DEVELOPMENT OF 1H-NMR SERUM PROFILING METHODS FOR HIGH-THROUGHPUT METABOLOMICS

Rubén Barrilero Regadera

ADVERTIMENT. L'accés als continguts d'aquesta tesi doctoral i la seva utilització ha de respectar els drets de la persona autora. Pot ser utilitzada per a consulta o estudi personal, així com en activitats o materials d'investigació i docència en els termes establerts a l'art. 32 del Text Refós de la Llei de Propietat Intel·lectual (RDL 1/1996). Per altres utilitzacions es requereix l'autorització prèvia i expressa de la persona autora. En qualsevol cas, en la utilització dels seus continguts caldrà indicar de forma clara el nom i cognoms de la persona autora i el títol de la tesi doctoral. No s'autoritza la seva reproducció o altres formes d'explotació efectuades amb finalitats de lucre ni la seva comunicació pública des d'un lloc aliè al servei TDX. Tampoc s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX (framing). Aquesta reserva de drets afecta tant als continguts de la tesi com als seus resums i índexs.

ADVERTENCIA. El acceso a los contenidos de esta tesis doctoral y su utilización debe respetar los derechos de la persona autora. Puede ser utilizada para consulta o estudio personal, así como en actividades o materiales de investigación y docencia en los términos establecidos en el art. 32 del Texto Refundido de la Ley de Propiedad Intelectual (RDL 1/1996). Para otros usos se requiere la autorización previa y expresa de la persona autora. En cualquier caso, en la utilización de sus contenidos se deberá indicar de forma clara el nombre y apellidos de la persona autora y el título de la tesis doctoral. No se autoriza su reproducción u otras formas de explotación efectuadas con fines lucrativos ni su comunicación pública desde un sitio ajeno al servicio TDR. Tampoco se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR (framing). Esta reserva de derechos afecta tanto al contenido de la tesis como a sus resúmenes e índices.

WARNING. Access to the contents of this doctoral thesis and its use must respect the rights of the author. It can be used for reference or private study, as well as research and learning activities or materials in the terms established by the 32nd article of the Spanish Consolidated Copyright Act (RDL 1/1996). Express and previous authorization of the author is required for any other uses. In any case, when using its content, full name of the author and title of the thesis must be clearly indicated. Reproduction or other forms of for profit use or public communication from outside TDX service is not allowed. Presentation of its content in a window or frame external to TDX (framing) is not authorized either. These rights affect both the content of the thesis and its abstracts and indexes.

UNIVERSITAT ROVIRA I VIRGILI
DEVELOPMENT OF 1H-NMR SERUM PROFILING METHODS FOR HIGH-THROUGHPUT METABOLOMICS
Rubén Barrilero Regadera

UNIVERSITAT ROVIRA i VIRGILI

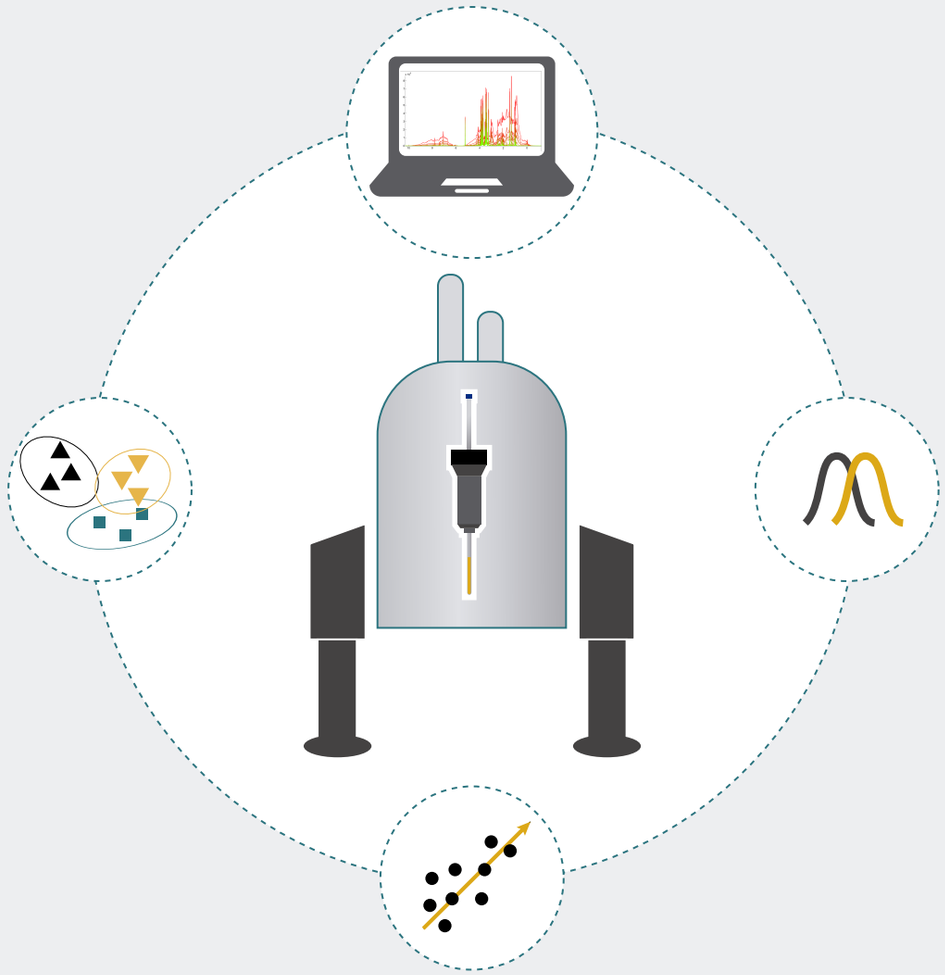
Development of ^{1}H-NMR Serum Profiling Methods for High-Throughput Metabolomics
Rubén Barrilero Regadera
2017

Doctoral Thesis

UNIVERSITAT ROVIRA i VIRGILI

Development of ^{1}H-NMR Serum Profiling Methods for High-Throughput Metabolomics

Rubén Barrilero Regadera



Doctoral Thesis
2017

UNIVERSITAT ROVIRA I VIRGILI
DEVELOPMENT OF 1H-NMR SERUM PROFILING METHODS FOR HIGH-THROUGHPUT METABOLOMICS
Rubén Barrilero Regadera

UNIVERSITAT ROVIRA I VIRGILI
DEVELOPMENT OF 1H-NMR SERUM PROFILING METHODS FOR HIGH-THROUGHPUT METABOLOMICS
Rubén Barrilero Regadera

Rubén Barrilero Regadera

Development of ^{1}H-NMR Serum Profiling Methods for High-Throughput Metabolomics

DOCTORAL THESIS

Supervised by Dr. Xavier Correig Blanchar

Departament d'Enginyeria Electrònica, Elèctrica i Automàtica

(DEEEA)

UNIVERSITAT ROVIRA I VIRGILI

Tarragona
2017

UNIVERSITAT ROVIRA I VIRGILI
DEVELOPMENT OF 1H-NMR SERUM PROFILING METHODS FOR HIGH-THROUGHPUT METABOLOMICS
Rubén Barrilero Regadera

UNIVERSITAT ROVIRA i VIRGILI

Escola Tècnica Superior d'Enginyeria
Departament d'Enginyeria Electrònica, Elèctrica i Automàtica
Av. Països Catalans 26
Campus Sescelades
43007 Tarragona

I STATE that the present study, entitled: "**Development of ^{1}H-NMR Serum Profiling Methods for High-Throughput Metabolomics**", presented by Rubén Barrilero Regadera for the award of the degree of Doctor, has been carried out under my supervision at the Department of Electronic, Electric and Automatic Engineering (DEEEA) of this university and meets the requirements to qualify for International Mention.

Tarragona, September 2017

Doctoral thesis supervisor

Dr. Xavier Correig Blanchar

UNIVERSITAT ROVIRA I VIRGILI
DEVELOPMENT OF 1H-NMR SERUM PROFILING METHODS FOR HIGH-THROUGHPUT METABOLOMICS
Rubén Barrilero Regadera

ACKNOWLEDGMENTS

En ocasiones resulta difícil comprender los motivos que dirigen nuestras acciones en la vida, lo cierto es que solo una conjunción de fuerzas hace que podamos llevarlas a cabo. Estos agradecimientos van dirigidos a todas esas fuerzas que han colaborado de una forma u otra en el desarrollo de esta tesis.

En primer lugar me gustaría dar las gracias a la persona que lo ha hecho materialmente posible. Gracias Xavier por concederme esta oportunidad y mantener tu confianza en mí durante estos años, y que me ha permitido descubrir un apasionante mundo que me era totalmente ajeno: la metabolómica. Tu dinamismo y entrega son un modelo a seguir, además, he de reconocer que sin tu fuerte convicción hoy no estaría escribiendo estas líneas.

Realizar esta tesis implicó también dejar atrás hogar, familia, pareja y amigos, un vacío que afortunadamente me han permitido sobrellevar otras tantas grandes personas. Gracias por tanto a Dídac por facilitarme la llegada a la escuela en esos primeros meses. A Núria, por ser tan acogedora y transparente, nunca olvidaré la cena de los idiotas y mi primer "al mar" de Manel. A Josep, Xavi y Daniel, gracias por la amistad, por el intercambio de conocimiento y por todas esas sesiones de "birroterapia" y desintoxicación laboral en la plaza mayor. A Lorena, Mabel, Salva y Pol, siempre os consideraré mi pandilla de Reus. Gracias también al resto de gente del departamento: Pere, Sonia, Serena, Nico y Jesús, por su predisposición a ayudar y por su amistad.

A la gente de Biosfer, en especial a Miriam y Rocío por la grandísima ayuda en el laboratorio. Ha sido un placer aprender de gente tan profesional. A Núria y Roger por sus múltiples colaboraciones y por allanar el camino a los nuevos doctorandos dejando su poso de sabiduría (y algún que otro código de Matlab).

A Miguel y a Mariona por ser ciencia en estado puro y trasmitir conocimiento de manera tan altruista. Para mí siempre seréis la M de NMR. Gracias también al resto de gente del COS: Miriam, Sara, Óscar y Jordi, por todos los buenos momentos compartidos.

I would like to thank Professor Bro and all the staff at the department of Food Science at the KU for the warm welcoming and for making my feel part of their team since the very first day. I enjoyed learning multivariate data analysis from the best. I would also like to thank to my "multicultural family" in Copenhagen: Jessica, Thomaz, Iuliana, Joe, Nunzia, Joana, Viola, Alex (x2) and Francesca. I had a wonderful time with you.

Por supuesto agradecer a la gente que siempre ha estado a mi lado en todo momento: a mis padres y hermana por vuestro apoyo incondicional, y a mis amigos, por creer en mí y por todas las vivencias de estos últimos años. Prometo recuperar el tiempo perdido…

Y por último a la persona que más merece mis agradecimientos: Vero. Gracias a tu apoyo personal ha sido posible que esta tesis haya salido adelante. Gracias por entender las decisiones tomadas estos años, por compartir las alegrías y por serenar mis momentos de crisis.

"Most of the fundamental ideas of science are essentially simple, and may, as a rule, be expressed in a language comprehensible to everyone"

Albert Einstein

UNIVERSITAT ROVIRA I VIRGILI
DEVELOPMENT OF 1H-NMR SERUM PROFILING METHODS FOR HIGH-THROUGHPUT METABOLOMICS
Rubén Barrilero Regadera

ABSTRACT

The irruption of metabolomics is replacing the traditional approach of clinical diagnostics focused on single biomarkers, such as glucose or cholesterol, with the profiling of complex metabolite patterns reflecting metabolic activity in multiple biological pathways. Blood serum/plasma is one of the main biological matrices used in NMR-based metabolomics as its collection is minimally invasive, requires minimal sample manipulation and provides hundreds of metabolites encoding multisystemic biological information.

High-throughput ^{1}H-NMR profiling of serum/plasma allows a quantitative multi-compound analysis including lipoprotein classes and constituent lipids, albumin, and a large variety of low-molecular-weight metabolites (LMWM), including amino acids, creatinine, glycolysis-related metabolites, and ketone bodies, with a cost similar to standard lipids. The large physicochemical heterogeneity of these compounds requires the acquisition of three ^{1}H-NMR measurements concerning the following molecular species: macromolecules, LMWM and lipids, where each measurement involves physical (sample extractions) and spectroscopic (editing NMR techniques) filters. However, molecular interactions and spectral complexity hamper a reliable metabolite profiling, which remains mostly manual. Developing robust and more automated methods of metabolite profiling is therefore desirable to consolidate high-throughput ^{1}H-NMR in the clinical practice.

In our first work, we calibrated and evaluated regression models to estimate the concentration of lipids used in the routine clinical practice (known as “lipid panel”). These lipids are still the main measurements and therapy targets of cardiovascular disease risk. Whereas most of the previous models have been calibrated using lipoprotein fractioning, our models were built using clinical enzymatic-colorimetric measurements in order to better reflect the clinical standards. Ultimately, these NMR-based regression models would lead to incorporate the standard clinical lipid panel in high-throughput ^{1}H-NMR profiling of serum and plasma. To do so, we developed and validated ^{1}H-NMR regression models of clinical measurements of total serum cholesterol and triglycerides, and cholesterol content of LDL, HDL and non-HDL particles, using 785 native serum/plasma samples comprising healthy subjects and subjects suffering from several dyslipidaemias. Different combinations of 1D and 2D ^{1}H-NMR experiments and chemometric techniques were evaluated. Moreover, our models used indistinctly plasma and serum samples, which were collected in four different clinical centres. Our lipid predictions performed similar to previous models based on

small and more homogeneous cohorts, but the diverse matrix and physiological conditions found in our samples made our models highly generalizable.

In our second work, we addressed the quantitative issues affecting the "NMR-invisible" low-molecular-weight metabolites (LMWM) in ^{1}H-NMR spectra of native serum. LMWM bind to proteins in native serum; consequently, their signals are totally or partially attenuated. These signal losses compromise absolute quantifications even if sophisticated signal deconvolution methods are used. In order to reduce protein binding effects on LMWM quantification, we developed a method to partially release bound LMWM from proteins. Our method relies on promoting competition for ligand-binding sites of proteins by the addition of a small quantity of deuterated trimethylsilylpropanoic acid (TSP). In order to precisely quantify the extent of these interactions, we performed our quantifications using a multidimensional CPMG approach, which avoids the signal attenuations due to T2 relaxations inserted with 1D CPMG filters. The application of both strategies showed that TSP addition increases in approximately 30% the signal for clinically-relevant binding metabolites phenylalanine, leucine and isoleucine. Moreover, competitive binding strategies are fully compatible with high-throughput analysis.

Finally, our third work addressed the quantitative profiling of serum lipids with ^{1}H-NMR. Whereas ^{1}H-NMR profiling of LMWM can be carried out with bioinformatics tools that allow automatic signal deconvolution based on specific metabolite signal patterns, similar solutions are not available for ^{1}H-NMR profiling of lipids. In this context we present LipSpin, a freely-distributed software for the semiautomatic profiling of ^{1}H-NMR spectra of lipids. Using a collection of signal patterns based on mathematical and reference spectral models, a constrained lineshape fitting analysis provides the quantification of 15 different lipid-related variables about major lipid classes in serum (fatty acids, triglycerides, phospholipids and cholesterols). Lipid quantifications obtained with LipSpin agreed with those from conventional techniques and were applied to a dietary intervention study.

LIST OF ABREVIATIONS

ARA	Arachidonic acid
ATP	Adult Treatment Panel
AUC	Area under the curve
BBPLED	Diffusion-editing pulse with bipolar gradients and longitudinal eddy-
CDC	Centers for disease control and prevention
CHD	Coronary heart disease
CIBERDEM	Centro de Investigación Biomédica en Red de Diabetes y enfermedades
COSY	Correlation spectroscopy
CPMG	Carr-Purcell-Meiboom-Gill
CSV	Comma-separated values file
CV	Coefficient of variation
DHA	Docosahexaenoic acid
DM2	Diabetes mellitus type 2
DSS	4,4-dimethyl-4-silapentane-1-sulfonic acid
DSTE	Double stimulated echo
EC	Esterified cholesterol
EDTA	Ethylenediaminetetraacetic acid
EPA	Eicosapentaenoic acid
ERETIC	Electronic reference to access in vivo Concentrations
FA	Fatty acids
FC	Free cholesterol
FFA	Free fatty acids
FID	Free induction decay
FT	Fourier transform
GC-FID	Gas chromatography – flame ionization detector
GlcNAc	N-Acetylglucosamine
GPL	Glycerophospholipids
HC	Hypercholesterolemia
HDL	High density lipoproteins
HDL-C	High density lipoprotein cholesterol
HMDB	Human Metabolome Database
HSA	Human serum albumin
HTG	Hypertriglyceridemia
IISPV	Institut d'Investigació Sanitària Pere Virgili
LDL	Low density lipoproteins
LDL-C	Low density lipoprotein cholesterol

LED	Longitudinal eddy-current delay
LMWM	Low-molecular-weight metabolites
LPC	Lysophosphatitylcholine
MCR-ALS	Multivariate curve resolution – alternating least squares
mRNA	Messenger ribonucleic acid
MS	Mass spectroscopy
MUFA	Monounsaturated fatty acids
NCEP	National Cholesterol Education Program
NMR	Nuclear magnetic resonance
N-PLS	N-way partial least squares
PARAFAC	Parallel factor analysis
PBS	Phosphate buffer solution
PC	Phosphatidylcholine
PCA	Principal component analysis
PE	Phosphatidylethanolamine
PL	Phospholipids
PLA	Plasmalogen
PLS	Partial least squares
PNG	Portable network graphics
PUFA	Polyunsaturated fatty acids
PULCON	Pulse Length based concentration determination
QUANTAS	Quantification by artificial signal
RF	Radiofrequency
ROC	Receiver operating characteristic
rRMSE	Relative root-mean-square error
RSD	Relative standard deviation
S/N	Signal-to-noise ratio
SD	Standard deviation
SFA	Saturated fatty acids
SIPOMICS	Signal processing for omics sciences
SM	Sphingomyelin
SOM	Self-organizing maps
STOCSY	Statistical total correlation spectroscopy
TC	Total cholesterol
TG	Triglycerides
TMS	Tetramethylsilane
TSP	Trimethylsilylpropanoic acid
VLDL	Very low density lipoproteins

LIST OF PUBLICATIONS

Rubén Barrilero*, Miriam Gil, Núria Amigó, Cintia Dias, Lisa G. Wood, Manohar L. Garg, Maria Vinaixa, Josep Ribalta, Mercedes Heras and Xavier Correig. *"LipSpin: a new bioinformatics tool for quantitative ^{1}H-NMR lipid profiling"*. [Submitted].

Rubén Barrilero*, Noelia Ramírez, Joan Carles Vallvé, Delia Taverner, Rocío Fuertes, Núria Amigó and Xavier Correig. *"Unravelling and Quantifying the "NMR-Invisible" Metabolites Interacting with Human Serum Albumin by Binding Competition and T2 Relaxation-Based Decomposition Analysis"*. Journal of Proteome Research , 2017, 16 (5), pp 1847–1856. DOI: 10.1021/acs.jproteome.6b00814

Rubén Barrilero*, Eduard Llobet, Roger Mallol, Jesús Brezmes, Lluis Masana, M. Ángeles Zulet, J. Alfredo Martínez, Josep Ribalta, Mònica Bulló and Xavier Correig. *"Design and evaluation of standard lipid prediction models based on ^{1}H-NMR spectroscopy of human serum/plasma samples"*. Metabolomics, 2015, 11 (5), pp 1394–1404. DOI: 10.1007/s11306-015-0796-5

UNIVERSITAT ROVIRA I VIRGILI
DEVELOPMENT OF 1H-NMR SERUM PROFILING METHODS FOR HIGH-THROUGHPUT METABOLOMICS
Rubén Barrilero Regadera

LIST OF CONGRESSES

Rubén Barrilero*, Roger Mallol and Xavier Correig. "Improving the Quantification of Amino Acids in Plasma/Serum by ^{1}H-NMR Spectroscopy, Considering their Interaction with Human Serum Albumin by ^{1}H-NMR Spectroscopy". Small Molecule NMR Conference (SMASH), Baveno, Italy (2015).

Josep Gómez*, **Rubén Barrilero**, Xavier Domingo, Xavier Correig and Nicolau Cañellas. "Evaluation of Multivariate Curve Resolution for Macromolecular Baseline Removal in ^{1}HNMR Spectra". Small Molecule NMR Conference (SMASH), Baveno, Italy (2015).

Rubén Barrilero* and Xavier Correig. "Enhancing the quantification of amino acids by ^{1}H-NMR spectroscopy, considering their interaction with human serum albumin". Metabolomics 2015, San francisco, USA (2015).

UNIVERSITAT ROVIRA I VIRGILI
DEVELOPMENT OF 1H-NMR SERUM PROFILING METHODS FOR HIGH-THROUGHPUT METABOLOMICS
Rubén Barrilero Regadera

TABLE OF CONTENTS

UNIVERSITAT ROVIRA I VIRGILI
DEVELOPMENT OF 1H-NMR SERUM PROFILING METHODS FOR HIGH-THROUGHPUT METABOLOMICS
Rubén Barrilero Regadera

CHAPTER 1

Introduction

UNIVERSITAT ROVIRA I VIRGILI
DEVELOPMENT OF 1H-NMR SERUM PROFILING METHODS FOR HIGH-THROUGHPUT METABOLOMICS
Rubén Barrilero Regadera

1.1. Metabolomics

The scientific techniques and approaches applied to molecular biology and biochemistry have experienced a dramatic change with the irruption of omics sciences. This revolution took place in the 1990s, and especially with the first determination of the human genome and the availability of automated micro-array methods [1]. The integration of omics sciences has motivated the new field of "systems biology" [2,3].

In the upper layers of the "omics cascade", genomics, transcriptomics and proteomics involve the global study of genes and proteins in a cell or organism, which are subject to epigenetic regulation and post-translational modifications [4]. The complete understanding of a biological system at this level is however uncertain; it is often difficult to relate observed gene expression changes to conventional end-points such as disease diagnosis or pharmaceutical evaluation, and proteomics technologies are still slow and labour-intensive [1]. Downstream, metabolomics deals with the comprehensive study of the metabolome. The metabolome can be defined as the complete complement of all small molecule (<1500 Da) metabolites found in an organism, cell system, tissue or biofluid [5]. These metabolites includes lipids, sugars, and amino acids. Metabolites serve as substrates and products of enzymatic reactions, and are influenced by gene and environmental factors, providing a bridge between genotype and phenotype [6]. Moreover, metabolic responses to changes in the microenvironment are extremely rapid compared with proteins or mRNA, reflecting the actual biological state, which is particularly important for the assessment of rapid and progressive diseases [7]. The closer relation with real-world end-points and the availability of low-cost high-throughput techniques have raised the interest in metabolomics. At the time of writing, the number of hits returned by the Web of Science citation indexing service containing "metabolomics" or "metabonomics" (interchangeable terms for the purpose of this study) was c.a. 25k. Fig. C1.1 illustrates the rapid growth of this emergent discipline.

Metabolomics have been applied to multiple research fields including disease diagnostics, biomarker discovery, drug discovery and development, toxicology, food science and nutritional studies [8–10]. The basis of most of these applications has a common aspect that alterations in metabolism due to functional responses of a biological system to any given condition result in changes in the abundance of groups of metabolites that form characteristic patterns, which can be used to derive insights into the underlying biological state [7].

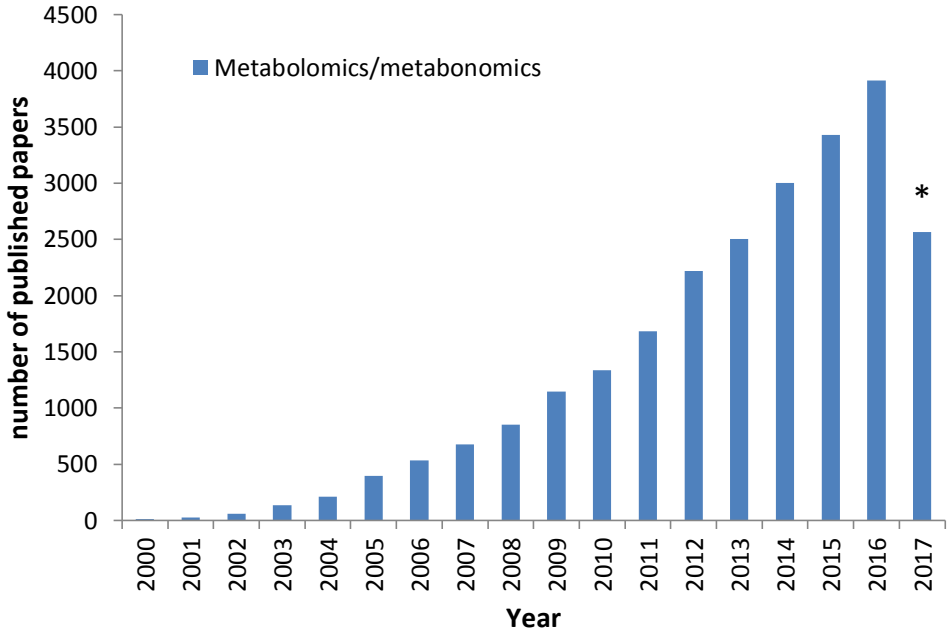


Fig. C1.1 Number of publications per year containing either metabolomics or metabonomics, returned by the Web Of Science (https://www.webofknowledge.com/). The asterisk indicates the number of papers published up to September 2017

The main analytical platforms applied to metabolomics are mass spectrometry (MS) and nuclear magnetic resonance spectroscopy (NMR). These techniques have different characteristics and they are usually combined to increase the detected metabolite coverage [11]. MS highlights by its outstanding sensitivity (in the range of femtomoles) and mass resolution [12], and is commonly coupled to gas or liquid chromatography to provide an extra dimension where the compounds are separated based on their different physicochemical properties. Because of these characteristics, the tandem chromatography and MS provides the identification and quantification of hundreds of metabolites in a single analysis. However, the need of multiple internal standards for quantitative analysis, intensive sample preparation and time-consuming chromatography hamper its use in high-throughput analysis and large-scale studies. On the contrary, NMR is quantitative in nature and extremely robust, providing high analytical reproducibility. Besides, NMR is non-destructive and requires minimal sample manipulation, keeping the sample intact for future analysis. As the main drawbacks of NMR, the low resolution and sensitivity compared with MS techniques, in the range of μmoles [13], limits the number of detected metabolites in biological samples.

1.2. High-throughput NMR-based metabolomics

As mentioned above, NMR has some interesting features that make this technique especially suitable for high-throughput analysis in large-scale metabolomics studies; the quantitative nature results from the fact that the peak areas in the NMR spectrum are directly related to the molar concentration of a specific nucleus (generally ^{1}H). Additionally, NMR allows the metabolic profiling of biofluids and intact tissues without metabolite extraction or separation.

1.2.1. Fundamentals of NMR

When the sample is introduced in the spectrometer, some nuclear spins in the sample are aligned with the surrounding constant magnetic field (low energy state) and the rest against it (high energy state). The distribution of spins between these two states can be altered by a radiofrequency (RF) pulse of a specific frequency known as Larmor frequency, which depends on the observed nucleus (^{1}H, ^{13}C, ^{31}P, etc.). Once the RF pulse is switched off, the energy loss of any excited spin to recover its equilibrium state, known as relaxation, is recorded. The obtained signal is known as the free induction decay (FID) and contains the sum of the relaxations of all the excited spins. Generally, several scans (i.e. FIDs) are recorded to increase sensitivity and cancel out random thermal noise and transients. Finally, the FID is Fourier transformed (FT) to the more informative frequency spectrum.

Importantly, spins in a molecule experience a slightly different magnetic environment (i.e. a slightly different Larmor frequency) depending on the surrounding nuclei. The different magnetic environments make a molecule show a set of signals dispersed along the frequency axis of an NMR spectrum, which are related to its functional groups (e.g. methyl, methylene, allyl, etc.). This spectral signature characterises the different molecular species and reveals their structural composition. Another important aspect is that relaxation depends on the molecular motion: the more rigid a molecule (or rather, a molecular moiety where the spin is located), the shorter the relaxation. Moreover, shorter relaxations imply broader peaks in the NMR spectrum.

One-dimensional (1D) NMR spectroscopy is the most common in NMR metabolomics as it can be carried out in few minutes and provides enough information from the different molecular species and their abundance, as previously mentioned. More complex RF pulse sequences allow modifying the observable spectral information based on physicochemical properties. It is usually referred as “NMR spectral editing”. In case of extensive signal overlapped that difficult the analysis using 1D

NMR spectra or if structural information is sought, 2D or 3D NMR experiments provide additional dimensions generally based on spin-spin coupling patterns or different motional properties. Fig. C1.2 shows proton 1D (90º pulse) and 2D (COSY) spectra of 3-hydroxybutyric acid in PBS.

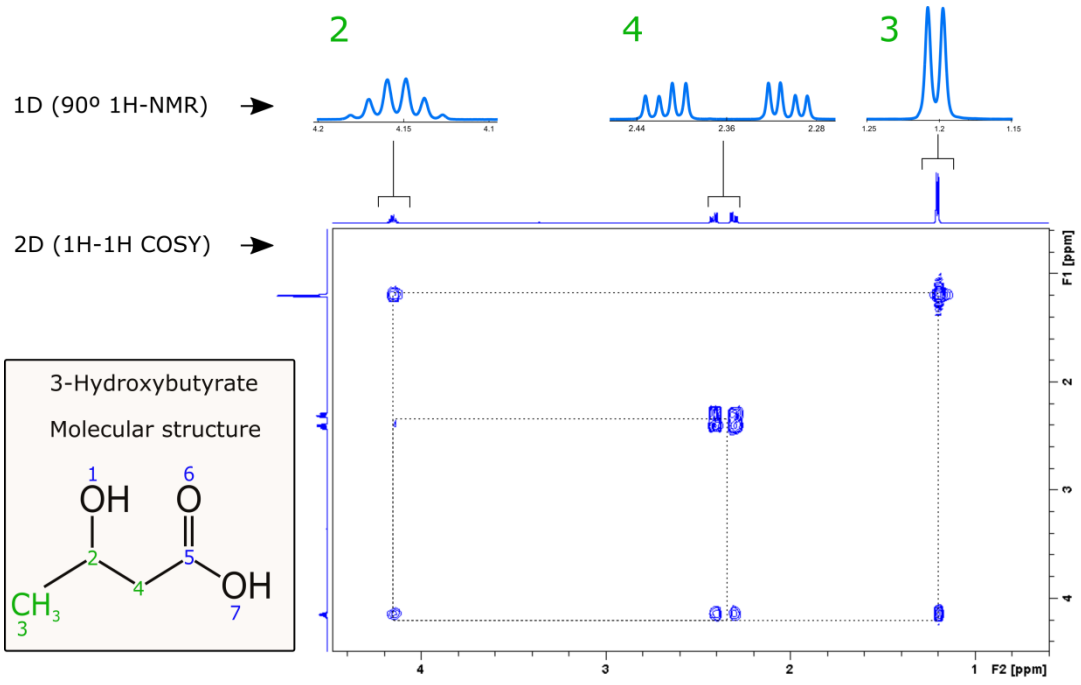


Fig. C1.2 Detail of signals in one-dimensional spectra obtained with a standard (90° pulse) ^{1}H-NMR experiment and two-dimensional ^{1}H-^{1}H COSY of 3-hydroxybutyric acid in PBS solution. Cross peaks in the 2D spectra (peaks out of the diagonal) represent coupled protons over 2 or 3 bonds. Signals numbered 2 to 4 in the 1D spectrum correspond with protons attached to labelled bonds in the molecular structure

The reader is referred to the following textbook for further information about NMR principles, main NMR experiments and their applications in metabolomics [14].

1.2.2. Spectral pre-processing

After NMR acquisition, there are some spectral corrections that should be applied in order to get reliable results from spectral analysis [15]. Prior the FT, zero-filling of FID provides high spectral resolution, i.e. smooth peaks, which is a critical step for reliable fingerprinting and spectral integration analysis. Then, window apodization is applied to the FID to increase the signal-to-noise ratio (S/N), usually with a Lorentzian function of 0.3-1 Hz line broadening [8], or to improve peak

resolution, using a shifted sine-bell or a Gaussian function [16]. After the FT, the spectral line has to be phase-corrected to provide a pure absorptive spectrum. Then, baseline corrections may still be needed to eliminate residual phase artifacts or broad background signals. Finally, spectral shifts in NMR acquisitions require referencing the whole spectrum to a signal of known chemical shift (frequency scale in ppm), usually an internal standard or any other signal not affected by sample conditions. Additionally, lineshape distortions produced by inhomogeneous magnetic fields can be corrected by reference deconvolution [17]. Examples of pre-processing corrections in spectral appearance are shown in Fig. C4.2. This spectral pre-processing workflow is commonly implemented in the software platforms of main NMR vendors. Alternatively, free software packages are available such as matNMR [18], NMRPipe [19] or MVAPACK [20].

1.2.3. ^{1}H-NMR profiling of biofluids

Biofluids are commonly used in metabolomics studies because they contain hundreds to thousands metabolites and samples can be obtained in a non-invasive (e.g., saliva, urine) or minimally invasive manner (e.g., blood plasma or serum, cerebrospinal fluid) [21]. Moreover, sample preparation for NMR experiments on biofluids only requires the addition of phosphate buffer in a small volume of deuterated solvent, and the addition of an internal standard for chemical shift reference and quantitative normalization. Commonly used internal standards are trimethylsilylpropanoic acid (TSP) or 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS) for aqueous solvents and tetramethylsilane (TMS) for organic solvents.

Among the possible nuclei that can be analysed with NMR, proton (^{1}H-NMR) is the most used in metabolomics because of its high sensitivity, fast relaxation, natural abundance, and its nearly ubiquitous presence in organic metabolites [21]. ^{1}H-NMR spectra of biofluids consists of a conglomerate of severely overlapped signals from a vast number of compounds at very different concentrations, making a reliable identification and quantification a challenging task. Additionally, the spectral complexity is amplified by chemical exchange processes. For instance, pH, ionic strength, and metal ion composition affect specific groups of metabolites, causing chemical shifts variations between samples [14]. Spectral misalignments can be reduced using one of the multiple algorithms available [22]. Chemical exchange also affects quantification, such as the decrease of urea signal due to proton exchange with water [8]. Similarly, small molecules binding to protein show severe attenuation and broadening of their signals [23].

With the aim of helping identification, several public libraries include lists of compound peaks, raw NMR files of standard compounds, and typical concentration ranges in common biofluids [24,25].

Concerning metabolite quantification, the first step comprises the quantification (in area units) of the signals assigned to known metabolites. Integrating isolated signals is the classical approach; however, this approach is very sensitive to baseline distortions and it is not recommended for overlapping peaks. Alternatively, spectral deconvolution with lineshape fitting analysis provides a more robust quantitative method, in which a ^{1}H-NMR spectrum is defined by a finite number of Lorentzian/Gaussian lineshapes following quantum mechanical rules (chemical shifts, coupling constants, etc.) and some baseline functions [26]. An example of lineshape fitting analysis is shown in Fig. C4.3b. Using lineshape fitting, overlapping signals can be efficiently resolved and baseline effects omitted. Lineshape fitting is commonly carried out with commercial software packages such as Chenomx NMR Suite [27], Mnova [28], and PERCH NMR software [29], although free solutions such as BATMAN [30], DOLPHIN [31] and BAYESIL [32] are also available. Finally, normalization of the signal areas with a reference compound of known concentration allows calculating molar concentrations of the detected compounds. The aforementioned internal standards TSP, DSS, and TMS are commonly used. However, TSP and DSS signals are affected by protein binding and they are usually placed in coaxial inserts in PBS solution, which comprises the quantitative precision due to media incompatibilities between the sample and the coaxial insert. Similarly, TMS is highly volatile and should be avoided for quantitative purposes. Different alternatives have been presented to overcome the problems of common internal standards [33,34]. Other strategies imply the use of a calibrated synthetic signal, such as ERETIC, QUANTAS or PULCON [15], which can be introduced artificially in NMR acquisition or after spectral pre-processing.

^{1}H-NMR profiling is a laborious process that requires intensive data manipulation (spectral pre-processing, identification and quantification). This process is still mainly carried out manually, even though analyst-dependent variations have been determined to be c.a. 20% [21]. In order to avoid this source of error and consolidate high-throughput NMR-based metabolomics, extensive automation of ^{1}H-NMR profiling workflow is required.

1.2.4. ^{1}H-NMR profiling of blood serum/plasma: the three molecular windows

Blood is the primary body fluid connected to systemic metabolism. Blood composition reflects even minimal changes in the whole metabolism and is therefore the natural choice for studies related to vascular and systemic diseases, as well as for nutritional assays [35,36]. Blood contains molecules of various size and mobility: proteins, lipids, lipoproteins, cholesterols, low-molecular-weight

metabolites and ions, and their concentrations range from nM to mM. Blood plasma is blood without blood cells, while blood serum is blood plasma without the blood clotting proteins (fibrinogens) [37]. Although plasma has a lower risk of uncontrolled and incomplete clotting [14], serum seems to be the preferred blood derivative for ^{1}H-NMR profiling, as serum spectra lack of interference signals assigned to clotting proteins and anti-coagulant additives. In order to ensure inter-laboratory reproducibility, protocols for serum/plasma sample preparation have been previously described [8]. In the following, the term serum will refer to both blood-derived matrices.

The complexity of ^{1}H-NMR spectra of serum, where sharp peaks from low-molecular-weight metabolites (LMWM) are severely overlapped with broad signals from macromolecules (mainly lipoproteins and albumin), prevents the use of a single ^{1}H-NMR experiment to fully characterise the biochemical diversity of blood. Instead, Ala-Korpela and co-workers proposed the implementation of a three molecular windows model involving different ^{1}H-NMR experiments and sample preparations [38]. The model allows the comprehensive high-throughput quantification of lipoprotein classes and constituent lipids, albumin, and a large variety of low-molecular-weight metabolites, including amino acids, creatinine, glycolysis-related metabolites, and ketone bodies, with costs comparable with standard lipid measurements [38]. In the following lines, this model will serve to illustrate the common strategies involving a comprehensive ^{1}H-NMR profiling of serum samples (Fig. C1.3).

1.2.4.1. Lipoprotein window

Lipoprotein window implies the acquisition of any ^{1}H-NMR experiment of native serum, in which the broad signals produced by macromolecules (proteins and lipoproteins) are visible. Water presaturation is required when using native serum to suppress the large residual signal from water protons. Water presaturation is typically applied in an NMR experiment known as NOESY-presat. This pulse sequence is identical to the 1st time increment of the 2D-NOESY experiment [39]. A NOESY-presat ^{1}H-NMR spectrum of fasting serum is dominated by a broad background signal from protein (mostly albumin) and several broad peaks assigned to lipid moieties from the lipoprotein subclasses VLDL, LDL and HDL [40], with minor contribution of LMWM peaks (Fig. C1.3).

Lipoprotein subclasses share the same constituents in different proportions; consequently, their spectral signatures are very similar showing large overlap. However, magnetic susceptibility anisotropy in the lipoprotein shell generates a subtle chemical shift dispersion of lipoprotein signals according to their size [41]. Lineshape fitting and regression methods have taken advantage of this

spectral dispersion to quantify the number of particles in each lipoprotein subclass and their lipid content (mainly cholesterol and triglycerides) [35]. Alternatively to a standard NOESY-presat experiments, diffusion-edited ^{1}H-NMR spectroscopy provides a filtered spectrum in which fast-diffusing LMWM signals are removed [42] (Fig. C1.3). This NMR strategy has been suggested to benefit lipoprotein analysis [43] and has been applied to some regression models of lipoprotein lipids [44,45]. Even more, recent studies have shown that the inclusion of a diffusion dimension with 2D ^{1}H-NMR diffusion experiments could provide more reliable deconvolution of lipoprotein signals and estimation of constituents lipids [46,47], on the basis that diffusion dimension provides a direct measure of lipoprotein sizes. These multidimensional data structures usually requires the application of multivariate curve resolution methods (MCR) [48] or multi-way techniques such as PARAFAC [49] or N-PLS [50].

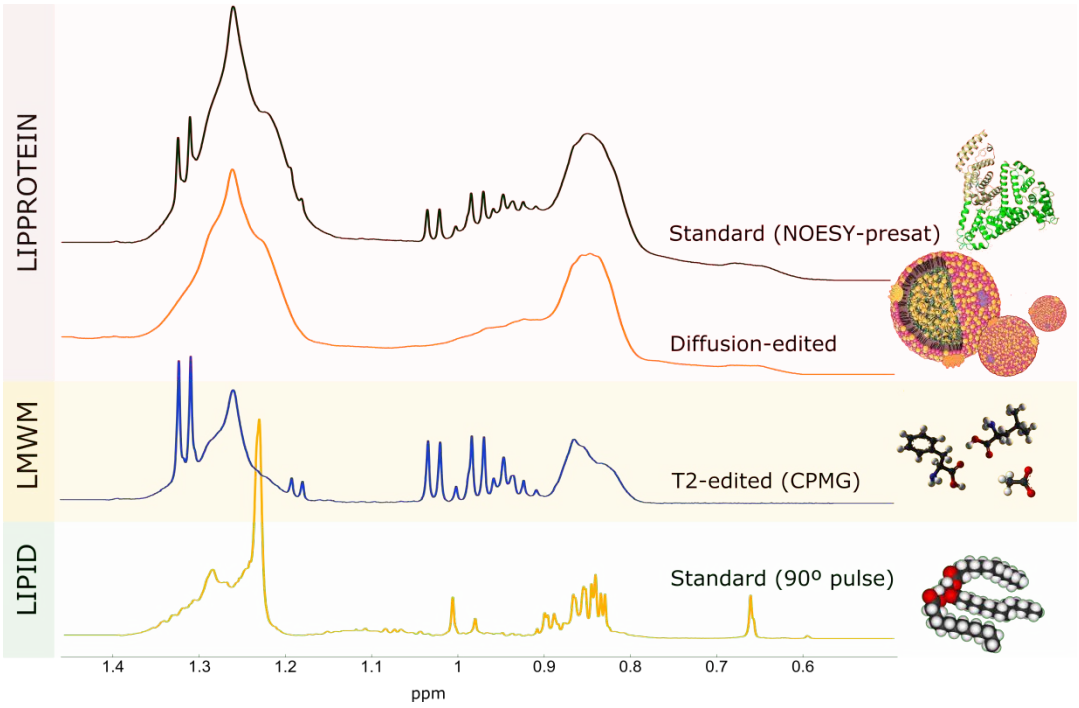


Fig. C1.3 Methyl and methylene regions in the three molecular window model and examples of molecular species that are analysed with each window

Other compounds that have been quantified using the lipoprotein window are albumin [38] (in signal area) and total glycoproteins using the composite signal at 2 ppm from N-acetyl methyl groups of mobile N-acetylglucosamine (GlcNAc) residues [16].

1.2.4.2. LMWM window

The dominant protein background with thousands of resonating protons per molecule, hampers the analysis of low-molecular-weight metabolites (LMWM) in a standard ^{1}H-NMR spectrum of native serum. The use of T2-edited ^{1}H-NMR experiments, such as the Carr-Purcell-Meiboom-Gill sequence (CPMG), improves the detection of LMWM by removing or decreasing the broad signals from fast-relaxing molecules such as proteins and lipoproteins [51] (Fig. C1.3) (note that T2-edited NMR can be understood as the reciprocal of diffusion-edited NMR). Then, signal deconvolution and quantification can be carried out with available software packages that implement automatic lineshape fitting algorithms based on LMWM signal libraries [30–32,52].

However, reliable quantifications are compromised by the fact that some LMWM are bound to albumin, consequently, their signals are "NMR-invisible" or significantly decreased [26,37,53–55]. Partial release of these metabolites from albumin can be achieved by strong acidification or twofold dilution of serum in D_2O [13,54,55], at the expense of modifying the native conditions. Deproteinization methods are usually applied to LMWM analysis [56]. Deproteinization avoids the use of T2-edited ^{1}H-NMR experiments and has been reported to increase the coverage of quantified LMWM from the approximately 30 compounds in native serum to 67 compounds in deproteinized serum [57]. It should be also noted that all the resonance intensities, including those of the LMWM, are reduced with a CPMG filter by its own spin-spin relaxation time (T2), adding additional (sometimes negligible) quantitative error [26].

Alternatively to lipoprotein window, LMWM window has been applied to the quantification of total glycoproteins [38] and albumin, although albumin concentration in LMWM window is indirectly derived from changes in chemical shift position of some LMWM signals caused by the albumin-induced bulk magnetic susceptibility [58].

1.2.4.3. Lipid window

A detail analysis of serum lipids requires the breakdown of protein and lipoprotein complexes and their lipid extraction. Contrary to the high-throughput dogma, current lipid extraction procedures are manual and require time-consuming centrifugation steps (the reader is referred to [59] for information about lipid extraction, storage and NMR sample preparation). Methods for combined LMWM and lipid extractions have also been proposed [8]. Lipid analysis is usually performed with a standard (90° pulse) ^{1}H-NMR experiment and its spectrum provides information about fatty acid families, free and esterified cholesterol, triglycerides, choline phospholipids and total

glycerophospholipids [60]. Contrary to the LMWM, free software packages for the automatic deconvolution and quantification of ^{1}H-NMR lipid extracts are still not available, consequently, most of the studies of ^{1}H-NMR lipids are carried out using spectral integration and fingerprinting analysis [61–65]. The lack of automatic tools is motivated by the complex signals arising from the multiple couplings patterns in the long carbon chains and the similarity of the spectra of lipid species with respect to the limited structural carbon chain information [66].

1.2.5. Applications of ^{1}H-NMR profiling of blood serum/plasma in clinical research

^{1}H-NMR profiling of blood serum has led to a deeper understanding of disease pathogenesis and the identification of metabolic biomarkers for disease diagnosis or treatment monitoring. Based on the reported findings, ^{1}H-NMR serum profiling could improve the clinical diagnosis of several types of cancer disease [67], inflammatory bowel diseases [67], inborn errors of metabolism [33], Alzheimer's disease [60], type-2 diabetes [38,67] and cardiovascular disease [38], among many others [68]. Besides, ^{1}H-NMR serum profiling has been found to reveal the metabolic effects of physical activity [38] and dietary interventions [69,70].

1.3. Thesis motivation and objectives

The doctoral thesis presented in this document is the result of the research conducted in the Signal Processing for Omic Sciences (SIPOMICS) research group, belonging to the Department of Electronic, Electrical and Automation Engineering at the Rovira i Virgili University (URV), and the Metabolomics Platform (http://metabolomicsplatform.com/), a joint research facility created by URV and the CIBER of Diabetes and Metabolic Diseases (CIBERDEM, http://www.ciberdem.org/). The Metabolomics Platform is also part of the Pere Virgili Health Research Institute (IISPV, http://www.iispv.cat/), a major medical research organization in the south of Catalonia that undertakes numerous research initiatives in the country.

The main objective of the Metabolomics Platform is to provide technical support to biomedical and clinical research groups in the field of metabolomics. Studies carried out by these groups usually comprise large sample cohorts and aim at the discovery of new metabolic biomarkers, metabolic disease patterns, metabolic changes associated to drugs, age, diets, nutritional supplements, etc. In an initial stage, most of these studies demand the analysis of biofluids, such as urine and blood serum, because of their low-cost and easy availability. The large number of samples and measured

metabolites, and the technical challenges of metabolomics methods and platforms motivate the automation of sample preparation and data analysis processes in order to generate reliable biological information.

Since 2012, the Metabolomics Platform has been developing a set of strategies to replace the time-consuming and error-prone manual ^{1}H-NMR profiling of serum samples with automatic or semiautomatic bioinformatics tools. These high-throughput strategies are based on the three molecular windows model previously described and comprise Liposcale, an advanced lipoprotein test based on 2D diffusion-ordered ^{1}H NMR spectroscopy and Dolphin, a tool for automated targeted LMWM profiling using 1D and 2D ^{1}H-NMR data. The present work aims at complementing the previous developments and design methodological and computational strategies to deal with issues affecting quantitative high-throughput ^{1}H-NMR serum profiling. More concretely, the main objectives of this thesis are the following:

- Develop prediction models for the quantitative estimation of standard lipids (also known as "lipid panel") using 1D and 2D ^{1}H-NMR spectra of native serum/plasma samples and linear regression methods, and evaluate their generalization in large-scale analysis including samples with lipid and lipoprotein abnormalities. This will ultimately replace the need of clinical biochemical measurements.
- Design methodological and computational strategies to improve the quantification of LMWM by ^{1}H-NMR that is affected by protein binding in native serum.
- Develop an open source bioinformatics package for the profiling of serum lipids using 1D ^{1}H-NMR and evaluate its functionality with conventional techniques and clinical studies.

1.4. Organization of the document

Chapter 1 provides a general background of the field of application and the common strategies applied to ^{1}H-NMR profiling of blood serum samples in metabolomics. This chapter also exposes the motivations for implementing robust data analysis workflows and introduces the multiple issues affecting reliable metabolic quantifications, which motivate the development of the multiple studies presented in this thesis.

Chapters 2 to 4 contain the three scientific articles published or submitted for publication during the realization of this thesis. Each article is related to each one of the objectives defined in Chapter 1. Therefore, Chapter 2 describes the development of prediction models of standard lipids based on

^{1}H-NMR spectra of native serum/plasma, i.e., total cholesterol and triglycerides in serum and cholesterol content of pro- and anti-atherogenic LDL and HDL lipoproteins, respectively. It evaluates if the inclusion of a second NMR dimension related to molecular sizes and N-way chemometrics methods could improve lipid estimations. The use of large heterogeneous cohorts comprising lipid and lipoprotein abnormalities allows the generalization of the results, which are compared with classical colorimetric-enzymatic measurements and evaluated in the classification of several dyslipidaemias. The results have been published in *Metabolomics* journal.

Chapter 3 explains the quantitative issues in ^{1}H-NMR profiling of serum derived from the "NMR-invisibility" of some LMWM binding to serum proteins. It also presents both a competitive binding and a multidimensional ^{1}H-NMR strategy to increase their "NMR-visibility" and achieve quantifications closer to their absolute concentrations in serum. These strategies are evaluated from synthetic models to human plasma cohorts. Finally, the benefits for quantitative high-throughput ^{1}H-NMR profiling of serum are discussed. This article has been published in *Journal of Proteome Research*.

Chapter 4 explains the current limitations of ^{1}H-NMR profiling of serum lipids and presents LipSpin, a new open source package that allows the semiautomatic profiling of lipophilic extracts of serum samples using ^{1}H-NMR. The article briefly describes the main software functionalities, its possibilities and limitations. Moreover, results of the different analytical and clinical validations with established methods and a dietary intervention study are presented. This article has been submitted to *Analytical Chemistry* journal.

Finally, Chapter 5 and 6 contain the general discussion and the conclusions of this thesis, respectively.

1.5. References

1. Nicholson, J. K., Holmes, E., & Lindon, J. C. (2007). Chapter 1 - Metabonomics and Metabolomics Techniques and Their Applications in Mammalian Systems BT - The Handbook of Metabonomics and Metabolomics (pp. 1–33). Amsterdam: Elsevier Science B.V. https://doi.org/10.1016/B978-044452841-4/50002-3

2. Greef, J. van der, Stroobant, P., & Heijden, R. van der. (2004). The role of analytical sciences in medical systems biology. *Current Opinion in Chemical Biology*, *8*(5), 559–565. https://doi.org/10.1016/j.cbpa.2004.08.013

3. Kell, D. B. (2004). Metabolomics and systems biology: making sense of the soup. *Current Opinion in Microbiology*, *7*(3), 296–307. https://doi.org/10.1016/j.mib.2004.04.012

4. Patti, G. J., Yanes, O., & Siuzdak, G. (2012). Innovation: Metabolomics: the apogee of the omics trilogy. *Nature Reviews Molecular Cell Biology*, *13*(4), 263–269. https://doi.org/10.1038/nrm3314

5. Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., … Querengesser, L. (2007). HMDB: the Human Metabolome Database. *Nucleic Acids Research*, *35*(Database issue), D521–D526. https://doi.org/10.1093/nar/gkl923

6. German, J. B., Hammock, B. D., & Watkins, S. M. (2005). Metabolomics: building on a century of biochemistry to guide human health. *Metabolomics*, *1*(1), 3–9. https://doi.org/10.1007/s11306-005-1102-8

7. Monteiro, M. S., Carvalho, M., Bastos, M. L., & Guedes de Pinho, P. (2013). Metabolomics analysis for biomarker discovery: advances and challenges. *Current Medicinal Chemistry*, *20*(2), 257–71. https://doi.org/CMC-EPUB-20121126-5 [pii]

8. Beckonert, O., Keun, H. C., Ebbels, T. M. D., Bundy, J., Holmes, E., Lindon, J. C., & Nicholson, J. K. (2007). Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature Protocols*, *2*(11), 2692–2703. https://doi.org/10.1038/nprot.2007.376

9. Wishart, D. S. (2016). Emerging applications of metabolomics in drug discovery and precision medicine. *Nature Reviews Drug Discovery*, *15*(7), 473–484. https://doi.org/10.1038/nrd.2016.32

10. Wishart, D. S. (2008). Metabolomics: applications to food science and nutrition research. *Trends in Food Science & Technology*, *19*(9), 482–493. https://doi.org/10.1016/j.tifs.2008.03.003

11. Psychogios, N., Hau, D. D., Peng, J., Guo, A. C., Mandal, R., Bouatra, S., … Wishart, D. S. (2011). The Human Serum Metabolome. *PLoS ONE*, *6*(2), e16957. https://doi.org/10.1371/journal.pone.0016957

12. Forcisi, S., Moritz, F., Kanawati, B., Tziotis, D., Lehmann, R., & Schmitt-Kopplin, P. (2013). Liquid chromatography–mass spectrometry in metabolomics research: Mass analyzers in ultra high pressure liquid chromatography coupling. *Journal of Chromatography A*, *1292*, 51–65. https://doi.org/10.1016/j.chroma.2013.04.017

13. Smolinska, A., Blanchet, L., Buydens, L. M. C., & Wijmenga, S. S. (2012). NMR and pattern recognition methods in metabolomics: From data acquisition to biomarker discovery: A review. *Analytica Chimica Acta*. https://doi.org/10.1016/j.aca.2012.05.049

14. Ross, A., Schlotterbeck, G., Dieterle, F., & Senn, H. (2007). Chapter 3 - NMR Spectroscopy Techniques for Application to Metabonomics BT - The Handbook of Metabonomics and Metabolomics (pp. 55–112). Amsterdam: Elsevier Science B.V. https://doi.org/10.1016/B978-044452841-4/50004-7

15. Bharti, S. K., & Roy, R. (2012). Quantitative 1H NMR spectroscopy. *TrAC Trends in Analytical Chemistry*, *35*, 5–26. https://doi.org/10.1016/j.trac.2012.02.007

16. Otvos, J. D., Shalaurova, I., Wolak-Dinsmore, J., Connelly, M. A., Mackey, R. H., Stein, J. H., & Tracy, R. P. (2015). GlycA: A Composite Nuclear Magnetic Resonance Biomarker of Systemic Inflammation. *Clinical Chemistry*, *61*(5), 714–723.

https://doi.org/10.1373/clinchem.2014.232918

17. Morris, G. A., Barjat, H., & Home, T. J. (1997). Reference deconvolution methods. *Progress in Nuclear Magnetic Resonance Spectroscopy*, *31*(2), 197–257. https://doi.org/10.1016/S0079-6565(97)00011-3

18. van Beek, J. D. (2007). matNMR: A flexible toolbox for processing, analyzing and visualizing magnetic resonance data in Matlab®. *Journal of Magnetic Resonance*, *187*(1), 19–26. https://doi.org/10.1016/j.jmr.2007.03.017

19. Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J., & Bax, A. (1995). NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *Journal of Biomolecular NMR*, *6*(3), 277–293. https://doi.org/10.1007/BF00197809

20. Worley, B., & Powers, R. (2014). MVAPACK: A Complete Data Handling Package for NMR Metabolomics. *ACS Chemical Biology*, *9*(5), 1138–1144. https://doi.org/10.1021/cb4008937

21. Larive, C. K., Barding, G. A., & Dinges, M. M. (2015). NMR Spectroscopy for Metabolomics and Metabolic Profiling. *Analytical Chemistry*, *87*(1), 133–146. https://doi.org/10.1021/ac504075g

22. Vu, T. N., & Laukens, K. (2013). Getting your peaks in line: A review of alignment methods for NMR spectral data. *Metabolites*. https://doi.org/10.3390/metabo3020259

23. Van, Q. N., Chmurny, G. N., & Veenstra, T. D. (2003). The depletion of protein signals in metabonomics analysis with the WET–CPMG pulse sequence. *Biochemical and Biophysical Research Communications*, *301*(4), 952–959. https://doi.org/10.1016/S0006-291X(03)00079-2

24. Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., … Scalbert, A. (2013). HMDB 3.0—The Human Metabolome Database in 2013. *Nucleic Acids Research*, *41*(Database issue), D801–D807. https://doi.org/10.1093/nar/gks1065

25. Ulrich, E. L., Akutsu, H., Doreleijers, J. F., Harano, Y., Ioannidis, Y. E., Lin, J., … Markley, J. L. (2007). BioMagResBank. *Nucleic Acids Research*, *36*(Database), D402–D408. https://doi.org/10.1093/nar/gkm957

26. Tiainen, M., Soininen, P., & Laatikainen, R. (2014). Quantitative Quantum Mechanical Spectral Analysis (qQMSA) of 1H NMR spectra of complex mixtures and biofluids. *Journal of Magnetic Resonance*, *242*, 67–78. https://doi.org/10.1016/j.jmr.2014.02.008

27. Chenomx Inc. http://www.chenomx.com/

28. Mestrelab Research S.L. http://mestrelab.com/

29. PERCH Solutions Ltd. http://perchnmrsoftware.com/

30. 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*(15), 2088–2090. https://doi.org/10.1093/bioinformatics/bts308

31. Gómez, J., Brezmes, J., Mallol, R., Rodríguez, M. A., Vinaixa, M., Salek, R. M., … Cañellas,

N. (2014). Dolphin: a tool for automatic targeted metabolite profiling using 1D and 2D 1H-NMR data. *Analytical and Bioanalytical Chemistry*, *406*(30), 7967–7976. https://doi.org/10.1007/s00216-014-8225-6

32. Ravanbakhsh, S., Liu, P., Bjordahl, T. C., Mandal, R., Grant, J. R., Wilson, M., … Wishart, D. S. (2015). Accurate, Fully-Automated NMR Spectral Profiling for Metabolomics. *PLOS ONE*, *10*(5), e0124219. https://doi.org/10.1371/journal.pone.0124219

33. Oostendorp, M. (2006). Diagnosing Inborn Errors of Lipid Metabolism with Proton Nuclear Magnetic Resonance Spectroscopy. *Clinical Chemistry*, *52*(7), 1395–1405. https://doi.org/10.1373/clinchem.2006.069112

34. Kriat, M., Confort-Gouny, S., Vion-Dury, J., Sciaky, M., Viout, P., & Cozzone, P. J. (1992). Quantitation of metabolites in human blood serum by proton magnetic resonance spectroscopy. A comparative study of the use of formate and TSP as concentration standards. *NMR in Biomedicine*, *5*(4), 179–184. https://doi.org/10.1002/nbm.1940050404

35. Mallol, R., Rodriguez, M. A., Brezmes, J., Masana, L., & Correig, X. (2013). Human serum/plasma lipoprotein analysis by NMR: Application to the study of diabetic dyslipidemia. *Progress in Nuclear Magnetic Resonance Spectroscopy*, *70*, 1–24. https://doi.org/10.1016/j.pnmrs.2012.09.001

36. Soininen, P., Kangas, A. J., Wurtz, P., Tukiainen, T., Tynkkynen, T., Laatikainen, R., … Ala-Korpela, M. (2009). High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism. *Analyst*, *134*(9), 1781–1785. https://doi.org/10.1039/B910205A

37. Jupin, M., Michiels, P. J., Girard, F. C., Spraul, M., & Wijmenga, S. S. (2013). NMR identification of endogenous metabolites interacting with fatted and non-fatted human serum albumin in blood plasma: Fatty acids influence the HSA–metabolite interaction. *Journal of Magnetic Resonance*, *228*, 81–94. https://doi.org/10.1016/j.jmr.2012.12.010

38. Soininen, P., Kangas, A. J., Würtz, P., Suna, T., & Ala-Korpela, M. (2015). Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Cardiovascular Epidemiology and Genetics. *Circulation: Cardiovascular Genetics*, *8*(1), 192–206. https://doi.org/10.1161/CIRCGENETICS.114.000216

39. Mckay, R. T. (2011). How the 1D-NOESY suppresses solvent signal in metabonomics NMR spectroscopy: An examination of the pulse sequence components and evolution. *Concepts in Magnetic Resonance Part A*, *38A*(5), 197–220. https://doi.org/10.1002/cmr.a.20223

40. Nicholson, J. K., Foxall, P. J. D., Spraul, M., Farrant, R. D., & Lindon, J. C. (1995). 750 MHz 1H and 1H-13C NMR Spectroscopy of Human Blood Plasma. *Analytical Chemistry*, *67*(5), 793–811. https://doi.org/10.1021/ac00101a004

41. Lounila, J., Ala-Korpela, M., Jokisaari, J., Savolainen, M. J., & Kesäniemi, Y. A. (1994). Effects of orientational order and particle size on the NMR line positions of lipoproteins. *Phys. Rev. Lett.*, *72*(25), 4049–4052. https://doi.org/10.1103/PhysRevLett.72.4049

42. Stejskal, E. O., & Tanner, J. E. (1965). Spin Diffusion Measurements: Spin Echoes in the Presence of a Time-Dependent Field Gradient. *The Journal of Chemical Physics*, *42*(1), 288–292. https://doi.org/doi:10.1063/1.1695690

43. Liu, M., Nicholson, J. K., & Lindon, J. C. (1996). High-resolution diffusion and relaxation edited one- and two-dimensional 1H NMR spectroscopy of biological fluids. *Analytical Chemistry*, *68*(19), 3370–3376. https://doi.org/10.1021/ac960426p

44. Mihaleva, V. V., Van Schalkwijk, D. B., De Graaf, A. A., Van Duynhoven, J., Van Dorsten, F. A., Vervoort, J., … Jacobs, D. M. (2014). A systematic approach to obtain validated partial least square models for predicting lipoprotein subclasses from serum nmr spectra. *Analytical Chemistry*, *86*(1), 543–550. https://doi.org/10.1021/ac402571z

45. Petersen, M., Dyrby, M., Toubro, S., Engelsen, S. B., Nørgaard, L., Pedersen, H. T., & Uyerberg, J. (2005). Quantification of lipoprotein subclasses by proton nuclear magnetic resonance-based partial least-squares regression models. *Clinical Chemistry*, *51*(8), 1457–1461. https://doi.org/10.1373/clinchem.2004.046748

46. Dyrby, M., Petersen, M., Whittaker, A. K., Lambert, L., Nørgaard, L., Bro, R., & Engelsen, S. B. (2005). Analysis of lipoproteins using 2D diffusion-edited NMR spectroscopy and multi-way chemometrics. *Analytica Chimica Acta*, *531*(2), 209–216. https://doi.org/10.1016/j.aca.2004.10.052

47. Mallol, R., Rodríguez, M. A., Heras, M., Vinaixa, M., Cañellas, N., Brezmes, J., … Correig, X. (2011). Surface fitting of 2D diffusion-edited 1H NMR spectroscopy data for the characterisation of human plasma lipoproteins. *Metabolomics*, *7*(4), 572–582. https://doi.org/10.1007/s11306-011-0273-8

48. de Juan, A., Jaumot, J., & Tauler, R. (2014). Multivariate Curve Resolution (MCR). Solving the mixture analysis problem. *Analytical Methods*, *6*(14), 4964. https://doi.org/10.1039/c4ay00571f

49. Bro, R. (1997). PARAFAC. Tutorial and applications. In *Chemometrics and Intelligent Laboratory Systems* (Vol. 38, pp. 149–171). https://doi.org/10.1016/S0169-7439(97)00032-4

50. Bro, R. (1996). Multiway calibration. Multilinear PLS. *Journal of Chemometrics*, *10*(1), 47–61. https://doi.org/10.1002/(SICI)1099-128X(199601)10:1<47::AID-CEM400>3.0.CO;2-C

51. Tang, H., Wang, Y., Nicholson, J. K., & Lindon, J. C. (2004). Use of relaxation-edited one-dimensional and two dimensional nuclear magnetic resonance spectroscopy to improve detection of small metabolites in blood plasma. *Analytical Biochemistry*, *325*(2), 260–272. https://doi.org/10.1016/j.ab.2003.10.033

52. Weljie, A. M., Newton, J., Mercier, P., Carlson, E., & Slupsky, C. M. (2006). Targeted Profiling: Quantitative Analysis of 1H NMR Metabolomics Data. *Analytical Chemistry*, *78*(13), 4430–4442. https://doi.org/10.1021/ac060209g

53. De Graaf, R. A., & Behar, K. L. (2003). Quantitative 1H NMR spectroscopy of blood plasma metabolites. *Analytical Chemistry*, *75*(9), 2100–2104. https://doi.org/10.1021/ac020782+

54. Nicholson, J. K., & Gartland, K. P. R. (1989). 1H NMR studies on protein binding of histidine, tyrosine and phenylalanine in blood plasma. *NMR in Biomedicine*, *2*(2), 77–82. https://doi.org/10.1002/nbm.1940020207

55. Bell, J. D., Brown, J. C. C., Kubal, G., & Sadler, P. J. (1988). NMR-invisible lactate in blood plasma. *FEBS Letters*, *235*(1–2), 81–86. https://doi.org/10.1016/0014-5793(88)81238-9

56. Daykin, C. A., Foxall, P. J. D., Connor, S. C., Lindon, J. C., & Nicholson, J. K. (2002). The comparison of plasma deproteinization methods for the detection of low-molecular-weight metabolites by (1)H nuclear magnetic resonance spectroscopy. *Analytical Biochemistry*, *304*(2), 220–30. https://doi.org/10.1006/abio.2002.5637

57. Nagana Gowda, G. A., Gowda, Y. N., & Raftery, D. (2015). Expanding the limits of human blood metabolite quantitation using NMR spectroscopy. *Analytical Chemistry*, *87*(1), 706–715. https://doi.org/10.1021/ac503651e

58. Jupin, M., Michiels, P. J., Girard, F. C., & Wijmenga, S. S. (2015). Magnetic susceptibility to measure total protein concentration from NMR metabolite spectra: Demonstration on blood plasma. *Magnetic Resonance in Medicine*, *73*(2), 459–468. https://doi.org/10.1002/mrm.25178

59. Kostara, C. E., & Bairaktari, E. T. (2013). Lipid Profiling in Health and Disease. In J. V Sweedler, N. W. Lutz, & R. A. Wevers (Eds.), *Methodologies for Metabolomics: Experimental Strategies and Techniques*. Cambridge: Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9780511996634.020

60. Tukiainen, T., Tynkkynen, T., Mäkinen, V.-P., Jylänki, P., Kangas, A., Hokkanen, J., … Ala-Korpela, M. (2008). A multi-metabolite analysis of serum by 1H NMR spectroscopy: Early systemic signs of Alzheimer's disease. *Biochemical and Biophysical Research Communications*, *375*(3), 356–361. https://doi.org/10.1016/j.bbrc.2008.08.007

61. Jiang, C., Yang, K., Yang, L., Miao, Z., Wang, Y., & Zhu, H. (2013). A 1H NMR-Based Metabonomic Investigation of Time-Related Metabolic Trajectories of the Plasma, Urine and Liver Extracts of Hyperlipidemic Hamsters. *PLoS ONE*, *8*(6), e66786. https://doi.org/10.1371/journal.pone.0066786

62. Vinaixa, M., Ángel Rodríguez, M., Rull, A., Beltrán, R., Bladé, C., Brezmes, J., … Correig, X. (2010). Metabolomic Assessment of the Effect of Dietary Cholesterol in the Progressive Development of Fatty Liver Disease. *Journal of Proteome Research*, *9*(5), 2527–2538. https://doi.org/10.1021/pr901203w

63. Kostara, C. E., Papathanasiou, A., Cung, M. T., Elisaf, M. S., Goudevenos, J., & Bairaktari, E. T. (2010). Evaluation of Established Coronary Heart Disease on the Basis of HDL and Non-HDL NMR Lipid Profiling. *Journal of Proteome Research*, *9*(2), 897–911. https://doi.org/10.1021/pr900783x

64. Beckonert, O., Monnerjahn, J., Bonk, U., & Leibfritz, D. (2003). Visualizing metabolic changes in breast-cancer tissue using 1H-NMR spectroscopy and self-organizing maps. *NMR in Biomedicine*, *16*(1), 1–11. https://doi.org/10.1002/nbm.797

65. Fernando, H., Bhopale, K. K., Kondraganti, S., Kaphalia, B. S., & Shakeel Ansari, G. A. (2011). Lipidomic changes in rat liver after long-term exposure to ethanol. *Toxicology and Applied Pharmacology*, *255*(2), 127–137. https://doi.org/10.1016/j.taap.2011.05.022

66. Hyötyläinen, T., Bondia-Pons, I., & Orešič, M. (2013). Lipidomics in nutrition and food research. *Molecular Nutrition & Food Research*, *57*(8), 1306–1318. https://doi.org/10.1002/mnfr.201200759

67. Emwas, A.-H. M., Salek, R. M., Griffin, J. L., & Merzaban, J. (2013). NMR-based

metabolomics in human disease diagnosis: applications, limitations, and recommendations. *Metabolomics*, *9*(5), 1048–1072. https://doi.org/10.1007/s11306-013-0524-y

68. Duarte, I. F., Diaz, S. O., & Gil, A. M. (2014). NMR metabolomics of human blood and urine in disease research. *Journal of Pharmaceutical and Biomedical Analysis*, *93*, 17–26. https://doi.org/10.1016/j.jpba.2013.09.025

69. Bondia-Pons, I., Abete, I., Cañellas, N., Abete, I., Rodríguez, M. Á., Perez-Cornago, A., … Martínez, J. A. (2013). Nutri-Metabolomics: Subtle Serum Metabolic Differences in Healthy Subjects by NMR-Based Metabolomics after a Short-Term Nutritional Intervention with Two Tomato Sauces. *Omics : A Journal of Integrative Biology*, *17*(12), 611–8. https://doi.org/10.1089/omi.2013.0027

70. Solanky, K. S., Bailey, N. J. C., Beckwith-Hall, B. M., Davis, A., Bingham, S., Holmes, E., … Cassidy, A. (2003). Application of biofluid 1H nuclear magnetic resonance-based metabonomic techniques for the analysis of the biochemical effects of dietary isoflavones on human plasma profile. *Analytical Biochemistry*, *323*(2), 197–204. https://doi.org/10.1016/j.ab.2003.08.028

CHAPTER 2

Design and Evaluation of Standard Lipid Prediction Models Based on 1H-NMR Spectroscopy of Human Serum/Plasma Samples

UNIVERSITAT ROVIRA I VIRGILI
DEVELOPMENT OF 1H-NMR SERUM PROFILING METHODS FOR HIGH-THROUGHPUT METABOLOMICS
Rubén Barrilero Regadera

2.1. Abstract

New approaches are increasingly being used for studying and evaluating coronary heart disease (CHD), especially since the irruption of metabolomics. The classical approach is to use enzymatically-measured standard lipids and these are still the main markers for assessing risk of CHD. Since metabolomics relies on advanced analytical technologies, such as MS and NMR, using them to estimate standard lipids would be of great interest because there is no need for additional biochemical measures. The present study evaluates partial least squares (PLS) and N-way partial least squares (N-PLS) regression models to predict standard lipid concentrations by using serum and plasma sample sets from various clinical centres. Information provided by editing NMR techniques and 2D diffusion NMR was incorporated in these models using four different data structures. Firstly, the models were calibrated and validated with three of the four sample sets (n=591) involved. Then the best estimation models were selected and applied to the left-out sample set. This evaluation of a new sample set gave correlation coefficients of predicted versus biochemical variables above 0.86 and %rRMSE lower than 18%. These values are similar to those found by other studies although, in our case, the results are more general because we used a higher number of samples (n=785) from different sample sets, different clinical centres and different blood matrices (serum and plasma). Finally, we compared the performance of NMR predicted lipids and enzymatically measured lipids in a clinical case study.

2.2. Introduction

Measurement of standard lipid concentrations in fasting blood is one of the main methods used for assessing risk of coronary heart disease (CHD) as indicated by the National Cholesterol Education Program (NCEP) in the guidelines of the Adult Treatment Panel III (ATP III) [1]. Conventionally, standard lipids are total plasma cholesterol (TC), total plasma triglycerides (TG), high-density lipoprotein cholesterol (HDL-C) and low-density lipoprotein cholesterol (LDL-C). The primary target for cholesterol-lowering therapy is LDL-C. However, in cases of high triglycerides (>200 mg/dL), non-HDL cholesterol (non-HDL-C) is preferred as very low-density lipoprotein cholesterol (VLDL-C) is added to LDL-C and better represents the concentrations of all atherogenic lipoproteins [2]. In routine biochemical assays, TC, TG and HDL-C are measured by enzymatic methods (in the case of HDL-C only after a process of precipitation), whereas LDL-C is calculated using the Friedewald equation [3] and non-HDL-C is calculated by subtracting HDL-C

from TC. LDL-C calculated using the Friedewald equation is valid only when the concentration of triglycerides is less than 400 mg/dL, whereas non-HDL-C is not influenced by this limitation.

Metabolomics offers further insights into the study of CHD by discovering new pathologic markers. The ability of these new markers to evaluate CHD risk is usually compared and complemented with standard lipids [4,5]. Therefore, it seems useful to estimate standard lipids with common analytical platforms used in metabolomics.

Lipid moieties of plasma lipoproteins are highly visible in NMR fingerprints [6,7]. However, lipoprotein subclasses have a similar lipid composition and their resonances in NMR spectra overlap. For this reason, spectral areas cannot be directly integrated to quantify the concentration of individual lipid classes. Differences in magnetic susceptibility associated with lipoprotein sizes cause subtle variations in chemical shift [8]. Multivariate analysis methods based on different ^{1}H-NMR spectra take advantage of those shifts found in lipid peaks to quantify lipid classes [9]. Editing NMR techniques can be used to simplify the spectral information and highlight a group of compounds on the basis of their physiochemical properties. Furthermore, systematic variation in the editing parameter yields a 2D spectrum, from which the observed physical property can be extracted and used to improve the estimation of a compound in a mixture [10,11].

Several studies have shown the feasibility of multivariate analysis, based on different ^{1}H-NMR pulses, to predict cholesterol and triglycerides in plasma/serum and the main lipoprotein fractions [10,12–15]. However, they used only one serum (or plasma) sample set obtained from one clinical centre and, in most cases, a limited number of samples. This raises concerns about the generalization of the reported models. Another factor to take into account is the blood matrix selected. Although serum is preferred in metabolomics because the common additives used in plasma (Li-heparin, EDTA or sodium citrate) can interfere, both matrices are used in metabolomic experiments.

The present study evaluates the performance of partial least squares (PLS) and N-way partial least squares (N-PLS) regression models to predict TC, TG, LDL-C, HDL-C and non-HDL-C. These models use a group of data structures comprising editing NMR techniques, 2D NMR (based on diffusion gradients) and the combination of 1D NMR and diffusion coefficients derived from 2D diffusion NMR in order to explore multi-way relationships. We include a total of 785 samples belonging to 4 sample sets (2 serums and 2 plasmas) from different clinical centres, with the aim of obtaining generalizable prediction models for standard lipids.

2.3. Materials and methods

2.3.1. Sample sets and biochemical analysis

Four blood-derived sample sets from different clinical centres were used to calibrate, validate and evaluate the regression models. A detailed explanation of sample collection and handling of each set is given in Table C2.1.

The first set comprises 325 serum samples from male and female subjects, aged 22-80. Of these, 75% were suffering from diabetes mellitus type 2 (DM2) and 25% were a control group of healthy volunteers [16]. The second set comprises 147 plasma samples from healthy male subjects aged 20-75 participating in a study on lipid changes with age [17]. The third set comprises 119 serum samples belonging to healthy male and female subjects (aged 33-45) undergoing a nutritional intervention [18]. The fourth set comprises 194 plasma samples belonging to male and female subjects (aged 39-64), most of them were healthy although some had abnormal lipid levels, and all of them were undergoing a four-month nutritional intervention. Fasting blood samples were collected at baseline and at the end of each four-month intervention period and lipid concentrations were measured using standard enzymatic automated methods and Friedewald equation.

Table C2.1 Collection and handling procedure of plasma/serum samples

Set	Matrix	Plasma/Serum extraction	Storage time	Frozen-thaw?	Aliquots of enzymatic lipids	LDL (TG>400 mg/dL)
set 1	Serum	25 ºC, 2000 rpm, 10 min	2 years (-80 ºC)	No	Cabre et al. 2012	Direct measure
Sample set 2	Plasma EDTA	4 ºC, 910 g, 15 min	8 years (biobank)	No	Sundl et al. 2007	No samples
Sample set 3	Serum	4 ºC, 2205 g (3500 rpm), 15 min	2 years (-80 ºC)	No	Abete et al. 2013	No samples
Sample set 4	Plasma EDTA	4 ºC, 2500 rpm, 15 min	< 2 years (-80 ºC)	No	Different from NMR aliquot	Friedewald

2.3.2. ^{1}H-NMR measurements

For the NMR measurements, a double tube system was used to lock the field and reduce convection currents [19]. First, 430 μL of either serum or plasma was transferred to a 5 mm NMR tube. Then, the inner tube (o.d. 2 mm, supported by a Teflon adapter) containing D_2O (99.9%) as the lock reference was placed coaxially in the NMR sample tube (o.d. 5 mm). Samples were kept at 4°C in the sample changer (SampleJet, Bruker®) until the moment of the analysis. At this point, the sample was pre-heated to 27°C for 1 minute and then to 37°C for 3 minutes inside the magnet. It was kept at that temperature during acquisition. All the samples were prepared by the same NMR technician.

A set of ^{1}H-NMR spectra was acquired for serum/plasma profiling of each sample [11,20]. The set consists of one standard spectrum and a group of edited NMR spectra: a T2-relaxation-edited CPMG spin-echo spectrum with total reduction of the protein background, a 1D diffusion-edited spectrum with suppression of the low molecular weight compounds and a 2D diffusion spectrum to improve separation in the diffusion dimension due to variety of molecular sizes in the plasma/serum mixture. All ^{1}H-NMR spectra were recorded on a Bruker Avance III 600 spectrometer operating at a proton frequency of 600.20 MHz using a 5 mm CPTCI triple resonance pulse field gradient cryoprobe.

Standard ^{1}H-NMR spectra were obtained using a 1D NOESY-presat sequence to suppress the residual water peak. The τ_1 time was set to 4 μs and τ_m (mixing time) to 100 ms. Relaxation-edited spectra were obtained using the CPMG spin-spin T2 relaxation filter, with a total filter time of 410 ms. 1D diffusion-edited spectra were measured using a diffusion-editing pulse sequence with bipolar gradients and longitudinal eddy-current delay (BBPLED). The diffusion time was 120 ms (Δ), the bipolar sine-shaped gradient pulses were 2.6 ms (δ) in length and had a strength of 34.5 G/cm (G1), and an eddy current delay (τ) of 5 ms. For 2D diffusion spectra, a double stimulated echo (2D DSTE) pulse program including bipolar gradients and LED delay was used. This pulse avoids convention artefacts present in diffusion measurements when sample temperature is different from ambient temperature [21]. The gradient pulse strength was increased linearly from 5 to 95% of the maximum strength of 53.5 G/cm (0.535 T/m) in 32 steps. An eddy current delay (τ) of 50 ms, a diffusion time (Δ) of 120 ms and bipolar sine shaped gradient pulses of 6 ms (δ) in length were applied to obtain a reasonable amount of diffusion of the lipoprotein signals in the raw serum or plasma (see Fig. C2.1a).

Shimming was performed for each sample before acquisition. At 50% height, the width of a single line at the glucose doublet (5.2 ppm) was reported to have a mean of 2.34 Hz (±SD: 0.59 Hz) in CPMG spectra without line broadening.

The 90° pulse length was calibrated for each sample [22] and varied from 9.64 µs to 12.44 µs. This length was applied in the four NMR experiments. The spectral width was 20 ppm, and a total of 64 transients were collected into 32k data points for each time signal. The total acquisition time per sample was 90 minutes and only 3 samples were discarded because of receiver overflow.

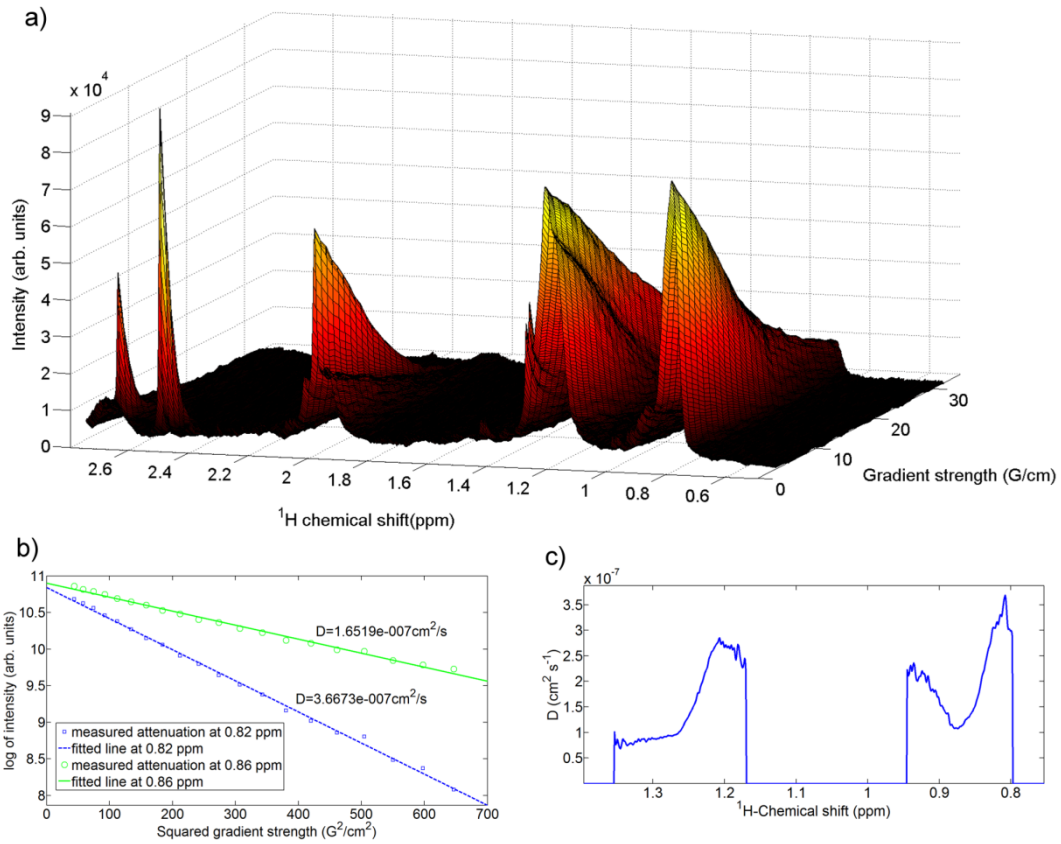


Fig. C2.1 (a) 2D DSTE spectrum of a plasma sample in the region from 0.5 to 2.8 ppm and 32 pulse gradients from 5 to 95% at the maximum strength of 53.5 G/cm, (b) fitting of the observed intensity attenuation (y-axis, logarithmic scale) when different gradients are applied (x-axis, quadratic scale) at two ppm's in the methyl region and their associated diffusion coefficients and (c) "diffusion-weighted" spectrum for methyl and methylene regions. Spectral regions where diffusion coefficients fitted with a mean Pearson's correlation coefficient lower than 0.95 were ignored (set to zero in all the samples)

After acquisition, free induction decays were Fourier-transformed by applying an exponential window function with 1 Hz line broadening, which reduces the presence of noise in the models without compromising the resolution of broad signals from lipids. Baseline correction and automatic phasing were done using in-house scripts written in MATLAB (MathWorks Inc.). The GlcNAc (N-Acetylglucosamine) peak at 2.04 ppm [7] was used as a reference for these spectra. 'Digital ERETIC' (Digital Electronic REference To access In vivo Concentrations, Bruker®) was used for normalization. This signal was calibrated against a sealed reference sample of known concentration before measuring every batch of samples.

In order to check the variation associated with the sample preparation and NMR analysis, five aliquots of a serum sample were measured sequentially.

2.3.3. Calculation of a "diffusion-weighted" NMR spectrum

As well as the aforementioned ^{1}H-NMR measurements, a "diffusion-weighted" NMR spectrum was obtained as a simplification of the second dimension of a 2D diffusion spectrum. It consists of a 1D pseudo-spectrum where diffusion coefficients (y-axis) are plotted versus ppm's (x-axis). The diffusion coefficient (D) is extracted for every point in the frequency axis of a 2D DSTE spectrum, where signal intensity follows an exponentially damped attenuation:

$$I = I_0 e^{-kDG^2}$$

where $k=(2a\gamma\delta)^2(\Delta-5\delta/4-\tau)$, $a=(2/\pi)$ is a gradient shape factor for the half-sine shape, G is the gradient strength in G/cm, δ is the gradient length and I and I_0 are the NMR signal intensities for G and 0 gradient strength, respectively. In our case, we obtained the diffusion coefficient from the slope of the linear regression of the logarithm of intensities when the square of the increasing gradient strength (Fig. C2.1b) is used. Fig. C2.1c depicts an example of a "diffusion-weighted" NMR spectrum for methyl and methylene regions. Spectral regions where diffusion coefficients fitted with a mean Pearson's correlation coefficient lower than 0.95 were ignored (set to zero in all the samples), as they were heavily influenced by noise.

2.3.4. Implementation of multivariate data analysis methods

Correlation heat-maps were calculated as part of the variable selection procedure before the prediction models were built. These heat-maps were obtained as correlations between all the points in the spectra of sample sets 1, 2 and 3 and every lipid class determined by standard enzymatic

methods: TC (ChEBI:50404), TG (ChEBI:17855, mostly ChEBI:47776), LDL-C (ChEBI:47774), HDL-C (ChEBI:47775) and non-HDL-C). The procedure was adapted from the one indicated for one-dimensional STOCSY [23]. Heat-maps were constructed for each 1D ^{1}H-NMR experiment. For the 2D diffusion experiment, the gradient of 14.71 G/cm was selected as a 1D spectrum (so that the double stimulated effect could be evaluated). This gradient shows optimal signal-to-noise ratio and little contribution from low molecular weight metabolites. Fig. C2.2 represents an example of heat-maps using TC concentrations. The regions strongly correlated to lipids are red (positive correlation) and blue (negative correlation).

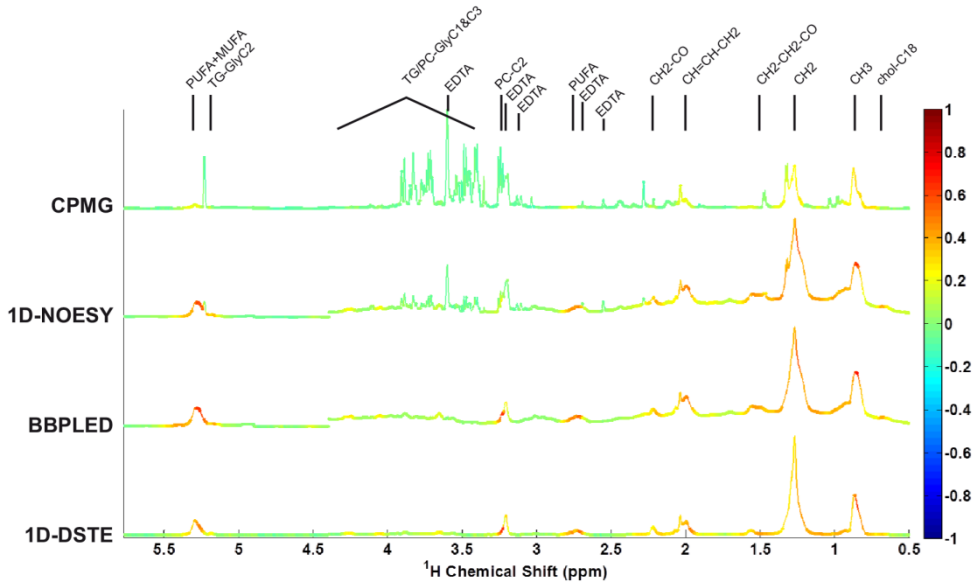


Fig. C2.2 Heat-maps showing the correlations between 1D ^{1}H-NMR spectra and biochemical measurements of TC where 0 means no correlation and 1 and -1 mean total correlation and total anticorrelation, respectively. Resonances from lipids are labelled in the graph: C18 from cholesterol (0.5-0.72 ppm), methyl from fatty acids (0.72-1.08 ppm), methylene from fatty acids (1.08-1.42 ppm), protons attached to beta and alpha carbons of fatty acids (1.42-1.63 and 2.16-2.25 ppm, respectively), allylic hydrogens from fatty acids (1.88-2.1 ppm), bisallylic hydrogens from fatty acids (2.64-2.84 ppm), choline (3.17-3.28 ppm), triglycerides and phosphatidylcholine glycerol backbone (3.62-3.73; 3.84-3.92; 4.02-4.09; 4.21-4.32 and 5.1-5.2 ppm) and olefinic hydrogens from fatty acids (5.2-5.5 ppm)

After the variables of interest had been selected, several data structures were designed to be the inputs of PLS and N-PLS models. These data structures aim to explore the efficiency of using spectral information by itself and in conjunction with extra information from diffusion coefficients for the prediction of lipids. Four data structures were used and PLS or N-PLS methods applied.

The first structure, known as “Dataset 1”, consists of 1D ^{1}H-NMR spectra. This structure is used as input for PLS. It is a two-way array in which each spectrum is regarded as an object and intensities as independent variables.

The second structure, known as “Dataset 2”, combines 1D ^{1}H-NMR and "diffusion weighted NMR spectra." It is a single two-way array, in which each object is a vector with the concatenation of both types of spectra. Because of significant differences in magnitude scales, 1D ^{1}H-NMR spectra and diffusion coefficient “diffusion-weighted” NMR spectra were normalized separately by the mean of their standard deviations. This structure is used as input for PLS.

The third structure, known as “Dataset 3”, is a three-way array in which each object is a 2D diffusion spectrum. Only 13 out of the 32 gradients were used. Low gradients were left out so that resonances due to low molecular weight metabolites would not interfere, whilst larger gradients were discarded because of their low signal-to-noise ratio. This structure is used as input for N-PLS [24].

The fourth structure, known as “Dataset 4”, considers the same inputs as Dataset 2 (normalized 1D ^{1}H-NMR spectra and “diffusion-weighted” NMR spectra) but the ppm scales of both spectra are aligned and arranged in a three-way array. This use of “diffusion-weighted” NMR spectra can be regarded as a compressed version of Dataset 3 in which the diffusion coefficients codify the intensity attenuation at each ppm. Again, this structure is used as input for N-PLS.

In order to test the effect of the most commonly used pre-processing methods in linear regressions [25], both autoscaling (AS) and mean-centring (MC) pre-processing were applied before the models were built.

To evaluate the performance of the prediction models, we used Pearson´s correlation coefficient (r) and the relative Root-Mean-Square Error expressed as a percentage (%rRMSE):

$$\%rRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_{pred}-y_{exp}}{y_{exp}}\right)^{2}} * 100$$

where y_{pred} and y_{exp} are the predicted and experimental lipid concentration, respectively, and n is the number of samples. Due to the large dynamic ranges of the dependent variables (i.e. lipid concentrations), this error is used as it penalizes large deviations at lower values.

Multivariate analysis was performed with PLS_Toolbox version 5.8.3 (Eigenvector Research Inc.) under MATLAB Version 7.10 (MathWorks Inc., Natick, MA).

2.4. Results

2.4.1. Implementation and validation of the prediction models

Correlation heat-maps reveal that those experiments based on diffusion NMR (BBPLED and DSTE) contain more information about standard lipids. This is a clear consequence of filtering signals from low molecular weight metabolites in diffusion-edited NMR experiments, some of whose resonances overlap lipid peaks. Consequently, models based on 1D NOESY-presat and CPMG were discarded.

Although other studies base their analysis almost exclusively on methyl and methylene peaks, correlation heat-maps show large correlations in other spectral regions, such as the choline peak from phospholipids (at 3.2 ppm). Consequently, 14 peaks attributable to lipids were included as 2772 spectral points, from the initial 32k points (listed in Fig. C2.2).

Sample sets 1, 2 and 3 (n=591) were used to calibrate and validate the models based on the data structures listed in section 2.3.4. A previous step of outlier detection was applied for these samples as follows: for each of the sample sets individually, 50% of the samples were used for calibration and the other 50% for the validation of a PLS based on BBPLED and an N-PLS based on 2D DSTE. Both types of model were built for each lipid. The halves were swapped and the procedure was repeated. This was done for 100 permutations of random halves and the RMSE of validation for each sample was averaged. Since errors were found to be normally distributed, Grubbs' test was used to detect outliers for a confidence interval of 95%. When detected in at least one model (PLS or N-PLS) and at least one lipid, samples were considered as outliers and were discarded for future calibration. No more than 5% (n=27) of the samples were considered outliers and excluded.

The aim of this part of the study was to establish the data structure that provides the best predictions for each lipid class. A total of 50% of the samples were used for calibration and the other 50% for validation so that the number of latent variables (LVs) could be chosen. LVs were established as a compromise between maximum explained variance and minimum RMSE in validation without over-fitting the model. This process was repeated 100 times with random

subsets of samples for each half. It ensures that the results were not biased by a single calibration or validation subset. It also enabled the statistical significance between model predictions to be tested.

Table C2.2 shows the mean error for validation samples over the 100 permutations. The lowest error for each lipid class is highlighted in bold. It can be observed that we cannot achieve the best prediction for all standard lipids using a single data structure. To clearly evaluate the differences in the mean errors between the models, we analysed the statistical significance of the results using Student's t-tests. Considering a significance level of 0.05, we found that PLS models based on Dataset 1 (BBPLED) best predicts all the cholesterol measures whereas TG is better adjusted by N-PLS models based on Dataset 3 (2D DSTE). This improvement in TG could be attributed to an enhancement in TG visibility due to longer relaxation effects appearing in the double stimulated sequence.

Since a 2D NMR spectrum requires longer acquisition times than a 1D spectrum (n-times longer than 1D, where n is the number of gradients) and only improves TG prediction slightly, Dataset 1 was established as the reference data structure to be used for the prediction of all the lipid classes in future samples. Moreover, differences in errors between pre-processing methods are still significant at $p<0.05$, but considering that TG and LDL-C show larger errors, we decided to select the mean-centred method as the reference pre-processing method because it predicted these lipids slightly better.

Table C2.2 Mean %rRMSE of 100 permutations of validation subsets of samples sets 1, 2 and 3 for the models based on Datasets 1-4 and their mean-centred (MC) and autoscaled (AS) versions

	%rRMSE Dataset 1 BBPLED (PLS)		%rRMSE Dataset 2 BBPLED+D (PLS)		%rRMSE Dataset 3 2D DSTE (NPLS)	%rRMSE Dataset 4 BBPLED+D (N-PLS)	
	MC	AS	MC	AS	MC	MC	AS
TC	7.28	**7.27**	8.30	7.88	10.04	8.17	7.56
TG	14.88	15.57	15.19	15.57	**14.03**	14.89	15.29
HDL-C	9.66	**9.35**	11.26	10.53	11.69	11.18	10.20
LDL-C	**13.24**	13.70	14.54	14.26	17.32	14.41	13.86
non-HDL-C	9.34	**9.31**	10.17	9.95	12.94	10.15	9.51

Best regression models for each lipid class (those with low %rRMSE) are given in bold.

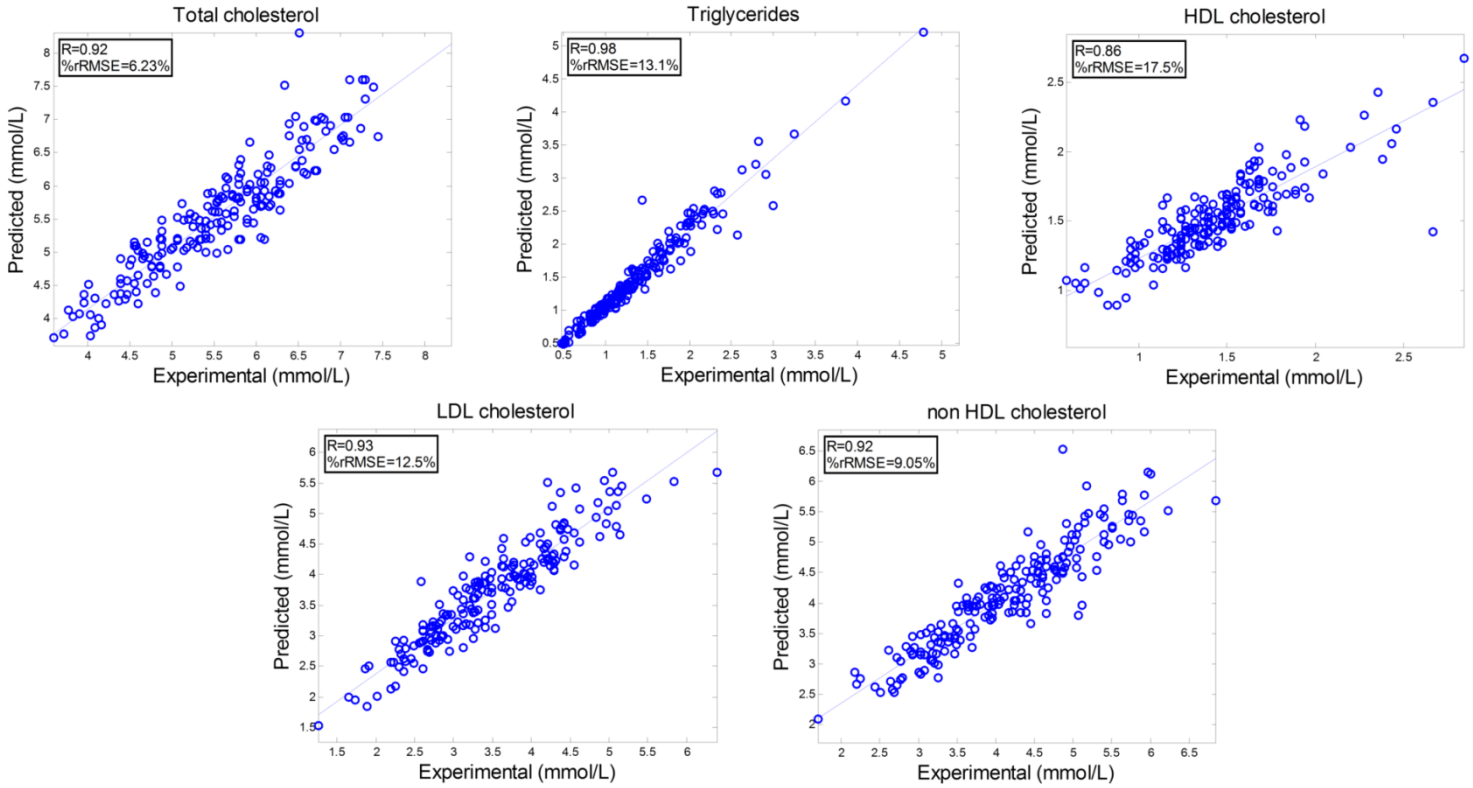


Fig. C2.3 Correlation plot of predicted versus experimental values for Total Cholesterol (TC), Triglycerides (TG), HDL Cholesterol, LDL Cholesterol and non-HDL Cholesterol for the test on sample set 4

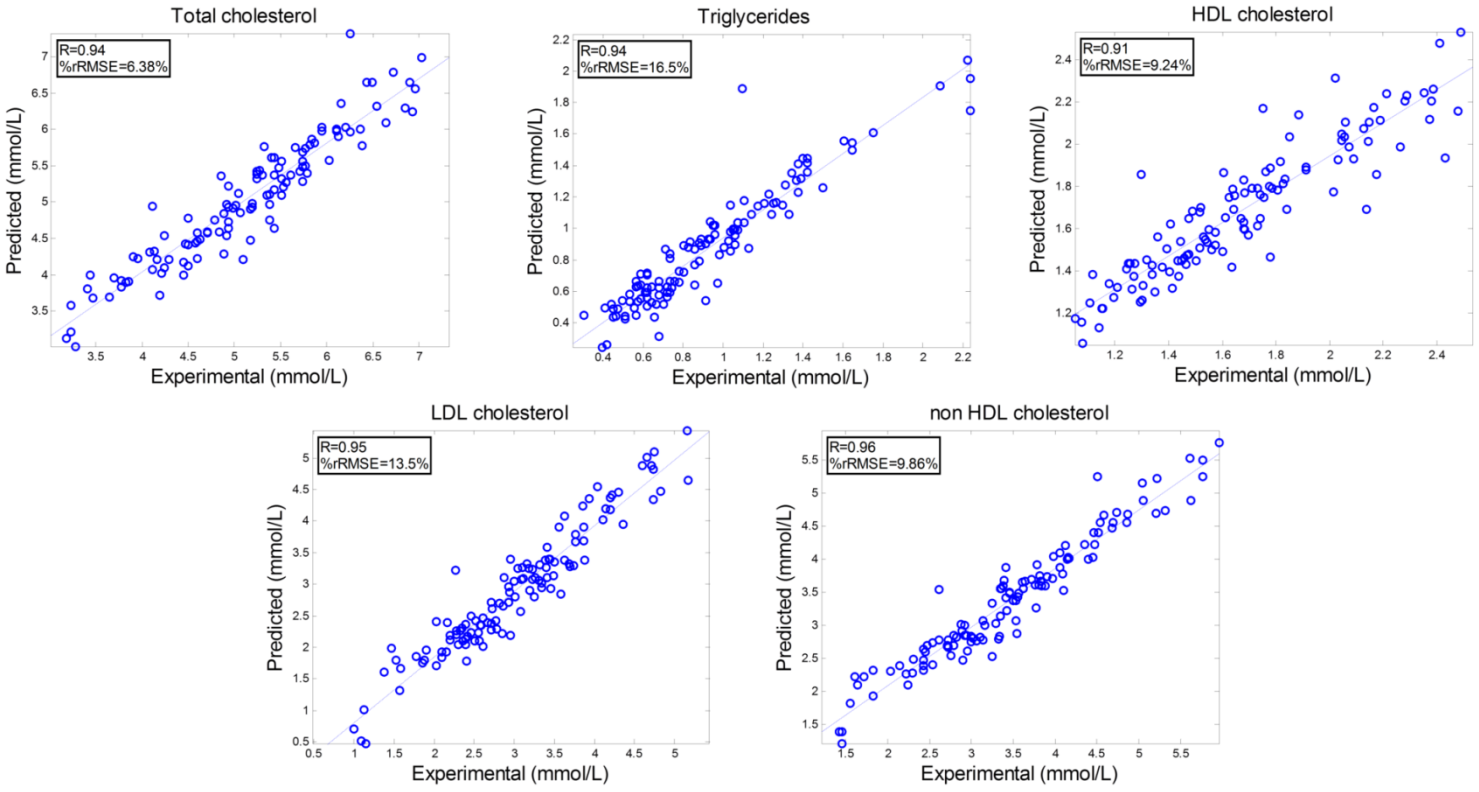


Fig. C2.4 Correlation plot of predicted versus experimental values for TC, TG, HDL-C, LDL-C and non-HDL-C for the test on sample set 3 (serum)

Once the data structure had been chosen, the reference models based on Dataset 1 were built with all the samples of sample sets 1, 2 and 3 and the number of LVs was selected as the mean LV value out of the 100 permutations.

2.4.2. Evaluation of prediction models with a new sample set

Sample set 4 (n=194) was used to evaluate the robustness of the reference models over a new sample set. Plots of predicted values versus reference concentrations are shown in Fig. C2.3.

Pearson's correlation coefficients (denoted as R in Fig. C2.3) between the concentrations measured by enzymatic methods and the predicted concentrations were 0.92, 0.98, 0.86, 0.93 and 0.92 whereas %RMSEs were 6.2%, 13.1%, 17.5%, 12.5% and 9.1% for TC, TG, HDL-C, LDL-C and non-HDL-C, respectively. In general, these results are in agreement with those in Table C2.2 except for HDL-C, which shows a larger deviation.

Because sample set 4 is a plasma set, in order to evaluate the performance in serum samples, we replicated the reference models, but this time they were trained with sample sets 1, 2 and 4 and the blind validation was carried out using the serum sample set 3 (Fig. C2.4).

The influence of sample preparation and NMR analysis was evaluated for the 5 aliquots of one serum sample (see section 2.3.2). The coefficient of variance (CV) for each lipid class was 1.28%, 7.28%, 5.67%, 2.12% and 3.15% for TC, TG, HDL-C, LDL-C and non-HDL-C, respectively. These values may partly explain the larger errors found for TG and HDL-C predictions and the considerable vulnerability of the models to variations in these lipids.

2.4.3. Example of a clinical application of predicted lipids

Since standard lipids are mainly used to characterize CHD risk, we evaluate the ability of our predicted lipid values to classify patients according to several lipid abnormalities involved in CHD.

Samples in sample set 4 were classified into groups according to the presence or absence of hyperlipidaemias. The concentrations limits for each group were taken from ATP III guidelines [1] and each sample was assigned to a group using the TG and TC blood levels measured by enzymatic methods. Four groups were considered: normolipidaemic (TC<200 mg/dL and TG<150 mg/dL), hypercholesterolaemic (HC) (TC>240 mg/dL), hypertriglyceridaemic (HTG) (TG>200 mg/dL) and both HC and HTG. Samples that did not belong to any of the above groups were

considered borderline. They were introduced into the model but not considered for the analysis of groups. A first Principal Component Analysis (PCA) was then carried out using enzymatic measurements of TC, TG, HDL-C, LDL-C and non-HDL-C as the descriptors. A biplot graph (Fig. C2.5a) shows clustering of groups and the distribution of these clusters around the descriptors. A statistical analysis of cluster separation was made using the Mahalanobis distance between centroids as described in [26] and significant cluster separation was found between the four groups ($p<0.001$).

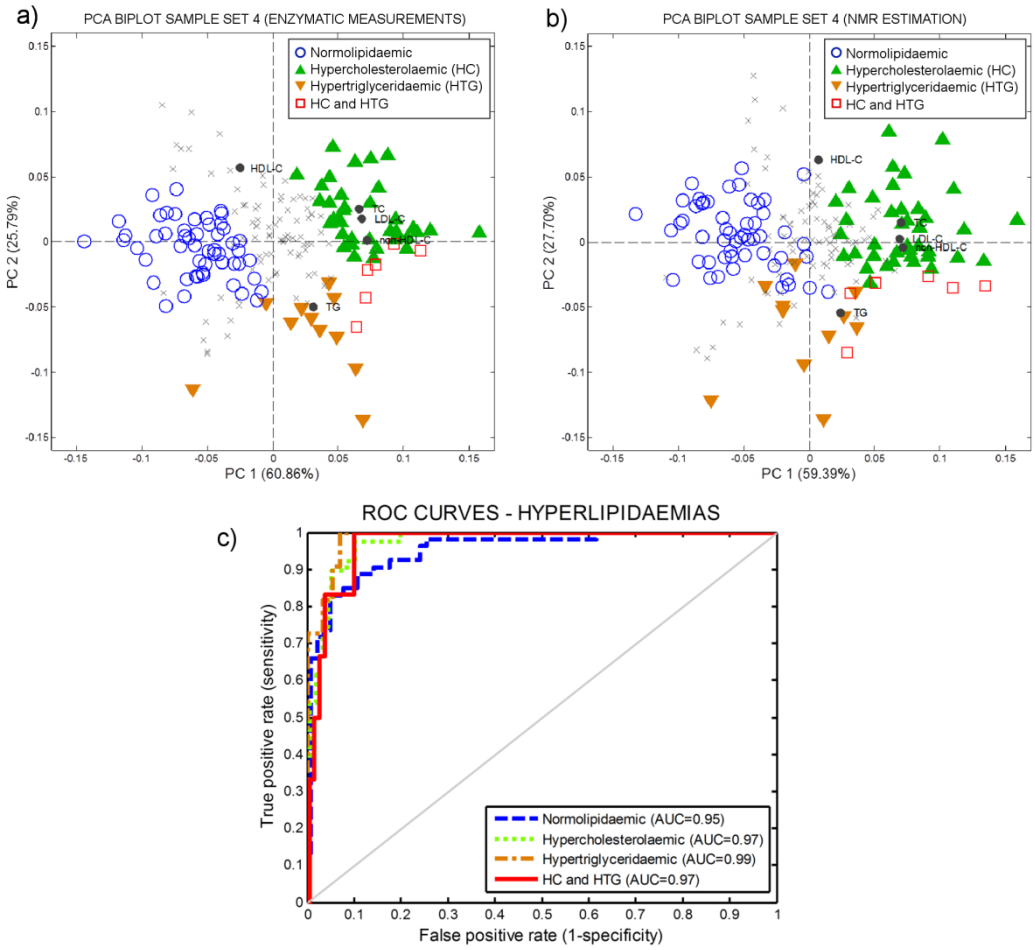


Fig. C2.5 PCA analysis of hyperlipidaemia clustering of patients in sample set 4 using lipids determined by enzymatic test (a) and predicted (b). Samples are classified as follows: normolipidaemic (TC<200 mg/dL and TG<150 mg/dL), HC (TC>240 mg/dL), HTG (TG>200 mg/dL) and both HC and HTG. Samples not included in the previous groups were considered borderline and they are indicated as grey

crosses. Significant cluster separation was found between the four groups in the left PCA ($p < 0.001$) and in the right PCA ($p < 0.01$). (c) ROC curves for hyperlipidaemias for $p < 0.05$

A second PCA (Fig. C2.5b) was performed with the same classified samples but using PLS-predicted lipids as PCA descriptors. At first glance, both biplot graphs depict a similar clustering of groups and distribution of variables. The computation of Mahalanobis distances gives as good a separation between groups ($p<0.01$) as that found for the PCA built with enzymatic measurements. This example highlights that both procedures for determining lipids can give similar clinical outcomes and suggests that PLS-predicted lipids could replace enzymatically measured lipids in metabolomic experiments.

The robustness of classifications using predicted lipids was also evaluated using ROC curves. The area under-the-curve (AUC) for the four groups was 0.95, 0.97, 0.99 and 0.97 for normolipidaemic, hypercholesterolaemic (HC), hypertriglyceridaemic (HTG) and both HC and HTG. These values show that the predicted lipids have a good capacity to diagnose each dyslipidaemia individually. Fig. C2.5c shows the ROC curve and the AUC of predicted lipids for each of the groups in PCA.

2.5. Discussion

Previous studies have shown the feasibility of using multivariate techniques to predict lipids in total plasma and in the main lipoprotein fractions. Table C2.3 summarizes the findings of these studies and compares them with the results reported herein. HDL-cholesterol prediction was slightly worse than in the aforementioned reports. This result could be attributable to the additional precipitation step because of the variability associated with the precipitation reagents used in HDL isolation [27]. Although LDL-C and non-HDL-C depend on HDL-C concentration, these lipids do not reflect this variability because of the greater contribution of TC to their numerical calculation. To improve HDL-C predictions, it could be better to calibrate regression models against direct HDL-C measurements which have shown less inter-laboratory variation [28]. To fully comply with NCEP recommendations, it has been suggested that the CDC reference procedure should be used to calibrate and validate the models for predicting HDL-C [29].

Table C2.3 Summary of the main characteristics in studies of predictions of standard lipids using multivariate methods and NMR spectroscopy

Reference	Matrix	Nº samples	^{1}H-NMR Experiment	Model	TC (Pearson's r)	TG (Pearson's r)	HDL-C (Pearson's r)	LDL-C (Pearson's r)
Bathen et al. (2000)	plasma	44 (calibration and full CV) + 8 (blind test)	simple 90° pulse	PLS	0.99 (blind)	0.98 (blind)	0.88 (blind)	0.97 (blind)
Petersen et al. (2005)	plasma	103 (calibration and full CV)	diffusion-edited	PLS	0.98 (CV)	0.91 (CV)	0.94 (CV)	0.9 (CV)
Dyrby et al. (2005)	plasma	11 (calibration and full CV)	2D diffusion-edited	NPLS	Not evaluated	Not evaluated	0.91 (CV)	0.82 (CV)
Vehtari et al. (2007)	plasma	75 (calibration and 10-fold CV)	simple 90° pulse	Bayesian	Not evaluated	Not evaluated	0.93 (CV)	0.94 (CV)
Mihaleva et al. (2014)	serum	190 (calibration and fold CV) + 100 (blind test)	diffusion-edited	PLS	Not evaluated	Not evaluated	0.98 (blind)	0.92 (blind)
Present article	Plasma and serum	591 (calibration and fold CV) + 194 (blind test)	diffusion-edited	PLS	0.92 (blind)	0.98 (blind)	0.86 (blind)	0.93 (blind)

Key: CV, cross-validation; PLS, partial least squares; NPLS, N-way partial least squares; TC, total cholesterol; TG, triglycerides; HDL-C, high density lipoprotein cholesterol; LDL-C, low density lipoprotein cholesterol.

Moreover, although the diffusion-edited technique is the best for standard lipid predictions, the inclusion of the diffusion dimension derived from 2D NMR experiments (here we used 2D DSTE) has not been found to improve these predictions (except in the case of TG). These results do not agree with those of [10] for standard lipids. They achieved slightly better results than those of previous studies by using 2D diffusion NMR. Unfortunately, direct comparison is not very reliable as different sample sets were used in each study. In the present study, the four data structures were evaluated with the same samples, so the results of the models are directly comparable. Considering the results presented here, we suggest using simpler and cheaper 1D NMR experiments instead of 2D diffusion NMR experiments in order to get better predictions of standard lipids.

Finally, the results show that prediction models of standard lipids perform well when sample sets from different clinical centres are involved, similar to models built and validated using samples from the same metabolomic study. Additionally, plasma and serum were used. Standard lipid prediction models prove to be robust against possible biases caused by different blood collection and treatment protocols applied in clinical practice and different blood matrices (serum and plasma). This validates the generalization of the method presented. The method is also very useful for metabolomic centres that handle samples from different clinical centres, and where mixed serum and plasma samples are received for the same experiment.

2.6. Concluding remarks

This study has evaluated the performance of standard lipid prediction models based on ^{1}H-NMR spectroscopy. We conclude that models based on diffusion-edited NMR yield the best results but that the inclusion of the diffusion dimension only improves predictions of TG. We recommend the use of 14 peaks as they have proved to correlate with all the standard lipids and, consequently, they are assumed to improve the estimation of lipids. The prediction results were similar to those of previous studies but, in this case, the use of sets from different clinical centres and both serum and plasma has led to generalizable results. In practice, this means that common regression models could be used as a general tool for quantifying the lipid panel in metabolomic studies, regardless of the blood matrix used or the origin of samples.

2.7. References

1. Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. (2002). *Circulation*, *106*(25), 3143–3421.

2. Blaha, M. J., Blumenthal, R. S., Brinton, E. A., & Jacobson, T. A. (2008). The importance of non-HDL cholesterol reporting in lipid management. *Journal of Clinical Lipidology*. https://doi.org/10.1016/j.jacl.2008.06.013

3. Friedewald, W. T., Levy, R. I., & Fredrickson, D. S. (1972). Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clinical Chemistry*, *18*(6), 499–502. https://doi.org/10.1177/107424840501000106

4. Mora, S., Otvos, J. D., Rifai, N., Rosenson, R. S., Buring, J. E., & Ridker, P. M. (2009). Lipoprotein particle profiles by nuclear magnetic resonance compared with standard lipids and apolipoproteins in predicting incident cardiovascular disease in women. *Circulation*, *119*(7), 931–939. https://doi.org/10.1161/CIRCULATIONAHA.108.816181

5. Wang, T. J., Larson, M. G., Vasan, R. S., Cheng, S., Rhee, E. P., McCabe, E., … Gerszten, R. E. (2011). Metabolite profiles and the risk of developing diabetes. *Nature Medicine*, *17*(4), 448–53. https://doi.org/10.1038/nm.2307

6. Ala-Korpela, M. (1995). 1H NMR spectroscopy of human blood plasma. *Progress in Nuclear Magnetic Resonance Spectroscopy*, *27*(5–6), 475–554. https://doi.org/10.1016/0079-6565(95)01013-0

7. Nicholson, J. K., Foxall, P. J. D., Spraul, M., Farrant, R. D., & Lindon, J. C. (1995). 750 MHz 1H and 1H-13C NMR Spectroscopy of Human Blood Plasma. *Analytical Chemistry*, *67*(5), 793–811. https://doi.org/10.1021/ac00101a004

8. Lounila, J., Ala-Korpela, M., Jokisaari, J., Savolainen, M. J., & Kesäniemi, Y. A. (1994). Effects of orientational order and particle size on the NMR line positions of lipoproteins. *Phys. Rev. Lett.*, *72*(25), 4049–4052. https://doi.org/10.1103/PhysRevLett.72.4049

9. Mallol, R., Rodriguez, M. A., Brezmes, J., Masana, L., & Correig, X. (2013). Human serum/plasma lipoprotein analysis by NMR: Application to the study of diabetic dyslipidemia. *Progress in Nuclear Magnetic Resonance Spectroscopy*, *70*, 1–24. https://doi.org/10.1016/j.pnmrs.2012.09.001

10. Dyrby, M., Petersen, M., Whittaker, A. K., Lambert, L., Nørgaard, L., Bro, R., & Engelsen, S. B. (2005). Analysis of lipoproteins using 2D diffusion-edited NMR spectroscopy and multi-way chemometrics. *Analytica Chimica Acta*, *531*(2), 209–216. https://doi.org/10.1016/j.aca.2004.10.052

11. Mallol, R., Rodríguez, M. A., Heras, M., Vinaixa, M., Cañellas, N., Brezmes, J., … Correig, X. (2011). Surface fitting of 2D diffusion-edited 1H NMR spectroscopy data for the characterisation of human plasma lipoproteins. *Metabolomics*, *7*(4), 572–582. https://doi.org/10.1007/s11306-011-0273-8

12. Bathen, T. F., Krane, J., Engan, T., Bjerve, K. S., & Axelson, D. (2000). Quantification of plasma lipids and apolipoproteins by use of proton NMR spectroscopy, multivariate and neural network analysis. *NMR in Biomedicine*, *13*(5), 271–288. https://doi.org/10.1002/1099-1492(200008)13:5<271::AID-NBM646>3.0.CO;2-7

13. Petersen, M., Dyrby, M., Toubro, S., Engelsen, S. B., Nørgaard, L., Pedersen, H. T., & Uyerberg, J. (2005). Quantification of lipoprotein subclasses by proton nuclear magnetic resonance-based partial least-squares regression models. *Clinical Chemistry*, *51*(8), 1457–1461. https://doi.org/10.1373/clinchem.2004.046748

14. Mihaleva, V. V., Van Schalkwijk, D. B., De Graaf, A. A., Van Duynhoven, J., Van Dorsten, F. A., Vervoort, J., … Jacobs, D. M. (2014). A systematic approach to obtain validated partial least square models for predicting lipoprotein subclasses from serum nmr spectra. *Analytical Chemistry*, *86*(1), 543–550. https://doi.org/10.1021/ac402571z

15. Vehtari, A., Mäkinen, V.-P., Soininen, P., Ingman, P., Mäkelä, S. M., Savolainen, M. J., … Ala-Korpela, M. (2007). A novel Bayesian approach to quantify clinical variables and to determine their spectroscopic counterparts in 1H NMR metabonomic data. *BMC Bioinformatics*, *8 Suppl 2*(Suppl 2), S8. https://doi.org/10.1186/1471-2105-8-S2-S8

16. Cabré, A., Babio, N., Lázaro, I., Bulló, M., Garcia-Arellano, A., Masana, L., & Salas-Salvadó, J. (2012). FABP4 predicts atherogenic dyslipidemia development. The PREDIMED study. *Atherosclerosis*, *222*(1), 229–234. https://doi.org/10.1016/j.atherosclerosis.2012.02.003

17. Sundl, I., Guardiola, M., Khoschsorur, G., Solà, R., Vallvé, J. C., Godàs, G., … Ribalta, J. (2007). Increased concentrations of circulating vitamin E in carriers of the apolipoprotein A5 gene −1131T>C variant and associations with plasma lipids and lipid peroxidation. *Journal of Lipid Research*, *48*(11), 2506–2513. https://doi.org/10.1194/jlr.M700285-JLR200

18. Abete, I., Perez-Cornago, A., Navas-Carretero, S., Bondia-Pons, I., Zulet, M. A., & Martinez, J. A. (2013). A regular lycopene enriched tomato sauce consumption influences antioxidant status of healthy young-subjects: A crossover study. *Journal of Functional Foods*, *5*(1), 28–35. https://doi.org/10.1016/j.jff.2012.07.007

19. Hedin, N., & Furo, I. (1998). Temperature imaging by 1H NMR and suppression of convection in NMR probes. *Journal of Magnetic Resonance (San Diego, Calif. : 1997)*, *131*(1), 126–30. https://doi.org/10.1006/jmre.1997.1352

20. Beckonert, O., Keun, H. C., Ebbels, T. M. D., Bundy, J., Holmes, E., Lindon, J. C., & Nicholson, J. K. (2007). Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature Protocols*, *2*(11), 2692–2703. https://doi.org/10.1038/nprot.2007.376

21. Jerschow, A., & Müller, N. (1997). Suppression of Convection Artifacts in Stimulated-Echo Diffusion Experiments. Double-Stimulated-Echo Experiments. *Journal of Magnetic Resonance*, *125*(2), 372–375. https://doi.org/10.1006/jmre.1997.1123

22. Wu, P. S. C., & Otting, G. (2005). Rapid pulse length determination in high-resolution NMR. *Journal of Magnetic Resonance*, *176*(1), 115–119. https://doi.org/10.1016/j.jmr.2005.05.018

23. Cloarec, O., Dumas, M. E., Craig, A., Barton, R. H., Trygg, J., Hudson, J., … Nicholson, J. (2005). Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic 1H NMR data sets. *Analytical Chemistry*, *77*(5), 1282–1289. https://doi.org/10.1021/ac048630x

24. Bro, R. (1996). Multiway calibration. Multilinear PLS. *Journal of Chemometrics*, *10*(1), 47–61. https://doi.org/10.1002/(SICI)1099-128X(199601)10:1<47::AID-CEM400>3.0.CO;2-C

25. Craig, A., Cloarec, O., Holmes, E., Nicholson, J. K., & Lindon, J. C. (2006). Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Analytical Chemistry*, *78*(7), 2262–2267. https://doi.org/10.1021/ac0519312

26. Goodpaster, A. M., & Kennedy, M. A. (2011). Quantification and statistical significance analysis of group separation in NMR-based metabonomics studies. *Chemometrics and Intelligent Laboratory Systems*, *109*(2), 162–170. https://doi.org/10.1016/j.chemolab.2011.08.009

27. Warnick, G. R., Nauck, M., & Rifai, N. (2001). Evolution of methods for measurement of HDL-cholesterol: From ultracentrifugation to homogeneous assays. *Clinical Chemistry*.

28. Rifai, N., Cole, T. G., Iannotti, E., Law, T., Macke, M., Miller, R., … Wiebe, D. A. (1998). Assessment of interlaboratory performance in external proficiency testing programs with a direct HDL-cholesterol assay. *Clinical Chemistry*, *44*(7), 1452–1458.

29. Warnick, G. R., & Wood, P. D. (1995). National Cholesterol Education Program recommendations for measurement of high-density lipoprotein cholesterol: executive summary. The National Cholesterol Education Program Working Group on Lipoprotein Measurement. *Clinical Chemistry*, *41*(10), 1427 LP-1433.

CHAPTER 3

Unravelling and Quantifying the "NMR-invisible" Metabolites Interacting with Human Serum Albumin by Binding Competition and T2 Relaxation-based Decomposition Analysis

UNIVERSITAT ROVIRA I VIRGILI
DEVELOPMENT OF 1H-NMR SERUM PROFILING METHODS FOR HIGH-THROUGHPUT METABOLOMICS
Rubén Barrilero Regadera

3.1. Abstract

Quantitative profiling of low-molecular-weight metabolites (LMWM) by ^{1}H-NMR is routinely used in high-throughput serum metabolomics. First, the protein background is attenuated using a T2 filter, then the LMWM signals are resolved by line-shape fitting. However, protein binding modifies the motional properties of LMWM and their signal partially attenuates with the T2 filter, along with the protein background. Consequently, the quantified LMWM signals do not reflect the total concentration in serum but the non-binding part. Here, we present a novel strategy based on binding competition to promote the release of the "NMR-invisible" metabolites from serum proteins and achieve quantifications closer to total concentrations. The study focuses in five clinically relevant amino acids with different binding properties (valine, isoleucine, leucine, tyrosine and phenylalanine). We analyzed their binding affinity to human serum albumin (HSA) in serum mimic samples and promoted the release of their bound fraction by TSP titration. Furthermore, we used a novel combination of pseudo-2D CPMG and multivariate curve resolution analysis, allowing the separation of LMWM and protein signals and providing LMWM quantifications corrected for transverse relaxation effects. We found that TSP concentrations larger than 3 mM released most of the bound fraction and validated these findings in real serum/plasma samples.

3.2. Introduction

High-throughput ^{1}H-NMR profiling of serum is extensively used in large-scale epidemiological studies since it offers dozens of identifiable metabolites measured with a cost per sample comparable to that of standard lipid measurements [1]. Profiling of low molecular weight metabolites (LMWM) commonly implies suppressing protein signals using a CPMG (Carr-Purcell-Meiboom-Gill) sequence [2,3], designed to take advantage of large differences in T2 relaxation times between macromolecules (proteins and lipoproteins) and LMWM [4]. Next, the LMWM signals (partially attenuated by the T2 filter depending on their transverse relaxation [5]) can be resolved based on known molecular characteristics [6–9]. However, complex biological samples, as in the case of serum and plasma, involve multiple sources of chemical exchange, such as metal ion chelation, protonation, proton exchange with water and molecular binding, leading to signal modifications, such as spectral shifts, amplitude decrease and line broadening [10].

In this respect, the ability of plasma proteins to bind small molecules potentially affects the quantitative NMR analysis of LMWM by two simultaneous exchange processes. For metabolites showing weak binding to proteins (K_d > 1 mM), the rate of exchange between bound and unbound ("free in solution") state is fast on the NMR timescale; the T2 relaxation time of the observable LMWM signal is composed by the weighted T2 relaxation times of both the "free" metabolite and the metabolite weakly-bound to protein [11]. Thus, the resulting T2 relaxation time is shorter than the "free" one and the attenuation introduced in the CPMG experiments is increased [12]. For metabolites showing strong binding to protein (K_d < 0.1 μM), the rate of exchange between bound and unbound state is slow on the NMR timescale and this is reflected in short T2 relaxation times and broad line-shapes for the bound part, similar to the proteins ones. Consequently, these "T2-shortened" signals are also suppressed in CPMG experiments [12]. This situation has been referred as the "NMR-invisible" pool of metabolites in serum [13,14]. For instance, previous studies demonstrated that lactate, 3-hydroxybutyrate, acetoacetate, pyruvate and 2-hydroxybutyrate were partly "invisible" in ^{1}H-NMR spectra of plasma and bovine serum albumin [14,15]. Nicholson, et al. [16] identified phenylalanine, tyrosine, histidine and citrate binding to plasma proteins in untreated plasma samples of normal individuals, causing a decrease in the expected signal. In both studies, a partial release of binding metabolites was obtained upon acidification. More recently, Jupin, et al. [17] confirmed the previous findings in a human serum albumin (HSA) model and added creatinine and lysine to the list of the known binding metabolites. They also found an increase of the signals of these metabolites in a fatted-HSA model. Moreover, protein binding could compromise inter-subject analysis, since regulating factors of protein binding such as those mentioned above (pH level and FFA/HSA ratio) are commonly altered under abnormal conditions, such as diabetes [18].

A common approach to avoid protein binding relies on serum deproteinization methods. Protein denaturation with organic solvents are commonly used for large metabolite recovery [19], at the expense of structural alteration of proteins. Conversely, ultrafiltration preserves the protein structure, but most of the bound metabolites coprecipitate with proteins [20]. Both deproteinization approaches are time-consuming, require moderate manipulation of samples and limit the characterization of macromolecular content such as lipoproteins and glycoproteins in the native environment.

An alternative way to reduce protein binding affecting LMWM relies in binding competition. This approach consists in the addition of an exogenous molecule that competes with endogenous metabolites for the binding domains in proteins. Moreover, using binding competition avoids the

sample pre-treatment and the separation of the protein content involved in deproteinization. Traditionally used in pharmacology, binding competition has also been evaluated in metabolomics to study the affinity of endogenous and exogenous molecules to serum proteins. For instance, Jupin, et al. [21] reported that HSA-binding metabolites can be released under the addition of fatty acids and Daykin, et al. [22] used ibuprofen to release and measure the citrate bound to HSA. These studies highlight the potential of binding competition in quantitative metabolomics and the need of further characterization of competitive compounds in order to selectively or non-selectively release the "NMR-invisible" pool of metabolites.

However, using competitive binding still requires the application of a T2 filter in order to suppress the protein background signal, at the expense of introducing T2-dependent attenuations in the "visible" LMWM signals as indicated above. In this sense, the use of an additional NMR dimension based on differences in T2 relaxation (e.g. by acquiring a series of CPMG or Hahn spin-echo experiments with increasing tau delays) and multivariate curve resolution (MCR) allows separating components by their different T2 relaxation times [23] and provides T2 relaxation decays, from which T2-corrected concentrations can be derived.

Hence, in this study we present two strategies to improve the "NMR-visibility" of metabolites involved in protein binding. The first strategy promotes the release of the binding metabolites by the addition of the exogenous compound 3-(trimethylsilyl)propionic-2,2,3,3-d4 acid (TSP) commonly used as internal standard in ^{1}H-NMR and with known affinity to HSA [24]. The second strategy consists of the application of MCR analysis to a series of CPMG experiments with fixed spin-echo delay and increasing number of loops (from now on, we refer to this concept as pseudo-2D CPMG), in order to extract the signal of the "NMR-visible" part of a metabolite and calculate its associated T2 relaxation time, which in turn allow the calculation of T2-corrected quantifications. This study mainly focuses on the quantification of five amino acids that are relevant in metabolic diseases (valine, isoleucine, leucine, tyrosine and phenylalanine) [25–27], some of them previously reported to bind to HSA and/or serum protein at different degrees [16,17]. The developed method has been optimized in a "serum mimic" composed by the mixture of 20 metabolites and HSA. HSA was used since it accounts for over 50% of total plasma protein content and its extraordinary ligand binding capacity due to its several low- and high-affinity ligand-binding sites dominated by hydrophobic and electrostatic interactions [28,29]. Finally, we validated our findings in human serum and plasma samples.

3.3. Experimental section

3.3.1. Materials

All chemical products were purchased from Sigma–Aldrich (Steinheim, Germany). The complete list of metabolite standards is listed in Table C3.1. We used two types of HSA in this study: globulin free (< 1 %) and fatty acid free (< 0.02 %, product number A3782) and globulin free (< 1 %) containing fatty acids (product number A8763). The lack of free metabolites in HSA and fHSA was visually confirmed by acquiring a standard one-dimensional (1D) ^{1}H NMR spectrum of each type in PBS, using a single 90° pulse experiment with water presaturation. Stock solutions were prepared in 50 mM phosphate buffer solution (PBS) at pH 7.4, consisting of 11 mM of NaH_2PO_4 and 39 mM of Na_2HPO_4 dissolved in 9 parts of MILLI-Q® H_2O and 1 part of D_2O (99.9 atom % D). 3-(trimethylsilyl)propionic-2,2,3,3-d_4 (TSP) sodium salt was prepared as a stock solution of 58 mM in PBS.

Table C3.1 Composition of serum mimic, reference T2-relaxation times and chemical shifts of quantified signals

Compound	CAS number	Concentration (mM)	T2 aqueous solution[a] (s)	Chemical shift (ppm)
3-hydroxybutyric acid	150-83-4	0.077	1.95	1.18 (d)
Alanine	56-41-7	0.427	1.97	1.47 (d)
Arginine	74-79-3	0.114	ND	-
Asparagine	70-47-3	0.082	1.80	2.81-2.96(m)
Citrate	77-92-9	0.114	1.02	2.65 (d)
Glucose (α, β)	50-99-7	4.971	1.58, 1.05	5.23 (d), 3.90(d)
Glutamic acid	56-86-0	0.097	ND	-
Glutamine	56-85-9	0.510	3.21	2.45 (m)
Histidine	71-00-1	0.131	3.47	7.06 (s)
Isoleucine	73-32-5	0.061	1.48	0.93 (t)
Lactate	79-33-4	1.489	2.65	4.1 (q)
Leucine	61-90-5	0.099	1.38	0.95 (dd)
Lysine	657-27-2	0.179	0.83	3.02 (t)
Phenylalanine	63-91-2	0.078	3.61	7.32-7.44 (m)
Proline	147-85-3	0.198	ND	-
Serine	56-45-1	0.160	ND	-
Threonine	72-19-5	0.128	2.95	3.58 (d)
Tryptophan	73-22-3	0.055	3.04	7.53 (d), 7.72 (d)
Tyrosine	60-18-4	0.055	3.03	6.89 (d)
Valine	72-18-4	0.212	1.61	1.25 (d)
HSA	70024-90-7	0.600	0.006-0.012	-

[a] From pure standards in PBS solution at pH 7.4 and 310K and measured using pseudo-2D CPMG. The T2-relaxation time was obtained by fitting an exponential function to the decaying intensity

3.3.2. Serum mimic and TSP titration

The serum mimic mixture was designed to characterize the binding properties of a set of LMWM in conditions that mimic the human serum, and was designed to take into account the naturally-occurring binding competition between the different serum metabolites. The serum mimic sample consisted of a mixture of the 20 most concentrated polar metabolites in serum, including the five target amino acids (L-Phenylalanine, L-Valine, L-Isoleucine, L-Leucine, L-Tyrosine), and HSA in PBS, all of them at the concentration range found in serum from healthy population [30]. The composition of the serum mimic, the chemical shifts of the signals used for quantification of each metabolite and a reference T2 of these metabolites in PBS solution are listed in Table C3.1. TSP was added to the serum mimic at concentrations ranging from 0.3 mM to 12 mM. The final volume for all samples was 700 μL.

3.3.3. Spiked human serum samples

Fasting blood from a healthy volunteer was collected and centrifuged immediately at 3000 rpm for 15 min at 21°C to obtain serum that was stored at -80°C until NMR analysis. Lipid, total albumin and total protein content in serum were determined using enzymatic and colorimetric assays, respectively (Spinreact S.A.U., Spain), adapted to a Cobas Mira Plus autoanalyzer (Roche Diagnostics, Spain). Lipid levels were 1.41 mM for triglycerides and 4.95 mM for total cholesterol, HSA (molecular weight: 66,437 Da) was 0.78 mM and total protein was 0.78 g/L. Spiked serum samples were prepared by the addition of 70 μL of diluted mixtures of the five target amino acids in PBS (L-Phenylalanine, L-Valine, L-Isoleucine, L-Leucine, L-Tyrosine) to 530 μL of serum. The spiked concentrations of these amino acids were 0 (only PBS was added to the serum), 0.5, 1, 2 and 4 times the concentration reported in healthy human serum [30] (see Table C3.1). The standard additions were repeated including TSP at a fixed concentration of 6 mM in all the standard solutions. In order to obtain robust calibration curves and evaluate the relative standard deviation (%RSD), we replicated three times the minimum and the maximum concentration points (0 and 4 respectively). Finally, we checked the effect of sample dilution on protein binding by a two-fold dilution of the samples without TSP in PBS.

3.3.4. Plasma samples for validation

The validation sample set comprised 83 plasma samples from male and female subjects, with a mean age of 58 (±12.2) suffering from rheumatoid arthritis. Blood samples were withdrawn from the antecubital vein of each participant at the time of recruitment after 12 h overnight fast. EDTA

plasma was prepared from venous blood collected into sterile, evacuated tubes (BD, Vacutainer®). Plasma was immediately separated by low-speed centrifugation at 4°C and frozen at -80°C until analysis. Previous to the analysis, 275 µL of each plasma sample was two-fold diluted in PBS to a final volume of 550 µL. After a first NMR analysis, each sample was mixed with 20 µL 6 mM TSP and analyzed again. The study was approved by the ethical committee of the Sant Joan University Hospital (Reus, Spain) and all participants gave written informed consent prior to their inclusion in the study.

3.3.5. NMR analysis

Samples were transferred into a 5-mm NMR tube before analysis. All the samples were measured at 310K in a Bruker Avance III 500 MHz spectrometer using ^{1}H-CPMG-presat experiment with T2 filters ranging from 0 to 2.1 s in 10 non-linearly distributed steps (0, 0.007, 0.013, 0.02, 0.046, 0.12, 0.21, 0.42, 1.05, and 2.1 s corresponding to 0, 8, 16, 24, 56, 144, 256, 512, 1280 and 2560 loops, respectively) with a fixed spin-echo delay of 400 µs. The relaxation delay between scans was 5 s in order to avoid most of the attenuation due to longitudinal relaxation. During this time, water signal was irradiated with a low-power RF pulse. The acquisition time was 3.5 s and the free induction decays (FIDs) consisted of 65536 complex data points. The spectral width was 20 ppm. For each spectrum 32 scans were recorded resulting in a total acquisition time of each sample of 50 min. After acquisition, the FIDs were zero-filled (131072 real data points) and apodized by an exponential window function with 1 Hz line broadening, prior to Fourier Transformation. LMWM signals were assigned based on bibliography [31] and previously confirmed in 1D and 2D NMR measurements of serum/plasma samples. Calculated signal areas were converted into molar concentrations using the PULCON procedure [32]. Briefly, a synthetic signal was added in the spectrum of a sealed reference sample containing 2 mM of sucrose (Part no. Z10036, Bruker BioSpin AG, Fällanden, Switzerland) and its area was converted into molar concentration. Next, this calibrated signal was inserted in the spectrum of all the samples to convert signal areas into molar concentrations compensating by the individual acquisition parameters. The reference sample was measured under the same conditions described above and immediately before the rest of the samples to avoid instrumental drifts.

3.3.6. T2 relaxation-based decomposition by multivariate curve resolution (MCR)

Each pseudo-2D CPMG spectrum consists of 10 spectra (one for each T2 filter) with 131072 spectral points each. For each signal to analyze, we created a matrix *D* of dimension 10 × n, where *n* is the number of points in the spectral window that contains the target LMWM signal. Since the target LMWM signal is normally superimposed on the broadband protein signal, the individual components were separated using multivariate curve resolution – alternating least squares (MCR-ALS) in the bilinear model [33]:

$$D = \sum_{i=1}^{k} c_i \, s_i^T + \varepsilon \quad (1)$$

where c_i is a 10 × 1 column vector containing the concentration profile (intensity decay due to T2 relaxation) for component i, s_i is a n × 1 column vector containing the pure spectral profile for component i and ε the non-modeled residual matrix (10 × n sized). We considered two components (k=2): one for the LMWM signal (i.e. the weighted averaged signal with the free and weakly-bound contribution of that metabolite) and the other for the protein signal (containing also the metabolite contribution in slow-exchange binding). The optimization process starts with an initial estimation of c_i, based on the equation of the transverse relaxation (T2 relaxation):

$$I_i(t) = I_i(0) * e^{-\frac{t}{T2_i}} \quad (2)$$

Where, $I_i(0)$ is the intensity at 0 s and $T2_i$ is the T2 relaxation time of the signal for component i, and *t* is the time elapsed, in seconds, after the 90° pulse of the CPMG sequence. For the initial estimation of c_i, $T2_{1,2}$ were set to 1 s and 0.01 s, since these values are expected to fall into the range of the T2 for LMWM and proteins, respectively, and $I_{1,2}(0)$ were set to 1. The concentration profile c_i results from taking the corresponding $I_i(t)$ at the experimental CPMG times defined in the "NMR analysis" section. Then, the optimization process runs an iterative sequence until the convergence criterion is satisfied: first, $s_{1,2}$ are calculated from the original matrix *D* and $c_{1,2}$ by least squares estimation. Secondly, $c_{1,2}$ are recalculated from *D* and $s_{1,2}$ by least squares estimation. The resulting $c_{1,2}$ are then adjusted to exponential decays by least squares approximation of Eq. (2). Finally, the residual ε is calculated to evaluate the convergence criterion and the sequence is repeated again. Once the optimization process converges, the MCR-ALS returns the bilinear model, $I_{1,2}(0)$, and $T2_{1,2}$ that provide the minimal residual solution, where $T2_{1,2}$ are the experimental relaxation times calculated for both the LMWM and the protein signal. The script was written in Matlab software (ver. 7.5.0. The Mathworks, Inc., Natick, MA, USA) and adapted from the MCR module in DOSY toolbox [34].

3.3.7. Calculation of T2-corrected concentrations

The T2-corrected area for each component *i* was obtained by scaling each individual spectral profile to the intensity at 0 s filter (multiplying s_i by $I_i(0)$). Conversion from areas to molar concentrations was made using the PULCON procedure mentioned above.

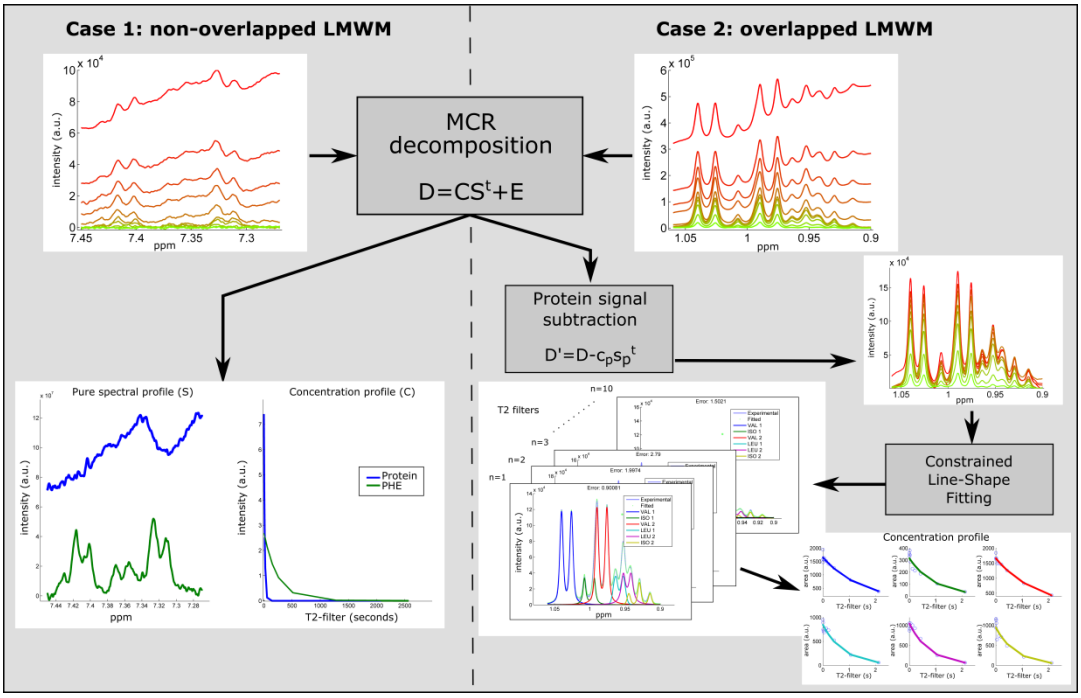


Fig. C3.1 Schemes of bilinear decomposition by MCR (left) and bilinear decomposition by MCR with additional line-shape fitting (right). In the left case, the original matrix was decomposed by MCR providing the two-ways (spectral profile "c" and relaxation decay "s") of each underlying component (LMWM and protein signals). In the right case, the protein background component obtained with MCR was first subtracted from the original matrix D to provide the resulting D' (i.e., D' does not have protein signal). Then, line-shape fitting was applied to each slide in the D' matrix (representing a 1D spectrum with a specific T2 filter) in order to deconvolve the spectral profile of each overlapping LMWM. Finally, the decrease in the area of these spectral profiles was plotted against the T2 filters and the relaxation decays (T2-relaxation times) obtained from the decaying rate

3.3.8. Line-shape fitting step for overlapped metabolites

In the case of overlapping resonances from different metabolites, such as leucine and isoleucine, MCR provides the component for protein background and a single component composed of both

metabolites due to their similar T2 relaxation times, where the apparent T2 relaxation time is the weighted average of their individual T2 relaxation times. Instead of using this component to characterize the resulting "composite" LMWM signal, the outcome of the MCR decomposition was used in a different manner. First, the MCR component of protein background was subtracted from the original matrix D. Then, the individual metabolite signals in every slice of the new "protein-removed" D' matrix were deconvolved using an in-house Matlab-based constrained lineshape fitting (CLS) algorithm of Lorentzian models (see Soininen, et al. [35] for additional information about the CLS method). The T2 relaxation time associated with each metabolite was calculated from the decaying areas using Eq. (2) and T2-corrected concentrations calculated as described before. These procedures are graphically shown in Fig. C3.1.

3.4. Results and discussion

3.4.1. Characterization of protein binding interactions in serum mimic under TSP titration

Fatty acid free HSA was initially used in the serum mimic in order to recreate the case with the largest amount of bound metabolites. For each point of the TSP titration, we acquired a pseudo-2D CPMG spectrum and calculated the T2-corrected concentration and the T2 relaxation time of each metabolite. Next, the recovery ratio (representing the "visible" fraction) was calculated as the ratio between T2-corrected concentration and the total concentration of each metabolite included in the serum mimic. Fig. C3.2 shows the recovery ratios and the T2 relaxation times of the five target amino acids (valine, isoleucine, leucine, phenylalanine and tyrosine) as a function of the TSP concentration. It is noteworthy to mention the significant decrease in the T2 relaxation times of LMWM in the serum mimic compared to the same LMWM in PBS (Table C3.1), as a consequence of the reduction of molecular mobility due to fast-exchange binding of LMWM to HSA and an increase in viscosity of the media [36]. At the initial point in Fig. C3.2a (no TSP addition), more than half of the phenylalanine signal was lost by slow-exchange binding to protein. Analogously, only 62%, 63% and 79% of the signals of tyrosine, leucine and isoleucine, respectively, were "NMR-visible". Conversely, no signal loss was observed for valine resonances. Due to the use of T2-corrected concentrations, these signal losses can only be imputed to the binding to HSA in slow-exchange regime. Despite quantitative discrepancies, the integrity of valine signal and the strong decrease of phenylalanine and tyrosine signals conform with previous findings [16,17]. The significant decrease of leucine signal in our study contrasts with the mild

decrease previously reported [17,21]. The quantitative discrepancies between our results and the reference literature could be attributed to dissimilarities in the composition of the serum mimic, the sample conditions (temperature and ionic strength) or the methodologies used for the analysis (integration vs. line-shape fitting, 1D CPMG vs. pseudo-2D CPMG and TSP normalization vs. PULCON). To our knowledge, the affinity of isoleucine to HSA binding is reported for the first time here. Fig. C3.2a also shows a continuous release of binding metabolites from HSA with the increase of TSP concentration. As seen in the figure, a TSP concentration of 3 mM in the sample released most of the five amino acids, with recoveries that remained nearly constant for larger concentrations of TSP, suggesting no further competition between TSP and the five amino acids for slow-exchange binding to HSA. Even after most of the signal of the five amino acids was recovered, the continuous increase of T2 relaxation times in Fig. C3.2b suggests further competition for fast-exchange binding to HSA. Nevertheless, the fast-exchange binding of the five amino acids to HSA appears to tend to an equilibrium state after the addition of 3 mM of TSP, as it is derived from the decrease of the slopes after this point.

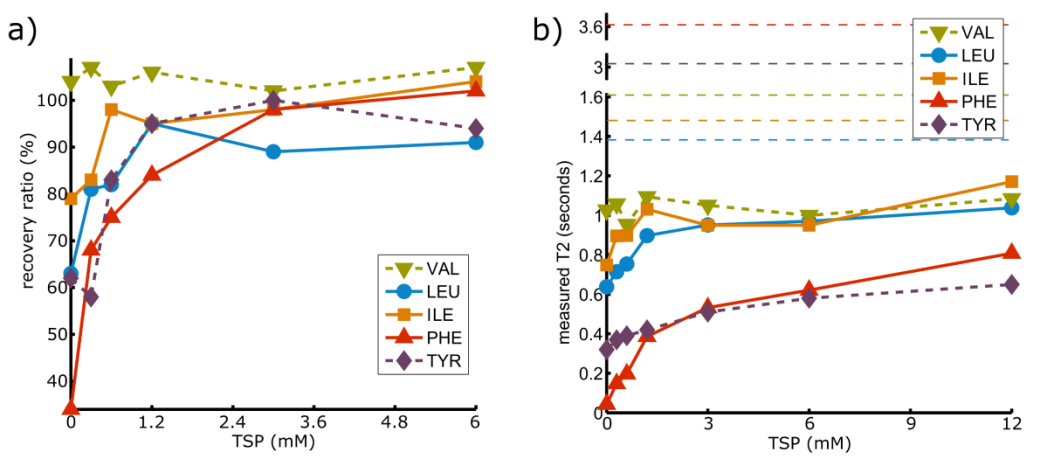


Fig. C3.2 (a) Recovery or "visible" ratios (calculated as the ratio between the T2-corrected concentrations and the total concentration in the serum mimic) for the five amino acids as a function of the TSP concentration. Most of the release was obtained at 3 mM TSP. The nuclei used for the quantification of each amino acid are indicated in Table C3.2. (b) T2 relaxation times as a function of the TSP concentration, showing the effect of TSP addition in fast-exchange interactions between metabolites and HSA. As a reference, the horizontal dashed-lines indicate the T2 relaxation times obtained from standard solutions of each metabolite in PBS (Table C3.1). Key: VAL, valine; ILE, isoleucine; LEU, leucine, TYR, tyrosine; PHE, phenylalanine

Table C3.2 summarizes the recoveries of 16 metabolites included in the serum mimic without the addition of TSP and at the TSP concentration of 3 mM. The recovery ratios without TSP allow classifying the 16 metabolites based on their degree of affinity to HSA in slow-exchange regime. At this point, citrate, lysine and tryptophan were mostly bound to HSA as confirmed by the lack of "visible" signals assignable to these metabolites. A second group of metabolites (lactate, 3-hydroxybutyrate, phenylalanine, tyrosine, leucine and isoleucine) were partly bound to HSA with recoveries between 13 and 80%. Finally, alanine, asparagine, threonine, glutamine, histidine, valine and glucose showed non-binding properties. This cluster pattern agrees with previous findings [17,21] for most of the metabolites in common, with the larger differences found for histidine and lysine. Nevertheless, direct comparison between studies should be taken prudently due to the multiple experimental dissimilarities exposed in previous lines.

Table C3.2 Recovery ratios for points at 0 and 3 mM of TSP (HSA) and 0 mM of TSP (fatted-HSA) from the titration series in the serum mimic

Compound	Quantified H's	HSA		Fatted HSA
		Recovery[1] at 0 mM TSP (%)	Recovery[1] at 3 mM TSP (%)	Recovery[1] at 0 mM TSP (%)
Citrate	half α(C**H**$_2$)	0	0	56
Lysine	ε(C**H**$_2$)	0	0	122
Tryptophan	H7, H6 ring	0	0	58
Lactate	β(C**H**$_3$)	13	96	89
3-hydroxybutirate	ɣ(C**H**$_3$)	26	108	94
Phenylalanine	H2-H6 ring	34	101	105
Tyrosine	H3, H5 ring	62	100	92
Leucine	δ(C**H**$_3$)	63	89	92
Isoleucine	δ(C**H**$_3$)	79	98	87
Alanine	β(C**H**$_3$)	95	97	105
Asparagine	half β(C**H**$_2$)	107	103	101
Glucose (α, β)	H1(α), half CH_2-C6 (β)	107	106	97
Glutamine	ɣ(C**H**$_2$)	103	103	107
Histidine	H2, H5 ring	130	136	108
Threonine	β(C**H**)	103	108	102
Valine	ɣ(C**H**$_3$)	104	102	104

[1]Recovery ratios calculated as the ratio between the T2-corrected concentration and the total added to the serum mimic

The capacity of HSA to selectively bind some LMWM could be explain by two types of interactions [16]: a first group of LMWM could bind due to hydrophobic characteristics, for

example, the aliphatic chain in the case of lysine, leucine and isoleucine, and the aromatic ring, such as the case of phenylalanine, tyrosine and tryptophan. The binding of a second group of LMWM could be explained by electrostatic forces and all these metabolites have in common that they cannot take a zwitterionic form. Examples of these metabolites are the alpha-hydroxy acids lactate or citrate. The results also denoted significant effects due to binding competition between metabolites. For instance, we have previously observed valine partially bound to HSA at the selected concentrations in a simple valine-HSA model in PBS (Fig. C3.3), probably displaced by binding competition with other metabolites in the serum mimic.

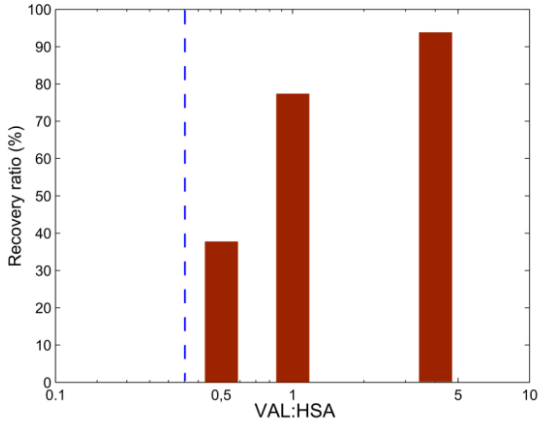


Fig. C3.3 Recovery ratios of valine as a function of valine addition to HSA solution. The recovery ratios were calculated as the ratio between the T2-corrected concentration and the total added. HSA concentration was fixed to 0.3 mM. Recoveries below 1 for lower ratios of valine-HSA mixture suggest slow-exchange binding between valine and HSA. Vertical dashed-line shows a reference value for a common VAL:HSA concentration ratio in serum of healthy population

Analogously to the results found for the five target amino acids, a TSP concentration of 3 mM was also found to release the maximum amount of most of the metabolites in the serum mimic. Nevertheless, lysine, citrate and tryptophan, reported to be released from HSA under the addition of other exogenous molecules [22,37–39], were not released by the TSP addition. The case of lysine and tryptophan could be explained by the non-polar nature of a long aliphatic chain and an indole ring, respectively. The case of citrate, however, could be explained by electrostatic forces arising from the three carboxylic groups. An accumulated error ±10 % along the whole workflow analysis was assumed; this value is broadly accepted in quantitative NMR and could explain recoveries exceeding 100 %.

Furthermore, since HSA is a free fatty acid transporter in serum, we replicated the experiment replacing the HSA by fatted-HSA to further study the interactions in the serum mimic. As seen in Table C3.2, in the serum mimic with fatted-HSA most of the studied metabolites were in "free" state before TSP addition, even for the strongest ligands, such as lysine, citrate and tryptophan, whose recoveries were above the 50%. These results are in line with previous studies [17,21], and confirm the higher binding affinity of HSA to circulating fatty acids compared to LMWM, that is likely explained by the hydrophobic interactions arising from their long aliphatic chain [40].

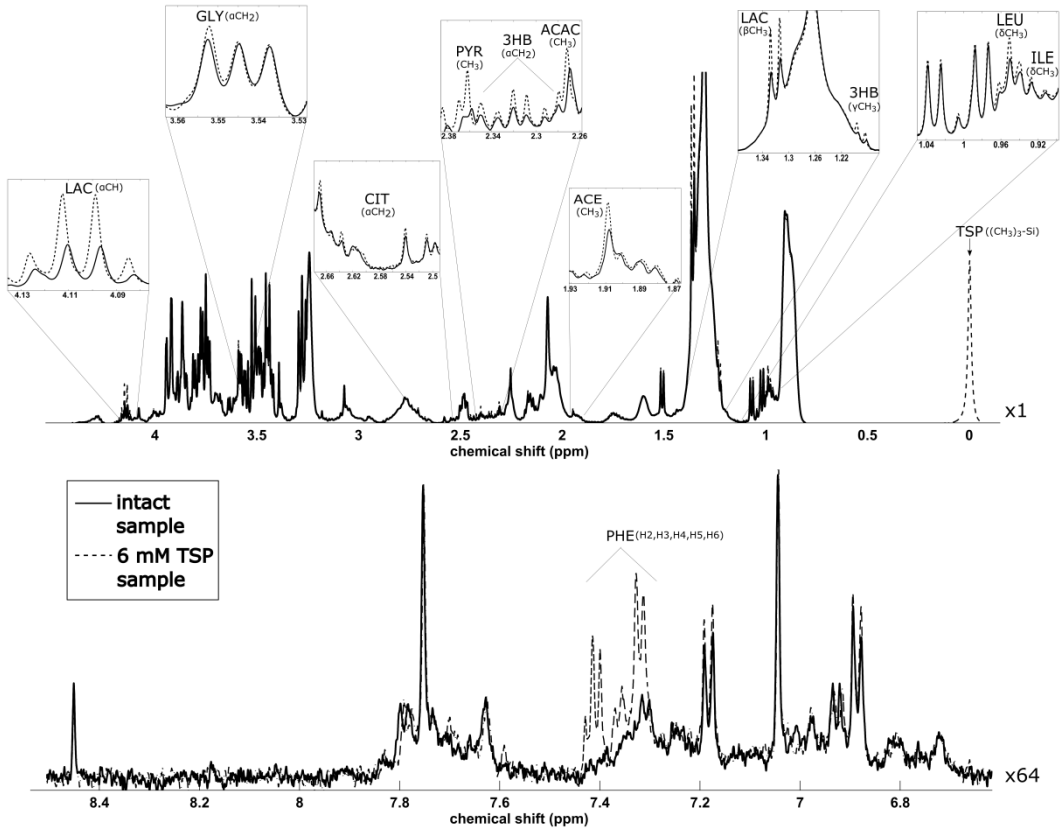


Fig. C3.4 Comparison between CPMG spectra (120 ms filter) of an intact serum sample and with the addition of TSP at 6 mM. To favour the visualization, spectra are split (top-aliphatic, bottom-aromatic) and appropriately scaled. Uninformative regions have been excluded. Signals showing clear differences between conditions are labelled. Key: ILE, isoleucine; LEU, leucine, 3HB, 3-hydroxybutyrate; LAC, lactate; ACE, acetate; ACAC, acetoacetate; PYR, pyruvate; CIT, citrate; GLY, glycine; PHE, phenylalanine

3.4.2. Quantitative analysis of LMWM release under TSP addition in real serum

Fig. C3.4 compares the spectra of intact serum and the same serum with 6 mM TSP, acquired using a CPMG sequence with a 120 ms T2 filter. TSP concentration was fixed at 6 mM considering the results of the serum mimic experiment and taking into account the additional HSA and the rest of the protein content (mostly globulins) in real serum. Signals showing clear differences in intensity are indicated in the figure. In accordance with the results obtained with the serum mimic, the signal intensities of phenylalanine, lactate, 3-hydroxybutyrate, leucine and isoleucine increased by the addition of TSP. Phenylalanine represents an extreme case in which its aromatic signal, barely detected in the intact sample, was clearly identified in the sample with TSP. Unlike the serum mimic, a slight release of citrate was observed in real serum. We also found other metabolites not included in the serum mimic, such as acetate, acetoacetate, pyruvate and glycine showing a similar behavior, some of them previously reported to have binding affinity to serum proteins [14,21]. Since TSP has a short aliphatic chain, it is not expected to compete with free fatty acids or promote large conformational changes in proteins [41]. This fact is illustrated in Fig. C3.4 by the intense sharp signal at 0 ppm corresponding to the excess of TSP not bound to HSA. Moreover, the deuterated form of TSP avoids interfering signals in ^{1}H-NMR spectra.

A detailed analysis based on T2-corrected concentrations and T2 relaxation times was carried out. We built calibration curves based on standard additions of the five target amino acids, with and without TSP (Fig. C3.5). The recovery ratios derived from the calibration curves and the T2 relaxation times are shown in Fig. C3.6. Significant increments in the recovery ratios ($p<0.001$, $n=9$, Welch's t-test) were found for isoleucine (79% to 109%), leucine (59% to 85%) and phenylalanine (41% to 75%) after the addition of TSP. In accordance with the results in serum mimic, valine signal presented minimal variation with TSP addition, indicating that most of the serum valine is in “free” state and thus revealing insignificant protein binding. Conversely to the serum mimic results, tyrosine was not completely released by TSP addition in the real serum, being the calculated “NMR-visible” fraction ca. 50% of its total content in the sample. This fact could be attributed to different binding mechanisms among the protein species in serum. Similarly, Jupin, et al. [17] observed dilution reducing the quantity of lysine weakly-bound to proteins in real serum but not in a HSA model. In the case of T2 relaxation times (Fig. C3.6c), all the metabolites except valine were significantly increased after TSP addition ($p<0.001$, $n=9$, Welch's t-test), suggesting a larger extent of fast-exchange interactions between LMWM and proteins in serum.

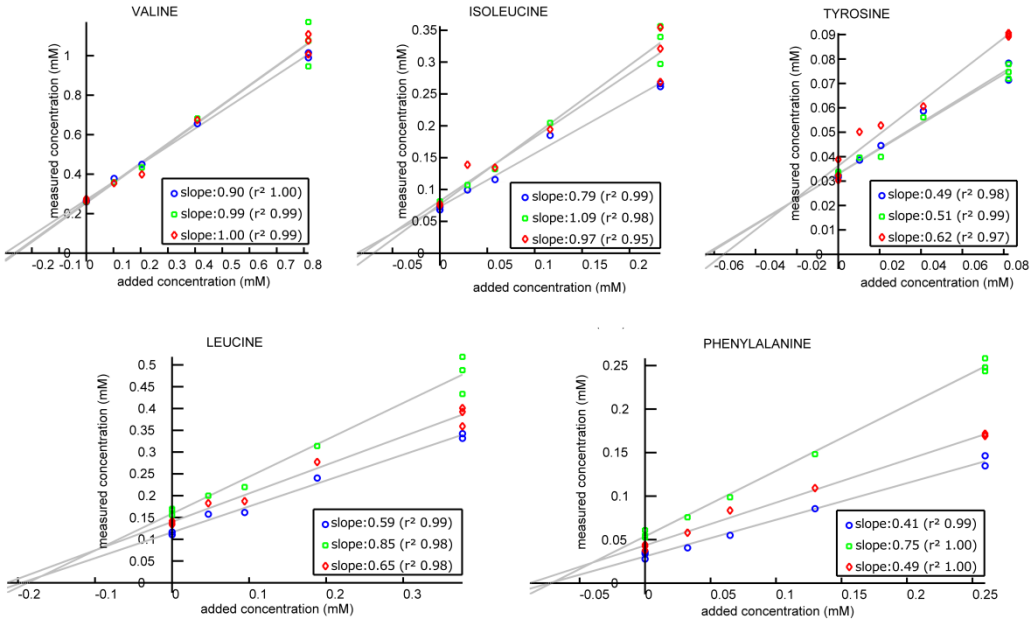


Fig. C3.5 Calibration curves of valine, isoleucine, tyrosine, leucine and phenylalanine in the following cases: intact serum (blue circles), 6 mM TSP serum (green squares) and two-fold diluted serum (red diamonds). The curves represent the T2-corrected concentrations as a function of the spiked concentration of each amino acid. The intersection in the y-axis indicates the observable concentration (i.e. the "NMR-visible"). The intersection in the x-axis represents the total concentration in sample (i.e. the "NMR-visible" plus the "NMR-invisible" part). The ratio between both concentrations (i.e. recovery ratio) is represented by the slope of the curve. Robust curves were calculated averaging the curves obtained by exchanging the replicates at the extremes (n=9). Using these 9 curves in each condition, statistical significance was tested using a parametric Welch's t-test and relative standard deviations (RSD) were 0.08 as average and below 0.15 in all the cases

3.4.3. Quantitative analysis of LMWM release under sample dilution in real serum

The effect of sample dilution in the release of binding metabolites was also investigated. Dilution is typically used in NMR-based metabolomics that demands large volumes of sample and has been reported to promote the release of bound ligands in serum [21]. As shown in Fig. C3.6a, although the amount of the "visible" fraction of all the metabolites increased with two-fold dilution of serum, it also provided lower release of binding metabolites than TSP addition, except for the case of tyrosine where the recovery increased from 49% to 62% and valine that showed similar recoveries using both approaches. Moreover, large dilution of the sample is not recommended

since it compromises the already poor sensitivity of NMR, as some metabolites could reach the acceptable quantification limit [42]. In the case of fast-exchange regime (Fig. C3.6c), dilution provided larger T2 relaxation times for valine (1.13 s) and tyrosine (0.65 s) than TSP addition (1.02 and 0.52 s respectively), similar for leucine and isoleucine and much lower in the case of phenylalanine (0.17 versus 0.4 s).

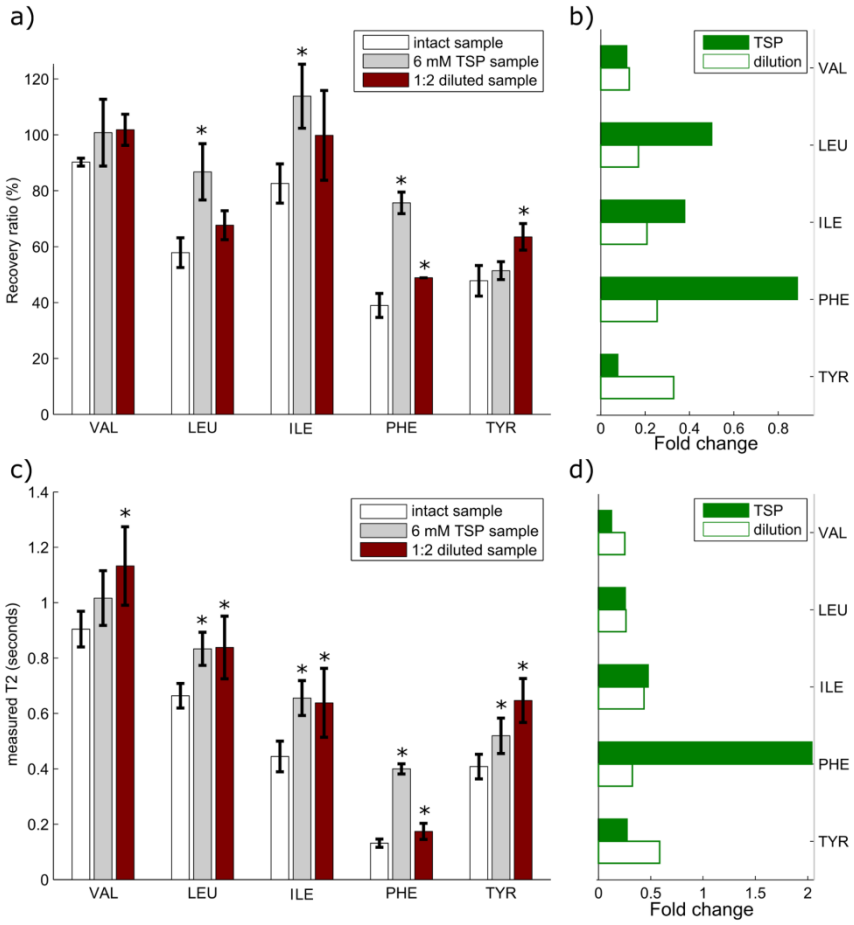


Fig. C3.6 Effects on the NMR signals of the five target amino acids under the addition of TSP at 6 mM and sample dilution compared with the intact serum samples. (a) Mean (±SD) recovery ratios. (b) Fold-changes comparing the mean recovery ratios in TSP-added (green) and diluted samples (white) with the mean recovery ratios in intact samples. (c) Mean (±SD) T2 relaxation times and (d) Fold-changes comparing the mean T2 relaxation times in TSP-added (green) and diluted samples (white) with the mean T2 relaxation times in intact samples. Asterisks indicate a significant difference respect to the intact case ($p<0.001$, n=9, Welch's t-test). Key: VAL, valine; LEU, leucine; ILE, isoleucine; PHE, phenylalanine; TYR, tyrosine

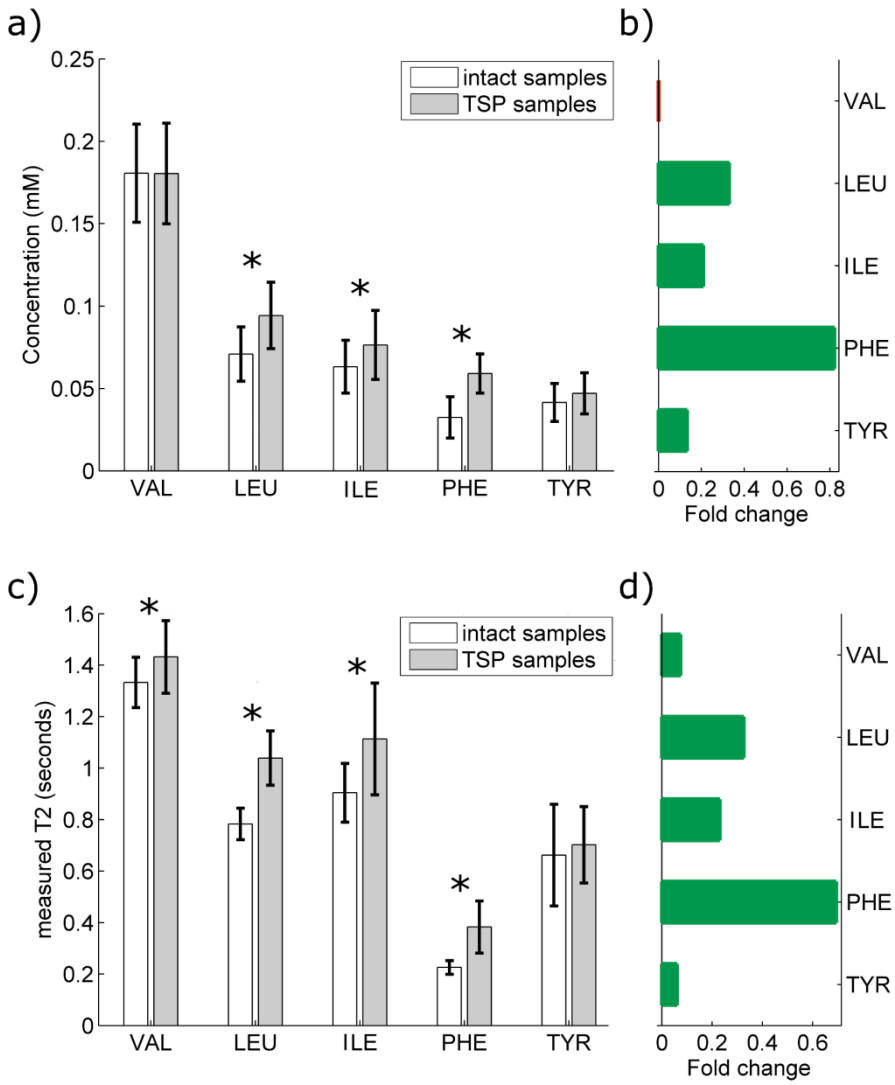


Fig. C3.7 Bar graphs representing (a) the mean (±SD) concentrations and (c) the mean (±SD) T2 relaxation times of valine (VAL), leucine (LEU), isoleucine (ILE), phenylalanine (PHE) and tyrosine (TYR) for intact plasma samples and the same samples with 6 mM TSP from the validation sample set (n=83). Asterisk indicates a significant difference in means for a p<0.001 (Student's t-test). (b, d) Fold changes between sample conditions for mean concentrations and mean T2 relaxation times, respectively

3.4.4. Validation in plasma samples

In order to evaluate the possible generalization of the previous findings, we compared the T2-corrected concentrations and T2 relaxation times of 83 plasma samples before and after the addition of TSP at 6 mM. Fig. C3.7 summarizes the results for the five target amino acids. Fig. C3.7a compares the mean T2-corrected concentrations (±SD) between the samples with and without TSP addition. Analogously with the results obtained with the spiked serum, significant increments ($p<0.001$, $n=83$, Student's t-test) were found for isoleucine, leucine and phenylalanine, but not for valine and tyrosine. Signal increases in the plasma samples (Fig. C3.7b) were similar to those found in the serum experiment (Fig. C3.6b), showing a high correlation of $r=0.93$ despite the different blood-derived matrices used in both experiments (Fig. C3.8). Concentration changes (as fold changes) for the rest of the profiled metabolites can be found in Table C3.3. In addition, the T2 relaxation times of the five amino acids, except for tyrosine, were significantly increased ($p<0.001$) after the TSP addition (Fig. C3.7c). These T2 relaxation times were larger than those in the serum experiment (Fig. C3.6c); however, the fold-changes were lower (Fig. C3.6d and Fig. C3.7d). These differences could be attributed to the dilution of plasma samples.

Table C3.3 Concentration change (as fold change) of quantified metabolites in the validation sample set after TSP addition compared to intact samples

Compound	fold change
Valine	-0.02
Leucine	0.33
Isoleucine	0.21
Phenylalanine	0.82
Lactate	0.44
Citrate	0.40
Alanine	-0.01
Tyrosine	0.13
Creatinine	0.05
Creatine	-0.02
Formate	-0.05
Acetate	0.45
Acetoacetate	-0.10
Pyruvate	0.43
Glucose α	0.02
Glucose β	0.04
Histidine	-0.07
Glutamine	-0.06
NAG 2.03 ppm	0.08
NAG 2.07 ppm	0.06

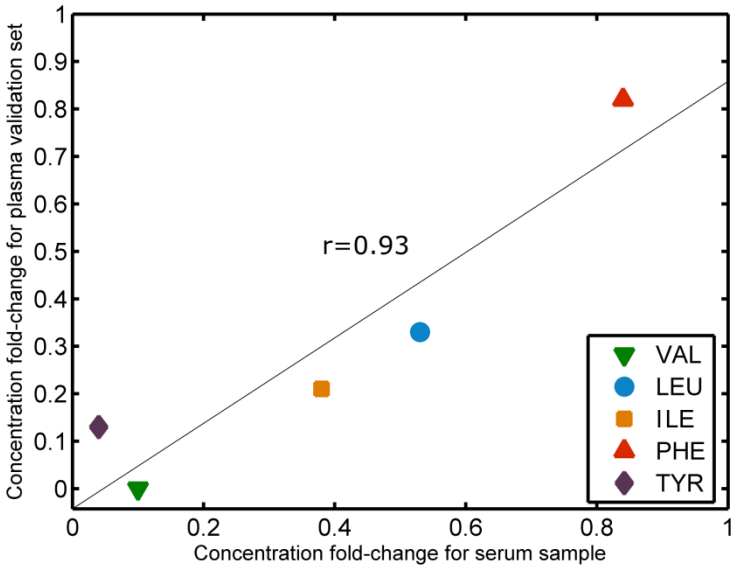


Fig. C3.8 Scatter plot and correlation between concentration changes (as fold changes) of real serum and plasma samples. Fold changes are extracted from Fig.C3.6b and Fig.C3.7b, respectively. Key: VAL, valine; LEU, leucine; ILE, isoleucine; PHE, phenylalanine; TYR, tyrosine

3.4.5. T2 relaxation effects and implications of T2 relaxation-based decomposition in protein binding monitoring

We have seen how T2 relaxation times varied depending on the conditions of the samples such as dilution and the presence of TSP. Ideally (i.e. not allowing transverse T2 relaxation) this dependency should not affect the quantification of the "visible" LMWM concentration, however, the arbitrary T2 filter of a CPMG sequence introduce this variation into the analysis in form of signal attenuation and can mislead the conclusions of recovery ratios under the effect of releasing agents. In order to illustrate how the effect of T2 attenuations could change the outcome of the study, we reproduced in Fig. C3.9 the calibration curves of phenylalanine in the human serum experiment using quantifications based on a 120 ms CPMG filter instead of T2-corrected quantifications. Using this approach, we found recoveries of 11% and 50% for samples without and with 6 mM of TSP respectively (fold-change of 3.54), instead of the 41% and 75% using T2-corrected quantifications (fold-change of 0.83). This fact would partially explain some numerical discrepancies between our results and previous results in HSA models [17], which reported 16% of "visible" phenylalanine in a HSA-based serum mimic, whereas this percentage increased up to 41% with our method.

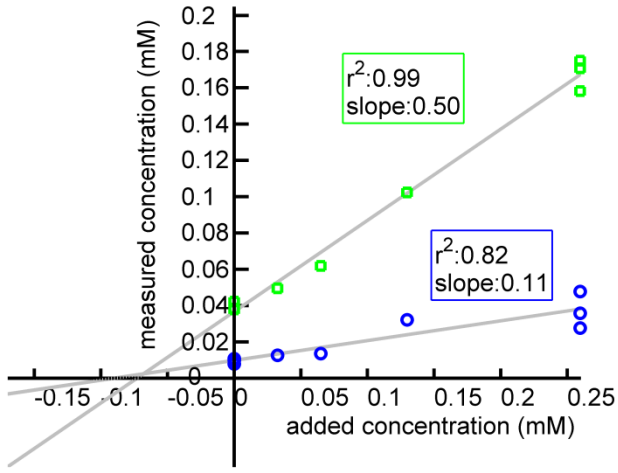


Fig. C3.9 Calibration curves of phenylalanine spiking to intact serum (blue circles) and serum including 6 mM TSP (green squares) calculated using a 120 ms T2-filter CPMG experiment. Phenylalanine signal was resolved using line-shape fitting and PULCON procedure. Comparison with curves in Fig. C3.5 denotes that the T2-filter introduces attenuations (seen as a general decrease of slopes), moreover, these attenuations are dependent on the conditions (intact and TSP-added sample), and increase the relative gap between slopes

So far, we have exposed that using pseudo-2D CPMG spectra would be desirable to obtain quantifications not affected by T2 attenuations. However, this approach is not cost and time-efficient for most applications, such as high-throughput ^{1}H-NMR metabolomics, in which one-dimensional CPMG spectra are normally used. In order to correct for the T2 attenuations found in a 1D ^{1}H-CPMG of serum, Bharti, et al. [43] suggested applying corrections based on T2 references to the LMWM concentrations, after acidifying the sample in order to increase the LMWM in free state. Adding TSP has been proved to reduce fast-exchange interactions between LMWM and serum proteins, making the relaxation properties of LMWM resemble more the ones in free state. Therefore, TSP addition could substitute the suggested acidification previous to T2 corrections, while avoiding the risk of protein denaturation related to acidification processes.

3.5. Concluding remarks

Competitive binding has been proved to be a valid strategy for reducing the impact of protein binding in quantitative ^{1}H-NMR profiling of low molecular weight metabolites in serum. In this

study, we have reported that the addition of a small quantity of TSP increases the "NMR-visibility" of most of the binding metabolites and also reduces the influence of T2 relaxations due to fast-exchange interactions. Based on our experiments, we observed 0.83, 0.45 and 0.38-fold increases of the measurable signal of phenylalanine, leucine and isoleucine, respectively, whereas valine and tyrosine "visibility" were barely modified. This method is compatible with high-throughput ^{1}H-NMR metabolomics, since it only requires the addition of TSP in the D_2O commonly used in NMR sample preparation workflows. The study also highlights the benefits of using pseudo-2D CPMG spectra and bilinear decomposition in order to characterize binding competition without the presence of T2 attenuations. We therefore suggest the use of this methodology to test other exogenous binding compounds, in order to develop a chemical "cocktail" that maximize the release of the "NMR-invisible" metabolome in serum. This methodology could also aid to investigate the clinical implications of the bound/unbound ratios of the LMWM in serum. Simultaneously, the clinical impact of increasing the "NMR-visibility" of metabolites and the evaluation of indirect effects of the addition of exogenous compounds in serum (both clinical and analytical) have to be investigated.

3.6. References

1. Soininen, P., Kangas, A. J., Würtz, P., Suna, T., & Ala-Korpela, M. (2015). Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Cardiovascular Epidemiology and Genetics. *Circulation: Cardiovascular Genetics*, *8*(1), 192–206. https://doi.org/10.1161/CIRCGENETICS.114.000216

2. Soininen, P., Kangas, A. J., Wurtz, P., Tukiainen, T., Tynkkynen, T., Laatikainen, R., … Ala-Korpela, M. (2009). High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism. *Analyst*, *134*(9), 1781–1785. https://doi.org/10.1039/B910205A

3. Beckonert, O., Keun, H. C., Ebbels, T. M. D., Bundy, J., Holmes, E., Lindon, J. C., & Nicholson, J. K. (2007). Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature Protocols*, *2*(11), 2692–2703. https://doi.org/10.1038/nprot.2007.376

4. Tang, H., Wang, Y., Nicholson, J. K., & Lindon, J. C. (2004). Use of relaxation-edited one-dimensional and two dimensional nuclear magnetic resonance spectroscopy to improve detection of small metabolites in blood plasma. *Analytical Biochemistry*, *325*(2), 260–272. https://doi.org/10.1016/j.ab.2003.10.033

5. Tiainen, M., Soininen, P., & Laatikainen, R. (2014). Quantitative Quantum Mechanical Spectral Analysis (qQMSA) of 1H NMR spectra of complex mixtures and biofluids. *Journal of Magnetic Resonance*, *242*, 67–78. https://doi.org/10.1016/j.jmr.2014.02.008

6. Gómez, J., Brezmes, J., Mallol, R., Rodríguez, M. A., Vinaixa, M., Salek, R. M., … Cañellas, N. (2014). Dolphin: a tool for automatic targeted metabolite profiling using 1D and 2D 1H-NMR data. *Analytical and Bioanalytical Chemistry*, *406*(30), 7967–7976. https://doi.org/10.1007/s00216-014-8225-6

7. 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*(15), 2088–2090. https://doi.org/10.1093/bioinformatics/bts308

8. Ravanbakhsh, S., Liu, P., Bjordahl, T. C., Mandal, R., Grant, J. R., Wilson, M., … Wishart, D. S. (2015). Accurate, Fully-Automated NMR Spectral Profiling for Metabolomics. *PLOS ONE*, *10*(5), e0124219. https://doi.org/10.1371/journal.pone.0124219

9. Weljie, A. M., Newton, J., Mercier, P., Carlson, E., & Slupsky, C. M. (2006). Targeted Profiling: Quantitative Analysis of 1H NMR Metabolomics Data. *Analytical Chemistry*, *78*(13), 4430–4442. https://doi.org/10.1021/ac060209g

10. Ross, A., Schlotterbeck, G., Dieterle, F., & Senn, H. (2007). Chapter 3 - NMR Spectroscopy Techniques for Application to Metabonomics BT - The Handbook of Metabonomics and Metabolomics (pp. 55–112). Amsterdam: Elsevier Science B.V. https://doi.org/10.1016/B978-044452841-4/50004-7

11. Fielding, L. (2007). NMR methods for the determination of protein-ligand dissociation constants. *Progress in Nuclear Magnetic Resonance Spectroscopy*. https://doi.org/10.1016/j.pnmrs.2007.04.001

12. Van, Q. N., Chmurny, G. N., & Veenstra, T. D. (2003). The depletion of protein signals in metabonomics analysis with the WET–CPMG pulse sequence. *Biochemical and Biophysical Research Communications*, *301*(4), 952–959. https://doi.org/10.1016/S0006-291X(03)00079-2

13. De Graaf, R. A., & Behar, K. L. (2003). Quantitative 1H NMR spectroscopy of blood plasma metabolites. *Analytical Chemistry*, *75*(9), 2100–2104. https://doi.org/10.1021/ac020782+

14. Bell, J. D., Brown, J. C. C., Kubal, G., & Sadler, P. J. (1988). NMR-invisible lactate in blood plasma. *FEBS Letters*, *235*(1–2), 81–86. https://doi.org/10.1016/0014-5793(88)81238-9

15. Chatham, J. C., & Forder, J. R. (1999). Lactic acid and protein interactions: implications for the NMR visibility of lactate in biological systems. *Biochimica et Biophysica Acta*, *1426*(1), 177–84.

16. Nicholson, J. K., & Gartland, K. P. R. (1989). 1H NMR studies on protein binding of histidine, tyrosine and phenylalanine in blood plasma. *NMR in Biomedicine*, *2*(2), 77–82. https://doi.org/10.1002/nbm.1940020207

17. Jupin, M., Michiels, P. J., Girard, F. C., Spraul, M., & Wijmenga, S. S. (2013). NMR identification of endogenous metabolites interacting with fatted and non-fatted human serum albumin in blood plasma: Fatty acids influence the HSA–metabolite interaction. *Journal of Magnetic Resonance*, *228*, 81–94. https://doi.org/10.1016/j.jmr.2012.12.010

18. Cistola, D. P., & Small, D. M. (1991). Fatty acid distribution in systems modeling the normal

and diabetic human circulation. A 13C nuclear magnetic resonance study. *Journal of Clinical Investigation*, *87*(4), 1431–1441. https://doi.org/10.1172/JCI115149

19. Nagana Gowda, G. A., & Raftery, D. (2014). Quantitating metabolites in protein precipitated serum using NMR spectroscopy. *Analytical Chemistry*, *86*(11), 5433–5440. https://doi.org/10.1021/ac5005103

20. Daykin, C. A., Foxall, P. J. D., Connor, S. C., Lindon, J. C., & Nicholson, J. K. (2002). The comparison of plasma deproteinization methods for the detection of low-molecular-weight metabolites by (1)H nuclear magnetic resonance spectroscopy. *Analytical Biochemistry*, *304*(2), 220–30. https://doi.org/10.1006/abio.2002.5637

21. Jupin, M., Michiels, P. J., Girard, F. C., Spraul, M., & Wijmenga, S. S. (2014). NMR metabolomics profiling of blood plasma mimics shows that medium- and long-chain fatty acids differently release metabolites from human serum albumin. *Journal of Magnetic Resonance*, *239*, 34–43. https://doi.org/10.1016/j.jmr.2013.11.019

22. Daykin, C. A., Bro, R., & Wulfert, F. (2012). Data handling for interactive metabolomics: Tools for studying the dynamics of metabolome-macromolecule interactions. *Metabolomics*, *8*, 52–63. https://doi.org/10.1007/s11306-011-0359-3

23. Vivó-Truyols, G., Ziari, M., Magusin, P. C. M. M., & Schoenmakers, P. J. (2009). Effect of initial estimates and constraints selection in multivariate curve resolution—Alternating least squares. Application to low-resolution NMR data. *Analytica Chimica Acta*, *641*(1–2), 37–45. https://doi.org/10.1016/j.aca.2009.03.020

24. Kriat, M., Confort-Gouny, S., Vion-Dury, J., Sciaky, M., Viout, P., & Cozzone, P. J. (1992). Quantitation of metabolites in human blood serum by proton magnetic resonance spectroscopy. A comparative study of the use of formate and TSP as concentration standards. *NMR in Biomedicine*, *5*(4), 179–184. https://doi.org/10.1002/nbm.1940050404

25. Würtz, P., Havulinna, A. S., Soininen, P., Tynkkynen, T., Prieto-Merino, D., Tillin, T., … Salomaa, V. (2015). Metabolite profiling and cardiovascular event risk: A prospective study of 3 population-based cohorts. *Circulation*, *131*(9), 774–785. https://doi.org/10.1161/CIRCULATIONAHA.114.013116

26. Suhre, K., Meisinger, C., Döring, A., Altmaier, E., Belcredi, P., Gieger, C., … Illig, T. (2010). Metabolic footprint of diabetes: A multiplatform metabolomics study in an epidemiological setting. *PLoS ONE*, *5*(11). https://doi.org/10.1371/journal.pone.0013953

27. Wang, T. J., Larson, M. G., Vasan, R. S., Cheng, S., Rhee, E. P., McCabe, E., … Gerszten, R. E. (2011). Metabolite profiles and the risk of developing diabetes. *Nature Medicine*, *17*(4), 448–53. https://doi.org/10.1038/nm.2307

28. Ghuman, J., Zunszain, P. A., Petitpas, I., Bhattacharya, A. A., Otagiri, M., & Curry, S. (2005). Structural basis of the drug-binding specificity of human serum albumin. *Journal of Molecular Biology*, *353*(1), 38–52. https://doi.org/10.1016/j.jmb.2005.07.075

29. Sugio, S., Kashima, A., Mochizuki, S., Noda, M., & Kobayashi, K. (1999). Crystal structure of human serum albumin at 2.5 A resolution. *Protein Engineering*, *12*(6), 439–46. https://doi.org/10.1093/PROTEIN/12.6.439

30. Psychogios, N., Hau, D. D., Peng, J., Guo, A. C., Mandal, R., Bouatra, S., … Wishart, D. S. (2011). The Human Serum Metabolome. *PLoS ONE*, *6*(2), e16957. https://doi.org/10.1371/journal.pone.0016957

31. Nicholson, J. K., Foxall, P. J. D., Spraul, M., Farrant, R. D., & Lindon, J. C. (1995). 750 MHz 1H and 1H-13C NMR Spectroscopy of Human Blood Plasma. *Analytical Chemistry*, *67*(5), 793–811. https://doi.org/10.1021/ac00101a004

32. Wider, G., & Dreier, L. (2006). Measuring protein concentrations by NMR spectroscopy. *Journal of the American Chemical Society*, *128*(8), 2571–2576. https://doi.org/10.1021/ja055336t

33. de Juan, A., Jaumot, J., & Tauler, R. (2014). Multivariate Curve Resolution (MCR). Solving the mixture analysis problem. *Analytical Methods*, *6*(14), 4964. https://doi.org/10.1039/c4ay00571f

34. Nilsson, M. (2009). The DOSY Toolbox: A new tool for processing PFG NMR diffusion data. *Journal of Magnetic Resonance*, *200*(2), 296–302. https://doi.org/10.1016/j.jmr.2009.07.022

35. Soininen, P., Haarala, J., Vepsäläinen, J., Niemitz, M., & Laatikainen, R. (2005). Strategies for organic impurity quantification by 1H NMR spectroscopy: Constrained total-line-shape fitting. *Analytica Chimica Acta*, *542*(2), 178–185. https://doi.org/10.1016/j.aca.2005.03.060

36. Rosenson, R. S., McCormick, A., & Uretz, E. F. (1996). Distribution of blood viscosity values and biochemical correlates in healthy adults. *Clinical Chemistry*, *42*(8), 1189 LP-1195.

37. Spano, P. F., Szyszka, K., Galli, C. L., & Ricci, A. (1974). Effect of clofibrate on free and total tryptophan in serum and brain tryptophan metabolism. *Pharmacological Research Communications*, *6*(2), 163–173. https://doi.org/10.1016/S0031-6989(74)80024-X

38. Smith, H. G., & Lakatos, C. (1971). Effects of acetylsalicylic acid on serum protein binding and metabolism of tryptophan in man. *Journal of Pharmacy and Pharmacology*, *23*(3), 180–189. https://doi.org/10.1111/j.2042-7158.1971.tb08639.x

39. Iwata, H., Okamoto, H., & Koh, S. (1975). Effects of Various Drugs on Serum Free and Total Tryptophan Levels and Brain Tryptophan Metabolism in Rats. *The Japanese Journal of Pharmacology*, *25*(3), 303–310. https://doi.org/10.1254/jjp.25.303

40. Spector, A. A. (1975). Fatty acid binding to plasma albumin. *Journal of Lipid Research*, *16*(3), 165–79.

41. Ashbrook, J. D., Spector, A. A., Santos, E. C., & Fletcher, J. E. (1975). Long chain fatty acid binding to human plasma albumin. *The Journal of Biological Chemistry*, *250*(6), 2333–8.

42. Barding, G. A., Salditos, R., & Larive, C. K. (2012). Quantitative NMR for bioanalysis and metabolomics. *Analytical and Bioanalytical Chemistry*. https://doi.org/10.1007/s00216-012-6188-z

43. Bharti, S. K., Sinha, N., Joshi, B. S., Mandal, S. K., Roy, R., & Khetrapal, C. L. (2008). Improved quantification from 1H-NMR spectra using reduced repetition times. *Metabolomics*, *4*(4), 367–376. https://doi.org/10.1007/s11306-008-0130-6

CHAPTER 4

LipSpin: a New Bioinformatics Tool for Quantitative ^{1}H-NMR Lipid Profiling

UNIVERSITAT ROVIRA I VIRGILI
DEVELOPMENT OF 1H-NMR SERUM PROFILING METHODS FOR HIGH-THROUGHPUT METABOLOMICS
Rubén Barrilero Regadera

4.1. Abstract

The structural similarity among lipid species, and the low sensitivity and spectral resolution of nuclear magnetic resonance (NMR) have traditionally hampered the routine use of ^{1}H-NMR lipid profiling, which remains mostly manual and lacks free bioinformatic tools. However, whereas other analytical platforms in metabolomics require multiple calibration steps, NMR is the only purely quantitative technique. ^{1}H-NMR lipid profiling provides fast screening of major lipid classes (fatty acids, glycerolipids, phospholipids and sterols) and some individual species, and has been used in numerous clinical and nutritional studies, leading to improved risk prediction models. In this article, we present LipSpin, a free and open-source bioinformatics tool for quantitative ^{1}H-NMR lipid profiling, based on constrained lineshape fitting analysis of lipid signals with voigt profiles and standard-based models. When given the optimal experimental conditions, LipSpin allows the characterization of severely overlapped spectral regions and complex coupling patterns. LipSpin provides the most detailed quantification of fatty acid families and choline phospholipids to date. Moreover, analytical and clinical results with LipSpin quantifications conform with other techniques commonly used for lipid analysis.

4.2. Introduction

Lipids play an important role in multiple cellular functions, including: membrane composition and anchoring, protein trafficking, signalling, and energy reservoirs [1]. The vast number of different species [2] and their influence in homeostatic processes and disease states have motivated the advent of lipidomics, a branch of metabolomics focused on the large-scale analysis of lipids in biological systems [3]. Lipid profiling provides a powerful means to monitor and understand lipid imbalance in pathophysiological conditions such as inflammatory disorders, metabolic syndrome, diabetes, cardiovascular diseases, neurodegenerative diseases and cancer, among others [4], leading to improved risk prediction models [5]. Similarly, lipid profiling has been applied to assess the health benefits of diets and nutritional supplements [5,6], and the effects of drug therapies in clinical trials [5]. Moreover, lipid profiling has become an important tool in food technology for the determination of nutritional and technological properties of foodstuff [4,6].

From the analytical perspective, techniques based on chromatography and mass spectrometry (MS) are the most widespread in lipidomics [7], as they provide a comprehensive characterization of all the constituent species based on their different physico-chemical properties. Contrary to MS,

the detailed characterization of lipid species by proton nuclear magnetic resonance spectroscopy (^{1}H-NMR) is unfeasible, as magnetically-equivalent molecular structures of lipids give largely overlapped resonances [6]. However, ^{1}H-NMR spectra of lipids from most biological matrices provide a fast overview of major lipid classes (fatty acids, glycerolipids, phospholipids and sterols) and some individual species [8]. Additionally, ^{1}H-NMR has some interesting features for high-throughput lipid profiling and large-scale metabolomics studies: no derivatization or compound separation is required, the spectral area is equivalent to the molecular abundance, and its spectral linearity avoids the use of multiple internal standards. In other words, ^{1}H-NMR is fully quantitative and requires minimal sample preparation. Another advantage is that NMR is non-destructive and intact lipids can be stored for further analysis. Complementary, ^{31}P-NMR spectroscopy is another technique commonly applied when further characterization of phospholipid species is sought [9].

The classical strategy to extract biochemical information from ^{1}H-NMR lipid spectra is fingerprinting analysis. ^{1}H-NMR fingerprinting usually implies a data reduction by spectral binning and the use of multivariate techniques, such as principal component analysis (PCA) or self-organizing maps (SOM), in order to reveal the underlying lipid patterns. This exploratory approach is partially valid as most of the signals from non-polar structures are aligned between samples, and has been largely applied to lipophilic extracts of serum, lipoproteins and tissues from human and animal models [10–13]. However, ^{1}H-NMR fingerprinting is subjected to unwanted variances from misalignments of polar signals and baseline distortions, and signal overlaps might obscure valuable information. More robust quantitative strategies have also been applied to ^{1}H-NMR spectra of lipids. These include: calibration curves [14], a combination of spectral subtraction and least-squares solution of linear systems with reference models [15], bucket integration [16], and lineshape fitting analysis based on Gaussian/Lorentzian models [17–19]. To our knowledge, there has been only one attempt to systematically apply lineshape fitting analysis in ^{1}H-NMR spectra of lipophilic extracts [17]. This solution consists of a constrained lineshape fitting strategy developed on PERCH NMR software and not publicly released. The solution has been implemented as a part of the high-throughput NMR workflow of serum in a metabolomics platform and run over numerous studies, revealing new biomarkers for early atherosclerosis, type 2 diabetes mellitus, diabetic nephropathy, coronary heart disease, and all-cause mortality [20].

In this article, we present LipSpin, a new freely-distributed and open-source software for semiautomatic profiling of ^{1}H-NMR spectra of lipid extracts. LipSpin integrates all the necessary steps to convert raw NMR data into quantitative information of lipid composition of a collection of

samples, without the need for additional software. Using a collection of signal patterns based on mathematical and reference spectral models, a constrained lineshape fitting analysis provides the quantification of 15 different lipid signals among major lipid classes (fatty acids, triglycerides, phospholipids and cholesterols). LipSpin has been optimized for serum and plasma, and validated in standard mixtures and lipophilic extracts of human plasma samples. Additionally, quantifications with LipSpin have been applied to a dietary intervention study.

4.3. Experimental section

4.3.1. Preparation of lipid mixtures

Lipid mixtures were designed to evaluate LipSpin quantifications in a set of calibrated samples and within a broad range of concentrations. We prepared ten different mixtures of varying concentrations of five standard lipids, namely cholesterol (CAS: 57-88-5), cholesteryl linoleate (CAS: 604-33-1), glyceryl trioleate (CAS: 122-32-7), phosphatidylcholine (18:0/18:0) (CAS: 816-94-4), and phosphatidylethanolamine (from bovine liver, CAS: 383907-31-1). Neutral lipids and phospholipids were purchased from Sigma-Aldrich (Steinheim, Germany) and Avanti Polar Lipids (Alabaster, AL), respectively. These standard lipids provide a representation of major lipid classes (cholesterols, triglycerides and phospholipids) and aliphatic structures in lipophilic extracts of biological samples. The composition of mixtures is detailed in Table C4.1. Standard stock solutions and mixtures were prepared in a solution of $CDCl_3$:CD_3OD:D_2O (16:7:1, v/v/v), immediately before NMR analysis.

Table C4.1 Composition of lipid mixtures

Compound	Concentration (mmol/l)									
	Mix1	Mix2	Mix3	Mix4	Mix5	Mix6	Mix7	Mix8	Mix9	Mix10
Cholesterol	0.57	0.70	0.39	0.87	0.30	0.52	0.65	0.70	0.48	0.35
Cholesteryl linoleate	0.37	0.15	0.44	0.17	0.49	0.29	0.42	0.44	0.22	0.56
Phosphatidylcholine (18:0/18:0)	0.42	0.64	0.30	0.58	0.61	0.19	0.55	0.25	0.22	0.17
Glyceryl trioleate	0.27	0.42	0.54	0.32	0.22	0.59	0.17	0.25	0.47	0.29
Phosphatidylethanolamine (bovine liver)	0.38	0.19	0.24	0.21	0.28	0.35	0.26	0.40	0.54	0.49

4.3.2. Preparation of plasma lipid extractions

Additional analytical and clinical validation were based on two sets of plasma samples: the first set consisted of 15 plasma samples from healthy adult volunteers in fasting state, recruited at Sant

Joan University Hospital (Reus, Spain). Venous blood was withdrawn into EDTA tubes and centrifuged immediately for 15 min at 4 °C and 1500g to obtain plasma. Total plasma cholesterol, triglycerides and phospholipids were determined using enzymatic and colorimetric/fluorimetric assays adapted to a COBAS 6000 autoanalyzer (Roche Diagnostics, Rotkreuz, Switzerland). The second set consisted of plasma samples from 26 healthy adults who participated in a dietary intervention as previously described [21]. Briefly, volunteers were randomised to receive a 6-week dietary intervention with either saturated fatty acids (SFA) or omega-6 polyunsaturated fatty acids (n-6PUFA), with all participants supplemented with 4x1g fish oil capsules (rich in omega-3 fats). Plasma samples were obtained at baseline and following intervention, resulting in a total of 52 plasma samples that were kept at -80°C in separate aliquots for ^{1}H-NMR and gas chromatography analyses. The fatty acid composition was determined using gas chromatography coupled with a flame ionization detector as described in [21], and total plasma cholesterol, phospholipids and triglycerides were determined using enzymatic colorimetric/fluorimetric methods.

Lipids were obtained from 100 μL of freshly thawed plasma aliquots using the BUME extraction method [22]. BUME was optimised for batch extractions with diisopropyl ether (DIPE) replacing heptane as the organic solvent, since the ^{1}H-NMR fingerprint of heptane highly overlaps fatty acid signals. After the extraction procedure, the lipophilic phase was completely dried in N_2 stream until evaporation of organic solvents and frozen at -80 °C until NMR analysis.

4.3.3. NMR sample preparation and data acquisition

Dried lipid extracts were reconstituted in a solution of $CDCl_3$:CD_3OD:D_2O (16:7:1, v/v/v) containing tetramethylsilane (TMS) at 2 mM as a chemical shift reference, and transferred into 5-mm NMR glass tubes. ^{1}H-NMR spectra were measured at 600.20 MHz using an Avance III-600 Bruker spectrometer equipped with a 5 mm CPTCI triple resonance pulse field gradient cryoprobe. A 90° pulse with water presaturation sequence (zgpr) was used. We performed measurements at 286 K, which shifts the residual water signal to the non-informative region at around 4.51 ppm. The relaxation delay between scans was set to 5 s to avoid most of the attenuation due to longitudinal relaxation [18]. During this time, water signal was irradiated with a low-power RF pulse. The acquisition time was 3 s and the free induction decays (FIDs) consisted of 64 k complex data points, leading to a spectral width of 18.6 ppm. For each spectrum, 128 scans were recorded resulting in a total acquisition time per sample of 17 min. After the acquisition, the FIDs were zero-filled to 128 k real data points and apodized by an exponential window function with 0.3 Hz line broadening prior to Fourier transformation.

4.3.4. ^{1}H-NMR lipid profiling and quantification

Quantification of lipid signals in ^{1}H-NMR spectra was carried out with LipSpin, an in-house software based on Matlab (ver. 7.5.0. The Mathworks, Inc., Natick, MA, USA) (see Results section for details). Briefly, samples were imported as FIDs and Fourier-transformed after zero-filling to increase peak resolution. Then, spectra were phase-corrected by flattening void spectral regions (regions: 9 to 7.8; 6.8 to 5.6; 0.5 to 0.2 and -0.2 to -1 ppm), baseline-removed using cubic Hermite interpolation with automatic detection of baseline points, and shift-referenced to TMS signal at 0 ppm. Finally, signals assigned to lipids in Fig. C4.1 were quantified with lineshape fitting analysis. After the quantification process, signal areas (in arbitrary units) can be converted into molar concentrations, for example, by using an internal standard of known concentration [19,23], a calibrated synthetic signal introduced in the spectrum [23] or normalising by external measurements. The last option is suggested, as it corrects the variability of recovery volumes in manual extractions and avoids errors caused by the high volatility of common internal standards, such as TMS [23]. In this study, signal areas were converted to mM before statistical analysis by using total cholesterol concentrations determined with other methods and applying the following equation:

$$M_x = A_x * \frac{M_{chol}}{A_{chol}}$$

Where A_{chol} and A_x refer to the ^{1}H-NMR areas of total cholesterol and signal x, respectively, and M_{chol} and M_x refer to the molar concentration of the externally-measured total cholesterol and ^{1}H-NMR signal x.

4.3.5. Statistical analysis

Analytical validation of ^{1}H-NMR lipid quantifications was performed by linear regression and Pearson's (r) correlation with analogous measurements obtained with other methods. Clinical validations were carried out using the dietary intervention study, comparing lipid concentrations at baseline and 6 weeks with paired t-test or Wilcoxon signed-rank test, for parametric and non-parametric data distributions, respectively. Changes with a p-value<0.05 were considered statistically significant. Statistical analyses were performed using Matlab (ver. 7.5.0. The Mathworks, Inc., Natick, MA, USA).

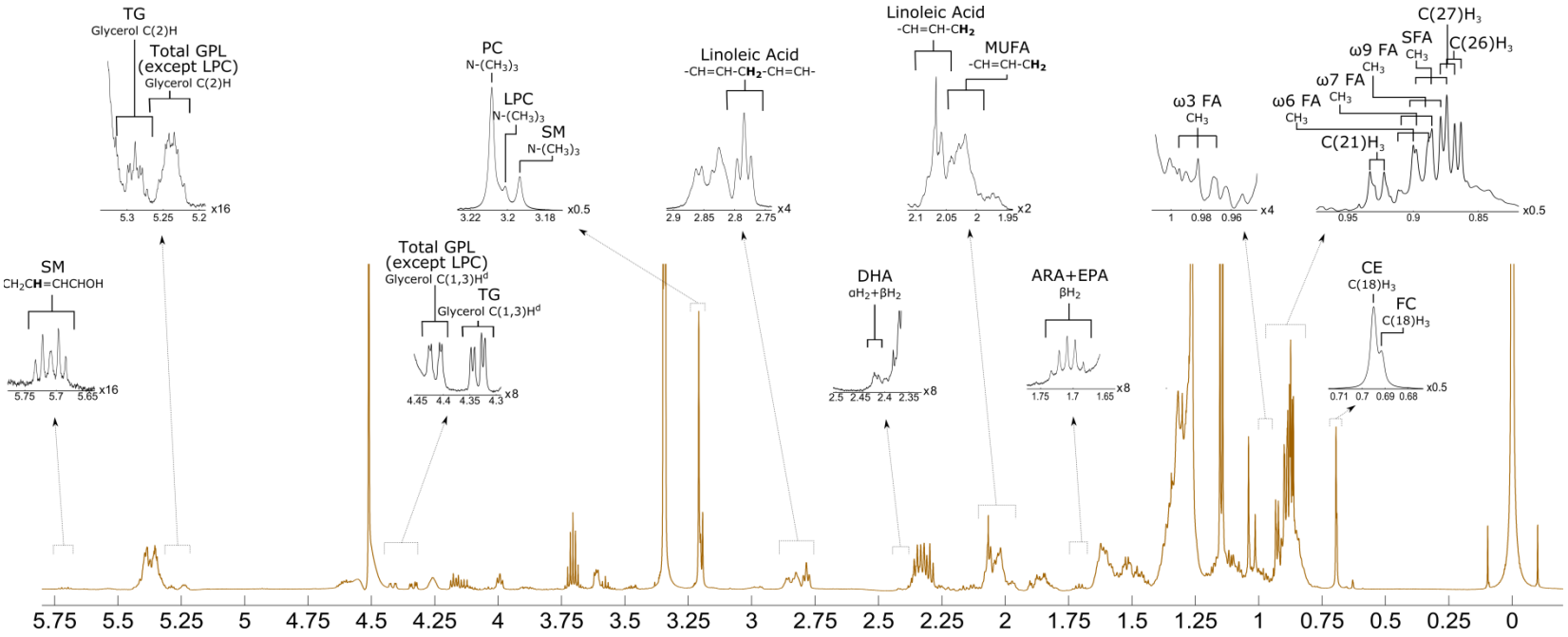


Fig. C4.1 ^{1}H-NMR spectrum of lipophilic extract of plasma with labelled signals used for quantification with LipSpin in the nutritional study. Other regions not included in the analysis but typically used in lipid analysis to estimate FA chain length and number of insaturations are (Kriat et. Al, 1993): methylene protons at 1.27 ppm, diallylic protons at 2.84 ppm (other PUFA than linoleic acid) and olefinic protons at 5.4 ppm (double bond

4.4. Results

4.4.1. Lipspin: a computational workflow for ^{1}H-NMR quantification of lipids

LipSpin is a graphical user interface (GUI) software that allows the quantitative profiling of ^{1}H-NMR spectra of lipid extracts for metabolomics assays in a user-friendly manner. The computational workflow covers all the necessary steps for preparing and analysing a batch of samples in a semiautomatic mode (Fig. C4.2), requiring minimal user intervention and no programming skills. Lipid quantifications relie on lineshape fitting analysis of spectral regions, from which individual signal areas are obtained. LipSpin has been written in Matlab (ver. 7.5.0. The Mathworks, Inc., Natick, MA, USA). The program source code and user manual can be freely downloaded from https://github.com/rbarri/LipSpin. A standalone version of LipSpin is provided on demand by contacting the corresponding author. The current release is provided with a set of signal patterns specifically optimized for blood serum lipids (Table C4.2), however, it can be easily adapted for lipophilic extracts of animal and plant tissues, cells or other biofluids. Hereinafter, a brief explanation of each of the main modules of LipSpin is given.

Data import. LipSpin imports data samples from either, raw FIDs or 1D spectra from Topspin or other third-party software exported in Bruker format. When FIDs are the input mode, LipSpin importer requires the root directory containing all the samples and the experiment number, following the Bruker folder structure. Using this import option, specific zero filling and window apodization can be applied prior to Fourier transformation. When 1D spectra are the input mode, a processing number must be provided (following Bruker nomenclature). After this step, users are required to select the samples to be loaded from the list of available samples in the indicated directory.

Spectral pre-processing. Once the samples are loaded and the spectra are displayed in the main screen (Fig. C4.3a), some spectral corrections may be needed derived from spectral misalignments, baseline and phase distortions. LipSpin integrates a set of routines that automatically correct by all this factors prior to lineshape fitting analysis: phase correction, baseline correction, chemical shift referencing, spectral alignment and reference deconvolution.

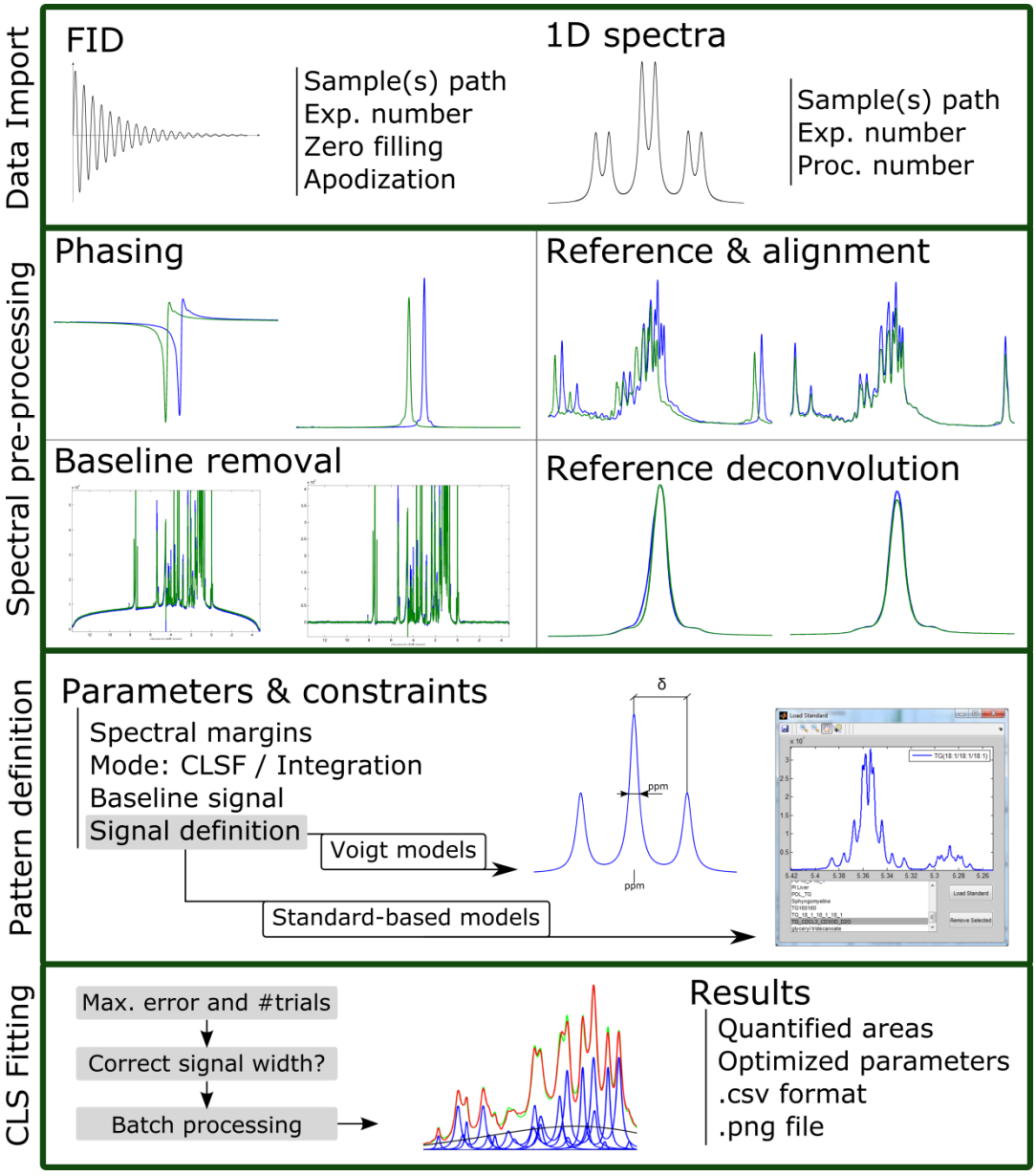


Fig. C4.2 Spectral analysis workflow in LipSpin

Phase correction should be applied before any other pre-processing step and it is an essential requirement for proper performance of the lineshape fitting algorithm. It sets the spectral line in pure absorptive mode. LipSpin provides two different methods to correct zero- and first-order phase. The first method seeks maximising the entropy of the spectrum [24] whereas the second is based on a least-squares problem; it minimises the residuals between a horizontal line (i.e. a flat

baseline) and a set of user-defined spectral regions, in which no signals are expected. If the spectral baseline presents drifts or rolls, the baseline correction tool removes these artifacts by interpolating functions (cubic spline, cubic hermite or polynomials) to a set of user- or automatically-defined points expected to lie in the x-axis. Then, spectra can be referenced to an internal peak and aligned between them, by maximising their correlation with a mean reference spectrum. Finally, signal shape distortions, such as those produced by magnetic field inhomogeneities, could be reduced by reference deconvolution [25] using a synthetic or best-shape peak (among all the spectra) as a reference.

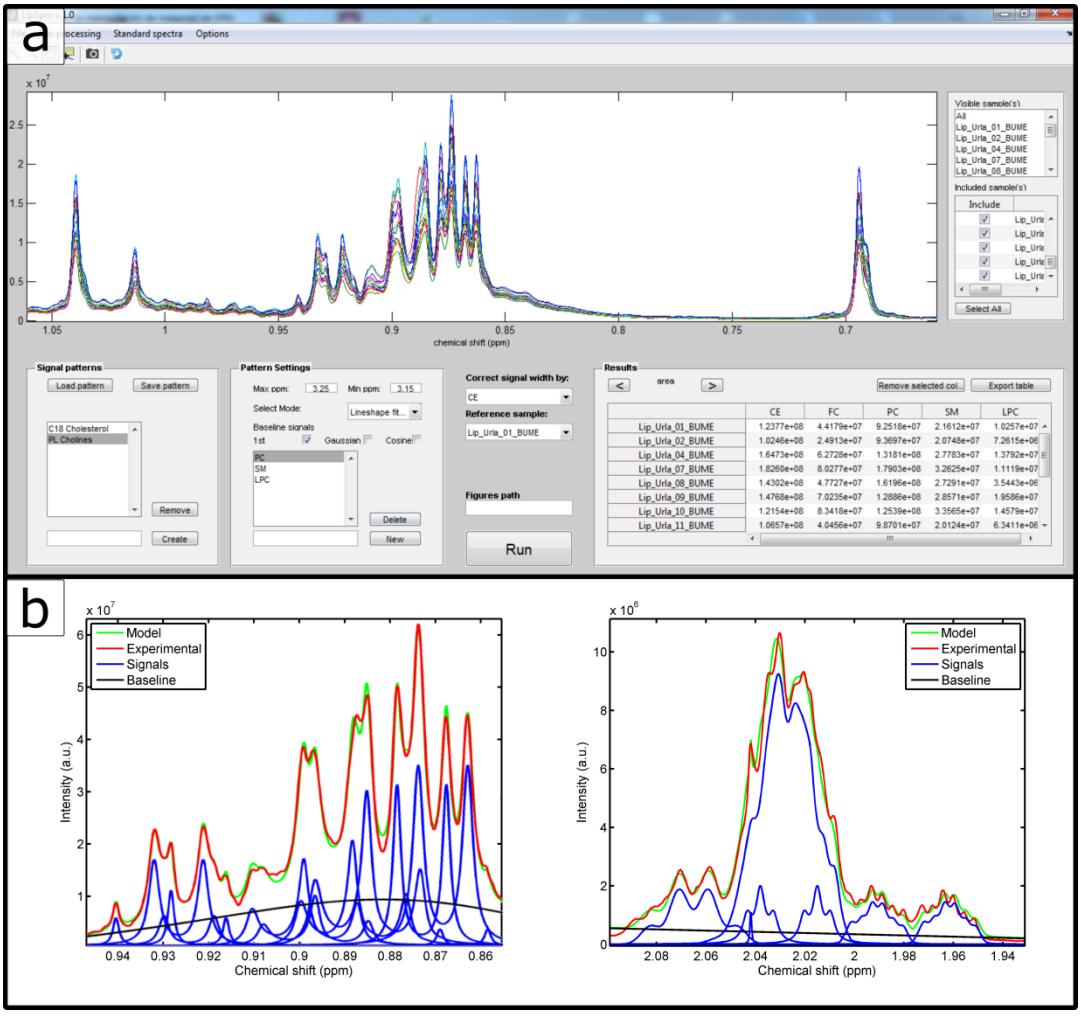


Fig. C4.3 (a) LipSpin main screen. (b) examples of graphical outcome of the fitting procedure for CH3 methyl region at 0.9 ppm (left) and allylic hydrogens from unsaturated fatty acids at 2.03 ppm (right)

Table C4.2 Spectral regions and signals included in the signal patterns

Signal pattern (ppm)	Signal assignment[a]	Model	Protons	Chemical shift (ppm)	J-Coupling (Hz)
C18 cholesterol (0.73 - 0.65)	**FC**	**voigt**	**3**	**0.694**	**s**
	CE	**voigt**	**3**	**0.691**	**s**
FA Methyl[b] (0.95 - 0.85)	chol C26	voigt	3	0.868	d (6.6)
	chol C27	voigt	3	0.873	d (6.6)
	SFA	**voigt**	**3**	**0.886**	**t (6.9)**
	ω9	**voigt**	**3**	**0.888**	**t (6.9)**
	ω7	**voigt**	**3**	**0.898**	**t (6.9)**
	ω6	**voigt**	**3**	**0.900**	**t (6.9)**
	chol C21	voigt	3	0.922	d (6.6)
FA ω3 Methyl (1.01 - 0.95)	**ω3**	**voigt**	**3**	**0.982**	**t (7.2)**
	FC	standard	-	0.982	m
	CE	standard	-	0.982	m
FA ARA+EPA β-methylene (1.75 - 1.66)	**EPA**	**standard**	**2**	**1.697**	**m**
	ARA	**standard**	**2**	**1.699**	**m**
FA Allylic[b] (2.10 - 1.93)	chol	standard	0.5	1.973	m
	chol	standard	0.5	2.001	m
	chol	standard	0.5	2.018	m
	MUFA	**standard**	**4**	**2.024**	**m**
	chol	standard	0.5	2.040	m
	linoleic	voigt	4	2.063	qua (7.05)
FA DHA α-methylene (2.45 - 2.40)	**DHA**	**standard**	**4**	**2.420**	**m**
FA Diallylic (2.94 - 2.70)	**linoleic**	**voigt**	**2**	**2.783**	**t (7.0)**
	PUFA	voigt	-	2.817	t (7.0)
	PUFA	voigt	-	2.827	t (7.0)
	PUFA	voigt	-	2.854	t (7.0)
	PUFA	voigt	-	2.862	t (7.0)
PE Alkyl (3.18 - 3.12)	**PE**[d]	**standard**	**2**	**3.152**	**m**
PL Cholines (3.25 - 3.18)	**SM**	**voigt**	**9**	**3.192**	**s**
	LPC	**voigt**	**9**	**3.201**	**s**
	PC	**voigt**	**9**	**3.208**	**s**
Glycerol backbone sn-1,3[c] (4.45 - 4.31)	**TG**	**standard**	**2**	**4.332**	**m**
	GPL (except LPC)	**standard**	**2**	**4.412**	**m**
Glycerol backbone sn-2 (5.33 - 5.15)	**TG**	**standard**	**1**	**5.237**	**m**
	GPL (except LPC)	**standard**	**1**	**5.288**	**m**
PL Olefinic (6 - 5.65)	**SM**	**standard**	**1**	**5.708**	**m**
	PLA[d]	**standard**	**1**	**5.922**	**m**

a. Signals used for lipid quantifications are given in bold.
b. BUME extraction could leave solvent residues: butanol triplet (7.35) at 0.929 ppm, diisopropyl ether (DIPE) doublet (6.7) at 0.8647 ppm and ethyl acetate (EtAc) singlet at 2.066 ppm.
c. Water suppression used in our NMR experiments affected signal intensities of glycerol sn-1,3 and therefore signals from glycerol backbone sn-2 were used for quantification of TG and GPL
d. Rarely observed in 1H-NMR spectra of human serum and plasma samples

Key: FC, free cholesterol; EC, ester cholesterol; FA, fatty acids; SFA, saturated fatty acids; EPA, eicosapentaenoic acid; ARA, arachidonic acid; DHA, docosahexaenoic acid; MUFA, monounsaturated fatty acids; PUFA, polyunsaturated fatty acids; PE, phosphatidylethanolamine; SM, sphingomyelin; LPC, lysophosphatidylcholine; PC, phosphatidylcholine; TG, triglycerides; GPL, glycerophospholipids; PLA, plasmalogen; s, singlet; d, doublet; t, triplet; m, multiplet

Pattern definition. The quantification procedure requires the definition of a set of parameters for each spectral region subject to analysis. Required parameters are spectral margins, baseline signal (to compensate for residual baseline distortions or spurious broad signals) and model definitions of underlying signals. Individual signals can be defined using either Voigt profiles (combination of Lorentzian and Gaussian functions) or spectral templates from spectra of standard lipids, whereas initial values and fitting constraints must be provided for each signal property (chemical shift, linewidth, Gaussian ratio and J-coupling). Besides, users are requested to choose between applying the constrained lineshape fitting analysis or bucket integration, in which case, only spectral margins are considered. Bucket integration should be restricted to isolated signals without baseline artifacts to guarantee quantitative results.

The released version of LipSpin includes a collection of signal patterns including the analysed regions in this study (see Table C4.2 and Fig. C4.1 for detailed information about regions and signals). These signal patterns were created for a 600 MHz spectrometer at 286 K and $CDCl_3$:CD_3OD:D_2O (16:7:1, v/v/v) solvent on the basis of literature values [16,26], and using spectra of standard lipids acquired in our lab or downloaded from public databases [27,28]. It is worth noting that NMR spectra of lipids could be subjected to inter-laboratory differences, consequently, users are encouraged to adjust these patterns or create new ones to meet their specific requirements. With the aim of defining the spectral templates, LipSpin includes several spectra of standard lipids representing most common lipid structures in serum, including palmitic acid (CAS: 57-10-3), stearic acid (CAS: 57-11-4), oleic acid (CAS: 143-19-1), linoleic acid (CAS: 60-33-3), eicosapentaenoic acid (CAS: 10417-94-4), arachidonic acid (CAS: 506-32-1), docosahexaenoic acid (CAS: 6217-54-5), phosphatidylcholine (CAS: 63-89-8), lyso-phosphatidylcholine (CAS: 17364-16-8), sphingomyelin (CAS: 383907-87-7), cholesterol (CAS: 57-88-5), cholesteryl linoleate (CAS: 604-33-1) and triglycerides (CAS: 122-32-7), among others.

CLS fitting. Once the spectra are prepared and the signal patterns are properly defined, the quantification procedure, consisting in a constrained lineshape fitting algorithm [29], is applied for every defined region in all the included samples. If bucket integration was chosen for a region, the sum of the points within the margins is computed instead. The option “Correct signal width by” allows compensating for linewidth variations within samples due to differences in shimming and viscosity, taking linewidth variations from a previously quantified signal as a reference (normally a well-resolved singlet or solvent signal). If a valid “Figures path” is provided, a .png file will be created for every quantified region and sample, so that users can visually inspect the algorithm performance. Finally, the result chart displays the quantified signal areas, which can be exported in

.csv format for further analysis with spreadsheet programs or statistical tools. Additional parameters from the fitting procedure are also provided in this chart so that users can refine the pattern settings if additional runs are required.

4.4.2. Analytical validation with lipid mixtures

Initially, we analysed ten mixtures of five standard lipids at different concentrations (Table C4.1), representing most of the common signals in ^{1}H-NMR spectra of lipophilic extracts of biological samples.

Fig. C4.4 shows the ^{1}H-NMR spectra of individual standards, highlighting the signals modelled for lineshape fitting analysis. Contribution of these signals to the total spectral area of each standard (in number of protons) is detailed in Table C4.3. Fig. C4.3b represents the graphical solution of applying lineshape fitting analysis to methyl and allylic FA regions at 0.89 and 2.02 ppm, respectively, for one of the mixtures. The left figure exemplifies a region with multiple overlapping resonances from methyl groups of fatty acids, which are grouped and spatially distributed according to the position of the first double bond. Since these signals follow the multiplicity rule of NMR, they were modelled as triplets of voigt profiles. Some doublets from cholesterol methyl groups are also observed. In the right figure, the complexity of the couplings hampers a mathematical definition and each signal was modelled using spectral templates from standard lipids.

Table C4.3 Contribution of assignable protons to the total spectral area (in number of protons) of each individual standard used in lipid mixtures

Compound (Supplier ref.)	Contribution to total spectral area (number of protons)												
	C18 (FC)	C18 (CE)	SFA CH3	ω9 FA CH3	ω6 FA CH3	ω3 FA CH3	ARA+ EPA	MUFA	Linoleic Acid	PE	PC	TG	Total GPL
Cholesterol (C8667)	3.0												
Cholesteryl linoleate (C0289)		3.0	0.6		2.7				2.0				
Phosphatidylcholine 18:0/18:0 (850365P)			6.0								9.0		0.9
Glyceryl trioleate (T7140)				9.1				11.0				1.0	
Phosphatidylethanolamine from bovine liver (840026P)			3.9		2.4	0.6	1.6	0.6	0.4	2.0			0.8

Key: FC, free cholesterol; EC, ester cholesterol; SFA, saturated fatty acids; EPA, eicosapentaenoic acid; ARA, arachidonic acid; MUFA, monounsaturated fatty acids; PE, phosphatidylethanolamine; PC, phosphatidylcholine; TG, triglycerides; GPL, glycerophospholipids

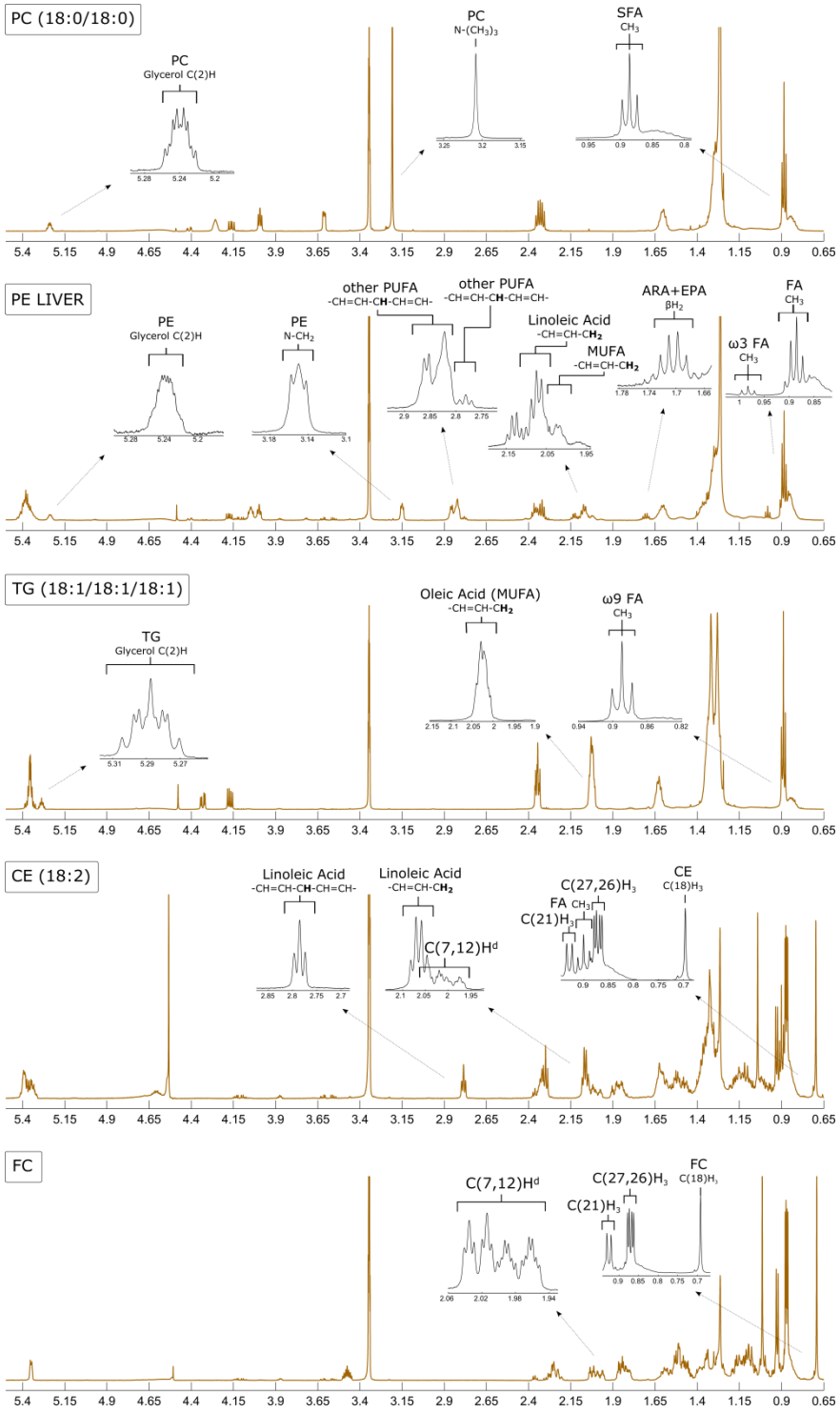


Fig. C4.4 ^{1}H-NMR spectra of standard lipids used in lipid mixtures with detail of regions used for quantification of individual lipids with LipSpin

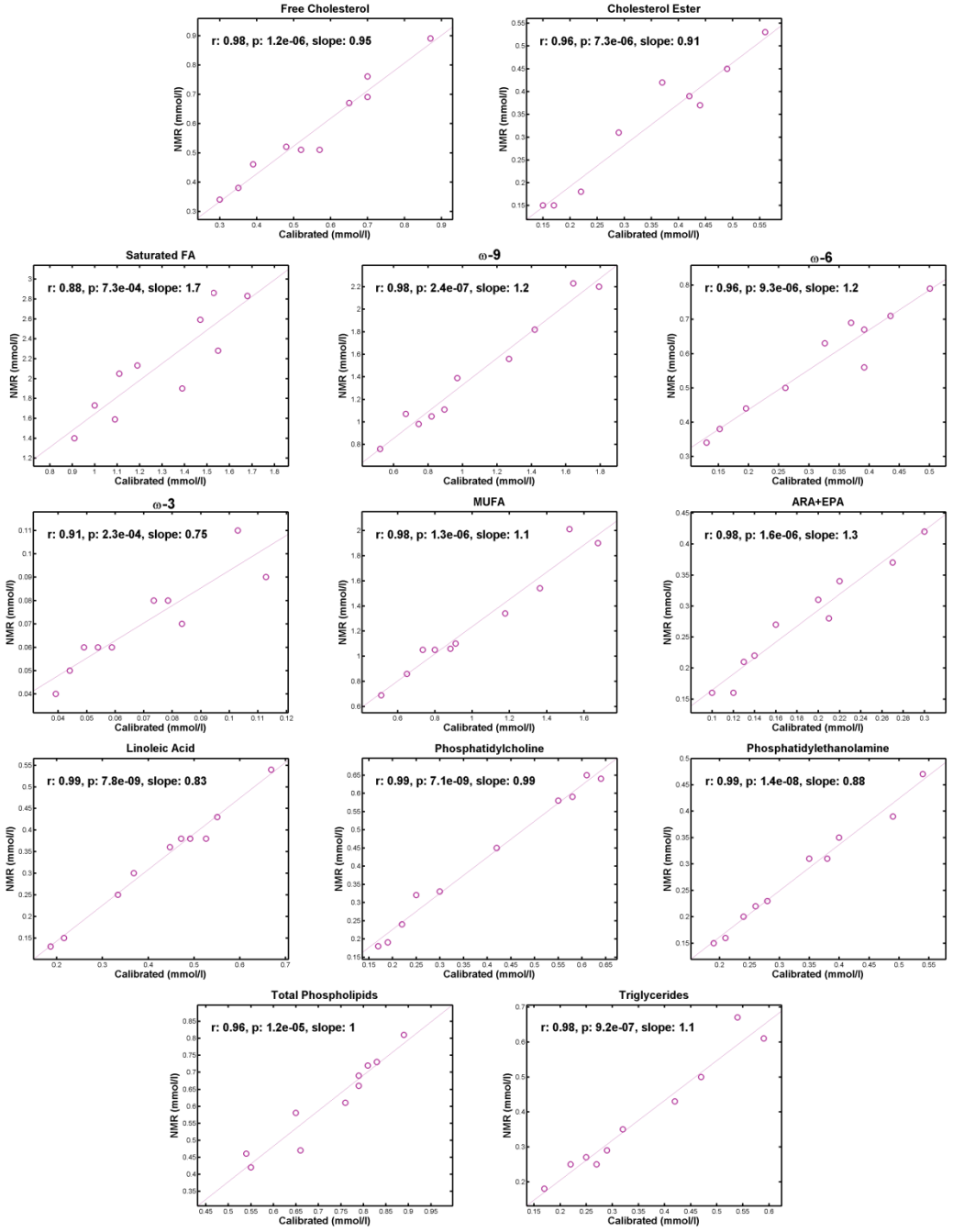


Fig. C4.5 Scatter plot and linear regressions of NMR quantification and lipid concentrations in table S1 for lipid mixtures

Table C4.4 lists the correlations between the quantified ^{1}H-NMR areas with LipSpin and the concentrations of the standard lipids listed in Table C4.1. The corresponding scatter plots and linear regressions are shown in Fig. C4.5. The excellent correlations ($r \geq 0.88$, $p < 1e\text{-}3$) and the linearity of the regressions (with slopes that approximate 1) demonstrate that LipSpin successfully extracts and quantifies each individual signal, preserving the quantitative nature of NMR spectroscopy. Only SFA quantification deviates from the expected values, with a slope of 1.8. This deviation could be attributed to sample contamination. We observed that the use of chloroform with plastic tubes generates residues from degradation that include visible signals at methyl and methylene regions.

Table C4.4 Pearson's r correlations and regression slopes of NMR quantifications and concentrations in Table C4.1 for lipid mixtures (n=10)

Lipid	Pearson's r	p-value	slope
Saturated FA	0.88	7.3e-04	1.8
ω-9 FA	0.98	2.4e-07	1.2
ω-6 FA	0.96	9.3e-06	1.2
ω-3 FA	0.91	2.3e-04	0.8
MUFA	0.98	1.3e-06	1.1
ARA+EPA	0.98	1.6e-06	1.3
Linoleic acid	0.99	7.8e-09	0.8
Free Cholesterol	0.98	1.2e-06	1.0
Cholesterol Ester	0.96	7.3e-06	0.9
Phosphatidylcholine	0.99	7.1e-09	1.0
Phosphatidylethanolamine	0.99	1.4e-08	0.9
Total Phospholipids	0.96	1.2e-05	1.0
Triglycerides	0.98	9.2e-07	1.1

Key: FA, fatty acids; MUFA, monounsaturated fatty acids; EPA, eicosapentaenoic acid; ARA, arachidonic acid

4.4.3. Analytical validation with plasma lipids

Further evaluation of ^{1}H-NMR quantifications with LipSpin was carried out using two sets of human plasma samples. For the first set comprising 15 samples of healthy adults, only enzymatically-measured lipids were available, i.e. total cholesterol, triglycerides and total phospholipids. These lipids are the primary constituents of lipoprotein structures and their distribution provides an overview of lipoprotein composition in blood [30]. First, ^{1}H-NMR cholesterol signals at 0.69 ppm from C18 protons of free and esterified forms were resolved with LipSpin, and added up to give total cholesterol area. Then, ^{1}H-NMR areas of total phospholipids

and triglycerides were obtained from C2 protons of glycerol backbone at 5.25 ppm. Correlation analysis in Fig. C4.6 showed an excellent agreement between techniques ($r \geq 0.89$, $p < 1e\text{-}5$), even though ^{1}H-NMR areas were not converted into molarity before comparison.

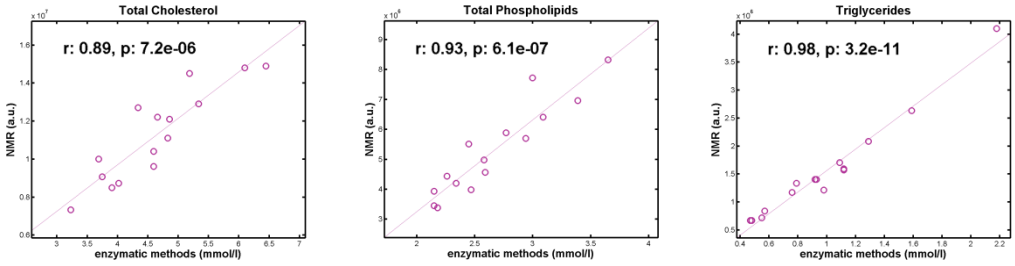


Fig. C4.6 Scatter plot and regression line of enzymatically-measured total cholesterol, phospholipids and triglycerides, and the same lipids using LipSpin quantifications in plasma of 15 healthy volunteers

For the second set, comprising 52 samples from a dietary intervention, fatty acid composition was available from GC-FID after methylation. ^{1}H-NMR is unable to provide information of individual fatty acid species due to their structural similarity. Instead, ^{1}H-NMR provides quantitative information on the main fatty acid families in lipid samples: the subtle separation between ω-6, ω-7, ω-9 and saturated (SFA) signals in the crowded methyl region at 0.89 ppm allows their individual quantification, ω-3 resonance appears downshifted to 0.98 ppm, and monounsaturated fatty acids (MUFA) can be resolved from its upshifted resonance in the allylic region at 2.02 ppm. In all cases, overlapping cholesterol residues need to be included in the signal patterns. Linoleic acid (18:2n-6) and docosahexaenoic acid (22:6n-3, DHA) can be uniquely determined from singular resonances at 2.78 and 2.42 ppm, respectively. Moreover, combined quantification of arachidonic and eicosapentaenoic acid (ARA+EPA) can be obtained from β-methylenes at 1.69 ppm. ^{1}H-NMR quantifications of cholesterol, triglycerides and phospholipids were done as previously mentioned. After the analysis with LipSpin, ω-6 and ω-7 areas were combined because of large ambiguity between these signals in the fitting process, whereas total fatty acids were computed as the sum of ω-3, ω-6, ω-7, ω-9 and SFA.

Table C4.5 shows the correlation between NMR quantifications (in molar concentration) and GC-FID lipids. Outlier detection was first applied to remove extreme and large residual samples to a final population no less than 44 samples in all cases. Most of the lipids correlated well ($r > 0.8$, $p < 0.01$), despite some discrepancies for SFA ($r = 0.62$, $p < 0.01$) and total phospholipids ($r = 0.68$, $p < 0.01$), attributed to contamination in lipid reconstitution, as previously mentioned. In general, NMR concentrations of fatty acids were larger than GC-FID due to discrepancies in quantitative

normalization. Nevertheless, the relative (mean) distribution of each fatty acid family agreed with typical values in human plasma for both techniques (data not shown) [31]. Scatter plots and linear regressions are available in Fig. C4.7.

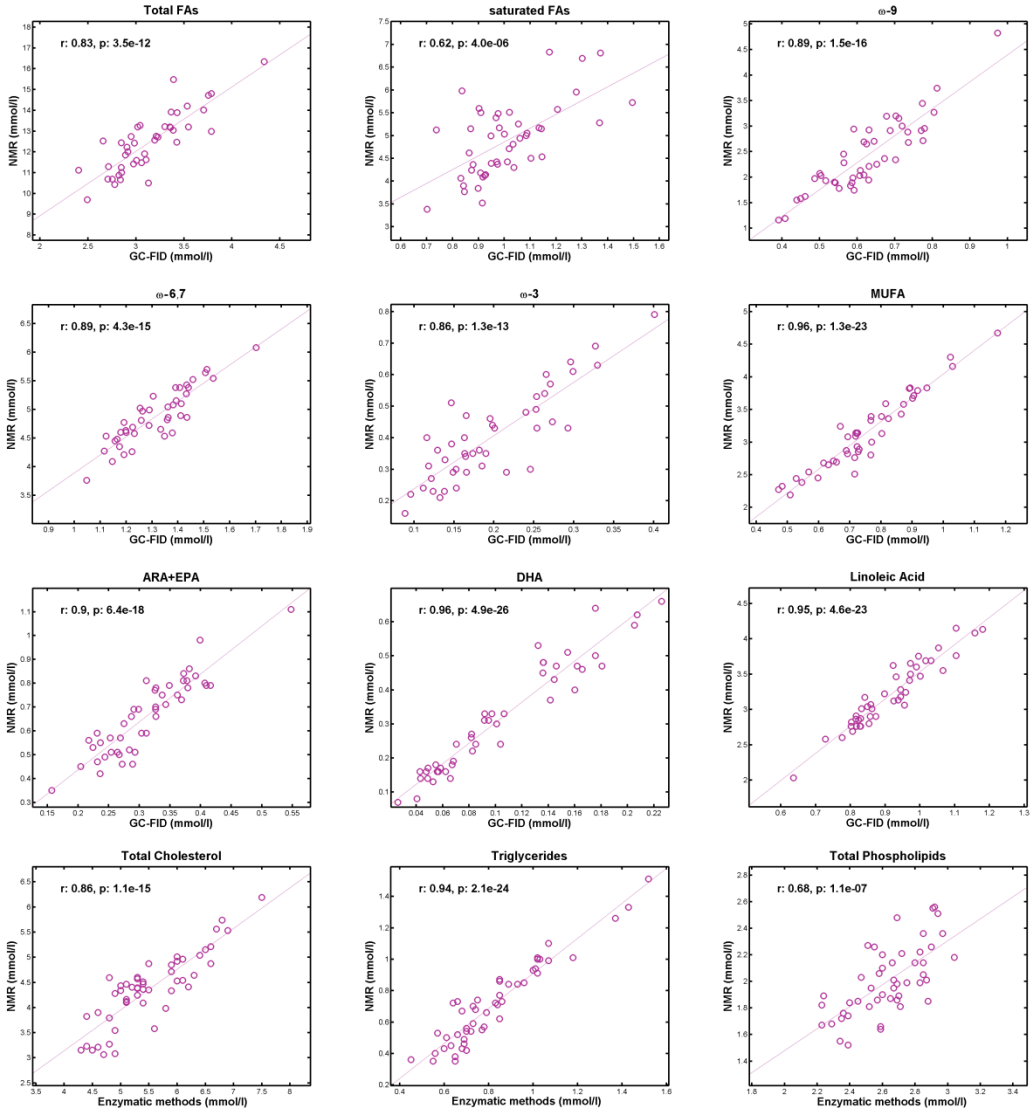


Fig. C4.7 Scatter plot and linear regressions of quantified lipids with LipSpin and the same lipids acquired with GC-FID and enzymatic methods from human plasma samples in the dietary intervention study

Table C4.5 Pearson's r correlations of NMR and GC-FID/enzymatic concentrations of lipids in the dietary intervention study (n≥44)

Lipid	Pearson's r	p-value
Total FA	0.83	3.5e-12
Saturated FA	0.62	4.0e-06
ω-9 FA	0.89	1.5e-16
ω-6 + ω-7 FA	0.89	4.3e-15
ω-3 FA	0.86	1.3e-13
MUFA	0.96	1.3e-23
ARA+EPA	0.90	6.4e-18
DHA	0.96	4.9e-26
Linoleic acid	0.95	4.6e-23
Total cholesterol	0.86	1.1e-15
Triglycerides	0.94	2.1e-24
Total phospholipids	0.68	1.1e-07

Key: FA, fatty acids; MUFA, monounsaturated fatty acids; EPA, eicosapentaenoic acid; ARA, arachidonic acid; DHA, docosahexaenoic acid

4.4.4. Application in a nutritional study

Since analytical validations based on linear regressions and correlations are subject to small sample deviations, we further evaluated the performance of NMR-quantified lipids in clinical samples. ^{1}H-NMR quantifications obtained from a dietary intervention study above were used to evaluate changes in plasma composition after consumption of two fatty acid-enriched diets, and compared with changes observed from GC-FID and enzymatic methods [21]. Fig. C4.8 shows the mean concentrations at baseline and post-intervention time points and their corresponding $\log_2$ fold-changes. Lipid modifications observed with ^{1}H-NMR agree with previous results from GC-FID and enzymatic methods, such as the significant increase in total cholesterol (TC) and eicosapentaenoic acid (EPA) after the diet enriched with SFA, and a similar change in docosahexaenoic acid (DHA) and triglycerides (TG) in both diets. Moreover, ^{1}H-NMR lipids confirm the lower incorporation of omega-3 fatty acids in the n-6PUFA diet, as expected from the reported metabolic competition between ω-3 and ω-6 [21].

4.5. Discussion

Metabolite profiling by ^{1}H-NMR has been largely investigated during the last decades after the first complete spectral assignments of biological matrices appeared. Most of these efforts have focused on the analysis of low-molecular-weight metabolites (LMWM), for which a variety of freely available bioinformatics tools have been recently developed [32–35]. Lipid profiling,

however, has been given less attention and, to date, there has been no freely available software solution. This preference for LMWM compounds relies on two principal reasons that favour high-throughput analysis: first, LMWM profiling can be performed directly on total serum and urine, reducing sample manipulation and inter-sample variability, and second, LMWM are structurally singular, which makes signal identification and modelling more straightforward.

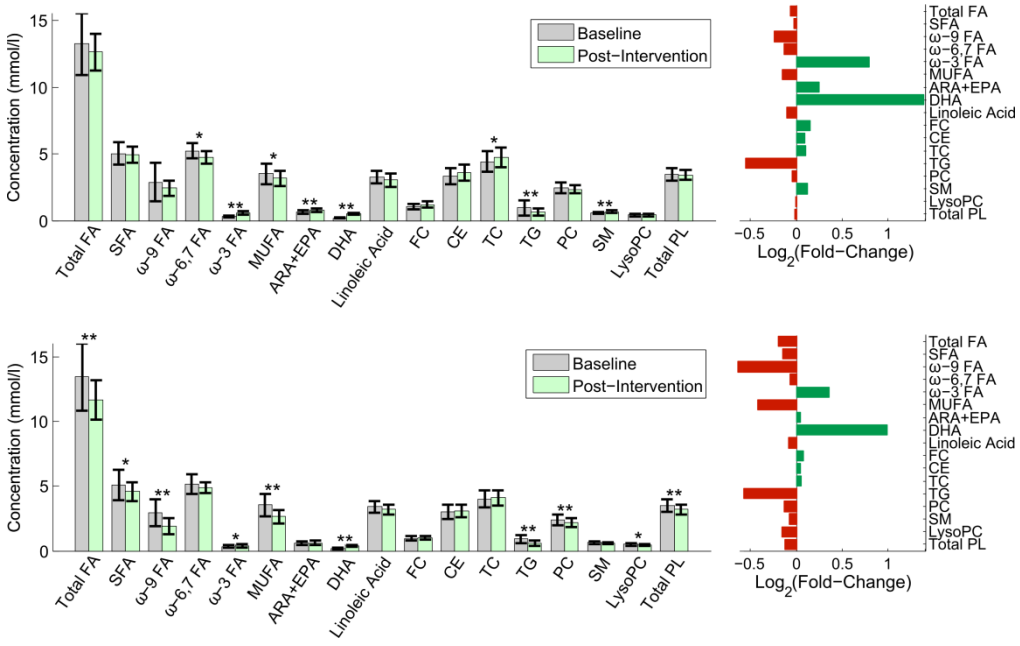


Fig. C4.8 Left vertical bar graphs: mean concentrations (± standard deviation) of NMR-quantified lipids before and after a dietary intervention enriched with SFA (top) and n-6PUFA (bottom), asterisks indicate significant changes between time points (* p-value < 0.05; ** p-value < 0.01). Right horizontal bar graphs: log_2 fold-change for each lipid in the SFA (top) and n-6PUFA (bottom) enriched diets. Fold-changes are calculated as the ratio between post-Intervention and baseline concentrations. Keywords: FA (fatty acids); SFA (saturated fatty acids); MUFA (monounsaturated fatty acids); ARA (arachidonic acid); EPA (eicosapentaenoic acid); FC (free cholesterol); CE (cholesterol ester); TC (total plasma cholesterol); TG (triglycerides); PC (phosphatidylcholine); SM (sphingomyelin); Total PL (total plasma phospholipids)

Concerning sample preparation, we used BUME, a method for lipid extraction of serum samples that can be automatized with liquid handling robots. Lipid recovery with BUME is comparable to Folch methods [36], but automatic extraction increases sample reproducibility and reduces contamination exposure. Besides, spontaneous phase separation in BUME avoids the long centrifugation times required with other methods. To our knowledge, this is the first time BUME

has been applied to NMR experiments. The excellent correlation between ^{1}H-NMR lipid areas and external concentrations in Fig. C4.6 reflects the low variability induced by BUME extractions. The main disadvantage is the long time required to completely evaporate the organic solvents; however, residual signals can be easily modelled and subtracted in the fitting process.

Regarding the ability to extract quantitative information, the success of lineshape fitting strategies relies on the discrimination of structurally-similar lipid compounds that, in turn, depends on the spectral dispersion achievable. It is well-known that spectral dispersion increases with spectrometer frequency. LipSpin has been optimized for a 600 MHz spectrometer and, from our experience, resolving highly overlapped regions becomes cumbersome with lower frequencies. The development of higher frequency spectrometers offers new perspectives on lineshape fitting of lipids. For instance, Soininen et al. [37] demonstrated that complete characterization of choline phospholipids at 3.2 ppm can be achieved for LDL fractions using a 800 MHz NMR, without the need of lipid extraction. Additionally, a proper optimization of sample solvents can also improve spectral dispersion. We have included a small volume of D_2O in lipid reconstitution. D_2O affects polar groups of phospholipids, maximises spectral separation between choline phospholipids at 3.2 ppm, and reduces the spectral overlap between glycerol backbone signals from phospholipids and glycerolipids at 5.25 ppm. Besides, D_2O has been previously used to avoid lipid aggregation [38].

LipSpin exploits the above advantages and, together with the constrained lineshape fitting analysis, increases the lipid coverage commonly obtained in previous studies of serum and plasma [12,17,19]. For instance, LipSpin allows a detailed characterization of saturated and unsaturated FA families from methyl peaks. Similarly, phosphocholine families (PC, LPC and SM) can be separated and quantified individually at the 3.2 ppm region. Although ω-6 and ω-7 FA were modelled separately, their low relative intensity compared with surrounding signals and large overlap made individual quantifications inaccurate and they have been considered together. To our knowledge, individual quantification of ω-9 FA and LPC are reported here for the first time. Moreover, the use of spectral templates allows modelling complex signals with high-order coupling patterns. Our results have shown the high agreement between ^{1}H-NMR quantifications of FA with LipSpin and traditional techniques such as GC-FID, even though the extraction methods were different and the same information was not compared in all cases; whereas GC-FID gives information on specific lipids above the detection limit, ^{1}H-NMR adds signals from magnetically-equivalent lipids (even the less concentrated). Hence, whereas quantification of a specific FA family with GC-FID contains the summation of the most concentrated species, ^{1}H-NMR quantification considers all the species in that FA family. Moreover, similar clinical outcomes in a

nutritional study have been reached using both techniques, whereas additional information provided with ^{1}H-NMR could support other previous findings [21]. For instance, the significant increase in SM for the SFA enriched diet compared with the n-6PUFA diet agrees with the reported increase in LDL-C, which is consistent with the higher SM content of LDL particles [39]. The decrease in total phospholipids (total PL) and total fatty acids (total FA) in the SFA diet could be related to the increase in HDL-C, as this lipoprotein type has lower lipid content [30].

Other signals not considered in our study and typically included in ^{1}H-NMR profiling of lipids can be easily quantified with LipSpin using bucket integration. For instance, the large methylene signal at 1.27 ppm, diallylic protons at 2.84 ppm and olefinic protons at 5.4 ppm from FA have been used to estimate FA chain length and degree of unsaturation [40]. Although these estimations are valid for relative comparison between samples, they are based on assumptions about serum composition in healthy subjects and could not be generalizable to all cases. Additionally, other reported signals in plasma lipids, such as 7-lathosterol [19], could be easily incorporated to LipSpin with the pattern creation tool, once proper signal assignments have been made.

Experimental time is another fundamental aspect for high-throughput ^{1}H-NMR lipid profiling. Despite the 128 scans used for each spectrum in this study, it is possible to reduce acquisition time to approximately 4 minutes using 32 scans, without loss of precision for most of the quantified signals. In the case of serum lipids, only the quantification of the low concentrated ARA+EPA and DHA (tens of mM) could be compromised by the reduction of signal-to-noise ratio (S/N). When these lipids need to be reliably quantified, a better compensation for intensity loss could be done by increasing the sample volume. In the case of BUME, linearity is preserved up to 0.1 mL of serum; one possibility is extracting and pooling several 0.1 mL aliquots. This approach could be a cheaper and more effective option when automatic extraction methods and enough sample volume are available, as S/N increases proportionally with volume, unlike the square root dependency on the number of scans.

4.6. Concluding remarks

The incorporation of methodologies in routine analysis requires the development of automatic procedures and tools to simplify their use. The inherent limitations of ^{1}H-NMR of lipid samples with the non-availability of bioinformatics tools have hampered the incorporation of ^{1}H-NMR profiling in lipidomics studies. LipSpin provides the first open-source semiautomatic tool for quantitative analysis of lipids based on constrained lineshape fitting, with the aim of being

included in metabolomics workflows. No programming skills are required and no additional software is needed, however, profound knowledge of lipid spectra and signal assignments is expected, since experimental conditions usually vary between laboratories and signal patterns need to be adapted to specific conditions. Experimental conditions used in this study (solvents, acquisition temperature and good shimming of samples) have been optimised to obtain the maximum spectral dispersion. This key factor allowed LipSpin to quantify 15 different lipid signals in plasma lipophilic extracts, some of them rarely reported before, providing results that agree with commonly used techniques. This collection of signals could be modified under different experimental conditions, biological matrices or sample populations, and users are encouraged to create and publish their specific signal pattern.

4.7. References

1. Watson, A. D. (2006). Thematic review series: Systems Biology Approaches to Metabolic and Cardiovascular Disorders. Lipidomics: a global approach to lipid analysis in biological systems. *Journal of Lipid Research* , *47*(10), 2101–2111. https://doi.org/10.1194/jlr.R600022-JLR200

2. Quehenberger, O., Armando, A. M., Brown, A. H., Milne, S. B., Myers, D. S., Merrill, A. H., … Dennis, E. A. (2010). Lipidomics reveals a remarkable diversity of lipids in human plasma,. *Journal of Lipid Research* , *51*(11), 3299–3305. https://doi.org/10.1194/jlr.M009449

3. German, J. B., Gillies, L. A., Smilowitz, J. T., Zivkovic, A. M., & Watkins, S. M. (2007). Lipidomics and lipid profiling in metabolomics. *Curr Opin Lipidol Current Opinion in Lipidology*, *18*(18), 66–7166. https://doi.org/10.1097/MOL.0b013e328012d911

4. Murphy, S. A., & Nicolaou, A. (2013). Lipidomics applications in health, disease and nutrition research. *Molecular Nutrition & Food Research*, *57*(8), 1336–1346. https://doi.org/10.1002/mnfr.201200863

5. Meikle, P. J., Wong, G., Barlow, C. K., & Kingwell, B. A. (2014). Lipidomics: Potential role in risk prediction and therapeutic monitoring for diabetes and cardiovascular disease. *Pharmacology & Therapeutics*, *143*(1), 12–23. https://doi.org/10.1016/j.pharmthera.2014.02.001

6. Hyötyläinen, T., Bondia-Pons, I., & Orešič, M. (2013). Lipidomics in nutrition and food research. *Molecular Nutrition & Food Research*, *57*(8), 1306–1318. https://doi.org/10.1002/mnfr.201200759

7. Sethi, S., & Brietzke, E. (2017). Recent advances in lipidomics: Analytical and clinical perspectives. *Prostaglandins & Other Lipid Mediators*, *128–129*, 8–16. https://doi.org/10.1016/j.prostaglandins.2016.12.002

8. Ala-Korpela, M. (1995). 1H NMR spectroscopy of human blood plasma. *Progress in Nuclear Magnetic Resonance Spectroscopy*, *27*(5–6), 475–554. https://doi.org/10.1016/0079-

6565(95)01013-0

9. Schiller, J., Muller, M., Fuchs, B., Arnold, K., & Huster, D. (2007). 31P NMR Spectroscopy of Phospholipids: From Micelles to Membranes. *Current Analytical Chemistry*, *3*(4), 283–301. https://doi.org/10.2174/157341107782109635

10. Kostara, C. E., Papathanasiou, A., Cung, M. T., Elisaf, M. S., Goudevenos, J., & Bairaktari, E. T. (2010). Evaluation of Established Coronary Heart Disease on the Basis of HDL and Non-HDL NMR Lipid Profiling. *Journal of Proteome Research*, *9*(2), 897–911. https://doi.org/10.1021/pr900783x

11. Fernando, H., Bhopale, K. K., Kondraganti, S., Kaphalia, B. S., & Shakeel Ansari, G. A. (2011). Lipidomic changes in rat liver after long-term exposure to ethanol. *Toxicology and Applied Pharmacology*, *255*(2), 127–137. https://doi.org/10.1016/j.taap.2011.05.022

12. Jiang, C., Yang, K., Yang, L., Miao, Z., Wang, Y., & Zhu, H. (2013). A 1H NMR-Based Metabonomic Investigation of Time-Related Metabolic Trajectories of the Plasma, Urine and Liver Extracts of Hyperlipidemic Hamsters. *PLoS ONE*, *8*(6), e66786. https://doi.org/10.1371/journal.pone.0066786

13. Beckonert, O., Monnerjahn, J., Bonk, U., & Leibfritz, D. (2003). Visualizing metabolic changes in breast-cancer tissue using 1H-NMR spectroscopy and self-organizing maps. *NMR in Biomedicine*, *16*(1), 1–11. https://doi.org/10.1002/nbm.797

14. Fauconnot, L., Robert, F., Villard, R., & Dionisi, F. (2006). Chemical synthesis and NMR characterization of structured polyunsaturated triacylglycerols. *Chemistry and Physics of Lipids*, *139*(2), 125–136. https://doi.org/10.1016/j.chemphyslip.2005.11.004

15. Sparling, M. L. (1990). Analysis of mixed lipid extracts using 1 H NMR spectra. *Bioinformatics*, *6*(1), 29–42. https://doi.org/10.1093/bioinformatics/6.1.29

16. Vinaixa, M., Ángel Rodríguez, M., Rull, A., Beltrán, R., Bladé, C., Brezmes, J., … Correig, X. (2010). Metabolomic Assessment of the Effect of Dietary Cholesterol in the Progressive Development of Fatty Liver Disease. *Journal of Proteome Research*, *9*(5), 2527–2538. https://doi.org/10.1021/pr901203w

17. Tukiainen, T., Tynkkynen, T., Mäkinen, V.-P., Jylänki, P., Kangas, A., Hokkanen, J., … Ala-Korpela, M. (2008). A multi-metabolite analysis of serum by 1H NMR spectroscopy: Early systemic signs of Alzheimer's disease. *Biochemical and Biophysical Research Communications*, *375*(3), 356–361. https://doi.org/10.1016/j.bbrc.2008.08.007

18. Srivastava, N. K., Pradhan, S., Mittal, B., & Gowda, G. A. N. (2010). High resolution NMR based analysis of serum lipids in Duchenne muscular dystrophy patients and its possible diagnostic significance. *NMR in Biomedicine*, *23*(1), 13–22. https://doi.org/10.1002/nbm.1419

19. Oostendorp, M. (2006). Diagnosing Inborn Errors of Lipid Metabolism with Proton Nuclear Magnetic Resonance Spectroscopy. *Clinical Chemistry*, *52*(7), 1395–1405. https://doi.org/10.1373/clinchem.2006.069112

20. Soininen, P., Kangas, A. J., Würtz, P., Suna, T., & Ala-Korpela, M. (2015). Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Cardiovascular Epidemiology and

Genetics. *Circulation: Cardiovascular Genetics*, *8*(1), 192–206. https://doi.org/10.1161/CIRCGENETICS.114.000216

21. Dias, C. B., Wood, L. G., & Garg, M. L. (2016). Effects of dietary saturated and n-6 polyunsaturated fatty acids on the incorporation of long-chain n-3 polyunsaturated fatty acids into blood lipids. *European Journal of Clinical Nutrition*, *70*(7), 812–818. https://doi.org/10.1038/ejcn.2015.213

22. Löfgren, L., Ståhlman, M., Forsberg, G.-B., Saarinen, S., Nilsson, R., & Hansson, G. I. (2012). The BUME method: a novel automated chloroform-free 96-well total lipid extraction method for blood plasma. *Journal of Lipid Research* , *53*(8), 1690–1700. https://doi.org/10.1194/jlr.D023036

23. Bharti, S. K., & Roy, R. (2012). Quantitative 1H NMR spectroscopy. *TrAC Trends in Analytical Chemistry*, *35*, 5–26. https://doi.org/10.1016/j.trac.2012.02.007

24. Chen, L., Weng, Z., Goh, L., & Garland, M. (2002). An efficient algorithm for automatic phase correction of NMR spectra based on entropy minimization. *Journal of Magnetic Resonance*, *158*(1), 164–168. https://doi.org/10.1016/S1090-7807(02)00069-1

25. Morris, G. A., Barjat, H., & Home, T. J. (1997). Reference deconvolution methods. *Progress in Nuclear Magnetic Resonance Spectroscopy*, *31*(2), 197–257. https://doi.org/10.1016/S0079-6565(97)00011-3

26. Subramanian, A., Shankar Joshi, B., Roy, A. D., Roy, R., Gupta, V., & Dang, R. S. (2008). NMR spectroscopic identification of cholesterol esters, plasmalogen and phenolic glycolipids as fingerprint markers of human intracranial tuberculomas. *NMR in Biomedicine*, *21*(3), 272–288. https://doi.org/10.1002/nbm.1191

27. Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., … Scalbert, A. (2013). HMDB 3.0—The Human Metabolome Database in 2013. *Nucleic Acids Research*, *41*(Database issue), D801–D807. https://doi.org/10.1093/nar/gks1065

28. Ulrich, E. L., Akutsu, H., Doreleijers, J. F., Harano, Y., Ioannidis, Y. E., Lin, J., … Markley, J. L. (2007). BioMagResBank. *Nucleic Acids Research*, *36*(Database), D402–D408. https://doi.org/10.1093/nar/gkm957

29. Soininen, P., Haarala, J., Vepsäläinen, J., Niemitz, M., & Laatikainen, R. (2005). Strategies for organic impurity quantification by 1H NMR spectroscopy: Constrained total-line-shape fitting. *Analytica Chimica Acta*, *542*(2), 178–185. https://doi.org/10.1016/j.aca.2005.03.060

30. Jonas, A., & Phillips, M. C. (2008). CHAPTER 17 – Lipoprotein structure. In *Biochemistry of Lipids, Lipoproteins and Membranes* (pp. 485–506). https://doi.org/10.1016/B978-044453219-0.50019-2

31. Psychogios, N., Hau, D. D., Peng, J., Guo, A. C., Mandal, R., Bouatra, S., … Wishart, D. S. (2011). The Human Serum Metabolome. *PLoS ONE*, *6*(2), e16957. https://doi.org/10.1371/journal.pone.0016957

32. Weljie, A. M., Newton, J., Mercier, P., Carlson, E., & Slupsky, C. M. (2006). Targeted Profiling: Quantitative Analysis of 1H NMR Metabolomics Data. *Analytical Chemistry*, *78*(13), 4430–4442. https://doi.org/10.1021/ac060209g

33. 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*(15), 2088–2090. https://doi.org/10.1093/bioinformatics/bts308

34. Gómez, J., Brezmes, J., Mallol, R., Rodríguez, M. A., Vinaixa, M., Salek, R. M., … Cañellas, N. (2014). Dolphin: a tool for automatic targeted metabolite profiling using 1D and 2D 1H-NMR data. *Analytical and Bioanalytical Chemistry*, *406*(30), 7967–7976. https://doi.org/10.1007/s00216-014-8225-6

35. Ravanbakhsh, S., Liu, P., Bjordahl, T. C., Mandal, R., Grant, J. R., Wilson, M., … Wishart, D. S. (2015). Accurate, Fully-Automated NMR Spectral Profiling for Metabolomics. *PLOS ONE*, *10*(5), e0124219. https://doi.org/10.1371/journal.pone.0124219

36. Folch, J., Lees, M., & Stanley, G. H. S. (1957). A simple method for the isolation and purification of total lipides from animal tissues. *Journal of Biological Chemistry*, *226*(1), 497–509.

37. Soininen, P., Öörni, K., Maaheimo, H., Laatikainen, R., Kovanen, P. T., Kaski, K., & Ala-Korpela, M. (2007). 1H NMR at 800MHz facilitates detailed phospholipid follow-up during atherogenic modifications in low density lipoproteins. *Biochemical and Biophysical Research Communications*, *360*(1), 290–294. https://doi.org/10.1016/j.bbrc.2007.06.058

38. Popkova, Y., Meusel, A., Breitfeld, J., Schleinitz, D., Hirrlinger, J., Dannenberger, D., … Schiller, J. (2015). Nutrition-dependent changes of mouse adipose tissue compositions monitored by NMR, MS, and chromatographic methods. *Analytical and Bioanalytical Chemistry*, *407*(17), 5113–5123. https://doi.org/10.1007/s00216-015-8551-3

39. Nilsson, Å., & Duan, R.-D. (2006). Absorption and lipoprotein transport of sphingomyelin. *Journal of Lipid Research* , *47*(1), 154–171. https://doi.org/10.1194/jlr.M500357-JLR200

40. Kriat, M., Vion-Dury, J., Confort-Gouny, S., Favre, R., Viout, P., Sciaky, M., … Cozzone, P. J. (1993). Analysis of plasma lipids by NMR spectroscopy: application to modifications induced by malignant tumors. *Journal of Lipid Research*, *34*, 1009–1019.

UNIVERSITAT ROVIRA I VIRGILI
DEVELOPMENT OF 1H-NMR SERUM PROFILING METHODS FOR HIGH-THROUGHPUT METABOLOMICS
Rubén Barrilero Regadera

CHAPTER 5

General Discussion

UNIVERSITAT ROVIRA I VIRGILI
DEVELOPMENT OF 1H-NMR SERUM PROFILING METHODS FOR HIGH-THROUGHPUT METABOLOMICS
Rubén Barrilero Regadera

Absolute metabolite concentration in molar units is crucial for comparison across multiple studies and even analytical platforms [1]. Epidemiology and large-scale studies with serum/plasma samples benefit from the quantitative nature of NMR spectroscopy and the minimal sample preparation required, providing a trade-off between metabolite coverage and measurement cost per sample [2]. The metabolite coverage provided by ^{1}H-NMR profiling of serum/plasma includes several routine clinical markers such as various cholesterol measurements, triglycerides, apolipoproteins A-I and B, creatinine, albumin, and glucose. This fact would ultimately avoid the use of multiple clinical assays [2]. Although multiple efforts in sample preparation protocols [3] and automatic data analysis have been made, quantitative ^{1}H-NMR profiling of serum/plasma is largely influenced by sample characteristics, molecular interactions and interlaboratory variability. These questions should be addressed to definitely consolidate NMR spectroscopy in clinical routine.

5.1. Lipoprotein analysis: a step towards generalization

Analysis of serum lipoprotein sizes, particle numbers and lipid content is one of the main applications in NMR metabolomics. It was boosted by the change of paradigm of cardiovascular risk assessment, where only standard lipids could not precisely reflect the development of a cardiovascular event in patients with diabetes or metabolic syndrome [4]. Besides, lipoprotein analysis with ^{1}H-NMR avoids the tedious physical separation by ultracentrifugation. It is usually carried out by decomposing the methyl peak into the individual lipoprotein peaks or by linear regression models. However, different methods have showed numerical discrepancies [5], probably because of the lack of consensus about the calibration reference, spectral regions included, NMR experiments and blood-derived matrix. Besides, most of these methods have been built with small and homogeneous sample sets, and previous studies have reported some differences between deconvolution models for normal lipid and dyslipidemic samples [6]. The question is then obvious: is it possible to generalise the analysis of lipoprotein and lipoprotein lipids by ^{1}H-NMR?

We tried to answer this question in chapter 2 by calibrating and evaluating prediction models of standard lipids (i.e. lipid panel). We chose these lipids as they are still the main measurements for coronary heart disease assessment and therapy targets, according to the European Atherosclerosis and Cardiology Societies, and The National Cholesterol Education Program (NCEP) through their

third report of the Adult Treatment Panel (ATP). We used 785 samples from four different clinical assays. With the aim of generalizing the results to different populations, samples contained a large representation of lipid-related abnormalities including several hyperlipidaemias, metabolic syndrome and diabetes mellitus type 2. Up to the date of publication of our study, the largest sample set used was 290 serum samples of healthy subjects [7]. Moreover, our study is the first applying indistinctly serum and plasma samples from different clinical centres. In line with our results, a recent study has demonstrated that matrix effect represents less than 2% of the total spectral variance in an inter-subject analysis [8].

The NMR experiments and the spectral regions that allow the best deconvolution and prediction models is another unresolved question. Whereas 1D ^{1}H-NMR models rely on the chemical shift separation of lipoprotein subclasses due to their different structural composition [9], 2D diffusion ^{1}H-NMR models include an extra dimension directly related to lipoprotein sizes [6]. Furthermore, the utility of other lipid signals than methyl is uncertain although some studies suggest that other regions such as choline could discriminate even better [10]. We evaluated different NMR experiments in several steps. First, we applied correlation analysis of different ^{1}H-NMR spectra with standard lipids to conclude that diffusion-based spectra best correlate with standard lipids. This is mainly explained because the lack of anticoagulant and LMWM signals improve the visibility of lipid signals. This fact could also explain the observed compatibility between plasma and serum samples in our models. Additionally, we observed strong correlations in 14 signals containing lipid information (fatty acid moieties, cholesterol methyl groups, glycerol from triglycerides and phospholipids and choline groups) that could potentially improve lipid regression models. Taking into account these considerations, 1D and 2D Diffusion ^{1}H-NMR-based regression models using different chemometrics methods and pre-processing modes were built and evaluated over the same samples. Interestingly, the best results were obtained with 1D diffusion-edited ^{1}H-NMR spectra, suggesting that diffusion dimension in 2D Diffusion ^{1}H-NMR could be redundant and that the high complexity introduced in the multi-way structures could hamper the ability to find the underlying explanatory models. The results showed in 2.4 agreed with previous studies although HDL-C was slightly worse [7,11–14], probably because HDL-C signal visibility was compromised in hyperlipidemic samples or because different precipitation reagents were used in HDL-C isolation.

Although our study only applies to standard lipoprotein lipids, it could indicate the ability to generalise more complex models of NMR-based lipoprotein analysis to different population cohorts. More importantly, the excellent estimation presented here would allow the inclusion of

these clinical lipids in the high-throughput ^{1}H-NMR measurements catalogue and avoid the multiple enzymatic-colorimetric assays. Moreover, it would allow the analysis of discordances between LDL-C and non-HDL-C with other LDL-related measurements, such as LDL-P, by using the same platform, and would make possible to better stratify cardiovascular risk and, consequently, select the more convenient therapy targets [15].

In parallel, it would be desirable to extend our study to other clinically-relevant lipoprotein variables such as apolipoproteins B and A1. Apolipoprotein B has been suggested a better risk marker than LDL-C. Our preliminary results (not reported in this thesis) using a subpopulation of approximately 500 samples gave Pearson's correlations (r) between immunoassay and NMR-derived measurements of r=0.95 and r=0.78 for apolipoproteins B and A1, respectively. Another important aspect would be to incorporate automatic and holistic methods of variable selection to reduce redundant or noisy-prone spectral regions and provide more robust regression models [16].

5.2. Unravelling the "NMR-invisible" metabolome

Along with lipoprotein analysis, low-molecular-weight metabolite (LMWM) profiling is the other main application of NMR in high-throughput metabolomics. This success relies on the ability to provide quantitative information about dozens of polar metabolites in intact serum with the only requirement of a CPMG filter for spectral protein removal. Moreover, some LMWM have been found to play an important role in disease development. For instance, high levels of some amino acids can be used to identify individuals at risk of developing type 2 diabetes [17] and cardiovascular disease [18]. Consequence of this has been the development of multiple automatic or semiautomatic tools for ^{1}H-NMR profiling of LMWM [19]. These tools provide reliable deconvolution and quantification of highly overlapped signals by fitting known molecular fingerprints.

However, native serum is governed by complex physicochemical interactions that affect quantification by 1H-NMR spectroscopy. Among them, protein binding of LMWM compromises the "NMR-visibility" of LMWM in a CPMG spectrum. This situation generates two potential problems: the first is the impossibility to profile largely bound metabolites adequately. The second implies that clinical outcomes using ^{1}H-NMR profiling could be misled if quantified LMWM deviate from their absolute content in serum. This would especially be the case if comparing

healthy subjects with subjects with abnormal albumin content or elevated free fatty acids (FFA), as FFA compete with LMWM to be carried by proteins through the bloodstream [20]. Although serum dilution is recommended to increase LMWM visibility in ^{1}H-NMR spectra [21], the only method that has shown to recover most of the LMWM content in proteins is deproteneization [22]. Besides, serum dilution accentuates the low sensitivity of NMR. On the contrary, a protein-free serum increases substantially the metabolite coverage, but the time-consuming processes required confronts with high-throughput metabolomics needs. Using native or deproteinized serum is still a matter of debate in most metabolomics platforms.

An unexplored alternative has been proposed in chapter 3. Trimethylsilylpropionic acid (TSP), a well-known binding molecule used in ^{1}H-NMR for spectral referencing, was used to compete for ligand-binding sites of proteins and promote the release of binding LMWM. Our approach is fully compatible with high-throughput NMR metabolomics: it does not involve additional sample manipulation (TSP can be added to the D_2O required for locking the NMR) and is a low-cost solution (less than 0.15 euros/sample). Section 3.4 demonstrates that adding TSP increases in approximately 40% the signal of some clinically-relevant amino acids when applying a typical 1D CPMG filter. Our method has been proved to be more effective than sample dilution that only provides a 10% of additional signal. It also performs similarly in serum and plasma, despite their different protein content. Additionally, competitive binding does not affect spectral S/N.

Previous studies of protein binding and “NMR-invisible” LMWM based their findings on one-dimensional T2-edited CPMG spectra. A CPMG experiment consists of a fixed T2 cutoff filter, where each signal is differently attenuated according to its own T2-relaxation. It is noteworthy to mention that T2-relaxation varies according to the free-to-bound ratios, i.e. T2 attenuations in CPMG spectra are lower for LMWM not binding to protein. Moreover, molar concentrations are normally calculated by normalising LMWM signal areas with the signal area of a reference compound, which can have very large T2-relaxation differences. Based on these considerations, it is expected a significant quantitative error if using common CPMG spectra that could lead to misinterpretations of the protein binding interactions. In this article, we originally used for the first time multivariate curve resolution with multispectral ^{1}H-NMR to precisely characterise the “NMR-invisibility” of LMWM at different degrees of protein binding. Our multispectral approach has shown an effective strategy to obtain reliable quantifications without this T2 variation.

Further research in competitive binding for ^{1}H-NMR profiling involves analytical and clinical considerations:

- It should be evaluated if the increase in "NMR-visibility" of disease biomarkers improves disease risk prediction/diagnosis. Protein binding phenylalanine, leucine and isoleucine in type 2 diabetes patients could be good candidates.
- It is required a comprehensive profiling of TSP-added samples in order to update the metabolite coverage under these conditions.
- It should be analysed the analytical impact of TSP addition in quantitative ^{1}H-NMR serum/plasma profiling, such as in lipoprotein deconvolution or prediction models, as lipoproteins could also bind TSP.
- The characterization of other competitive-binding reagents and the combination with alternative methods, i.e. sample dilution or soft acidification, should be explore in order to maximise the free LMWM and obtain quantification closer to absolute.

5.3. Automating ^{1}H-NMR lipid profiling

Contrary to lipoprotein and LMWM analysis, serum lipid analysis has been given less attention in high-throughput NMR metabolomics due to two important aspects. First, lipid analysis cannot be performed in native serum, which implies one extraction procedure and additional NMR measurements. Moreover, the long centrifugation times required and the fact that the lipophilic phase settle in the lower layer of the triphasic separation [23,24] hamper the automation of lipid extraction procedures. Second, NMR is unable to identify structural aspects of lipids, such as the specific sn position of a fatty acid residue in a glycerol moiety, and specific lipid species, such as PC(18:0/18:2). The catalogue of lipid species in mammalians comprises thousands of combinations of glycerol-, sphingosine- and cholesterol-based moieties with multiple fatty acids species, which spectral signatures are not as distinctive as in the case of LMWM. Moreover, the complexity of lipid structures gives NMR signals with high complex coupling patterns. These aspects discourage metabolomics groups to develop automatic tools for ^{1}H-NMR lipid profiling and favour MS platforms for lipidomics analysis.

Despite the inherent limitations of NMR-based lipids, ^{1}H-NMR lipid profiling benefit from its main analytical advantage respect to other platforms: the absolute quantification. NMR quantification brings the possibility of applying lipid analysis to large-scale studies between different laboratories and analytical platforms [1]. Moreover, ^{1}H-NMR spectra of serum lipids provides information of the abundance of major lipid families (fatty acids, glycerolipids, glycerophospholipids and sterols) without prior separation, that has proved valuable in clinical

research, such as the study of Alzheimer's disease [25], human pancreatic cancer [26] and inborn errors of metabolism [27]. In Chapter 4, we present several strategies to improve the automation of ^{1}H-NMR lipid profiling.

The limitation imposed by manual extraction procedures can be partially palliated using BUME, a recent lipid extraction method that can be automated with standard liquid handling 96-well robots [28]. Ultimately, automation would induce less variation in sample preparation. The main drawbacks of using BUME for ^{1}H-NMR are the long evaporation times and the multiple signals introduced by the organic solvents. Fortunately, most of these signals do not overlap lipid signals or are easily modelled in the lineshape fitting analysis, except for heptane that can be replaced with diisopropyl ether.

Respect to the complex signals, it is first interesting to increase spectral dispersion to facilitate spectral deconvolution. The use of high-magnetic field spectrometers is known to increase spectral dispersion that could aid to separate highly-overlapped signals [29]. However, cutting-edge spectrometers are still prohibitive. Instead, we optimised the solvent composition to obtain maximum separation for a common 600 MHz spectrometer. Reconstituting lipids in $CDCl_3$:CD_3OD:D_2O (16:7:1, v/v/v) solution provided enough signal separation to resolve choline families and increase separation in other overlapping signals. Once the spectral conditions are the best case scenario, we developed LipSpin, a tool for the semiautomatic deconvolution of the ^{1}H-NMR lipid spectra. The software is open source aimed to be complemented by other developers in order to boost lipid analysis with NMR in high-throughput metabolomics. LipSpin relies on the classical lineshape fitting approach of Lorentzian/Gaussian functions, but also reference spectral models for signals with complex coupling patterns. In section 4.4 we run analytical validations of the whole workflow against conventional techniques, showing large inter-platform agreement. Moreover, the clinical utility of NMR-based lipids was evaluated satisfactory in a dietary intervention study. In resume, LipSpin allows the reliable quantification of 15 different lipid-related variables in plasma: Cholesterol (free and esterified), fatty acids (saturated, omega-9, omega-6, omega-3, monounsaturated, arachidonic + eicosapentanoic, docosahexaenoic and linoleic), phospholipids (phosphatidylcholine, lysophosphatidylcholine, sphyngomyeline, total glycerophospholipids) and triglycerides. This list represents absolute quantitative data with the abundance of several lipid classes and some individual species. Additionally, qualitative information about mean FA chain length and number of double bonds fatty acid can also be extracted.

Although LipSpin has been optimised and proved with plasma lipids, the universality of the mathematical approach in NMR makes LipSpin be applicable to other biological matrices. In parallel, it would be desirable to characterise the biochemical composition of other biofluids (cerebrospinal fluid, bile, etc.) and tissues, as well as to increment the collection of signal patterns with new signal definitions and reference spectra. LipSpin is also suitable for plant and food lipids. Additionally, the lineshape fitting algorithm could benefit of more sophisticated quantum mechanical models that have been applied to other in-house constrained total lineshape fitting approaches [30], from probabilistic models with prior information of signal characteristics [31] and from parallel computing of lineshape fitting analysis in multiple samples [32].

5.4. References

1. Nagana Gowda, G. A., & Raftery, D. (2017). Recent advances in NMR-based metabolomics. *Analytical Chemistry*. https://doi.org/10.1021/acs.analchem.6b04420

2. Soininen, P., Kangas, A. J., Würtz, P., Suna, T., & Ala-Korpela, M. (2015). Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Cardiovascular Epidemiology and Genetics. *Circulation: Cardiovascular Genetics*, *8*(1), 192–206. https://doi.org/10.1161/CIRCGENETICS.114.000216

3. Beckonert, O., Keun, H. C., Ebbels, T. M. D., Bundy, J., Holmes, E., Lindon, J. C., & Nicholson, J. K. (2007). Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature Protocols*, *2*(11), 2692–2703. https://doi.org/10.1038/nprot.2007.376

4. Mallol, R., Rodriguez, M. A., Brezmes, J., Masana, L., & Correig, X. (2013). Human serum/plasma lipoprotein analysis by NMR: Application to the study of diabetic dyslipidemia. *Progress in Nuclear Magnetic Resonance Spectroscopy*, *70*, 1–24. https://doi.org/10.1016/j.pnmrs.2012.09.001

5. Mallol, R., Amigó, N., Rodríguez, M. A., Heras, M., Vinaixa, M., Plana, N., … Correig, X. (2015). Liposcale: a novel advanced lipoprotein test based on 2D diffusion-ordered 1H NMR spectroscopy. *Journal of Lipid Research*, *56*(3), 737–746. https://doi.org/10.1194/jlr.D050120

6. Mallol, R., Rodríguez, M. A., Heras, M., Vinaixa, M., Cañellas, N., Brezmes, J., … Correig, X. (2011). Surface fitting of 2D diffusion-edited 1H NMR spectroscopy data for the characterisation of human plasma lipoproteins. *Metabolomics*, *7*(4), 572–582. https://doi.org/10.1007/s11306-011-0273-8

7. Mihaleva, V. V., Van Schalkwijk, D. B., De Graaf, A. A., Van Duynhoven, J., Van Dorsten, F. A., Vervoort, J., … Jacobs, D. M. (2014). A systematic approach to obtain validated partial least square models for predicting lipoprotein subclasses from serum nmr spectra. *Analytical Chemistry*, *86*(1), 543–550. https://doi.org/10.1021/ac402571z

8. Monsonis Centelles, S., Hoefsloot, H. C. J., Khakimov, B., Ebrahimi, P., Lind, M. V, Kristensen, M., … Smilde, A. K. (2017). Toward Reliable Lipoprotein Particle Predictions from NMR Spectra of Human Blood: An Interlaboratory Ring Test. *Analytical Chemistry*, *89*(15), 8004–8012. https://doi.org/10.1021/acs.analchem.7b01329

9. Otvos, J. D., Jeyarajah, E. J., & Bennett, D. W. (1991). Quantification of plasma lipoproteins by proton nuclear magnetic resonance spectroscopy. *Clinical Chemistry*, *37*(3), 377–386.

10. Suna, T., Salminen, A., Soininen, P., Laatikainen, R., Ingman, P., Mäkelä, S., … Ala-Korpela, M. (2007). 1H NMR metabonomics of plasma lipoprotein subclasses: Elucidation of metabolic clustering by self-organising maps. *NMR in Biomedicine*, *20*(7), 658–672. https://doi.org/10.1002/nbm.1123

11. Vehtari, A., Mäkinen, V.-P., Soininen, P., Ingman, P., Mäkelä, S. M., Savolainen, M. J., … Ala-Korpela, M. (2007). A novel Bayesian approach to quantify clinical variables and to determine their spectroscopic counterparts in 1H NMR metabonomic data. *BMC Bioinformatics*, *8 Suppl 2*(Suppl 2), S8. https://doi.org/10.1186/1471-2105-8-S2-S8

12. Bathen, T. F., Krane, J., Engan, T., Bjerve, K. S., & Axelson, D. (2000). Quantification of plasma lipids and apolipoproteins by use of proton NMR spectroscopy, multivariate and neural network analysis. *NMR in Biomedicine*, *13*(5), 271–288. https://doi.org/10.1002/1099-1492(200008)13:5<271::AID-NBM646>3.0.CO;2-7

13. Petersen, M., Dyrby, M., Toubro, S., Engelsen, S. B., Nørgaard, L., Pedersen, H. T., & Uyerberg, J. (2005). Quantification of lipoprotein subclasses by proton nuclear magnetic resonance-based partial least-squares regression models. *Clinical Chemistry*, *51*(8), 1457–1461. https://doi.org/10.1373/clinchem.2004.046748

14. Dyrby, M., Petersen, M., Whittaker, A. K., Lambert, L., Nørgaard, L., Bro, R., & Engelsen, S. B. (2005). Analysis of lipoproteins using 2D diffusion-edited NMR spectroscopy and multi-way chemometrics. *Analytica Chimica Acta*, *531*(2), 209–216. https://doi.org/10.1016/j.aca.2004.10.052

15. Mora, S., Buring, J. E., & Ridker, P. M. (2014). Discordance of Low-Density Lipoprotein (LDL) Cholesterol With Alternative LDL-Related Measures and Future Coronary Events. *Circulation*, *129*(5), 553–561. https://doi.org/10.1161/CIRCULATIONAHA.113.005873

16. Kristensen, M., Savorani, F., Ravn-Haren, G., Poulsen, M., Markowski, J., Larsen, F. H., … Engelsen, S. B. (2010). NMR and interval PLS as reliable methods for determination of cholesterol in rodent lipoprotein fractions. *Metabolomics*, *6*(1), 129–136. https://doi.org/10.1007/s11306-009-0181-3

17. Wang, T. J., Larson, M. G., Vasan, R. S., Cheng, S., Rhee, E. P., McCabe, E., … Gerszten, R. E. (2011). Metabolite profiles and the risk of developing diabetes. *Nature Medicine*, *17*(4), 448–53. https://doi.org/10.1038/nm.2307

18. Würtz, P., Havulinna, A. S., Soininen, P., Tynkkynen, T., Prieto-Merino, D., Tillin, T., … Salomaa, V. (2015). Metabolite profiling and cardiovascular event risk: A prospective study of 3 population-based cohorts. *Circulation*, *131*(9), 774–785. https://doi.org/10.1161/CIRCULATIONAHA.114.013116

19. Spicer, R., Salek, R. M., Moreno, P., Cañueto, D., & Steinbeck, C. (2017). Navigating freely-

available software tools for metabolomics analysis. *Metabolomics*, *13*(9), 106. https://doi.org/10.1007/s11306-017-1242-7

20. Jupin, M., Michiels, P. J., Girard, F. C., Spraul, M., & Wijmenga, S. S. (2013). NMR identification of endogenous metabolites interacting with fatted and non-fatted human serum albumin in blood plasma: Fatty acids influence the HSA–metabolite interaction. *Journal of Magnetic Resonance*, *228*, 81–94. https://doi.org/10.1016/j.jmr.2012.12.010

21. Smolinska, A., Blanchet, L., Buydens, L. M. C., & Wijmenga, S. S. (2012). NMR and pattern recognition methods in metabolomics: From data acquisition to biomarker discovery: A review. *Analytica Chimica Acta*. https://doi.org/10.1016/j.aca.2012.05.049

22. Daykin, C. A., Foxall, P. J. D., Connor, S. C., Lindon, J. C., & Nicholson, J. K. (2002). The comparison of plasma deproteinization methods for the detection of low-molecular-weight metabolites by (1)H nuclear magnetic resonance spectroscopy. *Analytical Biochemistry*, *304*(2), 220–30. https://doi.org/10.1006/abio.2002.5637

23. Bligh, E. G., & Dyer, W. J. (1959). A RAPID METHOD OF TOTAL LIPID EXTRACTION AND PURIFICATION. *Canadian Journal of Biochemistry and Physiology*, *37*(8), 911–917. https://doi.org/10.1139/o59-099

24. Folch, J., Lees, M., & Stanley, G. H. S. (1957). A simple method for the isolation and purification of total lipides from animal tissues. *Journal of Biological Chemistry*, *226*(1), 497–509.

25. Tukiainen, T., Tynkkynen, T., Mäkinen, V.-P., Jylänki, P., Kangas, A., Hokkanen, J., … Ala-Korpela, M. (2008). A multi-metabolite analysis of serum by 1H NMR spectroscopy: Early systemic signs of Alzheimer's disease. *Biochemical and Biophysical Research Communications*, *375*(3), 356–361. https://doi.org/10.1016/j.bbrc.2008.08.007

26. Beger, R. D., Schnackenberg, L. K., Holland, R. D., Li, D., & Dragan, Y. (2006). Metabonomic models of human pancreatic cancer using 1D proton NMR spectra of lipids in plasma. *Metabolomics*, *2*(3), 125–134. https://doi.org/10.1007/s11306-006-0026-2

27. Oostendorp, M. (2006). Diagnosing Inborn Errors of Lipid Metabolism with Proton Nuclear Magnetic Resonance Spectroscopy. *Clinical Chemistry*, *52*(7), 1395–1405. https://doi.org/10.1373/clinchem.2006.069112

28. Löfgren, L., Ståhlman, M., Forsberg, G.-B., Saarinen, S., Nilsson, R., & Hansson, G. I. (2012). The BUME method: a novel automated chloroform-free 96-well total lipid extraction method for blood plasma. *Journal of Lipid Research* , *53*(8), 1690–1700. https://doi.org/10.1194/jlr.D023036

29. Soininen, P., Öörni, K., Maaheimo, H., Laatikainen, R., Kovanen, P. T., Kaski, K., & Ala-Korpela, M. (2007). 1H NMR at 800MHz facilitates detailed phospholipid follow-up during atherogenic modifications in low density lipoproteins. *Biochemical and Biophysical Research Communications*, *360*(1), 290–294. https://doi.org/10.1016/j.bbrc.2007.06.058

30. Tiainen, M., Soininen, P., & Laatikainen, R. (2014). Quantitative Quantum Mechanical Spectral Analysis (qQMSA) of 1H NMR spectra of complex mixtures and biofluids. *Journal of Magnetic Resonance*, *242*, 67–78. https://doi.org/10.1016/j.jmr.2014.02.008

31. Ravanbakhsh, S., Liu, P., Bjordahl, T. C., Mandal, R., Grant, J. R., Wilson, M., … Wishart, D. S. (2015). Accurate, Fully-Automated NMR Spectral Profiling for Metabolomics. *PLOS ONE*, *10*(5), e0124219. https://doi.org/10.1371/journal.pone.0124219

32. 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*(15), 2088–2090. https://doi.org/10.1093/bioinformatics/bts308

CHAPTER 6

General Conclusions

UNIVERSITAT ROVIRA I VIRGILI
DEVELOPMENT OF 1H-NMR SERUM PROFILING METHODS FOR HIGH-THROUGHPUT METABOLOMICS
Rubén Barrilero Regadera

This section summarizes the conclusions of the doctoral thesis.

Work 1: Design and evaluation of standard lipid prediction models based on ^{1}H-NMR spectroscopy of human serum/plasma samples

The first study has demonstrated the ability to quantify standard lipids used in the routine clinical practice (total cholesterol, triglycerides, HDL cholesterol, LDL cholesterol and non-HDL cholesterol) by PLS regression models based on diffusion-edited ^{1}H-NMR spectra of serum and plasma samples. Correlations between NMR-predicted lipids and enzymatically-measured lipids were similar to previous studies with small and homogeneous cohorts. However, our models are more generalizable as they were calibrated and validated using large and heterogeneous sample sets, including subjects with normal and abnormal lipid and lipoprotein profiles, different blood-derived matrices (plasma and serum), and samples obtained at different clinical centres. Other findings of this study include:

- 14 lipids signals in ^{1}H-NMR spectra of plasma/serum showed large correlation with standard lipids; consequently, they were used to build the regression models.
- Lipids signals in diffusion ^{1}H-NMR spectra showed larger correlation with standard lipids than lipid signals in non-editing or T2-editing ^{1}H-NMR spectra.
- For most standard lipids, regression models based on 1D diffusion-edited ^{1}H-NMR spectra gave the best predictions. Regression models including the additional diffusion dimension (i.e. those using 2D diffusion ^{1}H-NMR spectra) only improved triglycerides predictions. Consequently, there is no need for increasing measurement times using 2D NMR experiments.
- Plasma and serum samples can be indistinctly used for standard lipid predictions with ^{1}H-NMR regression models.
- Scaling spectral data in regression models by mean-centring and auto-scaling gave similar results.

Work 2: Unravelling and quantifying the "NMR-invisible" metabolites interacting with human serum albumin by binding competition and T2 relaxation-based decomposition analysis

The second study has quantitatively determined the impact of human serum protein binding in the "NMR-visibility" of five low-molecular-weight metabolites (LMWM) that have been found to early predict the development of diabetes mellitus type 2. Only c.a. 90%, 80%, 50%, 60% and

40% of valine, isoleucine, tyrosine, leucine and phenylalanine signal, respectively, could be quantified in native serum, whereas the remaining signal was merged with the protein background signal due to slow-exchange binding with serum protein. The addition of 6 mM TSP allowed the quantification of c.a. 99%, 109%, 51%, 85%, 75% of valine, isoleucine, tyrosine, leucine and phenylalanine signal, respectively. These results assumed a quantitative error of approximately 10%. Our findings with TSP highlight that competitive protein binding in human serum (i.e. forced competition for protein ligand-binding sites between endogenous low-molecular-weight metabolites and an exogenous compound) could be an alternative to other methods based on serum deproteneization to improve LMWM quantifications. Moreover, the competitive binding approach is fully compatible with high-throughput NMR. Other findings of this study include:

- Isoleucine has been found to bind to serum protein for the first time.
- Competitive binding with TSP showed to be more effective at LMWM signal recovery than two-fold sample dilution in most of the cases.
- Multivariate curve resolution of multidimensional spectra (pseudo 2D CPMG) provides T2-corrected quantifications, which allows determining only the LMWM "NMR-invisible" signal due to slow-exchange with protein. Moreover, information about fast-exchange binding with protein can be extrapolated from the T2 decays of the LMWM.
- LMWM T2-relaxations are lengthened as fast-exchange of LMWM with protein is reduced with TSP addition. It ultimately implies less T2-attenuation in 1D CPMG spectra.
- Results in human serum were consistent with results in synthetic serum models with human serum albumin, except for the case of tyrosine, which remained the same after TSP addition in the human serum. Similarly, tyrosine was the only signal increased with sample dilution but not with TSP addition.
- Similar mean effects in "NMR-invisibility" before and after TSP addition were observed in 85 plasma samples.

Work 3: LipSpin: a new bioinformatics tool for quantitative ^{1}H-NMR lipid profiling

The third study has presented LipSpin, a new bioinformatics tool for quantitative ^{1}H-NMR lipid profiling. This tool allows the quantification of 15 lipid-related variables of major lipid classes in lipophilic serum extracts (fatty acids, triglycerides, phospholipids and cholesterols). Both IDE independent standalone and open source versions of LipSpin are publicly available. LipSpin includes all the required steps to convert raw NMR data into quantitative lipid variables in a

semiautomatic (only requires minimal parameter adjustments before algorithm execution) and user-friendly way (no programming skills are required). It also allows batch processing of multiple spectra and is versatile, since signal patterns collection can be modified or expanded. Other findings of this study include:

- BUME, the lipid extraction method used in this study, allows automation of plasma lipid extraction with liquid handling robots. Solvent signals from the extraction procedure have low interference with lipid signals.
- Lipid reconstitution in $CDCl_3$:CD_3OD:D_2O (16:7:1, v/v/v) solution maximises the spectral dispersion (especially for lipids with polar head groups).
- Lipid quantifications obtained with LipSpin have shown large correlation with conventional techniques such as GC-FID and enzymatic-colorimetric measurements. Moreover, the clinical utility of these quantifications have been validated in a dietary intervention study.

UNIVERSITAT ROVIRA I VIRGILI
DEVELOPMENT OF 1H-NMR SERUM PROFILING METHODS FOR HIGH-THROUGHPUT METABOLOMICS
Rubén Barrilero Regadera

ANNEXES

LipSpin user manual

UNIVERSITAT ROVIRA I VIRGILI
DEVELOPMENT OF 1H-NMR SERUM PROFILING METHODS FOR HIGH-THROUGHPUT METABOLOMICS
Rubén Barrilero Regadera

Installation

LipSpin is released as original m-files and standalone versions. In order to execute LipSpin, follow the instructions depending on the version you have received.

M-files

M-files coding LipSpin can be downloaded from github repositories and loaded in the MATLAB IDE. Matlab scripts and functions have been developed with MATLAB v7.10 and compatibility with other versions cannot be guaranteed. Note that some functionalities of LipSpin require external toolboxes commonly supplied with most MATLAB versions such as Statistics, Optimization and Signal toolboxes. In order to execute LipSpin in the MATLAB IDE follow the next steps:

1. Download LipSpin folder and copy it in your local computer.
2. Copy the above directory in the MATLAB path variable (include subdirectories).
3. Type “lipspin” in the MATLAB command window.

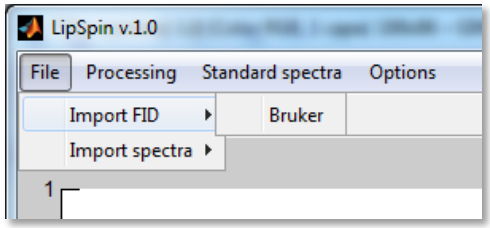
Standalone version

LipSpin can also be provided as a standalone version on demand, which can be run in Microsoft Windows OS without having MATLAB installed. Standalone LipSpin only requires Matlab Compiler Runtime (MCR) to be installed in your local computer. MCR contains all the necessary MATLAB libraries called by LipSpin. MCR version is optimised for each LipSpin compilation, consequently, be sure of using the right version of the MCR by asking the developer of your compiled LipSpin standalone version.

1. Verify if the required version of the MATLAB Compiler Runtime (MCR) is installed in your computer.
2. If the MCR is not installed, run MCRInstaller.exe provided with your LipSpin version. Now, MCR should have created a folder in “Program Files” folder and a new variable in the windows PATH environment variable.
3. Run LipSpin.exe.

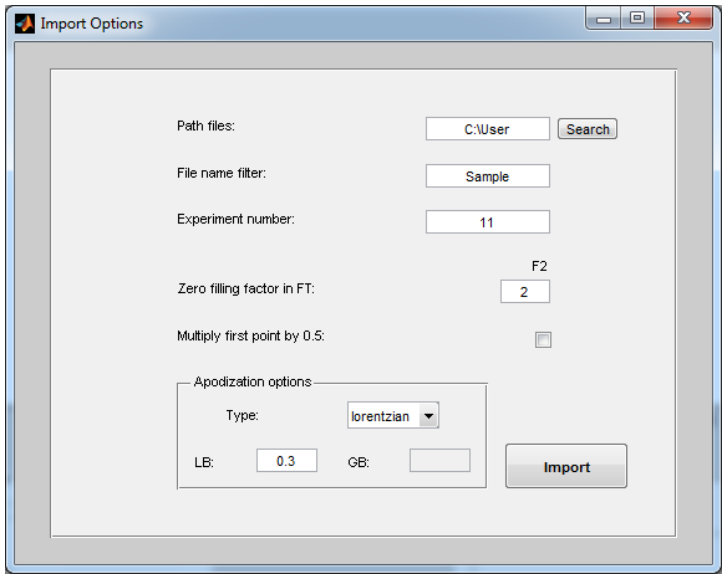
Import NMR data

NMR data can be imported as either time-domain “free induction decays” (FID) or Fourier transformed 1D NMR spectra from the tab “File” in the menu bar of the main screen. LipSpin allows loading multiple NMR data files at once, providing that all share the same time or spectral axis scale. Only Bruker files are supported in current versions of LipSpin but other NMR manufacturer formats are expected to be included in future releases.

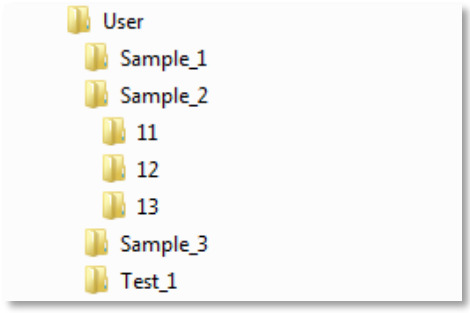


Import FID

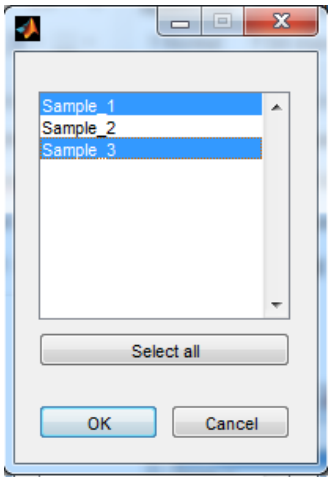
By converting NMR spectra from FID, the user can load raw data directly from NMR acquisition while avoiding the use of additional software other than LipSpin for data processing. The Import window includes the following options:



- ***Path files***: full path of the directory where the sample folders are located. In the example below, it refers to the full path of *User*, (for example *C:\User*).



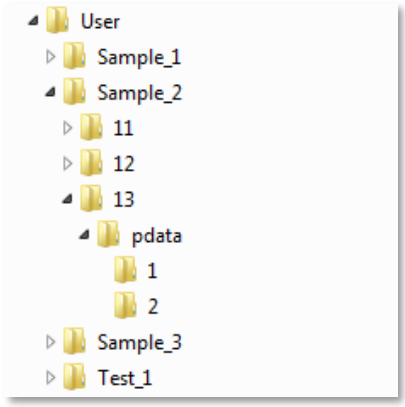
- ***File name filter:*** filters the list of samples to those having this string in their folder names. In the example above, entering "Sample" will remove *Test_1* from latter selection.
- ***Experiment number:*** following the Bruker folder structure, LipSpin requires the experiment number where the *fid* file is located (in the example above: *11, 12* or *13*).
- ***Zero filling:*** increases the spectral resolution of the NMR spectra by multiplying the data length by 2^n, where n is a natural number included in the "F2" field. ***Note: Increasing the data size will increase RAM demand and could slow down the program and the OS execution. Recommended values: 0-2.*
- ***Multiply first point by 0.5:*** this option is aimed to reduce the DC offset in the NMR spectra. In most of the cases, it produces negligible effects.
- ***Apodization:*** allows applying none, Gaussian or Lorentzian windowing to FID before Fourier transformation (FT). Gaussian is aimed to increase peak resolution and could be suitable for peak identification in large overlapped regions. If Gaussian window is applied, LB should be a value close to the negative of the peak width in Hz (measured at half height, e.g: -2) and a good starting point for GB could be 0.2 to 0.4. Lorentzian is aimed to increase the S/N ratio and is the common choice for quantitative spectral analysis. Common values for LB using Lorentzian window range between 0.3 and 1.
- ***Import:*** opens a selection window that lists the samples in "Path files" after filtering by "File name filter". Selected samples will be loaded into the main window.



Import spectra

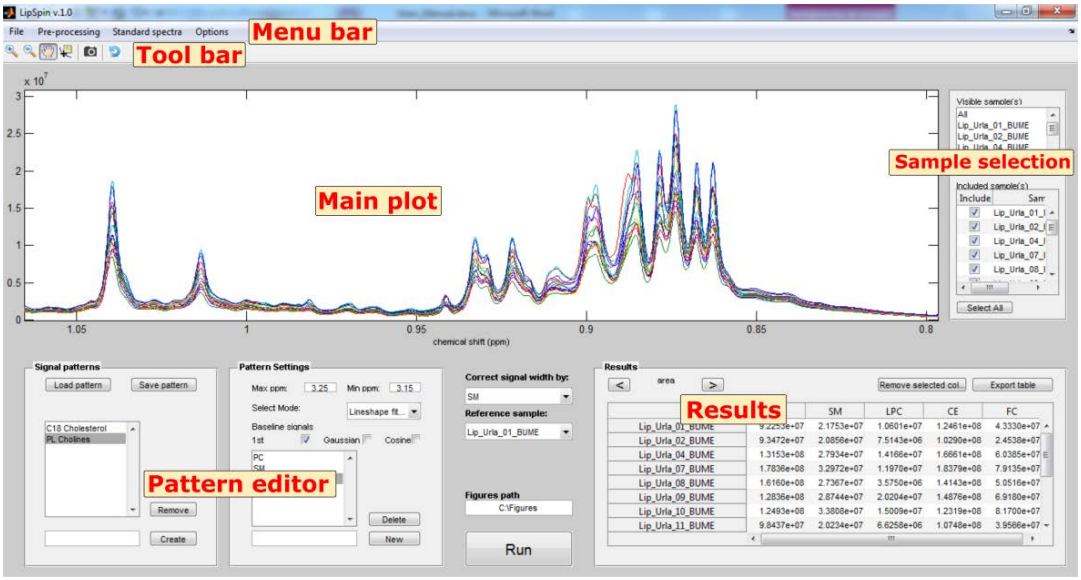
This option allows loading spectra that have been previously converted from FID using third-party software. Refer to the previous section "Import FID" for explanation of ***Path files***, ***File name filter*** and ***Experiment number*** fields. Additionally, this window includes:

- ***Processing number:*** folder name of processed spectra (*1D* file). In the example below, folders *1* and *2* contains different processed spectra from the same FID. Note that LipSpin requires the processing folders to be in a *pdata* folder.

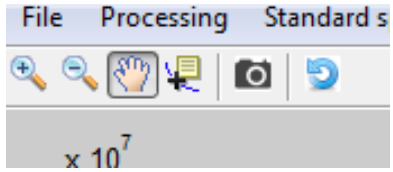


Main LipSpin Window

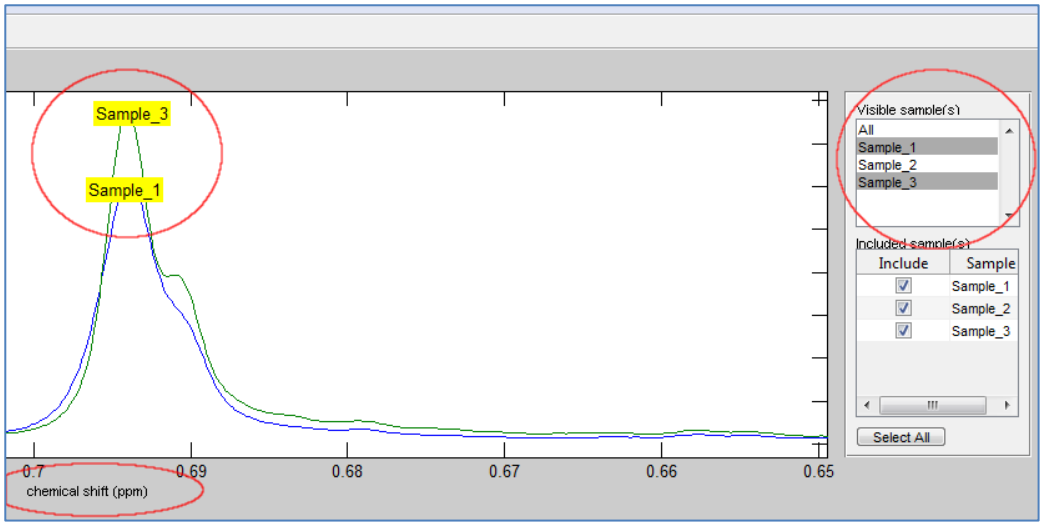
Once the spectra have been imported, they are displayed in the main LipSpin window.



The toolbar contains typical MATLAB navigation tools that allow zooming, panning and displaying data point coordinates. A png snapshot can also be saved from the “Save Figure” tool. The undo button allows reverting last action (only once).



For the aim of aiding sample identification, right-clicking in a spectral line will show/hide a tag with its sample name. Users can also restrict visualization to only selected samples in the “Visible sample(s)” list.



Right-clicking the tag “chemical shift (ppm)” below the x-axis will swap the axis scale in ppm for the axis scale in Hz. This feature is interesting for determining J-coupling constants and line widths.

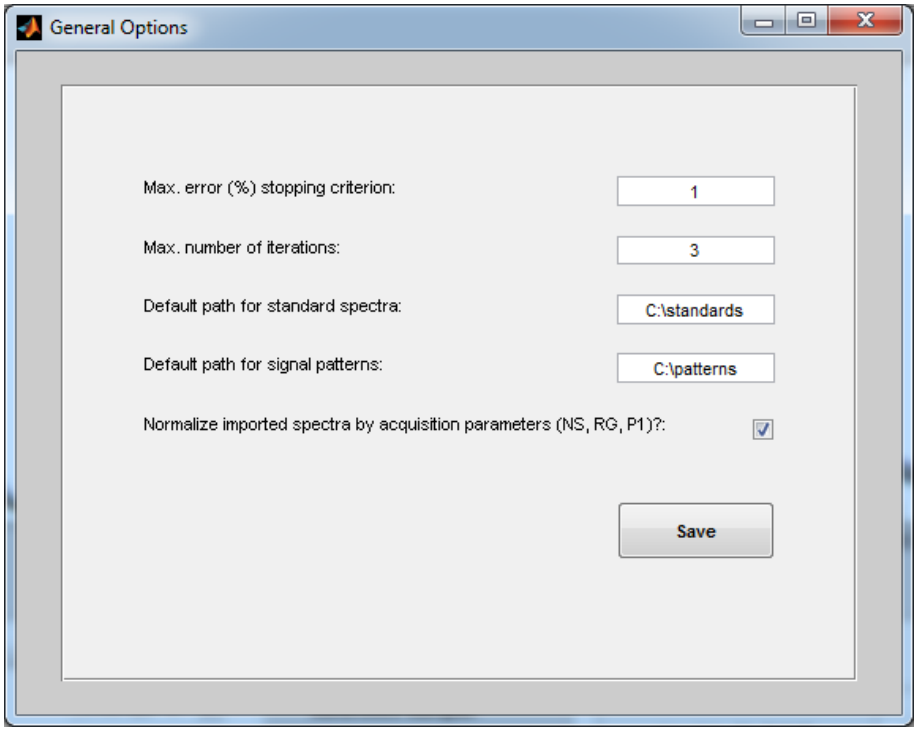
Finally, the checkbox list in the “included sample(s)” panel allows indicating the samples that will be included if a spectral pre-processing or lineshape fitting is carried out.

General options

The General options window is accessible from the menu bar and allows setting the general options of LipSpin. These options are permanently saved in the *options.nmrcfg* file which should be located in the same folder of LipSpin script (m-files) or LipSpin.exe (standalone version).

- ***Max. error (%) stopping criterion:*** threshold that will stop lineshape fitting iterations based on minimum %RMSE.
- ***Max. number of iterations:*** maximum allowed iterations of lineshape fitting algorithm without fulfilling the maximum %RMSE stopping criterion.
- ***Default path for standard spectra:*** directory with “.nmrstd” files to be automatically loaded in the current session.
- ***Default path for signal patterns:*** directory with “.nmrsgnl” files.
- ***Normalise imported spectra by acquisition parameters (NS, RG, P1):*** checking this box will normalise spectra base on their specific acquisition conditions: number of scans

(NS), receiver gain (RG) and 90º pulse length (P1). ***Note: this is mandatory for quantitative inter-sample comparison if the spectra are not normalised using internal standards.*



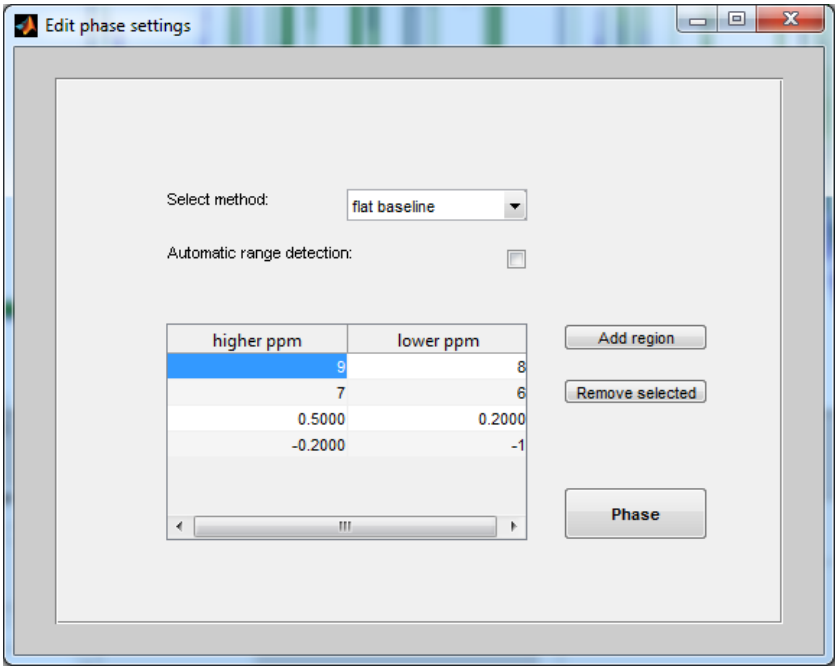
Preparing the NMR spectra

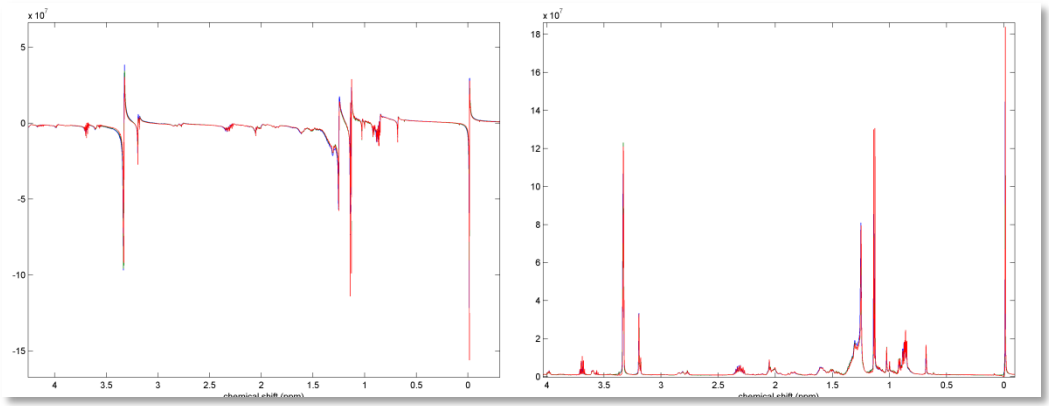
Preparing the NMR spectra for quantitative analysis implies several processing steps including phase correction, baseline correction, shift reference, spectral alignment and line-shape enhancement.

Phase correction

Phase correction is available from the tab “Autophase” in the “Pre-processing” option of the menu bar. Phase correction should be applied before any other processing step and it is an essential requirement for proper performance of the lineshape fitting algorithm. It sets the spectral line in pure absorptive mode.

- ***Select method***: LipSpin provides two different methods to correct zero- and first-order phase:
    - ***Entropy*****:** maximises the entropy of the spectrum. This method provides modest results but it works well for standards and spectra with few peaks. More info in: Chen et al. (2002) (https://doi.org/10.1016/S1090-7807(02)00069-1)
    - ***Flat baseline*****:** minimises the least-squares differences between a horizontal line and the spectral line for the defined regions, considered to have no peaks. An automatic version of region selection is included for this method (an example of the performance in the figure below). ***Tip: for the best performance, select regions disperse along the whole axis scale and of similar length. Flat regions closer to the intense and separated peaks (chloroform and TMS) are the ones more affected by phase distortions, selecting them will provide the best phase correction.*



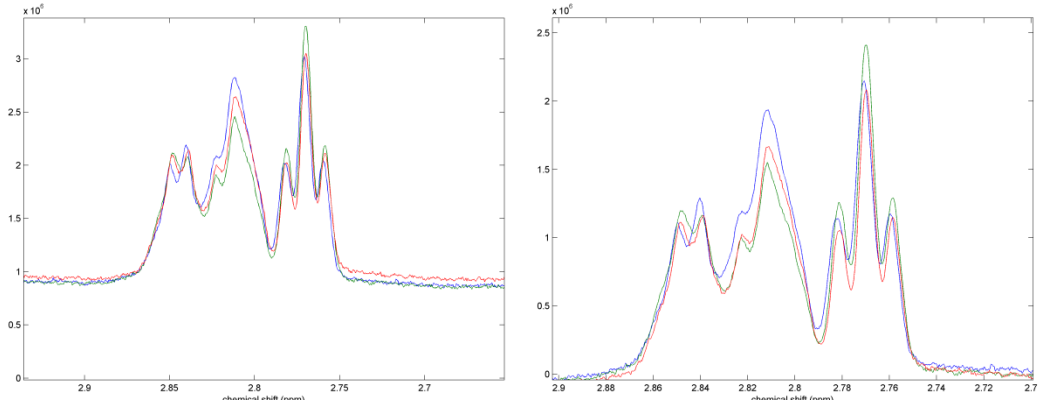


- ***Regions:*** list of regions defined as [high_ppm low_ppm]. This regions are used to find the points used for baseline interpolation.

Baseline correction

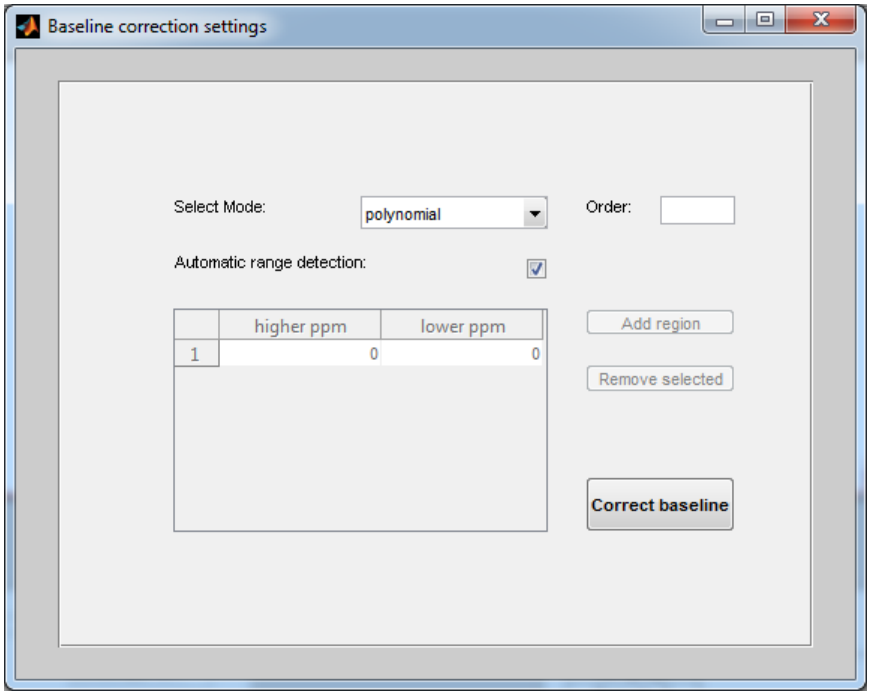
Baseline correction is available from the tab "Baseline" in the "Pre-processing" option of the menu bar. Baseline correction interpolates polynomial functions to a set of points within the user- or automatically-defined regions. Baseline correction can be skipped for lineshape fitting as it includes baseline functions, but it should be thoroughly applied when quantification by bucket integration is used.

- ***Select mode***: select between four interpolation methods, all admitting automatic range detection.
    - ***Median subtraction:*** in the strict sense, this mode is not an interpolation method. It simply subtracts the median of a set of data points to all spectral data points.
    - **Cubic spline** and **cubic Hermite:** spline interpolations implemented with spline and pchip MATLAB functions where each piece is a third-degree polynomial. **Tip: Cubic Hermite provides best results for baseline correction as it reduces oscillation between data points. The example below shows the spectra before (left) and after (right) the baseline correction showing the intensity differences for the red spectrum and the general offset elimination.



- ***Polynomial:*** fits a polynomial of user-defined order to a set of points. Recommended orders: > 5.

- ***Ranges:*** list of chemical shift ranges defined as [high_ppm low_ppm]. These regions are used to find the points used for baseline interpolation.



Shift reference

Shift reference is available from the tab “Chemical shift reference” in the “Pre-processing” option of the menu bar. This function shifts the spectra to align the most intense peak within a region to

the centre of that region. Spectra should be chemical shift referenced so that signal patterns for lineshape fitting are valid between different samples. It requires two parameters:

- ***Chemical shift reference:*** reference ppm where the highest intensity peak has to be positioned.
- ***Tolerance:*** ± ppm around the "chemical shift reference" within the peak is sought.

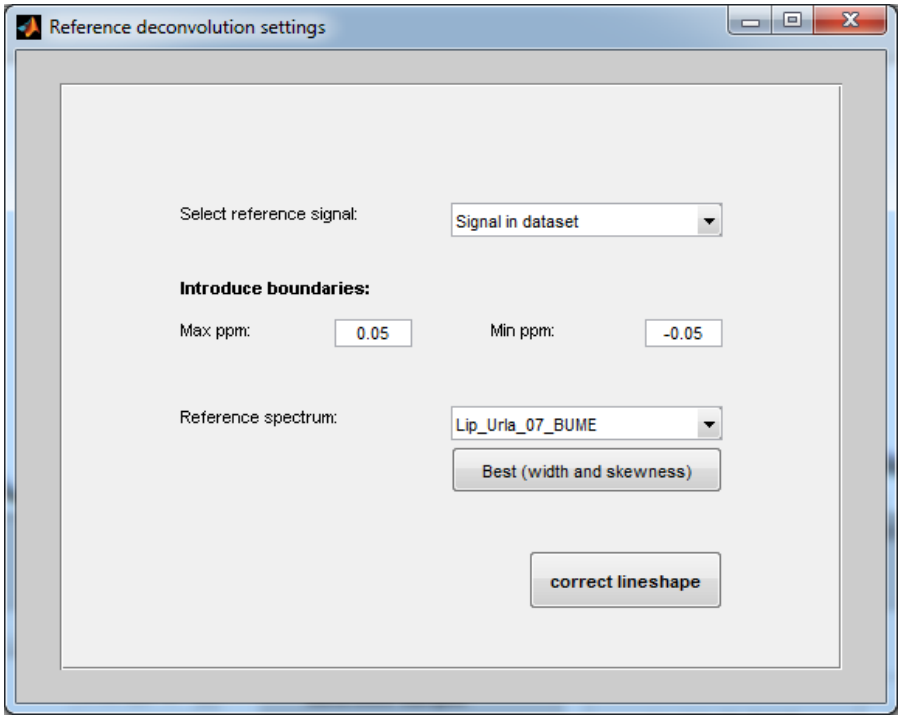
Spectral alignment

Spectral alignment is available from the tab "Align" in the "Pre-processing" option of the menu bar. This function shifts the spectra to maximise correlation (alignment) between samples by using cross-correlation MATLAB function. It allows correcting spectral misalignments of signals from polar groups due to pH or ionic strength discrepancies among samples. It requires three parameters:

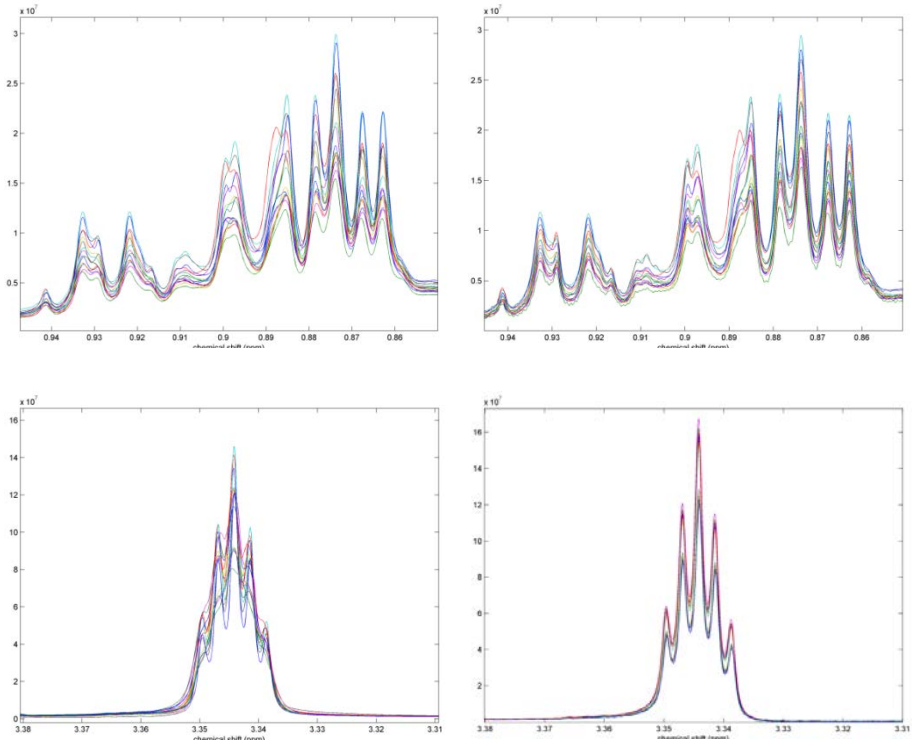
- ***Max ppm:*** maximum chemical shift used in spectral alignment.
- ***Min ppm:*** minimum chemical shift used in spectral alignment.
- ***Shift all spectrum [X] / Only region [ ]:*** If checked the calculated shifts are applied to the whole spectrum. Otherwise the algorithm only shifts the region between "Max" and "Min ppm".

Line-shape enhancement with reference deconvolution

Line-shape enhancement is available from the tab "Reference deconvolution" in the "Pre-processing" option of the menu bar. This optional function allows correcting lineshape distortions due to magnetic inhomogeneities or poor shimming by using a reference peak and providing that this peak has been previously aligned among spectra. ***Note: reference deconvolution could not be recommended for low intensity signal as it decrease S/N and introduce unwanted wiggles in the spectral lines.* More about reference deconvolution in Morris et al. (1997) https://doi.org/10.1016/S0079-6565(97)00011-3.



- Select reference signal:
    - ***Signal in dataset:*** select this option if spectral lineshapes will be corrected by using a signal (preferably solvent signals or singles) within a specific spectrum a as a reference. In such a case, select the spectrum with the signal of lowest line width and symmetrical shape in the "Reference spectrum" list after defining ppm boundaries.
    - ***Synthetic TMS signal:*** generates a synthetic TMS signal to be used as a reference peak. Figures below show the effect of using reference deconvolution with synthetic TMS signal in the methyl region (top) and the chloroform signal (bottom).
- ***Boundaries:*** chemical shift limits (in ppm) for the spectral region used for reference deconvolution.
- ***Reference spectrum (only for "signal in dataset" mode):*** spectrum in the dataset used as reference. The button below allows the automatic selection of the spectrum with the best signal based on lowest line width and skewness.



Standard library

The standard library comprises spectra of chemical standards of lipids that have been previously conditioned and saved with LipSpin to be used as reference templates in lineshape fitting. Using templates is recommended for complex signals such as high-order coupling patterns and multiplets that do not follow the multiplicity rules of first-order coupling (i.e. singlets, doublets, triplets, etc.)

Saving a standard spectrum

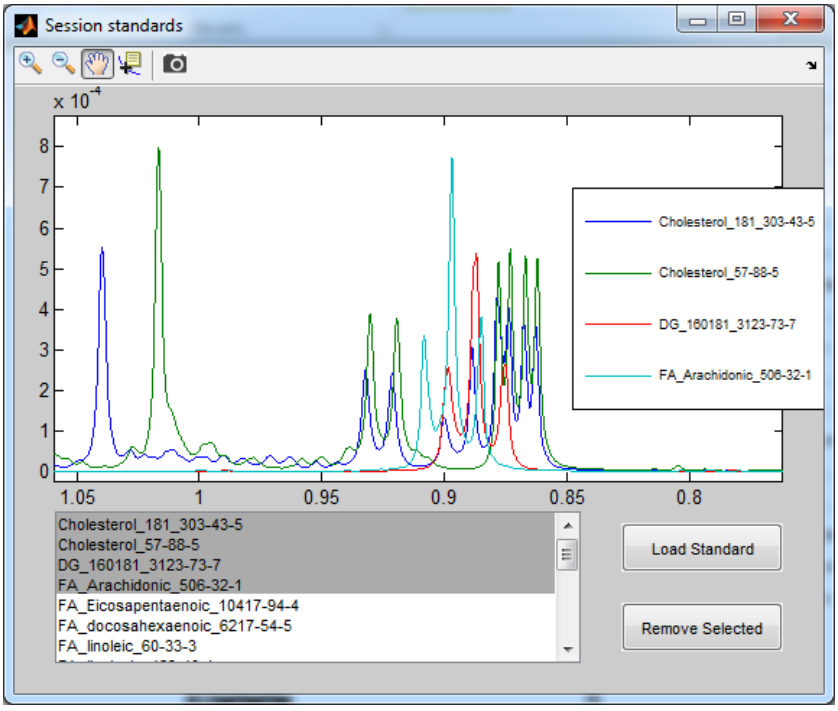
Any spectrum imported in LipSpin can be saved as a standard spectrum by pressing “Save standard” in the tab “Standard spectra” of the menu bar. It requires only one spectrum included in the “Included sample(s)” list.

**Tip: Prepare carefully your standard spectra. Be sure that your new standard spectrum is well phased and does not have baseline offsets in the signals that will be used as templates in lineshape fitting (otherwise baseline will be computed in signal quantification). Be sure they are well-referenced; a good practice is using an internal standard such as TMS in all your standards. Finally, keep in mind that templates from standards are mostly valid for samples acquired under the same experimental conditions (temperature, spectrometer frequency and solvent); otherwise the templates could be no longer valid.

Load standards to current session

Standard spectra need to be loaded before being used in lineshape fitting. By default, all the standards in the "Default path for standard spectra" of the "General options" will be loaded automatically after successfully importing sample spectra. These standards can be inspected from the "Load session standards" in the tab "Standard spectra" of the menu bar.

Additionally, more standard spectra can be loaded and removed from the current session.



Editing signal patterns

Create/load signal patterns

Users can create their own patterns or use patterns from the library of patterns that are supplied with released versions of LipSpin. These patterns were created to be used with lipophilic extracts of human serum and plasma samples but they could be applied to other lipid samples. The "Signal patterns" panel in the main window lists all the patterns that will be applied to the fitting process after clicking the "Run" button. Users can create a new pattern by typing the name in the textbox and clicking the "Create" button.

Setting pattern parameters

The "Pattern settings" panel contains all the parameters that define a signal pattern.

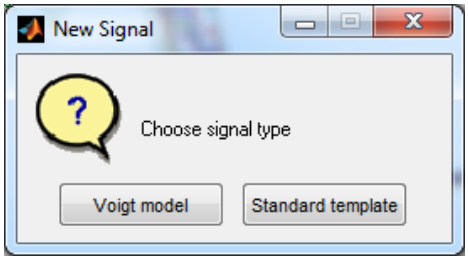
- ***Max ppm:*** left limit (in chemical shift) of the fitting region.
- ***Min ppm:*** right limit (in chemical shift) of the fitting region.
- ***Select mode:***
    - ***Lineshape fitting:*** this mode uses the lsqcurvefit function from the optimization toolbox to adjust the signals defined in the signal pattern to the real spectral line. Baseline parameters are considered in this mode. Use this mode for overlapped signals or isolated signals that may benefit from baseline reduction.
    - ***Integration:*** sums all the point of the spectral line between "Max" and the "Min ppm".
- ***Baseline signals:*** check the signals to add the models to the baseline signal:
    - ***1st poly:*** first-order polynomial (can take negative values).
    - ***Gaussian:*** spatially-distributed broad gaussian lines across the region (only positive).
    - ***Cosine:*** cosine series up to 12th term (only positive).

      ***Note: be careful when using Gaussian and cosine baseline functions as they could model part of non-baseline signals and mislead other signal's quantification. In this case, visual inspection of fitting solutions is a good way of getting insights.*

- ***Signals:*** list with models (Voigt and templates) used for curve fitting the real spectra in the defined region.

Create and edit signals

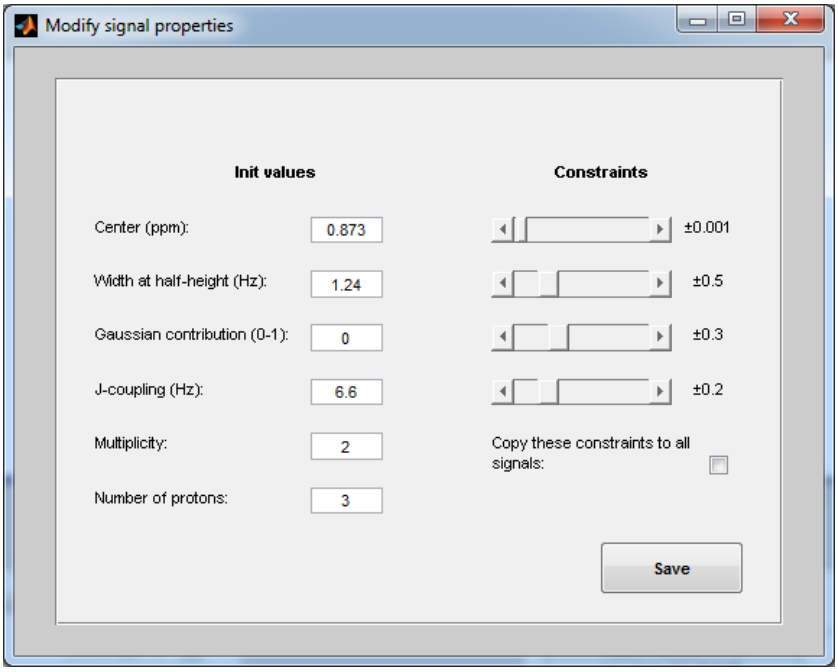
Typing a name (not previously used) and pressing "New" will create a new signal. First, a small window will let choosing between creating a new signal based on "Voigt Model" (following first-order coupling patterns) or a "Standard template".



- ***Voigt models***

Voigt models are based on complex peak structures of Voigt profiles (combination of Lorentzian and Gaussian profiles) following multiplicity patterns of first-order NMR couplings (i.e. singlets, doublets, triplets, etc), with intensity ratios that follow the Pascal's triangle relations. "Init values" defines the initial values used in the optimization process. "Constraints" sets the limits between which each parameter can oscillate in the optimization process.

  - ***Center (ppm):*** Center of the Voigt profile in ppm units.
  - ***Width at half height (Hz):*** full width of the Voigt peak measured at its half height (FWHH). This value usually lies about 1 or 2 Hz. ***Tip: switching the axis scale to Hz will help measuring this parameter.*
  - ***Gaussian contribution (0-1):*** ratio of Gaussian shape in the Voigt profile.
  - ***J-Coupling (Hz):*** distance between peaks in the multiplet.
  - ***Multiplicity:*** number of peaks that form the multiplet.
  - ***Number of protons:*** number of H's that raise the signal.
  - ***Copy these constraints to all signals:*** check and present constraints will be copied to the rest of the signals after pressing "Save".

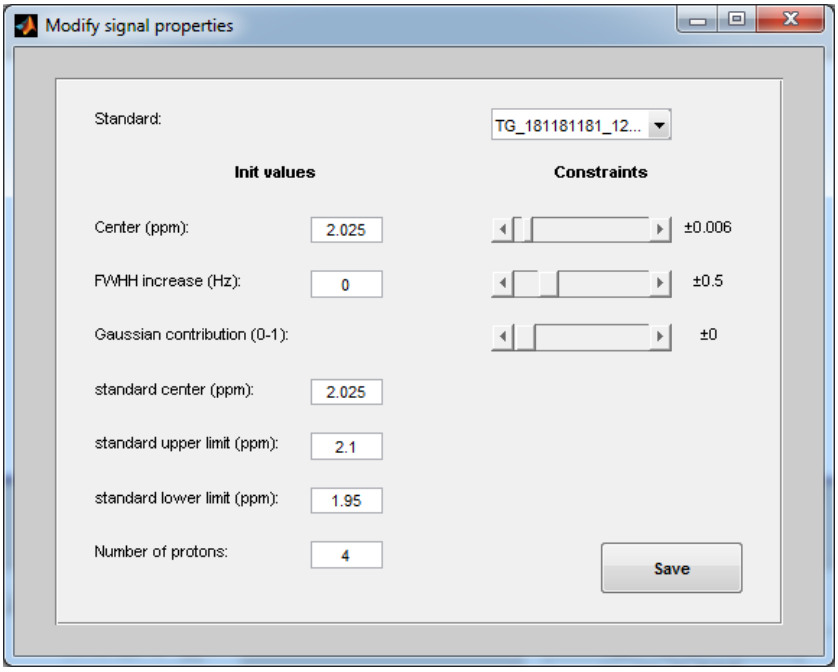


- ***Standard templates***

Standard templates allow using signals from standard spectra as fitting models. This choice is suitable for resonances with complex coupling patterns not following fist-order multiplicity rules.

- ***Standard:*** select the standard spectrum from the list of standards in the current session. ***Note: if this field is void after loading a saved pattern indicates that the standard is not loaded in the current session. Press "Run" button in the main window will show the name of the standard for this signal.*
- ***Center (ppm):*** signal position in sample spectra.
- ***FWHH increase (Hz):*** increase the width at half height of the peaks in the template by the indicated factor in Hz. Typical values: 0-1.
- ***Gaussian contribution (0-1):*** allow increasing the Gaussian shape of the spectral template.
- ***Standard center (ppm):*** signal position in the standard spectrum used as a template.
- ***Standard upper limit (ppm):*** left limit of the region of the standard spectrum used as a template

- ***Standard lower limit (ppm):*** right limit of the region of the standard spectrum used as a template.
- ***Number of protons:*** number of H's that raise the signal.



Signal quantification

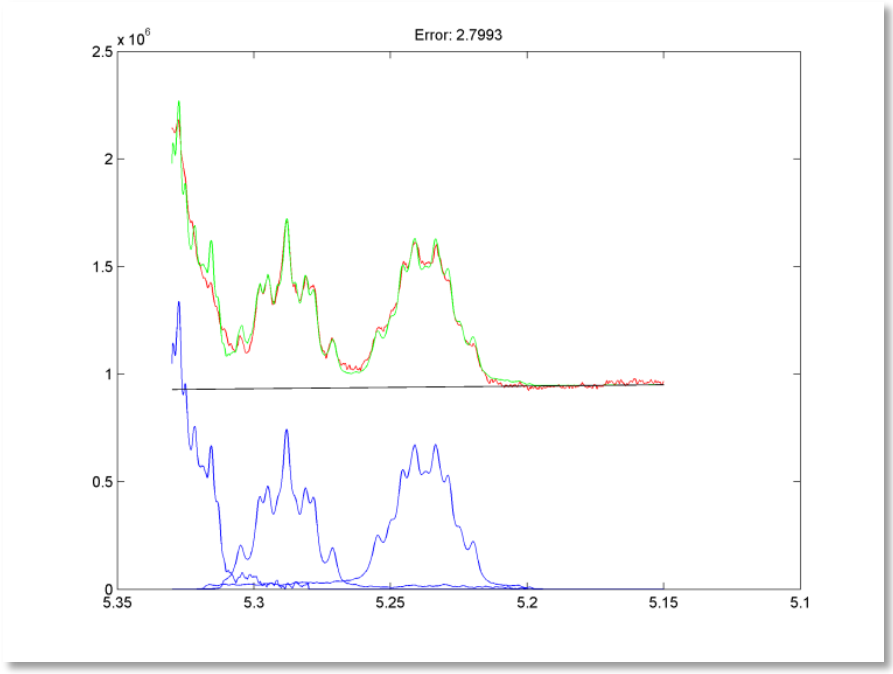
Once the sample spectra are loaded and properly pre-processed, and when signal patterns have been properly configured and loaded, LipSpin is ready for quantifying the signals defined in the signal patterns for the included samples in the "Included sample(s)" list. Pressing "Run" button runs the quantification process in batch mode comprising all the included samples and for each sample all the signal patterns in the "Signal patterns" list.

With "***Correct signal width by***", the quantification process can adapt the "Width at half height" of all the signals for each sample according to the width variations of a previously fitted signal obtained from the "Results" table. This tool can be useful for correcting variations of linewidth

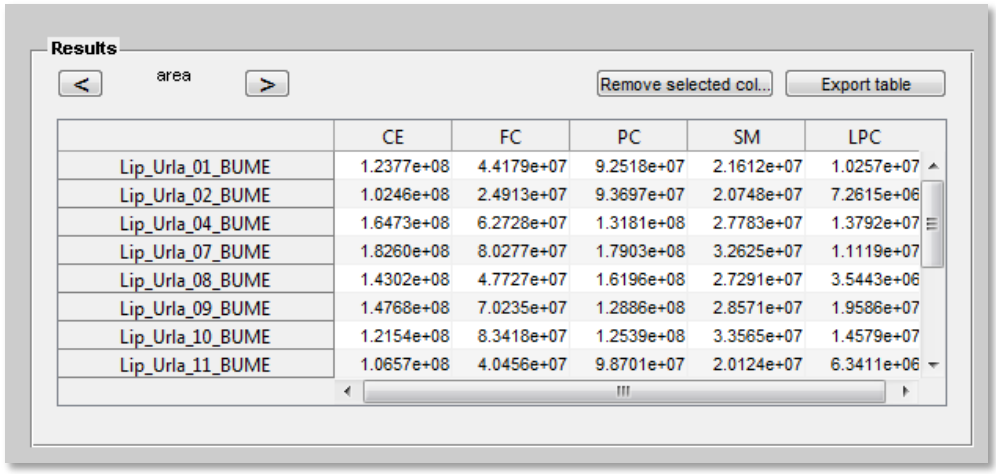
between samples that affect the whole spectrum (i.e. all the peaks in a spectrum). For instance, if linewidth varies severely between samples it could be better to fit first an isolated singlet (e.g. TMS peak at 0 ppm) and then apply FWHH variations of this signal to the rest of the signals, instead of setting large boundaries in the FWHH constraints. This feature needs a sample to be used as a reference, which should be the sample with narrowest signal.

Inspecting quantification results

If an existing "***Figures path***" is indicated, a .png file containing the graphical solution of the fitting process will be saved. In the example below, the red line indicates the sample spectrum, the green line the fitted spectrum, the blue lines the individual fitted signals and the black line the fitted baseline. The graph also shows the %RMSE of fitting as a goodness of fit indicator.



Once all the samples and patterns have been analysed, the "Results" table is updated reflecting the quantified areas (normalised by the number of resonating protons) and several parameters related to the fitted solution for each signal (intensity, centre, FWHH, Gaussian contribution and J-Coupling). Inspecting these parameters could help to optimise subsequent analysis if needed.



	CE	FC	PC	SM	LPC
Lip_Urla_01_BUME	1.2377e+08	4.4179e+07	9.2518e+07	2.1612e+07	1.0257e+07
Lip_Urla_02_BUME	1.0246e+08	2.4913e+07	9.3697e+07	2.0748e+07	7.2615e+06
Lip_Urla_04_BUME	1.6473e+08	6.2728e+07	1.3181e+08	2.7783e+07	1.3792e+07
Lip_Urla_07_BUME	1.8260e+08	8.0277e+07	1.7903e+08	3.2625e+07	1.1119e+07
Lip_Urla_08_BUME	1.4302e+08	4.7727e+07	1.6196e+08	2.7291e+07	3.5443e+06
Lip_Urla_09_BUME	1.4768e+08	7.0235e+07	1.2886e+08	2.8571e+07	1.9586e+07
Lip_Urla_10_BUME	1.2154e+08	8.3418e+07	1.2539e+08	3.3565e+07	1.4579e+07
Lip_Urla_11_BUME	1.0657e+08	4.0456e+07	9.8701e+07	2.0124e+07	6.3411e+06

Finally, the current table can be saved in a .csv file by pressing the ***"Export table"*** option.

UNIVERSITAT ROVIRA I VIRGILI
DEVELOPMENT OF 1H-NMR SERUM PROFILING METHODS FOR HIGH-THROUGHPUT METABOLOMICS
Rubén Barrilero Regadera

UNIVERSITAT ROVIRA I VIRGILI
DEVELOPMENT OF 1H-NMR SERUM PROFILING METHODS FOR HIGH-THROUGHPUT METABOLOMICS
Rubén Barrilero Regadera