization of Biomarker Quantification Using NMR-Based Metabolomics with Particular Focus on Urinary Analysis. *Journal of Proteome Research* **15**: 360-373.
65. Jackson F, Georgakopoulou N, Kaluarachchi M, Kyriakides M, Andreas N, *et al.* 2016. Development of a Pipeline for Exploratory Metabolic Profiling of Infant Urine. *Journal of Proteome Research* **15**: 3432-3440.
66. Tiziani S, Emwas A-H, Lodi A, Ludwig C, Bunce CM, *et al.* 2008. Optimized metabolite extraction from blood serum for ¹H nuclear magnetic resonance spectroscopy. *Analytical Biochemistry* **377**: 16-23.
67. Deda O, Chatziioannou AC, Fasoula S, Palachanis D, Raikos N, *et al.* 2017. Sample preparation optimization in fecal metabolic profiling. *Journal of Chromatography B* **1047**: 115-123.
68. Schmitke JL, Wescott CR, Klibanov AM. 1996. The Mechanistic Dissection of the Plunge in Enzymatic Activity upon Transition from Water to Anhydrous Solvents. *Journal of the American Chemical Society* **118**: 3360-3365.
69. Fiehn O. 2002. Metabolomics - the link between genotypes and phenotypes. *Plant Molecular Biology* **48**: 155-171.
70. Verpoorte R, Choi YH, Mustafa NR, Kim HK. 2008. Metabolomics: back to basics. *Phytochemistry Reviews* **7**: 525-537.
71. Wolfender JL, Rudaz S, Choi YH, Kim HK. 2013. Plant metabolomics: from holistic data to relevant biomarkers. *Current Medicinal Chemistry* **20**: 1056-1090.
72. Ejsing CS, Sampaio JL, Surendranath V, Duchoslav E, Ekroos K, *et al.* 2009. Global analysis of the yeast lipidome by quantitative shotgun mass spectrometry. *Proceedings of the National Academy of Sciences* **106**: 2136-2141.
73. Wu H, Southam AD, Hines A, Viant MR. 2008. High-throughput tissue extraction protocol for NMR- and MS-based metabolomics. *Analytical Biochemistry* **372**: 204-212.

74. Sekiyama Y, Chikayama E, Kikuchi J. 2011. Evaluation of a Semipolar Solvent System as a Step toward Heteronuclear Multidimensional NMR-Based Metabolomics for ^{13}C -Labeled Bacteria, Plants, and Animals. *Analytical Chemistry* **83**: 719-726.
75. Del Coco L, De Pascali SA, Iacovelli V, Cesari G, Schena FP, *et al.* 2014. Following the olive oil production chain: 1D and 2D NMR study of olive paste, pomace, and oil. *European Journal of Lipid Science and Technology* **116**: 1513-1521.
76. Kim HK, Choi YH, Verpoorte R. 2010. NMR-based metabolomic analysis of plants. *Nature Protocols* **5**: 536-549.
77. Salem MA, Jüppner J, Bajdzienko K, Giavalisco P. 2016. 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**: 45.
78. Nowick JS, Khakshoor O, Hashemzadeh M, Brower JO. 2003. DSA: A New Internal Standard for NMR Studies in Aqueous Solution. *Organic Letters* **5**: 3511-3513.
79. Giraudeau P, Silvestre V, Akoka S. 2015. Optimizing water suppression for quantitative NMR-based metabolomics: a tutorial review. *Metabolomics* **11**: 1041-1055.
80. Marchev A, Yordanova Z, Alipieva K, Zahmanov G, Rusinova-Videva S, *et al.* 2016. Genetic transformation of rare *Verbascum eriophorum* Godr. plants and metabolic alterations revealed by NMR-based metabolomics. *Biotechnology Letters* **38**: 1621-1629.
81. Sobolev AP, Mannina L, Proietti N, Carradori S, Daglia M, *et al.* 2015. Untargeted NMR-based methodology in the study of fruit metabolites. *Molecules* **20**: 4088-4108.
82. Jézéquel T, Deborde C, Maucourt M, Zhendre V, Moing A, *et al.* 2015. Absolute quantification of metabolites in tomato fruit extracts by fast 2D NMR. *Metabolomics* **11**: 1231-1242.
83. Izrayelit Y, Robinette SL, Bose N, von Reuss SH, Schroeder FC. 2013. 2D NMR-Based Metabolomics Uncovers Interactions between Conserved Biochemical Pathways in the Model Organism *Caenorhabditis elegans*. *ACS Chemical Biology* **8**: 314-319.
84. Huang Y, Zhang Z, Chen H, Feng J, Cai S, *et al.* 2015. A high-resolution 2D *J*-resolved NMR detection technique for metabolite analyses of biological samples. *Scientific Reports* **5**: 8390.
85. Allen PJ, Wise D, Greenway T, Khoo L, Griffin MJ, *et al.* 2015. Using 1-D ^1H and 2-D ^1H *J*-resolved NMR metabolomics to understand the effects of anemia in channel catfish (*Ictalurus punctatus*). *Metabolomics* **11**: 1131-1143.
86. Chylla RA, Van Acker R, Kim H, Azapira A, Mukerjee P, *et al.* 2013. Plant cell wall profiling by fast maximum likelihood reconstruction (FMLR) and region-of-interest (ROI) segmentation of solution-state 2D ^1H - ^{13}C NMR spectra. *Biotechnology for Biofuels* **6**: 45.
87. Chae YK, Kim SH, Nam YK. 2013. Application of Two-Dimensional NMR Spectroscopy to Metabotyping Laboratory *Escherichia coli* Strains. *Chemistry & Biodiversity* **10**: 1816-1827.
88. Kang WY, Kim SH, Chae YK. 2012. Stress adaptation of *Saccharomyces cerevisiae* as monitored via metabolites using two-dimensional NMR spectroscopy. *FEMS Yeast Res* **12**: 608-616.
89. Hansen AL, Li D, Wang C, Brüsweiler R. 2017. Absolute Minimal Sampling of Homonuclear 2D NMR TOCSY Spectra for High-Throughput Applications of Complex Mixtures. *Angewandte Chemie International Edition* **56**: 8149-8152.

90. Robinette SL, Ajredini R, Rasheed H, Zeinomar A, Schroeder FC, *et al.* 2011. Hierarchical Alignment and Full Resolution Pattern Recognition of 2D NMR Spectra: Application to Nematode Chemical Ecology. *Analytical Chemistry* **83**: 1649-1657.
91. Ludwig C, Viant MR. 2010. Two-dimensional *J*-resolved NMR spectroscopy: review of a key methodology in the metabolomics toolbox. *Phytochemical Analysis* **21**: 22-32.
92. Liang YS, Kim HK, Lefeber AWM, Erkelens C, Choi YH, *et al.* 2006. Identification of phenylpropanoids in methyl jasmonate treated *Brassica rapa* leaves using two-dimensional nuclear magnetic resonance spectroscopy. *Journal of Chromatography A* **1112**: 148-155.
93. Nakabayashi R, Kusano M, Kobayashi M, Tohge T, Yonekura-Sakakibara K, *et al.* 2009. Metabolomics-oriented isolation and structure elucidation of 37 compounds including two anthocyanins from *Arabidopsis thaliana*. *Phytochemistry* **70**: 1017-1029.
94. Cooper JW. 1976 *The Computer in Fourier Transform NMR*. In: Levy GC, editor. Topics in Carbon-13 NMR Spectroscopy. New York: Wiley.
95. Siegel MM. 1981. The use of the modified simplex method for automatic phase correction in fourier-transform nuclear magnetic resonance spectroscopy. *Analytica Chimica Acta* **133**: 103-108.
96. Babij NR, McCusker EO, Whiteker GT, Canturk B, Choy N, *et al.* 2016. NMR Chemical Shifts of Trace Impurities: Industrially Preferred Solvents Used in Process and Green Chemistry. *Organic Process Research & Development* **20**: 661-667.
97. Akoka S, Barantin L, Trierweiler M. 1999. Concentration Measurement by Proton NMR Using the ERETIC Method. *Analytical Chemistry* **71**: 2554-2557.
98. Xi Y, Rocke DM. 2008. Baseline Correction for NMR Spectroscopic Metabolomics Data Analysis. *BMC Bioinformatics* **9**: 324.
99. Brown DE. 1995. Fully Automated Baseline Correction of 1D and 2D NMR Spectra Using Bernstein Polynomials. *Journal of Magnetic Resonance, Series A* **114**: 268-270.
100. Whittaker ET. 2009. On a New Method of Graduation. *Proceedings of the Edinburgh Mathematical Society* **41**: 63-75.
101. Zolnai Z, Macura S, Markley JL. 1989. Spline method for correcting baseplane distortions in two-dimensional NMR spectra. *Journal of Magnetic Resonance (1969)* **82**: 496-504.
102. Savorani F, Tomasi G, Engelsen SB. 2010. icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal of Magnetic Resonance* **202**: 190-202.
103. Alonso A, Rodríguez MA, Vinaixa M, Tortosa R, Correig X, *et al.* 2014. Focus: A Robust Workflow for One-Dimensional NMR Spectral Analysis. *Analytical Chemistry* **86**: 1160-1169.
104. Nielsen N-PV, Carstensen JM, Smedsgaard J. 1998. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A* **805**: 17-35.
105. Vu TN, Laukens K. 2013. Getting Your Peaks in Line: A Review of Alignment Methods for NMR Spectral Data. *Metabolites* **3**: 259-276.
106. De Meyer T, Sinnaeve D, Van Gasse B, Tsiportkova E, Rietzschel ER, *et al.* 2008. NMR-Based Characterization of Metabolic Alterations in Hypertension Using an Adaptive, Intelligent Binning Algorithm. *Analytical Chemistry* **80**: 3783-3790.
107. Sousa SAA, Magalhães A, Ferreira MMC. 2013. Optimized bucketing for NMR spectra: Three case studies. *Chemometrics and Intelligent Laboratory Systems* **122**: 93-102.

108. Li M-H, Du HZ, Kong G-J, Liu LB, Li XX, *et al.* 2017. Nuclear Magnetic Resonance-Based Metabolomics Approach to Evaluate the Prevention Effect of *Camellia nitidissima* Chi on Colitis-Associated Carcinogenesis. *Frontiers in Pharmacology* **8**: 447.
109. Bernier M, Catazaro J, Singh NS, Wnorowski A, Boguszevska-Czubara A, *et al.* 2017. GPR55 receptor antagonist decreases glycolytic activity in PANC-1 pancreatic cancer cell line and tumor xenografts. *International Journal of Cancer* **141**: 2131-2142.
110. Karaman I, Ferreira DLS, Boulangé CL, Kaluarachchi MR, Herrington D, *et al.* 2016. Workflow for Integrated Processing of Multicohort Untargeted ¹H NMR Metabolomics Data in Large-Scale Metabolic Epidemiology. *Journal of Proteome Research* **15**: 4188-4194.
111. Puchades-Carrasco L, Palomino-Schätzlein M, Pérez-Rambla C, Pineda-Lucena A. 2016. Bioinformatics tools for the analysis of NMR metabolomics studies focused on the identification of clinically relevant biomarkers. *Briefings in Bioinformatics* **17**: 541-552.
112. Kohl SM, Klein MS, Hochrein J, Oefner PJ, Spang R, *et al.* 2012. State-of-the art data normalization methods improve NMR-based metabolomic analysis. *Metabolomics* **8**: 146-160.
113. Worley B, Powers R. 2014. Simultaneous Phase and Scatter Correction for NMR Datasets. *Chemometrics and intelligent laboratory systems* **131**: 1-6.
114. Dieterle F, Ross A, Schlotterbeck G, Senn H. 2006. Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in ¹H NMR Metabolomics. *Analytical Chemistry* **78**: 4281-4290.
115. Lourenço AB, Roque FC, Teixeira MC, Ascenso JR, Sá-Correia I. 2013. Quantitative ¹H-NMR-Metabolomics Reveals Extensive Metabolic Reprogramming and the Effect of the Aquaglyceroporin FPS1 in Ethanol-Stressed Yeast Cells. *PLOS ONE* **8**: e55439.
116. Ji HG, Lee YR, Lee MS, Hwang KH, Kim EH, *et al.* 2017. Metabolic phenotyping of various tea (*Camellia sinensis* L.) cultivars and understanding of their intrinsic metabolism. *Food Chemistry* **233**: 321-330.
117. Guitton Y, Tremblay-Franco M, Le Corguillé G, Martin J-F, Pétéra M, *et al.* 2017. 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. *The International Journal of Biochemistry & Cell Biology* **93**: 89-101.
118. Puig-Castellví F, Alfonso I, Piña B, Tauler R. 2016. ¹H NMR metabolomic study of auxotrophic starvation in yeast using Multivariate Curve Resolution-Alternating Least Squares for Pathway Analysis. *Scientific Reports* **6**: 30982.
119. Puig-Castellví F, Pérez Y, Piña B, Tauler R, Alfonso I. 2018. Comparative analysis of ¹H NMR and ¹H-¹³C HSQC NMR metabolomics to understand the effects of medium composition in yeast growth. Submitted.
120. Kruger NJ, Troncoso-Ponce MA, Ratcliffe RG. 2008. ¹H NMR metabolite fingerprinting and metabolomic analysis of perchloric acid extracts from plant tissues. *Nature Protocols* **3**: 1001-1012.
121. Savorani F, Rasmussen MA, Mikkelsen MS, Engelsen SB. 2013. A primer to nutritional metabolomics by NMR spectroscopy and chemometrics. *Food Research International* **54**: 1131-1145.
122. van den Berg RA, Hoefsloot HCJ, Westerhuis JA, Smilde AK, van der Werf MJ. 2006. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* **7**: 142-142.

123. Madrid-Gambin F, Llorach R, Vázquez-Fresno R, Urpi-Sarda M, Almanza-Aguilera E, *et al.* 2017. Urinary ^1H Nuclear Magnetic Resonance Metabolomic Fingerprinting Reveals Biomarkers of Pulse Consumption Related to Energy-Metabolism Modulation in a Subcohort from the PREDIMED study. *Journal of Proteome Research* **16**: 1483-1491.
124. Liu Y, Huang R, Liu L, Peng J, Xiao B, *et al.* 2010. Metabonomics study of urine from Sprague–Dawley rats exposed to Huang-yao-zi using ^1H NMR spectroscopy. *Journal of Pharmaceutical and Biomedical Analysis* **52**: 136-141.
125. Puig-Castellví F, Bedia C, Alfonso I, Piña B, Tauler R. 2018. Deciphering the underlying metabolomic and lipidomic patterns linked to thermal acclimation in *Saccharomyces cerevisiae*. *Journal of Proteome Research*. DOI: 10.1021/acs.jproteome.7b00921.
126. Meiler J, Meusinger R, Will M. 2000. Fast Determination of ^{13}C NMR Chemical Shifts Using Artificial Neural Networks. *Journal of Chemical Information and Computer Sciences* **40**: 1169-1176.
127. Binev Y, Marques MMB, Aires-de-Sousa J. 2007. Prediction of ^1H NMR Coupling Constants with Associative Neural Networks Trained for Chemical Shifts. *Journal of Chemical Information and Modeling* **47**: 2089-2097.
128. Biemann K, Pretsch E, Clerc T, Seibl J, Simon W. 2013. *Tables of Spectral Data for Structure Determination of Organic Compounds*. Springer Berlin Heidelberg.
129. Hansen PE. 1981. Carbon—hydrogen spin-spin coupling constants. *Progress in Nuclear Magnetic Resonance Spectroscopy* **14**: 175-295.
130. Balci M. 2005 *Basic ^1H - and ^{13}C -NMR Spectroscopy*. Elsevier Science.
131. Kaplan JI, Fraenkel G. 1980 Chapter 7 - NMR of Exchanging Systems at High rf Fields. In *NMR of Chemically Exchanging Systems* (pp. 130-137). Academic Press.
132. Jacobsen NE. 2007 *Interpretation of Proton (^1H) NMR Spectra. NMR Spectroscopy Explained: Simplified Theory, Applications and Examples for Organic Chemistry and Structural Biology*. Hoboken, New Jersey: John Wiley & Sons, Inc.
133. Jacob D, Deborde C, Moing A. 2013. An efficient spectra processing method for metabolite identification from ^1H -NMR metabolomics data. *Analytical and Bioanalytical Chemistry* **405**: 5049-5061.
134. Everett JR. 2015. A New Paradigm for Known Metabolite Identification in Metabonomics/Metabolomics: Metabolite Identification Efficiency. *Computational and Structural Biotechnology Journal* **13**: 131-144.
135. Larive CK, Barding GA, Dinges MM. 2015. NMR Spectroscopy for Metabolomics and Metabolic Profiling. *Analytical Chemistry* **87**: 133-146.
136. Moreira AS, Lourenço AB, Sá-Correia I. 2016. ^1H -NMR-Based Endometabolome Profiles of *Burkholderia cenocepacia* Clonal Variants Retrieved from a Cystic Fibrosis Patient during Chronic Infection. *Frontiers in Microbiology* **7**: 2024.
137. El Ghazi I, Sheng WS, Hu S, Reilly BG, Lokensgard JR, *et al.* 2010. Changes in the NMR Metabolic Profile of Human Microglial Cells Exposed to Lipopolysaccharide or Morphine. *Journal of neuroimmune pharmacology: the official journal of the Society on NeuroImmune Pharmacology* **5**: 574-581.
138. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, *et al.* 2007. HMDB: the human metabolome database. *Nucleic Acids Research* **35**: D521-D526.
139. Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, *et al.* 2013. HMDB 3.0—The Human Metabolome Database in 2013. *Nucleic Acids Research* **41**: D801-D807.

140. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, *et al.* 2006. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research* **34**: D668-D672.
141. Guo AC, Jewison T, Wilson M, Liu Y, Knox C, *et al.* 2013. ECMDB: the *E. coli* Metabolome Database. *Nucleic Acids Research* **41**: D625-D630.
142. Jewison T, Knox C, Neveu V, Djoumbou Y, Guo AC, *et al.* 2012. YMDB: the Yeast Metabolome Database. *Nucleic Acids Research* **40**: D815-D820.
143. Bouatra S, Aziat F, Mandal R, Guo AC, Wilson MR, *et al.* 2013. The human urine metabolome. *PLOS ONE* **8**: e73076.
144. Cui Q, Lewis IA, Hegeman AD, Anderson ME, Li J, *et al.* 2008. Metabolite identification via the Madison Metabolomics Consortium Database. *Nature Biotechnology* **26**: 162-164.
145. Akiyama K, Chikayama E, Yuasa H, Shimada Y, Tohge T, *et al.* 2008. PRIME: a Web site that assembles tools for metabolomics and transcriptomics. *In Silico Biology* **8**: 339-345.
146. Ludwig C, Easton JM, Lodi A, Tiziani S, Manzoor SE, *et al.* 2012. Birmingham Metabolite Library: a publicly accessible database of 1-D ^1H and 2-D ^1H *J*-resolved NMR spectra of authentic metabolite standards (BML-NMR). *Metabolomics* **8**: 8-18.
147. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, *et al.* 2008. BioMagResBank. *Nucleic Acids Research* **36**: D402-D408.
148. Steinbeck C, Kuhn S. 2004. NMRShiftDB – compound identification and structure elucidation support through a free community-built web database. *Phytochemistry* **65**: 2711-2717.
149. Ellinger JJ, Chylla RA, Ulrich EL, Markley JL. 2013. Databases and Software for NMR-Based Metabolomics. *Current Metabolomics* **1**: 28-40.
150. Hao J, Liebeke M, Astle W, De Iorio M, Bundy JG, *et al.* 2014. Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN. *Nature Protocols* **9**: 1416-1427.
151. Dashti H, Westler WM, Tonelli M, Wedell JR, Markley JL, *et al.* 2017. Spin System Modeling of Nuclear Magnetic Resonance Spectra for Applications in Metabolomics and Small Molecule Screening. *Analytical Chemistry* **89**: 12201-12208.
152. Bruce SD, Higinbotham J, Marshall I, Beswick PH. 2000. An Analytical Derivation of a Popular Approximation of the Voigt Function for Quantification of NMR Spectra. *Journal of Magnetic Resonance* **142**: 57-63.
153. Nadal-Desbarats L, Aidoud N, Emond P, Blasco H, Filipiak I, *et al.* 2014. Combined ^1H -NMR and ^1H - ^{13}C HSQC-NMR to improve urinary screening in autism spectrum disorders. *Analyst* **139**: 3460-3468.
154. Chae YK, Kim SH. 2016. Discrimination of Rice Products by Geographical Origins and Cultivars by Two-Dimensional NMR Spectroscopy. *Bulletin of the Korean Chemical Society* **37**: 1612-1617.
155. Chae YK, Kim SH, Markley JL. 2017. Relationship between recombinant protein expression and host metabolome as determined by two-dimensional NMR spectroscopy. *PLOS ONE* **12**: e0177233.
156. Kang C-M, Seong Hyeon J, Ra Kim S, Kyeong Lee E, Jin Yun H, *et al.* 2015. Application of NMR Spectroscopy in the Assessment of Radiation Dose in Human Primary Cells. *Chemistry & Biodiversity* **12**: 1696-1705.

157. Guerrini M, Rudd TR, Mauri L, Macchi E, Fareed J, *et al.* 2015. Differentiation of Generic Enoxaparins Marketed in the United States by Employing NMR and Multivariate Analysis. *Analytical Chemistry* **87**: 8275-8283.
158. Chylla RA, Hu K, Ellinger JJ, Markley JL. 2011. Deconvolution of two-dimensional NMR spectra by fast maximum likelihood reconstruction: application to quantitative metabolomics. *Analytical Chemistry* **83**: 4871-4880.
159. Puig-Castellví F, Pérez Y, Piña B, Tauler R, Alfonso I. 2018. Compression of multidimensional NMR spectra allows a faster and more accurate analysis of complex samples. *Chemical Communications*. **54**: 3090-3093.
160. Jaumot J, Marchán V, Gargallo R, Grandas A, Tauler R. 2004. Multivariate Curve Resolution Applied to the Analysis and Resolution of Two-Dimensional [^1H , ^{15}N] NMR Reaction Spectra. *Analytical Chemistry* **76**: 7094-7101.
161. Castellanos ERR, Wist J. 2010. Decomposition of mixtures' spectra by multivariate curve resolution of rapidly acquired TOCSY experiments. *Magnetic Resonance in Chemistry* **48**: 771-776.
162. Snyder DA, Zhang F, Robinette SL, Bruschiweiler-Li L, Bruschiweiler R. 2008. Non-negative matrix factorization of two-dimensional NMR spectra: application to complex mixture analysis. *The Journal of Chemical Physics* **128**: 052313.
163. Koskela H. 2009 Chapter 1 - Quantitative 2D NMR Studies. In *Annual Reports on NMR Spectroscopy* (pp. 1-31). Academic Press.
164. Mutzenhardt P, Guenneau F, Canet D. 1999. A Procedure for Obtaining Pure Absorption 2D *J*-Spectra: Application to Quantitative Fully *J*-Decoupled Homonuclear NMR Spectra. *Journal of Magnetic Resonance* **141**: 312-321.
165. Hu KF, Westler WM, Markley JL. 2011. Simultaneous quantification and identification of individual chemicals in metabolite mixtures by two-dimensional extrapolated time-zero ^1H - ^{13}C HSQC (HSQC₀). *Journal of the American Chemical Society* **133**: 1662-1665.
166. Braunschweiler L, Ernst RR. 1983. Coherence transfer by isotropic mixing: Application to proton correlation spectroscopy. *Journal of Magnetic Resonance (1969)* **53**: 521-528.
167. Bharti SK, Roy R. 2012. Quantitative ^1H NMR spectroscopy. *TrAC Trends in Analytical Chemistry* **35**: 5-26.
168. McKay RT. 2011. How the 1D-NOESY suppresses solvent signal in metabolomics NMR spectroscopy: An examination of the pulse sequence components and evolution. *Concepts in Magnetic Resonance Part A* **38A**: 197-220.
169. Brereton RG. 2018 *Chemometrics: Data Driven Extraction for Science*. Wiley.
170. Valdivielso AM, Puig-Castellví F, Atcher J, Solà J, Tauler R, *et al.* 2017. Unraveling the Multistimuli Responses of a Complex Dynamic System of Pseudopeptidic Macrocycles. *Chemistry – A European Journal* **23**: 10789-10799.
171. Mocák J. 2012. Chemometrics in Medicine and Pharmacy. *Nova Biotechnologica et Chimica*. **11**: 11-26.
172. Folch-Fortuny A, Tortajada M, Prats-Montalbán JM, Llaneras F, Picó J, *et al.* 2015. MCR-ALS on metabolic networks: Obtaining more meaningful pathways. *Chemometrics and Intelligent Laboratory Systems* **142**: 293-303.
173. Challa S, Potumarthi R. 2013. Chemometrics-based process analytical technology (PAT) tools: applications and adaptation in pharmaceutical and biopharmaceutical industries. *Applied Biochemistry and Biotechnology* **169**: 66-76.

174. Trygg J, Lundstedt T. 2007 Chapter 6 - Chemometrics Techniques for Metabonomics. In *The Handbook of Metabonomics and Metabolomics* (pp. 171-199). Amsterdam: Elsevier Science B.V.
175. Legeret B, Schulz-Raffelt M, Nguyen HM, Auroy P, Beisson F, *et al.* 2016. Lipidomic and transcriptomic analyses of *Chlamydomonas reinhardtii* under heat stress unveil a direct route for the conversion of membrane lipids into storage lipids. *Plant, Cell & Environment* **39**: 834-847.
176. McLennan F, Kowalski BR. 1995 *Process Analytical Chemistry*. Springer Netherlands.
177. Massart DL, Vandeginste BG, Buydens LMC, Lewi PJ, Smeyers-Verbeke J, *et al.* 1997 *Handbook of Chemometrics and Qualimetrics: Part A*. Elsevier Science Inc.
178. Ballabio D. 2006 *Chemometric characterisation of physical-chemical fingerprints of food products*. Milano: Università degli Studi di Milano.
179. Tauler R, Kowalski B, Fleming S. 1993. Multivariate Curve Resolution Applied to Spectral Data from Multiple Runs of an Industrial-Process. *Analytical Chemistry* **65**: 2040-2047.
180. Karakach TK, Knight R, Lenz EM, Viant MR, Walter JA. 2009. Analysis of time course ¹H NMR metabolomics data by multivariate curve resolution. *Magnetic Resonance in Chemistry* **47**: S105-S117.
181. Smilde AK, Jansen JJ, Hoefsloot HCJ, Lamers R-JAN, van der Greef J, *et al.* 2005. ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics* **21**: 3043-3048.
182. Becerra-Martínez E, Florentino-Ramos E, Pérez-Hernández N, Gerardo Zepeda-Vallejo L, Villa-Ruano N, *et al.* 2017. ¹H NMR-based metabolomic fingerprinting to determine metabolite levels in serrano peppers (*Capsicum annum L.*) grown in two different regions. *Food Research International* **102**: 163-170.
183. Tomita S, Saito K, Nakamura T, Sekiyama Y, Kikuchi J. 2017. Rapid discrimination of strain-dependent fermentation characteristics among *Lactobacillus* strains by NMR-based metabolomics of fermented vegetable juice. *PLOS ONE* **12**: e0182229.
184. Scholz M. 2006 Approaches to analyse and interpret biological profile data. Methoden zur Analyse und Interpretation biologischer Profildaten. Potsdam: Universität Potsdam.
185. Bro R, Smilde AK. 2014. Principal component analysis. *Analytical Methods* **6**: 2812-2831.
186. Abdi H, Williams LJ. 2010. Principal component analysis. Wiley Interdisciplinary Reviews: *Computational Statistics* **2**: 433-459.
187. Golub GH, Van Loan CF. 1996 *Matrix Computations*. Johns Hopkins University Press.
188. David FN, 1966 *Research papers in statistics: festschrift for J. Neyman*. London: Wiley.
189. Cattell RB. 1966. The Scree Test For The Number Of Factors. *Multivariate Behavioral Research* **1**: 245-276.
190. Kaiser HF. 1960. The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement* **20**: 141-151.
191. Frontier S. 1976. Étude de la décroissance des valeurs propres dans une analyse en composantes principales: Comparaison avec le modèle du bâton brisé. *Journal of Experimental Marine Biology and Ecology* **25**: 67-75.
192. Huo Y, Kamal GM, Wang J, Liu H, Zhang G, *et al.* 2017. ¹H NMR-based metabolomics for discrimination of rice from different geographical origins of China. *Journal of Cereal Science* **76**: 243-252.

193. Cappello T, Maisano M, Mauceri A, Fasulo S. 2017. ¹H NMR-based metabolomics investigation on the effects of petrochemical contamination in posterior adductor muscles of caged mussel *Mytilus galloprovincialis*. *Ecotoxicology and Environmental Safety* **142**: 417-422.
194. Mazzei P, Spaccini R, Francesca N, Moschetti G, Piccolo A. 2013. Metabolomic by ¹H NMR Spectroscopy Differentiates "Fiano Di Avellino" White Wines Obtained with Different Yeast Strains. *Journal of Agricultural and Food Chemistry* **61**: 10816-10822.
195. Singh A, Sharma RK, Chagtoo M, Agarwal G, George N, *et al.* 2017. ¹H NMR Metabolomics Reveals Association of High Expression of Inositol 1, 4, 5 Trisphosphate Receptor and Metabolites in Breast Cancer Patients. *PLOS ONE* **12**: e0169330.
196. Mediani A, Abas F, Khatib A, Tan CP, Ismail IS, *et al.* 2015. Phytochemical and biological features of *Phyllanthus niruri* and *Phyllanthus urinaria* harvested at different growth stages revealed by ¹H NMR-based metabolomics. *Industrial Crops and Products* **77**: 602-613.
197. Awin T, Mediani A, Maulidiani, Shaari K, Faudzi SMM, *et al.* 2016. Phytochemical profiles and biological activities of *Curcuma* species subjected to different drying methods and solvent systems: NMR-based metabolomics approach. *Industrial Crops and Products* **94**: 342-352.
198. Qiu Y. 2016. Serum metabolomic analysis of rats with cisplatin-induced nephrotoxicity and panax notoginseng saponins treatment. *International journal of clinical and experimental medicine* **9**: 19291-19301.
199. Lamego I, Duarte IF, Marques MPM, Gil AM. 2014. Metabolic Markers of MG-63 Osteosarcoma Cell Line Response to Doxorubicin and Methotrexate Treatment: Comparison to Cisplatin. *Journal of Proteome Research* **13**: 6033-6045.
200. Zhao H, Xu J, Ghebrezadik H, Hylands PJ. 2015. Metabolomic quality control of commercial Asian ginseng, and cultivated and wild American ginseng using ¹H NMR and multi-step PCA. *Journal of Pharmaceutical and Biomedical Analysis* **114**: 113-120.
201. Wold S, Sjöström M, Eriksson L. 2001. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* **58**: 109-130.
202. Abdi H. 2010. Partial least squares regression and projection on latent structure regression (PLS Regression). *WIREs Computational Statistics* **2**: 97-106.
203. Ballabio D, Consonni V. 2013. Classification tools in chemistry. Part 1: linear models. PLS-DA. *Analytical Methods* **5**: 3790-3798.
204. Wold S, Johansson E, Cocchi M. 1993 PLS – Partial Least Squares Projection to Latent Structures. In *3D QSAR in Drug Design: Volume 1: Theory Methods and Applications* (pp. 523-550). Leiden. Springer.
205. Chong IG, Jun CH. 2005. Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems* **78**: 103-112.
206. Gorrochategui E, Casas J, Porte C, Lacorte S, Tauler R. 2015. Chemometric strategy for untargeted lipidomics: Biomarker detection and identification in stressed human placental cells. *Analytica Chimica Acta* **854**: 20-33.
207. Szymanska E, Saccenti E, Smilde AK, Westerhuis JA. 2012. Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics* **8**: 3-16.
208. Wold S, Antti H, Lindgren F, Öhman J. 1998. Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems* **44**: 175-185.

209. Fulcher YG, Fotso M, Chang CH, Rindt H, Reinero CR, *et al.* 2016. Noninvasive Recognition and Biomarkers of Early Allergic Asthma in Cats Using Multivariate Statistical Analysis of NMR Spectra of Exhaled Breath Condensate. *PLOS ONE* **11**: e0164394.
210. Motta A, Paris D, D'Amato M, Melck D, Calabrese C, *et al.* 2014. NMR Metabolomic Analysis of Exhaled Breath Condensate of Asthmatic Patients at Two Different Temperatures. *Journal of Proteome Research* **13**: 6107-6120.
211. Li Y, Yan GY, Zhou JQ, Bu Q, Deng PC, *et al.* 2012. ¹H NMR-based metabonomics in brain nucleus accumbens and striatum following repeated cocaine treatment in rats. *Neuroscience* **218**: 196-205.
212. Sadykov MR, Zhang B, Halouska S, Nelson JL, Kreimer LW, *et al.* 2010. Using NMR metabolomics to investigate tricarboxylic acid cycle-dependent signal transduction in *Staphylococcus epidermidis*. *The Journal of Biological Chemistry* **285**: 36616-36624.
213. Toya Y, Nakahigashi K, Tomita M, Shimizu K. 2012. Metabolic regulation analysis of wild-type and *arcA* mutant *Escherichia coli* under nitrate conditions using different levels of omics data. *Molecular Biosystems* **8**: 2593-2604.
214. Sayqal A, Xu Y, Trivedi DK. 2016. Metabolic analysis of the response of *Pseudomonas putida* DOT-T1E strains to toluene using Fourier transform infrared spectroscopy and gas chromatography mass spectrometry. *Metabolomics* **12**: 112.
215. Pitard FF. 2009 Chapter 1.01 - An Introduction to the Theory of Sampling: An Essential Part of Total Quality Management A2 - Brown, Steven D. In Tauler R, Walczak B, editors. *Comprehensive Chemometrics* (pp. 1-16). Oxford: Elsevier.
216. Stahle L, Wold S. 1990. Multivariate analysis of variance (MANOVA). *Chemometrics and Intelligent Laboratory Systems* **9**: 127-141.
217. Jansen JJ, Hoefsloot HCJ, van der Greef J, Timmerman ME, Westerhuis JA, *et al.* 2005. ASCA: analysis of multivariate data obtained from an experimental design. *Journal of Chemometrics* **19**: 469-481.
218. Berge JMFt, Kiers HAL, Stel VVd. 1992. Simultaneous Component Analysis. *Statistica Applicata* **4**: 377-392.
219. Timmerman ME, Hoefsloot HCJ, Smilde AK, Ceulemans E. 2015. Scaling in ANOVA-simultaneous component analysis. *Metabolomics* **11**: 1265-1276.
220. Hoefsloot HCJ, Vis DJ, Westerhuis JA, Smilde AK, Jansen JJ. 2009 Chapter 2.23 - Multiset Data Analysis: ANOVA Simultaneous Component Analysis and Related Methods A2 - Brown, Steven D. In Tauler R, Walczak B, editors. *Comprehensive Chemometrics* (pp. 453-472). Oxford: Elsevier.
221. Lawton WH, Sylvestre EA. 1971. Self Modeling Curve Resolution. *Technometrics* **13**: 617-633.
222. Tauler R, Smilde A, Kowalski B. 1995. Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. *Journal of Chemometrics* **9**: 31-58.
223. Tauler R, Juan Ad. 2006 Multivariate Curve Resolution. In: Gemperline P, editor. *Practical Guide To Chemometrics*: CRC Press.
224. Næs T. 2002 *A User-friendly Guide to Multivariate Calibration and Classification*. NIR Publications.
225. Mark H. 1991 *Principles and Practice of Spectroscopic Calibration*. Wiley.

226. Ahmadi G, Tauler R, Abdollahi H. 2015. Multivariate calibration of first-order data with the correlation constrained MCR-ALS method. *Chemometrics and Intelligent Laboratory Systems* **142**: 143-150.
227. Bauza MC, Ibanez GA, Tauler R, Olivieri AC. 2012. Sensitivity equation for quantitative analysis with multivariate curve resolution-alternating least-squares: theoretical and experimental approach. *Analytical Chemistry* **84**: 8697-8706.
228. Windig W, Stephenson DA. 1992. Self-modeling mixture analysis of second-derivative near-infrared spectral data using the SIMPLISMA approach. *Analytical Chemistry* **64**: 2735-2742.
229. Keller HR, Massart DL. 1991. Evolving factor analysis. *Chemometrics and Intelligent Laboratory Systems* **12**: 209-224.
230. Lawson C, Hanson R. 1995 *Solving Least Squares Problems*. Society for Industrial and Applied Mathematics.
231. Bro R, De Jong S. 1999. A fast non-negativity-constrained least squares algorithm. *Journal of Chemometrics* **11**: 393-401.
232. Ghosh S, Sengupta A, Sharma S, Sonawat HM. 2011. Multivariate modelling with ¹H NMR of pleural effusion in murine cerebral malaria. *Malaria Journal* **10**: 330.
233. Fernández C, Larrechi MS, Callao MP. 2009. Study of the influential factors in the simultaneous photocatalytic degradation process of three textile dyes. *Talanta* **79**: 1292-1297.
234. Peré-Trepát E, Tauler R. 2006. Analysis of environmental samples by application of multivariate curve resolution on fused high-performance liquid chromatography–diode array detection mass spectrometry data. *Journal of Chromatography A* **1131**: 85-96.
235. Vives M, Gargallo R, Tauler R, Moreno V. 2001. Study of the interaction of cis-dichloro-(1,2 diethyl-3-aminopyrrolidine)Pt(II) complex with poly(I), poly(C) and poly(I)·poly(C). *Journal of Inorganic Biochemistry* **85**: 279-290.
236. Fernández C, Pilar Callao M, Soledad Larrechi M. 2013. UV-visible-DAD and ¹H-NMR spectroscopy data fusion for studying the photodegradation process of azo-dyes using MCR-ALS. *Talanta* **117**: 75-80.
237. Gorrochategui E, Jaumot J, Lacorte S, Tauler R. 2016. Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow. *TrAC Trends in Analytical Chemistry* **82**: 425-442.
238. Szeto SW, Reinke S, Lemire B. 2011. ¹H NMR-based metabolic profiling reveals inherent biological variation in yeast and nematode model systems. *Journal of Biomolecular NMR* **49**: 245-254.
239. Monakhova YB, Tsikin AM, Kuballa T, Lachenmeier DW, Mushtakova SP. 2014. Independent component analysis (ICA) algorithms for improved spectral deconvolution of overlapped signals in ¹H NMR analysis: application to foods and related products. *Magnetic Resonance in Chemistry* **52**: 231-240.
240. Motegi H, Tsuboi Y, Saga A, Kagami T, Inoue M, *et al.* 2015. Identification of Reliable Components in Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS): a Data-Driven Approach across Metabolic Processes. *Scientific Reports* **5**: 15710.
241. Eads CD, Furnish CM, Noda I, Juhlin KD, Cooper DA, *et al.* 2004. Molecular Factor Analysis Applied To Collections of NMR Spectra. *Analytical Chemistry* **76**: 1982-1990.
242. Hyvärinen A, Oja E. 2000. Independent component analysis: algorithms and applications. *Neural Networks* **13**: 411-430.

243. Puig-Castellví F, Alfonso I, Tauler R. 2017. Untargeted assignment and automatic integration of ^1H NMR metabolomic datasets using a multivariate curve resolution approach. *Analytica Chimica Acta* **964**: 55-66.
244. Noothalapati H, Iwasaki K, Yamamoto T. 2017. Biological and Medical Applications of Multivariate Curve Resolution Assisted Raman Spectroscopy. *Analytical Sciences* **33**: 15-22.
245. Jaumot J, Piña B, Tauler R. 2010. Application of multivariate curve resolution to the analysis of yeast genome-wide screens. *Chemometrics and Intelligent Laboratory Systems* **104**: 53-64.
246. Navarro-Reig M, Jaumot J, Baglai A, Vivó-Truyols G, Schoenmakers PJ, et al. 2017. Untargeted Comprehensive Two-Dimensional Liquid Chromatography Coupled with High-Resolution Mass Spectrometry Analysis of Rice Metabolome Using Multivariate Curve Resolution. *Analytical Chemistry* **89**: 7675-7683.
247. Olmos V, Marro M, Loza-Alvarez P, Raldúa D, Prats E, et al. 2017. Combining hyperspectral imaging and chemometrics to assess and interpret the effects of environmental stressors on zebrafish eye images at tissue level. *Journal of Biophotonics*: e201700089.
248. Olmos V, Benítez L, Marro M, Loza-Alvarez P, Piña B, et al. 2017. Relevant aspects of unmixing/resolution analysis for the interpretation of biological vibrational hyperspectral images. *TrAC Trends in Analytical Chemistry* **94**: 130-140.
249. Bedia C, Tauler R, Jaumot J. 2017. Analysis of multiple mass spectrometry images from different *Phaseolus vulgaris* samples by multivariate curve resolution. *Talanta* **175**: 557-565.
250. Jaumot J, Tauler R. 2015. Potential use of multivariate curve resolution for the analysis of mass spectrometry images. *Analyst* **140**: 837-846.
251. Montoliu I, Martin Fo-PJ, Collino S, Rezzi S, Kochhar S. 2009. Multivariate Modeling Strategy for Intercompartmental Analysis of Tissue and Plasma ^1H NMR Spectrotypes. *Journal of Proteome Research* **8**: 2397-2406.
252. Bundy JG, Davey MP, Viant MR. 2008. Environmental metabolomics: a critical review and future perspectives. *Metabolomics* **5**: 3.
253. Walker GM. 1998 *Yeast Physiology and Biotechnology*. Wiley.
254. Gardner JM, Jaspersen SL. 2014 Manipulating the Yeast Genome: Deletion, Mutation, and Tagging by PCR. In Smith JS, Burke DJ, editors. *Yeast Genetics: Methods and Protocols* (pp. 45-78). New York, NY. Springer.
255. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, et al. 1996. Life with 6000 Genes. *Science* **274**: 546-567.
256. Couvillion MT, Soto IC, Shipkovenska G, Churchman LS. 2016. Synchronized mitochondrial and cytosolic translation programs. *Nature* **533**: 499-503.
257. Treco DA, Winston F. 2001 *Growth and Manipulation of Yeast*. *Current Protocols in Molecular Biology*. John Wiley & Sons, Inc.
258. Werner-Washburne M, Braun EL, Crawford ME, Peck VM. 1996. Stationary phase in *Saccharomyces cerevisiae*. *Molecular Microbiology* **19**: 1159-1166.
259. Caudy AA, Mülleder M, Ralser M. 2017. *Metabolomics in Yeast*. Cold Spring Harbor Protocols 2017.

260. Carneiro S, Pereira R, Rocha I. 2014 Yeast Metabolomics: Sample Preparation for a GC/MS-Based Analysis. In Mapelli V, editor. *Yeast Metabolic Engineering: Methods and Protocols* (pp. 197-207). New York, NY. Springer New York.
261. Canelas AB, ten Pierick A, Ras C, Seifar RM, van Dam JC, *et al.* 2009. Quantitative Evaluation of Intracellular Metabolite Extraction Techniques for Yeast Metabolomics. *Analytical Chemistry* **81**: 7379-7389.
262. Ewald Jennifer C, Kuehne A, Zamboni N, Skotheim Jan M. 2016. The Yeast Cyclin-Dependent Kinase Routes Carbon Fluxes to Fuel Cell Cycle Progression. *Molecular Cell* **62**: 532-545.
263. Laporte D, Lebaudy A, Sahin A, Pinson B, Ceschin J, *et al.* 2011. Metabolic status rather than cell cycle signals control quiescence entry and exit. *The Journal of Cell Biology* **192**: 949-957.
264. Li X, Snyder MP. 2016. Yeast longevity promoted by reversing aging-associated decline in heavy isotope content. *Npj Aging And Mechanisms Of Disease* **2**: 16004.
265. Mulleder M, Calvani E, Alam MT, Wang RK, Eckerstorfer F, *et al.* 2016. Functional Metabolomics Describes the Yeast Biosynthetic Regulome. *Cell* **167**: 553-565.
266. Lopez-Martinez G, Borrull A, Poblet M, Roy NR, Cordero-Otero R. 2014. Metabolomic characterization of yeast cells after dehydration stress. *International Microbiology* **17**: 131-139.
267. Boer VM, Amini S, Botstein D. 2008. Influence of genotype and nutrition on survival and metabolism of starving yeast. *Proceedings of the National Academy of Sciences* **105**: 6930-6935.
268. Boer VM, Crutchfield CA, Bradley PH, Botstein D, Rabinowitz JD. 2010. Growth-limiting intracellular metabolites in yeast growing under diverse nutrient limitations. *Molecular Biology of the Cell* **21**: 198-211.
269. Petti AA, Crutchfield CA, Rabinowitz JD, Botstein D. 2011. Survival of starving yeast is correlated with oxidative stress response and nonrespiratory mitochondrial function. *Proceedings of the National Academy of Sciences* **108**: E1089–E1098.
270. Ohta E, Nakayama Y, Mukai Y, Bamba T, Fukusaki E. 2016. Metabolomic approach for improving ethanol stress tolerance in *Saccharomyces cerevisiae*. *Journal of Bioscience and Bioengineering* **121**: 399-405.
271. Ortiz-Villanueva E, Jaumot J, Benavente F, Piña B, Sanz-Nebot V, *et al.* 2015. Combination of CE-MS and advanced chemometric methods for high-throughput metabolic profiling. *Electrophoresis* **36**: 2324-2335.
272. Farrés M, Piña B, Tauler R. 2015. Chemometric evaluation of *Saccharomyces cerevisiae* metabolic profiles using LC–MS. *Metabolomics* **11**: 210-224.
273. Chae YK, Kim SH, Ellinger JE, Markley JL. 2013. Dosage effects of salt and pH stresses on *Saccharomyces cerevisiae* as monitored via metabolites by using two dimensional NMR spectroscopy. *Bulletin of the Korean Chemical Society* **34**: 3602.
274. Sévin DC, Stählin JN, Pollak GR, Kuehne A, Sauer U. 2016. Global Metabolic Responses to Salt Stress in Fifteen Species. *PLOS ONE* **11**: e0148888.
275. Farres M, Pina B, Tauler R. 2016. LC-MS based metabolomics and chemometrics study of the toxic effects of copper on *Saccharomyces cerevisiae*. *Metallomics* **8**: 790-798.
276. Boone CHT, Grove RA, Adamcova D, Seravalli J, Adamec J. 2017. Oxidative stress, metabolomics profiling, and mechanism of local anesthetic induced cell death in yeast. *Redox Biology* **12**: 139-149.

277. Madigan MT, Bender KS, Buckley DH, Sattley WM, Stahl DA. 2017 *Brock Biology of Microorganisms*, Global Edition. Pearson Education Limited.
278. van der Greef J, Smilde AK. 2005. Symbiosis of chemometrics and metabolomics: past, present, and future. *Journal of Chemometrics* **19**: 376-386.
279. Müllleder M, Capuano F, Pir P, Christen S, Sauer U, *et al.* 2012. A prototrophic deletion mutant collection for yeast metabolomics and systems biology. *Nature Biotechnology* **30**: 1176-1178.
280. Almeida P, Barbosa R, Zalar P, Imanishi Y, Shimizu K, *et al.* 2015. A population genomics insight into the Mediterranean origins of wine yeast domestication. *Molecular Ecology* **24**: 5412-5427.
281. Rikhvanov EG, Varakina NN, Sozinov DY, Voinikov VK. 1999. Association of Bacteria and Yeasts in Hot Springs. *Applied and Environmental Microbiology* **65**: 4292-4293.
282. Buzzini P, Branda E, Goretti M, Turchetti B. 2012. Psychrophilic yeasts from worldwide glacial habitats: diversity, adaptation strategies and biotechnological potential. *FEMS Microbiology Ecology* **82**: 217-241.
283. Los DA, Murata N. 2004. Membrane fluidity and its roles in the perception of environmental signals. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1666**: 142-157.
284. Hassan N, Rafiq M, Hayat M, Shah AA, Hasan F. 2016. Psychrophilic and psychrotrophic fungi: a comprehensive review. *Reviews in Environmental Science and Bio/Technology* **15**: 147-172.
285. Elbein AD, Pan YT, Pastuszak I, Carroll D. 2003. New insights on trehalose: a multifunctional molecule. *Glycobiology* **13**: 17R-27R.
286. Aguilera J, Randez-Gil F, Prieto JA. 2007. Cold response in *Saccharomyces cerevisiae*: new functions for old mechanisms. *FEMS Microbiology Reviews* **31**: 327-341.
287. Sahara T, Goda T, Ohgiya S. 2002. Comprehensive Expression Analysis of Time-dependent Genetic Responses in Yeast Cells to Low Temperature. *Journal of Biological Chemistry* **277**: 50015-50021.
288. Duman JG, Olsen TM. 1993. Thermal Hysteresis Protein Activity in Bacteria, Fungi, and Phylogenetically Diverse Plants. *Cryobiology* **30**: 322-328.
289. Menonides FIC, Hellingwerf KJ, de Mattos MJT, Brul S. 2013. Multiphasic adaptation of the transcriptome of *Saccharomyces cerevisiae* to heat stress. *Food Research International* **54**: 1103-1112.
290. Attfield PV. 1987. Trehalose accumulates in *Saccharomyces cerevisiae* during exposure to agents that induce heat shock response. *FEBS Letters* **225**: 259-263.
291. Causton HC, Ren B, Koh SS, Harbison CT, Kanin E, *et al.* 2001. Remodeling of Yeast Genome Expression in Response to Environmental Changes. *Molecular Biology of the Cell* **12**: 323-337.
292. Groušl T, Ivanov P, Frydlová I, Vašicová P, Janda F, *et al.* 2009. Robust heat shock induces eIF2 α -phosphorylation-independent assembly of stress granules containing eIF3 and 40S ribosomal subunits in budding yeast, *Saccharomyces cerevisiae*. *Journal of Cell Science* **122**: 2078-2088.
293. Morano KA, Grant CM, Moye-Rowley WS. 2012. The Response to Heat Shock and Oxidative Stress in *Saccharomyces cerevisiae*. *Genetics* **190**: 1157-1195.

294. Widmann C, Gibson S, B. Jarpe M, Johnson GL. 1999. Mitogen-Activated Protein Kinase: Conservation of a Three-Kinase Module From Yeast to Human. *Physiological Reviews* **79**: 143-180.
295. Levin DE. 2005. Cell Wall Integrity Signaling in *Saccharomyces cerevisiae*. *Microbiology and Molecular Biology Reviews* **69**: 262-291.
296. Sugiyama K, Kawamura A, Izawa S, Inoue Y. 2000. Role of glutathione in heat-shock-induced cell death of *Saccharomyces cerevisiae*. *Biochemical Journal* **352**: 71-78.
297. Taymaz-Nikerel H, Cankorur-Cetinkaya A, Kirdar B. 2016. Genome-Wide Transcriptional Response of *Saccharomyces cerevisiae* to Stress-Induced Perturbations. *Frontiers in Bioengineering and Biotechnology* **4**: 17.
298. Roche B, Arcangioli B, Martienssen R. 2017. Transcriptional reprogramming in cellular quiescence. *RNA Biology* **14**: 843-853.
299. Cooper GM. 2000 *The Eukaryotic Cell Cycle. The Cell: A Molecular Approach*. ASM Press.
300. De Virgilio C. 2012 The essence of yeast quiescence. *FEMS Microbiology Reviews* **36**: 306-339.
301. Gray JV, Petsko GA, Johnston GC, Ringe D, Singer RA, *et al.* 2004. "Sleeping Beauty": Quiescence in *Saccharomyces cerevisiae*. *Microbiology and Molecular Biology Reviews* **68**: 187-206.
302. Smets B, Ghillebert R, De Snijder P, Binda M, Swinnen E, *et al.* 2010. Life in the midst of scarcity: adaptations to nutrient availability in *Saccharomyces cerevisiae*. *Current Genetics* **56**: 1-32.
303. Dhawan J, Laxman S. 2015. Decoding the stem cell quiescence cycle – lessons from yeast for regenerative biology. *Journal of Cell Science* **128**: 4467-4474.
304. Klosinska MM, Crutchfield CA, Bradley PH, Rabinowitz JD, Broach JR. 2011. Yeast cells can access distinct quiescent states. *Genes & Development* **25**: 336-349.
305. Rodkaer SV, Pultz D, Bruschi M, Bennetzen MV, Falkenby LG, *et al.* 2014. Quantitative proteomics identifies unanticipated regulators of nitrogen and glucose starvation. *Molecular Biosystems* **10**: 2176-2188.
306. Rødkær SV, Færgeman NJ. 2014. Glucose- and nitrogen sensing and regulatory mechanisms in *Saccharomyces cerevisiae*. *FEMS Yeast Research* **14**: 683-696.
307. Natarajan K, Meyer MR, Jackson BM, Slade D, Roberts C, *et al.* 2001. Transcriptional Profiling Shows that Gcn4p Is a Master Regulator of Gene Expression during Amino Acid Starvation in Yeast. *Molecular and Cellular Biology* **21**: 4347-4368.
308. Takeshige K, Baba M, Tsuboi S, Noda T, Ohsumi Y. 1992. Autophagy in yeast demonstrated with proteinase-deficient mutants and conditions for its induction. *The Journal of Cell Biology* **119**: 301-311.
309. Cebollero E, Reggiori F. 2009. Regulation of autophagy in yeast *Saccharomyces cerevisiae*. *BBA Molecular Cell Research* **1793**: 1413-1421.
310. González A, Hall MN. 2017. Nutrient sensing and TOR signaling in yeast and mammals. *The EMBO Journal* **36**: 397-408.
311. Laxman S, Sutter BM, Tu BP. 2013. Methionine is a signal of amino acid sufficiency that inhibits autophagy through the methylation of PP2A. *Autophagy* **10**: 386-387.
312. Sutter BM, Wu X, Laxman S, Tu BP. 2013. Methionine Inhibits Autophagy and Promotes Growth by Inducing the SAM-Responsive Methylation of PP2A. *Cell* **154**: 403-415.

313. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, *et al.* 1998. Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization. *Molecular Biology of the Cell* **9**: 3273-3297.
314. Klose C, Surma MA, Gerl MJ, Meyenhofer F, Shevchenko A, *et al.* 2012. Flexibility of a Eukaryotic Lipidome – Insights from Yeast Lipidomics. *PLOS ONE* **7**: e35063.
315. Hunter K, Rose AH. 1972. Lipid composition of *Saccharomyces cerevisiae* as influenced by growth temperature. *BBA - Lipids and Lipid Metabolism* **260**: 639-653.
316. Okuyama H, Saito M, Joshi VC, Gunsberg S, Wakil SJ. 1979. Regulation by temperature of the chain length of fatty acids in yeast. *Journal of Biological Chemistry* **254**: 12281-12284.
317. Torija MaJ, Beltran G, Novo M, Poblet M, Guillamón JM, *et al.* 2003. Effects of fermentation temperature and *Saccharomyces* species on the cell fatty acid composition and presence of volatile compounds in wine. *International Journal of Food Microbiology* **85**: 127-136.
318. Chatterjee MT, Khalawan SA, Curran BPG. 1997. Alterations in cellular lipids may be responsible for the transient nature of the yeast heat shock response. *Microbiology* **143**: 3063-3068.
319. Shimizu I, Katsuki H. 1975. Effect of temperature on ergosterol biosynthesis in yeast. *Journal of Biochemistry* **77**: 1023-1027.
320. Rossi M, Buzzini P, Cordisco L, Amaretti A, Sala M, *et al.* 2009. Growth, lipid accumulation, and fatty acid composition in obligate psychrophilic, facultative psychrophilic, and mesophilic yeasts. *FEMS Microbiology Ecology* **69**: 363-372.
321. Arthur H, Watson K. 1976. Thermal adaptation in yeast: growth temperatures, membrane lipid, and cytochrome composition of psychrophilic, mesophilic, and thermophilic yeasts. *Journal of Bacteriology* **128**: 56-68.
322. Řezanka T, Kolouchová I, Sigler K. 2016. Lipidomic analysis of psychrophilic yeasts cultivated at different temperatures. *BBA - Molecular and Cell Biology of Lipids* **1861**: 1634-1642.
323. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, *et al.* 2000. Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. *Molecular Biology of the Cell* **11**: 4241-4257.
324. Gasch AP, Werner-Washburne M. 2002. The genomics of yeast responses to environmental stress and starvation. *Functional & Integrative Genomics* **2**: 181-192.
325. Wu J, Zhang N, Hayes A, Panoutsopoulou K, Oliver SG. 2004. Global analysis of nutrient control of gene expression in *Saccharomyces cerevisiae* during growth and starvation. *Proceedings of the National Academy of Sciences* **101**: 3148-3153.
326. Neklesa TK, Davis RW. 2009. A Genome-Wide Screen for Regulators of TORC1 in Response to Amino Acid Starvation Reveals a Conserved Npr2/3 Complex. *PLOS Genetics* **5**: e1000515.
327. Oda A, Takemata N, Hirata Y, Miyoshi T, Suzuki Y, *et al.* 2015. Dynamic transition of transcription and chromatin landscape during fission yeast adaptation to glucose starvation. *Genes to Cells* **20**: 392-407.
328. Gresham D, Boer VM, Caudy A, Ziv N, Brandt NJ, *et al.* 2011. System-Level Analysis of Genes and Functions Affecting Survival During Nutrient Starvation in *Saccharomyces cerevisiae*. *Genetics* **187**: 299-317.

329. Strassburg K, Walther D, Takahashi H, Kanaya S, Kopka J. 2010. Dynamic transcriptional and metabolic responses in yeast adapting to temperature stress. *OMICS* **14**: 249-259.
330. López-Malo M, Querol A, Guillamon JM. 2013. Metabolomic Comparison of *Saccharomyces cerevisiae* and the Cryotolerant Species *S. bayanus* var. *uvarum* and *S. kudriavzevii* during Wine Fermentation at Low Temperature. *PLOS ONE* **8**: e60135.
331. Brauer MJ, Yuan J, Bennett BD, Lu W, Kimball E, *et al.* 2006. Conservation of the metabolomic response to starvation across two divergent microbes. *Proceedings of the National Academy of Sciences* **103**: 19302-19307.
332. Cloarec O, Dumas M-E, Craig A, Barton RH, Trygg J, *et al.* 2005. Statistical Total Correlation Spectroscopy: An exploratory approach for latent biomarker identification from metabolic ¹H NMR data sets. *Analytical Chemistry* **77**: 1282-1289.
333. Sandusky P, Raftery D. 2005. Use of Selective TOCSY NMR Experiments for Quantifying Minor Components in Complex Mixtures: Application to the Metabonomics of Amino Acids in Honey. *Analytical Chemistry* **77**: 2455-2463.
334. Shen B, Hohmann S, Jensen RG, Bohnert, Hans J. 1999. Roles of Sugar Alcohols in Osmotic Stress Adaptation. Replacement of Glycerol by Mannitol and Sorbitol in Yeast. *Plant Physiology* **121**: 45-52.
335. Siderius M, Van Wuytswinkel O, Reijenga KA, Kelders M, Mager WH. 2000. The control of intracellular glycerol in *Saccharomyces cerevisiae* influences osmotic stress response and resistance to increased temperature. *Molecular Microbiology* **36**: 1381-1390.
336. Hans M, Heinzle E, Wittmann C. 2001. Quantification of intracellular amino acids in batch cultures of *Saccharomyces cerevisiae*. *Applied Microbiology and Biotechnology* **56**: 776-779.
337. Jules M, Beltran G, François J, Parrou JL. 2008. New Insights into Trehalose Metabolism by *Saccharomyces cerevisiae*: NTH2 Encodes a Functional Cytosolic Trehalase, and Deletion of TPS1 Reveals Ath1p-Dependent Trehalose Mobilization. *Applied and Environmental Microbiology* **74**: 605-614.
338. de Graaf RA, Brown PB, McIntyre S, Nixon TW, Behar KL, *et al.* 2006. High magnetic field water and metabolite proton T₁ and T₂ relaxation in rat brain in vivo. *Magnetic Resonance in Medicine* **56**: 386-394.
339. Bloembergen N, Purcell EM, Pound RV. 1948. Relaxation Effects in Nuclear Magnetic Resonance Absorption. *Physical Review* **73**: 679-712.
340. Jacobsen NE. 2007 *Biological NMR Spectroscopy. NMR Spectroscopy Explained*. John Wiley & Sons, Inc.
341. Maris MA, Brown CW, Lavery DS. 1983. Nonlinear multicomponent analysis by infrared spectrophotometry. *Analytical Chemistry* **55**: 1694-1703.
342. Maeder M. 1987. Evolving factor analysis for the resolution of overlapping chromatographic peaks. *Analytical Chemistry* **59**: 527-530.
343. Havel TF, Najfeld I, Yang JX. 1994. Matrix decompositions of two-dimensional nuclear magnetic resonance spectra. *Proceedings of the National Academy of Sciences* **91**: 7962-7966.
344. Windig W, Antalek B. 1997. Direct exponential curve resolution algorithm (DECRA): A novel application of the generalized rank annihilation method for a single spectral mixture data set with exponentially decaying contribution profiles. *Chemometrics and Intelligent Laboratory Systems* **37**: 241-254.

345. Winning H, Larsen FH, Bro R, Engelsen SB. 2008. Quantitative analysis of NMR spectra with chemometrics. *Journal of Magnetic Resonance* **190**: 26-32.
346. Antalek B, Hewitt JM, Windig W, Yacobucci PD, Mourey T, *et al.* 2002. The use of PGSE NMR and DECRA for determining polymer composition. *Magnetic Resonance in Chemistry* **40**: S60-S71.
347. Dyrby M, Petersen M, Whittaker AK, Lambert L, Nørgaard L, *et al.* 2005. Analysis of lipoproteins using 2D diffusion-edited NMR spectroscopy and multi-way chemometrics. *Analytica Chimica Acta* **531**: 209-216.
348. Nilsson M, Khajeh M, Botana A, Bernstein MA, Morris GA. 2009. Diffusion NMR and trilinear analysis in the study of reaction kinetics. *Chemical Communications* **10**: 1252-1254.
349. Nilsson M, Botana A, Morris GA. 2009. T_1 -Diffusion-Ordered Spectroscopy: Nuclear Magnetic Resonance Mixture Analysis Using Parallel Factor Analysis. *Analytical Chemistry* **81**: 8119-8125.
350. Khajeh M, Botana A, Bernstein MA, Nilsson M, Morris GA. 2010. Reaction Kinetics Studied Using Diffusion-Ordered Spectroscopy and Multiway Chemometrics. *Analytical Chemistry* **82**: 2102-2108.
351. Björnerås J, Botana A, Morris GA, Nilsson M. 2014. Resolving complex mixtures: trilinear diffusion data. *Journal of Biomolecular NMR* **58**: 251-257.
352. Alam TM, Alam MK. 2003. Effect of non-exponential and multi-exponential decay behavior on the performance of the direct exponential curve resolution algorithm (DECRA) in NMR investigations. *Journal of Chemometrics* **17**: 583-593.
353. Antalek B. 2007. Using PGSE NMR for chemical mixture analysis: Quantitative aspects. *Concepts in Magnetic Resonance Part A* **30A**: 219-235.
354. Rodríguez-Vázquez N, Amorín M, Alfonso I, Granja JR. 2016. Anion Recognition and Induced Self-Assembly of an α,γ -Cyclic Peptide To Form Spherical Clusters. *Angewandte Chemie International Edition* **55**: 4504-4508.
355. Xu Q, Sachs JR, Wang TC, Schaefer WH. 2006. Quantification and Identification of Components in Solution Mixtures from 1D Proton NMR Spectra Using Singular Value Decomposition. *Analytical Chemistry* **78**: 7175-7185.
356. Shen H, Airiau CY, Brereton RG. 2002. Resolution of on-flow LC/NMR data by multivariate methods — a comparison. *Journal of Chemometrics* **16**: 469-481.
357. Wasim M, Brereton RG. 2005. Application of multivariate curve resolution methods to on-flow LC-NMR. *Journal of Chromatography A* **1096**: 2-15.
358. Jaumot J, Vives M, Gargallo R, Tauler R. 2003. Multivariate resolution of NMR labile signals by means of hard- and soft-modelling methods. *Analytica Chimica Acta* **490**: 253-264.
359. Llamas NE, Garrido M, Nezio MSD, Band BSF. 2009. Second order advantage in the determination of amaranth, sunset yellow FCF and tartrazine by UV-vis and multivariate curve resolution-alternating least squares. *Analytica Chimica Acta* **655**: 38-42.
360. Azzouz T, Tauler R. 2008. Application of multivariate curve resolution alternating least squares (MCR-ALS) to the quantitative analysis of pharmaceutical and agricultural samples. *Talanta* **74**: 1201-1210.

361. Ebrahimi P, Larsen FH, Jensen HM, Vogensen FK, Engelsen SB. 2016. Real-time metabolomic analysis of lactic acid bacteria as monitored by in vitro NMR and chemometrics. *Metabolomics* **12**: 77.
362. Engelsen S, Larsen F, Jensen H, Ebrahimi P. 2016. microPAT: A protocol for direct in vitro NMR observation of lactic acid bacteria fermentations. *Proceedings of the XIII International Conference on the Applications of Magnetic Resonance in Food Science*: 1-5.
363. Abdi H. 2007 Singular Value Decomposition (SVD) and Generalized Singular Value Decomposition (GSVD). In Salkind NJ, editor. *Encyclopedia of measurement and statistics* (pp. 907-912). SAGE Publications.
364. Sharma R, Gogna N, Singh H, Dorai K. 2017. Fast profiling of metabolite mixtures using chemometric analysis of a speeded-up 2D heteronuclear correlation NMR experiment. *RSC Advances* **7**: 29860-29870.
365. Hedenström M, Wiklund S, Sundberg B, Edlund U. 2008. Visualization and interpretation of OPLS models based on 2D NMR data. *Chemometrics and Intelligent Laboratory Systems* **92**: 110-117.
366. Hansen AL, Brüschweiler R. 2016. Absolute Minimal Sampling in High-Dimensional NMR Spectroscopy. *Angewandte Chemie International Edition* **55**: 14169-14172.
367. Lewis IA, Schommer SC, Hodis B, Robb KA, Tonelli M, *et al.* 2007. Method for Determining Molar Concentrations of Metabolites in Complex Solutions from Two-Dimensional ^1H - ^{13}C NMR Spectra. *Analytical Chemistry* **79**: 9385-9390.
368. Castañar L, Parella T. 2015 Chapter 4 - Recent Advances in Small Molecule NMR: Improved HSQC and HSQMBC Experiments. In Webb GA, editor. *Annual Reports on NMR Spectroscopy* (pp. 163-232). Academic Press.