UNIVERSITAT ROVIRA i VIRGILI

DESIGN OF LIPID PREDICTION PLS MODELS BASED ON 1H-NMR SPECTROSCOPY OF HUMAN SERUM SAMPLES

TRABAJO DE FIN DE GRADO

Grado de Bioquímica y Biología Molecular

Autor: Santiago Gómez Coso

Tutor académico: Montserrat Pinent Armengol

Tutor profesional: Enrique Ozcariz García

Universitat Rovira i Virgili (URV)

Tarragona, 2021

Content

Final degree project associated with an internship at the company Biosfer Teslab S.L. under the supervision of Enrique Ozcariz García

Acknowledgments

First, I would like to sincerely thank all the members of the Biosfer Teslab team, as each of them has contributed in their own way to my academic development. They have made me feel as if I was one of the family, despite being a trainee student, and that is the most important thing in a work environment.

Specifically, I would like to thank Enrique Ozcariz for all the effort he has put into me, thanks to you I have been able to learn new concepts and to be able to say that I am no longer a "rookie" in statistics. I wish you all the best in your future, and I know that you will be able to achieve whatever you set your mind to, as you are an exemplary person. I would also like to thank Daniel Rodriguez for these last months because you are always willing to help others, and you have helped me whenever I have needed it.

Finally, I would also like to thank my academic tutor Montse Pinent for her advice and recommendations. And for always taking the time to answer my questions and getting involved in my work.

Abstract

Cardiovascular diseases (CVD) currently have a high influence on the world's population and this trend is expected to increase over the coming years, so it is of particular interest to developing new methods to enable an optimal assessment of the cardiovascular profile. Since the irruption of metabolomics, new approaches based on advanced analytical technologies such as nuclear magnetic resonance (NMR) have been applied to predict standard lipids, which are strongly related to cardiovascular risk. The current study shows the development and evaluation of $_1$H-NMR-based partial least squares regression (PLS) models to predict sphingomyelin (SM), polyunsaturated fatty acids, phosphatidylcholine, and total phospholipids in human serum samples. First, the regions with the highest correlation with the concentrations of each lipid variable were selected. Then, after an outlier cleaning process, 70% of the samples involved (n=139) were used to train the models, while the remaining 30% served to validate them. In addition, a z-score standardisation was performed to decrease biological variability between samples, which led to the best estimation models giving correlation coefficients of predicted versus biochemical variables above 0.88 and cross-validation errors (RMSECVs) lower than 0.7. However, the selected model for SM gave an r=0.82, which is consistent with the low correlation of the regions of its NMR spectrum with its concentration. To conclude, based on the optimal performance of the PLS models in this study, it can be stated that NMR allows a robust prediction of lipids in serum matrixes, which could be used to better assess the cardiovascular risk.

Keywords

Cardiovascular disease (CVD)

NMR spectroscopy

Partial least squares regression (PLS)

Metabolomics

Standard lipids

Abbreviations

AA: Arachidonic acid

ALA: Alpha-linoleic acid

BUME: Butanol: methanol

CE: Cholesteryl ester

CVD: Cardiovascular disease

DHA: Docosahexaenoic acid

DOSY: Diffusion-ordered ^{1}H NMR spectroscopy

EPA: Eicosapentaenoic acid

FIDs: Free induction decays

HDL: High-density lipoprotein

HTGL: Hepatic triglyceride lipase

IDL: Intermediate-density lipoprotein

LA: Linoleic acid

LCAT: Lecithin-cholesterol acyltransferase

LDL: Low-density lipoprotein

LPC: Lyshophosphatidylcholine

LPL: Lipoprotein lipase

LV: Latent Variables

MS: Mass spectroscopy

NA: Non-acquired

NIH: National Institutes of Health

NMR: Nuclear magnetic resonance

PC: Phosphatidylcholine

PL: Phospholipids

PLS: Partial least squares regression

PUFA’s: Polyunsaturated fatty acids

RMSECV: Root Mean Square Error of Cross-Validation

SM: Sphingomyelin

SMase: Sphingomyelinase

STOCSY: Statistical total correlation spectroscopy

TG: Triglycerides

TMAO: Trimethylamine N-oxide

TMS: Tetramethylsilane

VLDL: Very low-density lipoprotein

WHO: World Health Organization

1. Introduction

1.1. Lipid importance in CVD

1.1.1. Incidence of Cardiovascular disease (CVD)

Cardiovascular disease encompasses heart and blood vessels disorders, such as ischemic heart disease and stroke (manifestations of atherosclerosis). They are the major cause of mortality, assuming 85% of cardiovascular deaths (World Health Organization 2021).

In accordance with health organizations like the World Health Organization (WHO) and the National Institutes of Health (NIH), cardiovascular disease (CVD) is the leading cause of mortality and morbidity worldwide. Latest studies released by WHO show that 17.9 million people died in 2019, representing 32% of global deaths. In Europe, CVDs are responsible for over 4 million deaths per year (Townsend et al. 2015). Due to the large number of risk factors associated with the disease, this trend is expected to continue, and the annual mortality will grow to almost 24 million by 2030 (Mozaffarian et al. 2016). Some of the behavioural risk factors are cigarette smoking, hypertension, obesity, physical inactivity, and modified lipid metabolism (Sabzmakan et al. 2014).

1.1.2. Lipid classification and metabolism

Lipids play a key role in many essential processes in the body, such as acting like structural components (membrane), energy storage sources, signalling, and protein trafficking (Fahy et al. 2011). Considering that there are two groups from which lipids are formed: ketoacyl and isoprene groups, lipids can be divided into eight categories. These are: (1) fatty acyls, (2) glycerolipids, (3) glycerophospholipids, (4) sphingolipids, (5) saccharolipids, (6) polyketides, (7) sterol lipids, and (8) prenol lipids (Fahy et al. 2011) (*Table 1*).

Table 1. *Lipids categories and examples.* [Adapted from Fahy et al. 2011]

Fatty Acyls	FA	Arachidonic acid
Glycerolipids	**GL**	1-hexadecanoyl-sn-glycerol
Glycerophospholipids	**GP**	Phosphatidyl serine
Sphingolipids	**SP**	Sphingosine
Sterol Lipids	**ST**	Cholesterol
Prenol Lipids	**PR**	Retinol
Saccharolipids	**SL**	Kdo2-lipid A
Polyketides	**PK**	Epothilone D

On the other hand, lipoproteins are a complex consisting of a non-polar lipid core -mostly cholesterol esters (CE) and triglycerides (TG)- surrounded by an hydrophilic membrane. This membrane consists of phospholipids (PL), free cholesterol, and apolipoproteins, which enables tissues to recognize and take up the lipoprotein particle (Engelking 2015). The later, together with the size and lipid composition are the premises that distinguish lipoproteins into 6 main classes (*Table 2*) (Feingold and Grunfeld 2021). The function of these lipoproteins is to transport lipids from the liver to surrounding tissues and vice versa. Lipids are transported in the hydrophobic core of the complex, the surface of which is polar. Therefore, they can be transported in the bloodstream (Feingold and Grunfeld 2021).

Table 2. *Lipoprotein classes.* [Obtained from Feingold and Grunfeld 2021]

Lipoprotein	Density (g/ml)	Size (nm)	Major Lipids	Major Apoproteins
Chylomicrons	<0.930	75-1200	Triglycerides	Apo B-48, Apo C, Apo E, Apo A-I, A-II, A-IV
Chylomicron Remnants	0.930- 1.006	30-80	Triglycerides Cholesterol	Apo B-48, Apo E
VLDL	0.930- 1.006	30-80	Triglycerides	Apo B-100, Apo E, Apo C
IDL	1.006- 1.019	25-35	Triglycerides Cholesterol	Apo B-100, Apo E, Apo C
LDL	1.019- 1.063	18- 25	Cholesterol	Apo B-100
HDL	1.063- 1.210	5- 12	Cholesterol Phospholipids	Apo A-I, Apo A-II, Apo C, Apo E
Lp (a)	1.055- 1.085	~30	Cholesterol	Apo B-100, Apo (a)

Lipid metabolism divides into two main pathways: exogenous and endogenous. In the first one, lipids derive from the diet and are packaged as chylomicrons in the small intestine, carried as TG. Also, lipids provide from endogenous synthesis by the liver (Feingold and Grunfeld 2021). Lipoprotein (a) and VLDL are the first synthesized particles and their interaction with lipoprotein lipase (LPL) and hepatic triglyceride lipase (HTGL) form IDL and LDL particles, consecutively. The later transports cholesterol to peripheral tissues as it has high concentrations of both free and esterified forms (Ouimet, Barrett, and Fisher 2019). On the contrary, HDL particles transport cholesterol from peripheral tissues to the liver -reverse cholesterol transport, shown in *Figure* 1-.

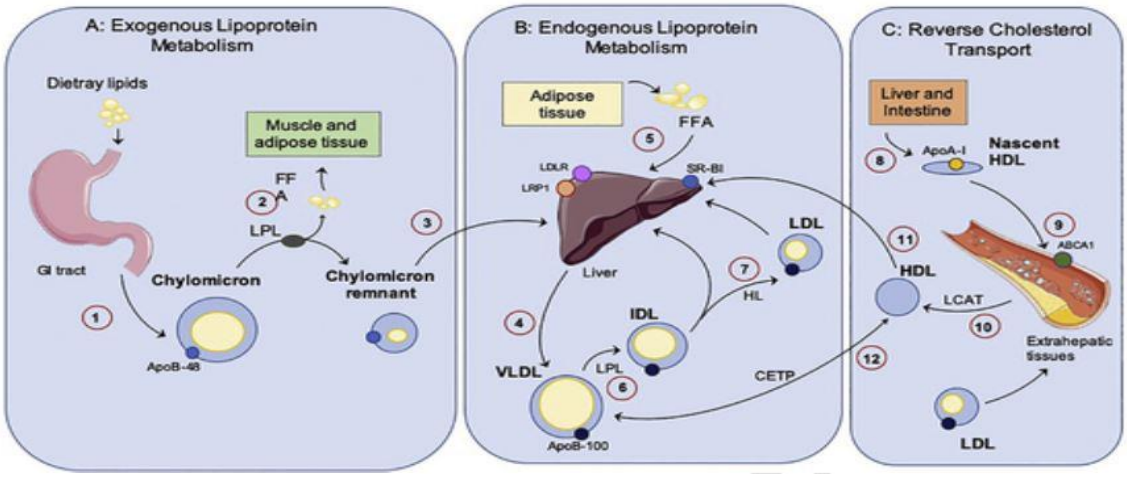

Figure 1. *Exogenous and endogenous lipoprotein metabolism and reverse cholesterol transport*. **A)** Exogenous pathway: Digested lipids enter the intestinal enterocytes to be assembled into chylomicron particles. With LPL hydrolysis, FFAs from the TG core are released for mitochondrial energy sustaining in muscle or for the storage in adipose tissue. **B)** Endogenous pathway: VLDL is formed in the liver from TG and CE and are hydrolysed by LPL to form IDL or VLDL. With HL, LDL particles are formed and internalized by hepatic and nonhepatic tissues and it also can be taken up by macrophages which can lead to plaque formation. **C)** Reverse cholesterol transport: Reverse cholesterol transport is a mechanism by which the body removes excess cholesterol from peripheral tissues and delivers them to the liver, where it will be redistributed to other tissues or removed from the body by the gallbladder. The main lipoprotein involved in this process is the HDL-c. [Obtained from Stemmer et al. 2020]

Lipids and lipoproteins are heavily involved in inflammation as well as in other pathological processes, altering the function of leukocytes, vascular and cardiac cells, contributing to the development of several dysfunctions like obesity and insulin resistance (Frostegård 2013). Furthermore, these dysfunctions are associated with atherosclerosis, a chronic build-up of lipids deposits -mainly cholesterol-, and fibrous elements in large arteries (Frostegård 2013). *Figure 2* shows the steps of atherosclerosis progression.

Many studies have shown that LDL-C and its oxidized modification (oxLDL) drive the initiation and progression of atherosclerosis and arterial thrombosis through diverse ways, like the accumulation of oxLDL in macrophages, triggering pro-inflammatory responses and endothelial activation (Leiva et al. 2015; Poznyak et al. 2021). Likewise, the different subclasses of LDL particles influence cardiovascular risk. Small dense LDL has a higher capacity to infiltrate into the vessel wall; therefore, it is considered more influential on cardiovascular risk than larger particles (Ivanova et al. 2017). HDL-C has been associated with cardioprotective, anti-oxidative, and anti-inflammatory effects as it reduces LDL oxidation as well as LDL/oxLDL accumulation in macrophages (Assmann and Gotto

2004). Consequently, lipids metabolism stands out for its influence on the cardiovascular risk and, not least, on other pathologies currently present.

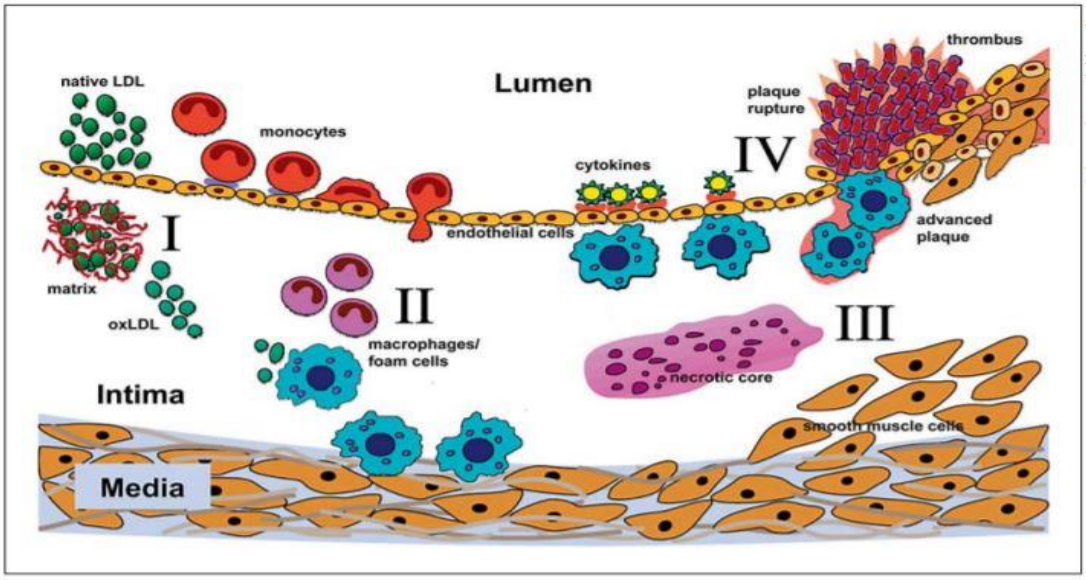

Figure 2. *Atherosclerosis stages.* **I**: LDL particles accumulate in the subendotelial cell layer and are oxidized (oxLDL) by cell-derieved reactive oxygen species. **II**: Growth factors stimulate monocytes differentiation into macrophages and the transmigration across the endotelial cell layer. **III**: Macrophages engulf the oxLDL, leading to foam cell formation, which also leads to the production of mediators that undergo apoptosis or necrosis. A necrotic core is formed. **IV**: The advanced plaque becomes unstable and prone to rupture, resulting in a thrombus or cardiovascular event. [Obtained from Tymchuk et al. 2006]

1.1.3. Lipids' role in CVD

As said before, many lipid families play a key role in CVD. To start with, polyunsaturated fatty acids (PUFAs) are a class of lipids composed of a hydrocarbon skeleton with a carboxyl group at one end. They have more than one double bond per molecule, being classified as ω–3, ω –6, ω –7, or ω –9 depending on the location of the first double bond from the methyl end of the molecule. The first two types must be obtained from the diet, as humans are unable to synthesise them themselves. ω –9, on the other side, are not considered essential, as our body can produce them through an unsaturation in this position (Allayee, Roth, and Hodis 2009). Linoleic acid (18:2ω6, LA) and alpha-linoleic acid (18:3ω3, ALA) are the main PUFAs in the diet. Delta-6,5 desaturases can convert them into longer-chained PUFAs, such as arachidonic acid (20:4ω6, AA), eicosapentaenoic acid (20:5ω3, EPA) and docosahexaenoic acid (22:6ω3, DHA).

According to several studies, the ω–3 fatty acids EPA and DHA are the most relevant in humans as they confer cardioprotective properties with their implications on various mechanisms shown in *Table 3* (Allayee, Roth, and Hodis 2009; Richard et al. 2009; Endo and Arita 2016).

Table 3. *Cardioprotective properties of EPA/DHA.* [Adapted from Endo and Arita 2016]

Effect	Proposed mechanism
Lowered serum triglycerides	reduction in hepatic triglyceride production and lipoprotein assembly
Anti-arrhythmic	modulation of electrophysiological properties of cardiac myocytes
Lowered blood pressure	improved endothelial cell function, vascular relaxation, and arterial compliance
Decreased platelet aggregation	reduction in prothrombotic prostanoids through competition with AA
Decreased inflammation	reduction in 4-series leukotriene production and signaling through competition with AA and leukotriene receptor antagonism, respectively

Both compete with AA to produce eicosanoids, known to have a huge variety of anti-inflammatory and immune-modulating actions. Even though eicosanoids from AA are thought to be considered proinflammatory and prothrombotic, a study (Marklund et al. 2019) has supported the potential benefits of the main dietary ω–6 fatty acids -i.e., linoleic acid- for CVD prevention in various populations. It also has shown that AA is not associated with cardiovascular mortality. In summary, higher *in vivo* circulating, and tissue levels of LA, and possibly AA were associated with a lower risk of cardiovascular events (Marklund et al. 2019).

Also, one clinical trial demonstrated that elevated intakes of PUFAs decrease LDL and increase HDL levels (Ding and Rexrode 2020). Based on these insights, inflammation can be asserted as a link between PUFAs and cardiovascular disease, and, in general terms, both ω–3 and ω –6 fatty acids confer cardioprotective properties which reduce cardiovascular risk.

Sphingolipids and cholines have also a relevant role in CVD. Firstly, sphingomyelin (SM) is a type of sphingolipid composed of two building blocks -long chained bases, usually sphingosine, and fatty acids- using a glycerol-based backbone to attach acyl chains (Kikas, Chalikias, and Tziakas 2018).

Although further studies are needed to understand the molecular pathways by which sphingolipids species are associated with some pathologies, dysregulation of sphingomyelin synthesis and transport is associated with CVD, diabetes, and valvular disease (Kikas, Chalikias, and Tziakas 2018).

Furthermore, lipoprotein aggregation in the vessel wall -one of the precursor steps to atherosclerosis- may be dependent on LDL enzymatic modification induced by the local sphingomyelinase (SMase). More specifically, the SM carried into the arterial wall on atherogenic lipoproteins is transformed into ceramide by the SMase, leading to lipoprotein aggregation (Posio et al. 2005). Ceramides, particularly glucosylceramide and lactosylceramide, accumulate in atherosclerotic plaques and accelerate the development of the anomaly (*Figure 3).*

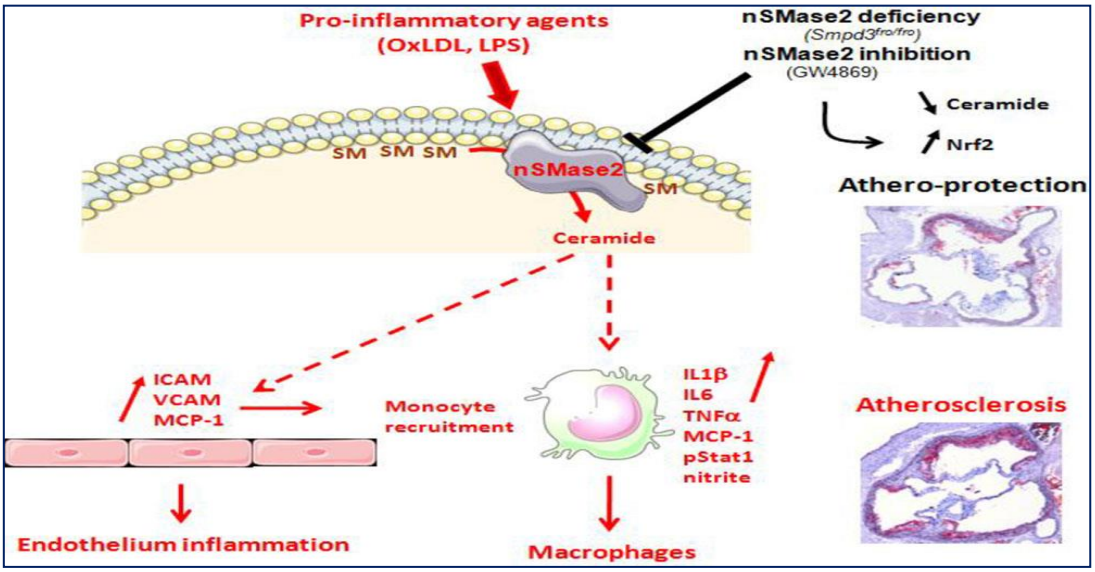

Figure 3. *Involvement of SMase in the development of atherosclerosis.* The inhibition of nSmase2 strongly decreases the development of atherosclerotic lesions by reducing inflammatory responses through a mechanism involving the Nrf2 patwhay. This enzyme converts SM into ceramide, which promotes endothelium inflammation and macrophages accumulation. [Obtained from Lallemand et al. 2018]

In addition, (Subbaiah et al. 2012) demonstrated that SM inhibits LCAT activity, triggering lower HDL-C levels and a more saturated cholesteryl ester (CE) profile, which has been shown to be a risk factor for atherosclerosis and CVD in humans and animals. In contrast to plasma phosphatidylcholine (PC), SM is not degraded by plasma enzymes like lipases or LCAT, but its degradation is dependent on hepatic clearance mechanisms, such as LDL receptor-related protein or proteoglycan pathways (Jiang et al. 2000).

Finally, SM becomes enriched in triglyceride-rich lipoprotein residues, considered proatherogenic. These remnants can be measured by RMN, allowing plasma SM levels to be considered as a marker of atherogenic remnant accumulation (Jiang et al. 2000).

PCs are characterised by having two fatty acids attached to a glycerol backbone with a phosphate group and choline head group like the SMs. Also, they are the main component of mammalian membranes and lipoproteins. A large variety of studies support a link between choline metabolism and major cardiovascular disease (Guasch-Ferré et al. 2017, among others). Dietary PC (meat, egg yolks, etc.) is the primary source of choline, and its catabolism by intestinal microbes leads to trimethylamine N-oxide (TMAO) production, which stands out for being associated with atherosclerosis and thrombosis (Lee et al. 2021). This metabolite can inhibit reverse cholesterol transport and affect platelet activation (Zhu et al. 2016).

Moreover, lysophosphatidylcholine (LPC) is derived from polar surface PC of lipoproteins or cell membrane-derived PC thanks to the reaction catalysed by phospholipase A2. In plasma, high levels of LPC are formed by endothelial lipase and LCAT (Schmitz and Ruebsaamen 2010). This molecule is a component of oxLDL and contributes significantly to its atherogenic activity (Law et al. 2019), discussed in *section 1.1.2*. Thus, PC is relevant as a biomarker of cardiovascular risk, by being a precursor of metabolites that possess proatherogenic and proinflammatory properties.

In view of the above, it can be stated that lipid metabolism is clearly involved in the development of cardiovascular diseases. Therefore, advances in metabolomics are important to correctly understand this relationship and to be able to detect cardiovascular risk earlier.

1.2. Metabolomics

1.2.1. Introduction to metabolomics

It is known that biological processes operate through complex interactions between genes, RNA, proteins, and metabolites. This interaction network is defined as the interactome (*Figure 4*), which is modified by environmental factors and, consequently, disease processes are also modified.

Firstly, advances in genomics made it possible to carry out genome sequencing and led to other fields like transcriptomics and proteomics, which consist of the study of all RNA

molecules in a cell and the quantification of the overall protein present content of a cell or tissue, respectively. Metabolomics is the study of endogenous and exogenous metabolites in biological systems, and it aims to provide information on metabolite abundances in a specific moment (Tan et al. 2016).

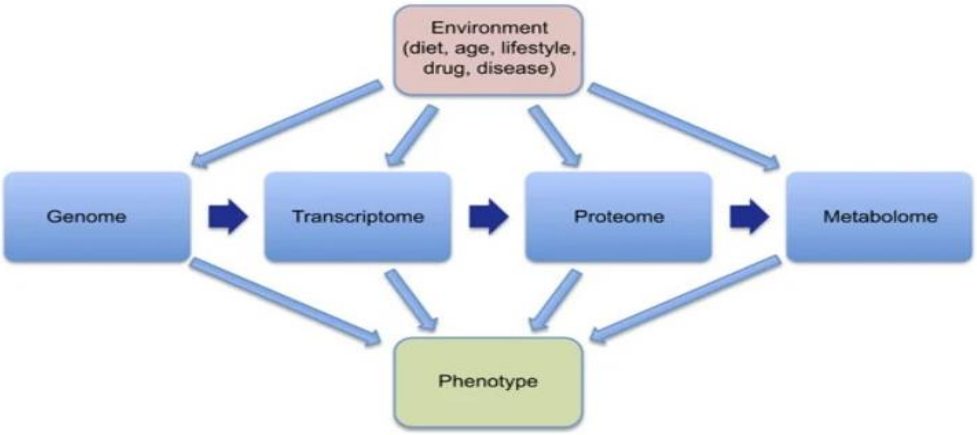

Figure 4. *Interactome.* Interaction between environment and the different biological fields of study. [Obtained from Tan et al. 2016]

Metabolites are usually measured in diverse biological matrixes, such as urine, faeces, or blood. In many cases, metabolites have been described as proximal reporters of disease due to their connection with the pathogenic mechanism. Therefore, to prevent the risk of cardiovascular diseases, a premature arrest of the lipid profile is of great importance. As a branch of metabolomics, lipidomics studies lipid species and their biological roles concerning health and diseases, discovering new pathologic markers (Li et al. 2014).

Regarding lipidomics, chromatography techniques and mass spectrometry (MS) are the most widespread analytical tools because they provide a great characterization of all lipid species, based on their different physicochemical properties (Khoury et al. 2018). Although it is not as sensitive as those mentioned above, proton nuclear magnetic resonance spectroscopy (^{1}H NMR) has some unique features which make this technique interesting for the quantification of the main lipid families in biological matrices. Firstly, the ^{1}H NMR lipid spectrum from most biological matrices offers a fast overview of their major lipid classes (Barrilero et al. 2018). Also, the spectral area is equivalent to the molecular abundance, so it is a quantitative technique and does not require extra steps for sample preparation, like separation or derivatization (Emwas 2015). Moreover, NMR is fast, reproducible, automatable, high-throughput, robust, and non-destructive, so intact

samples can be stored for further analysis. In summary, metabolomics offers insights into the study of CVD by discovering new pathologic markers (Emwas 2015).

1.2.2. NMR spectroscopy

Nuclear magnetic resonance (NMR) is based on the interaction of the nuclei of those atoms presenting an odd number of protons or neutrons (H, F, P, C, etc.) with a magnetic field. During this interaction, the atom has the property of spin under the influence of an external magnetic field B_0. The nuclear spins can adopt two different orientations: one corresponds to the lowest energy level of the nucleus aligned to the magnetic field (-1/2), and the other is associated with the highest energy level of the aligned against one (+1/2).

When the nuclei are irradiated with a radiofrequency pulse (RF) -perpendicular to the magnetic field-, there are transitions between energy levels, which lead to changes in the orientation of the nuclear spins. When the radiofrequency irradiation ends, the nuclei return to their natural state. This change in frequency is collected as Free Induction Decay (FID) signal. Finally, the Fourier Transformation is applied to move from the time domain to the frequency domain and results in a frequency vs intensity spectrum (*Figure 5*). The intensity is proportional to the number of atoms and, therefore, to the molecules contributing to it.

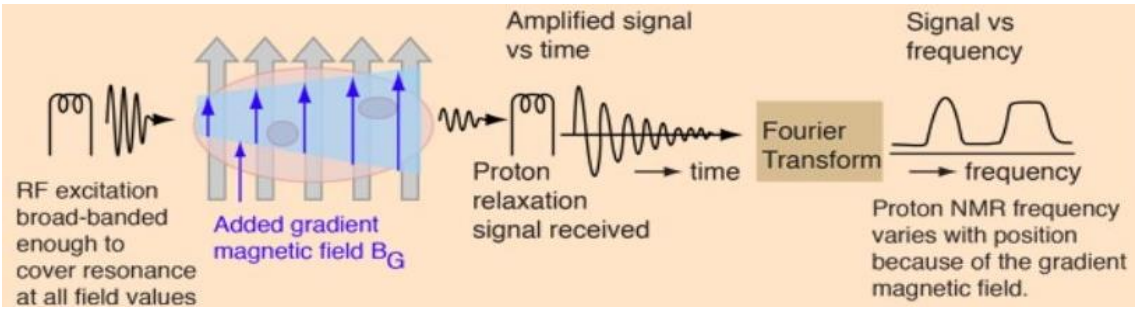

Figure 5. *RF excitation, capturing and Fourier transform proton signals.* [Obtained from http://hyperphysics.phy-astr.gsu.edu/hbase/Nuclear/mri.html, 2022]

1.2.3. Liposcale

Liposcale® test is a novel advanced lipoprotein test based on 2D diffusion-ordered ^{1}H NMR spectroscopy (DOSY). This new approach allows measuring the diffusion coefficients, which are each associated with the different subclasses of lipoproteins, and directly calculating their sizes through the Stokes-Einstein equation (Mallol et al. 2015).

In addition, this technique not only estimates the particle number of the main classes (VLDL, IDL, LDL, and HDL) but also quantifies the concentration of cholesterol and triglycerides of the latter.

As said before, some techniques that allow lipoprotein quantification have some disadvantages compared to ^{1}H NMR. For example, lipoproteins cannot be measured by MS and ultracentrifugation is not only tedious and time-consuming but the highly labile lipoproteins can be altered by high salt concentrations and centrifugal forces (Mora et al. 2009). However, to obtain the concentration of the main lipid families, a lipid extraction process is required as well as another ^{1}H-NMR measurement step applying a pre-saturation (ZGPR) pulse.

2. Hypothesis and objectives

If the profiling of some lipid variables such as PUFAs, SM, or PC could be obtained from the same ^{1}H NMR spectrum used for the lipoprotein profile obtention, more complete and accurate information could be known about the patient's cardiovascular status, as well as the patient's cardiovascular risk.

Then, the main objective of this study was to design generalizable regression models to predict PC, SM, PUFAs, and total PL from the LED resonance spectrum (^{1}H NMR) of human serum samples. Consequently, from the same NMR spectrum, both lipoprotein profiling and the concentrations of PUFAs and PL, which are strongly related to cardiovascular risk, could be obtained.

3. Materials and methods

3.1. Samples and ^{1}H NMR measurements

3.1.1. Samples sets

An initial number of 538 serum-derivated samples were selected from 8 sets belonging to the Biosfer Teslab database. After a process of cleaning outliers, as well as an analysis of which sets were comparable to each other, 139 samples were chosen to be used to calibrate, validate, and optimally evaluate the regression models. In *section 3.2.1*, the process of creating the database is discussed in more detail, as well as the cleaning of the latter.

3.1.2. *Sample processing for lipoprotein analysis*

Sample preparation. 200 µl of plasma sample were diluted with 300 µl of 50 Mm (pH=7.4) and 50 µl of deuterated water (D_2O) were added. All the process was carried out with the liquid handler Gilson robot, which also triggers a mixing process to homogenise the sample optimally.

^{1}H-NMR measurement. ^{1}H-NMR spectra were recorded on a Bruker Avance III 600 spectrometer, with a proton frequency of 600.20 MHz (14.1 T), at 305,95 K and applying a longitudinal eddy current delay (LED) pulse.

Liposcale Test®. Lipoprotein analysis was made by using Liposcale® Test, a novel advanced lipoprotein test based on 2D diffusion-ordered 1H-NMR spectroscopy (Mallol et al. 2015). The methyl signal was deconvoluted by using 9 lorentzian functions to determine the lipid concentration of the large, medium, and small subclasses of the main lipoprotein classes (VLDL, LDL, and HDL), and their size associated diffusion coefficients. Then, the lipid concentration was combined with their associated particle volume to quantify the number of particles required to transport the measured lipid concentration of each lipoprotein subclass. Finally, weighted average VLDL, LDL, and HDL particle sizes were calculated from various subclass concentrations by summing the known diameter of each subclass multiplied by its relative percentage of subclass particle number. The variation coefficients for particle number were between 2% and 4%, and for the particle sizes were lower than 0.3%.

3.1.3. *Lipid extraction for 1H NMR analysis*

Sample retrieval. Once the lipoprotein analysis was performed, the diluted serum was lyophilized for 24 hours. The lyophilized samples were then resuspended in 100 µl PBS (50mM and pH=7.4) in glass tubes. This avoids the obtaining of NMR spectra with overlapping peaks, due to the corrosion of the plastic by the organic solvents which were subsequently used.

BUME method optimised with diisopropyl ether (DIPE). The lipid extraction process with the BUME method (Löfgren, Forsberg, and Ståhlman 2016) was performed with the liquid handler robot -“Agilent Bravo Automated Liquid Handling Platform”-. The first step was to add to each sample tube 300 µl butanol:methanol 3:1 (v/v), 300 µl DIPE:ethyl acetate 3:1 (v/v), and 300 µl deionised water. The samples were then centrifuged at 4ºC for 5 minutes to achieve phase separation, with the organic phase at the top and the

aqueous phase at the bottom, both separated by an interface. This process was repeated twice by adding 300 µl DIPE:EtAc and centrifuging under the same conditions and, finally, the same robot recovered the organic phase in an automated way.

The tubes with the organic phase and the organic solvents were frozen. The later were dried with the SpeedVac for 4 hours to avoid lipid overlap in the spectrum. Manually, the samples were resuspended with 700 µl of deuterated chloroform, methanol, and water solution in a 16:7:1 (v/v/v) ratio. This solution also contained tetramethylsilane (TMS), which serves as a reference for the NMR spectrum.

The ^{1}H-NMR spectra were recorded on a Bruker Avance III 600 spectrometer, with a proton frequency of 600.20 MHz (14.1 T), at 286 K and applying a pre-saturation (ZGPR) pulse. From this spectrum, concentrations of the main lipid families in serum samples are obtained using a software developed in-house by Biosfer Teslab. These include the concentrations of PC, PUFAs, SM, and PL, which will be used to train PLS models.

3.2. Statistical analysis

3.2.1. Database development and outliers' analysis

Database development. The database was created in Excel® (ver. 2202) and then imported into MATLAB version 7.10.0.499 R2010a (MathWorks), where the statistical analysis was performed, as well as the creation of the PLS models. The database was composed of 2 sections. Firstly, one containing the points that make up the LED spectrum, for each sample collected from the ^{1}H-NMR analysis. The second one contained the concentrations of the target variables -PC, SM, PUFAs, and PL- calculated using the in-house-developed lipid profiling software discussed previously.

Outlier analysis. As stated above, this essay initially started with 538 samples. The first step was to remove the samples in which SM and PC were not present/detected and those whose spectra had a pre-processing error. Furthermore, various distribution plots were drawn up on the concentrations of the variables to be predicted, to see which sets could be compared with each other.

3.2.2. Selection of interest regions (inputs)

Statistical total correlation spectroscopy (STOCSY). This technique consists of correlating the variable of interest -concentration of lipids to be predicted- with all the coordinates of the spectrum (Cloarec et al. 2005). Therefore, it was performed to

determine which regions of the spectrum showed the strongest association with the variables to be predicted. The selection of these regions was complemented with bibliography to observe the moieties for the lipids of interest and the correspondence between moieties and their NMR spectra signals (*Figure 6*).

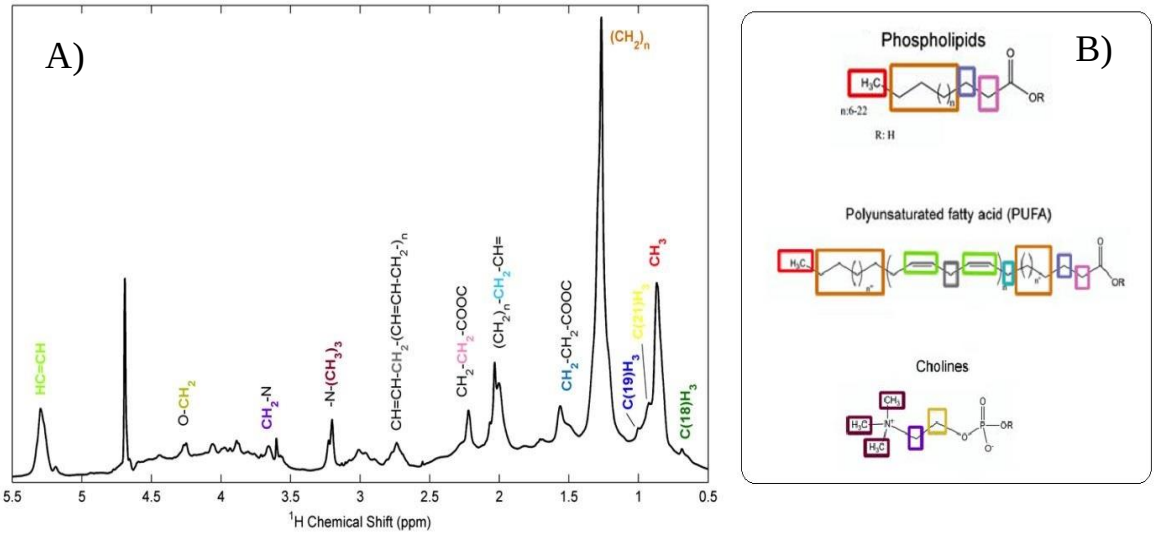

Figure 6. A) *Moieties for principal lipids found in lipoproteins and* ***B)*** *Correspondence between moieties and their NMR spectra signals.* [Obtained from Mallol, R. (2014). Development and evaluation of a novel advanced lipoprotein test based on 2d diffusion-ordered ^{1}H NMR spectroscopy. Universitat Rovira i Virgili]

3.2.3. Partial least square regression models

Partial least square regression (PLS). With the selected regions for each variable as an input, a training set (n=97) including 70 % of the samples was generated randomly, and the remaining 30 % was used to generate the test set (n=42). The first was used to train, strengthen, and teach more diversity to various PLS regression models, so that they could later be validated with the test set. In order to test the effect of the most commonly used pre-processing methods in linear regressions, both autoscaling and mean-centering pre-processing were applied before building the models. In addition, to compare and evaluate the performance of the prediction models, Pearson´s correlation coefficient (r), cumulative variance captured in predictors (X), and Root Mean Square Error of Cross-Validation (RMSECV) were considered. Multivariate analysis was performed with PLS_Toolbox version 5.8.3 (Eigenvector Research Inc.).

Z-score. To reduce the biological variability of the samples so that they can all have the same weight in the models, a standard score (z-score) was carried out for both the points

of the spectrum and the concentrations of the variables per each set. This technique allows us to uniformly transform data of different magnitudes into the same magnitude, called z-score to ensure comparability of the data.

$$Z = \frac{x - \mu}{\sigma}$$

(x: data; μ: mean per set; σ: standard deviation per set)

Ecuation 1. *Z-score calculation.*

Finally, a STOCSY was performed on the normalized data for each variable to confirm an improvement in correlation.

4. Results

4.1. Database development and outliers' analysis

A series of distribution plots concerning the concentrations for each variable per set were performed. This analysis was useful to check which samples were comparable to each other, as it was an important factor to train the model optimally and achieve a correct prediction. As can be seen in *Figure 7*, higher concentrations of both PC, SM, and total PL were observed in sets 5 and 8. These differing results are most likely due to minor modifications of the different protocols used for lipid extraction. Going into more detail, while the remaining sets were extracted following the protocol described in section *3.1.3.*, the samples belonging to these two sets were not lyophilised, so 100 µl of intact plasma was used for the quantification of lipid variables. It appears that lyophilisation may influence lipid quantification, causing an increase in serum phospholipid concentrations. This could be because it is an aggressive method that breaks down lipoproteins and releases phospholipids into the serum medium. Regarding PUFAs, lyophilisation does not have a huge impact on the concentration of the lipidic variable. However, some samples belonging to the sets in question that were outside the concentration range can be seen. Consequently, the samples belonging to sets 5 and 8 were eliminated as they are not comparable with the others, avoiding possible failures in the prediction of the models by training them with concentrations in the same range.

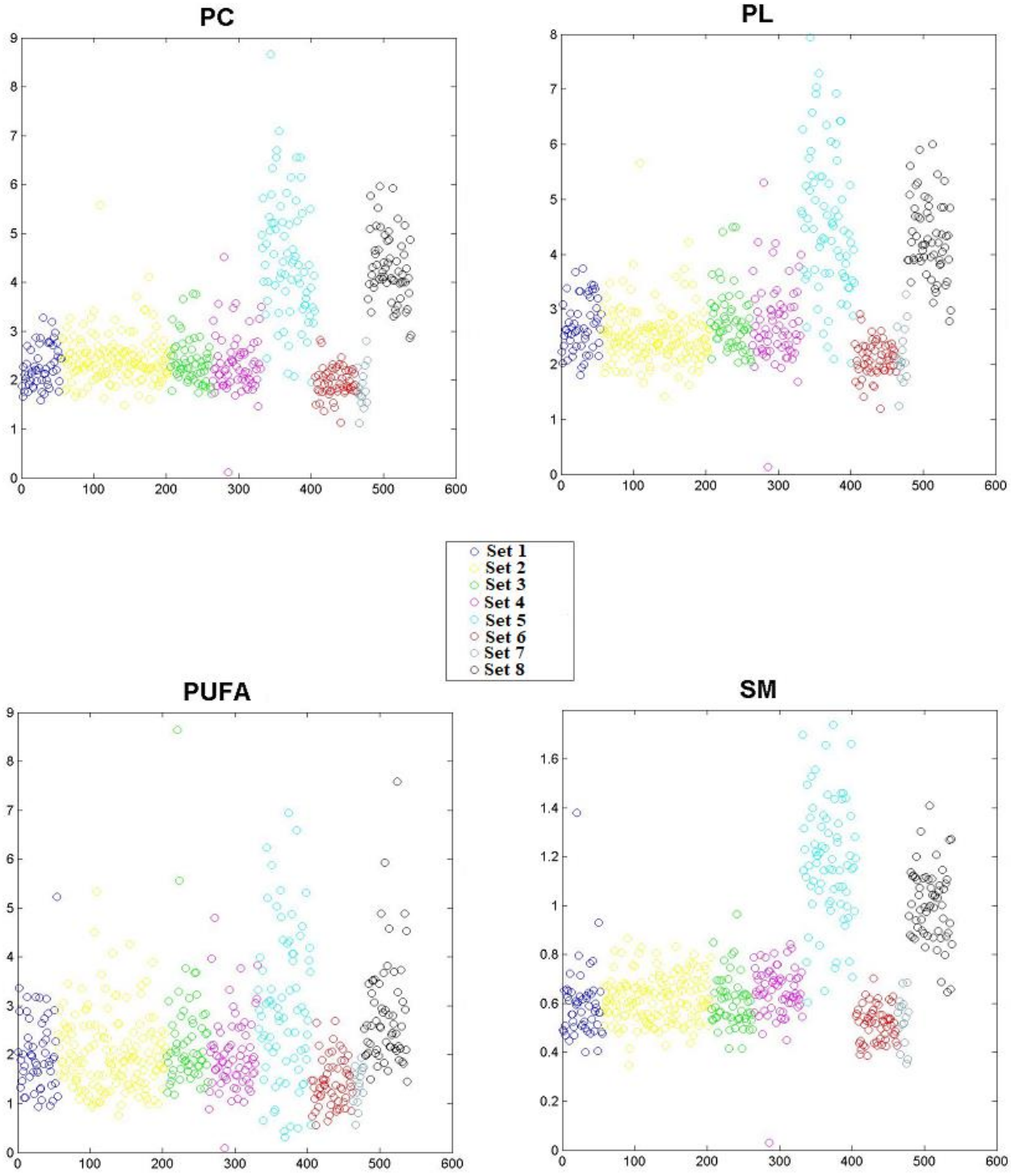

Figure 7. *Distribution plot for target lipidic variables (PC, PL, PUFAs and SM).* The x-axis shows the sample number while the y-axis shows the concentration (mmol/L). In the legend on the middle, we find the colours to identify each set.

4.2. Selection of interest regions (inputs)

Correlation heat maps STOCSY were created as part of the variable selection procedure before the building of the prediction models. As already mentioned, STOCSY provides graphical information on the correlation between the regions of the NMR spectrum and the variables to be predicted. Their interpretation was therefore key to selecting the points in the spectrum that delineate the regions with the strongest correlation and thus using these regions as predictors for more accurate results. The correlation ranges from 1 to -1, where the reddest areas show the highest correlation with lipids, while the bluest areas show a negative correlation.

Taking as a reference the bibliography consulted -shown in *Figure 6*-, it can be stated that, in general terms, the regions with the highest correlation agreed with the moieties of the target lipids. However, some signals showed a very strong correlation but were not associated with protons belonging to the characteristic groups of the different lipids. As an example, despite being a region associated with the signal of the protons forming the HC=CH double bond of PUFAs (5.25 ppm), it showed a very strong correlation in the STOCSY of PC and PL. For this reason, to develop the different models, the regions with the highest correlation were selected rather than only those associated with characteristic groups. As for the protons associated with the most highly correlated signals, we find the following: 1) $(C\mathbf{H_2})_n$; 1.31 ppm (2) $C\mathbf{H_3}$; 0.8 ppm (3) $(CH_2)_n$-$C\mathbf{H_2}$-CH=; 2.05 ppm (4) CH_2-$C\mathbf{H_2}$-COOC; 2.2 ppm (5) **H**C=C**H**; 5.25 ppm.

Looking at the STOCSY of PC, it was initially thought that it would be the variable for which the strongest prediction could be achieved, as it showed the highest correlation intensity. On the contrary, SM was the variable that showed the worst correlation between the points of the spectrum and the concentration. The left of *Figure 8* shows a STOCSY for the target lipid concentration, performed on the data before the z-score was carried out. On the right, we have the STOCSY after normalization, which confirms the improved correlation based on the standardised data for all target lipid variables. Despite being an increase in which a very large correlation is not achieved, the improvement can be appreciated. Consequently, once the z-score was performed, PLS models with a better predictive ability were obtained, as will be shown later in the next section.

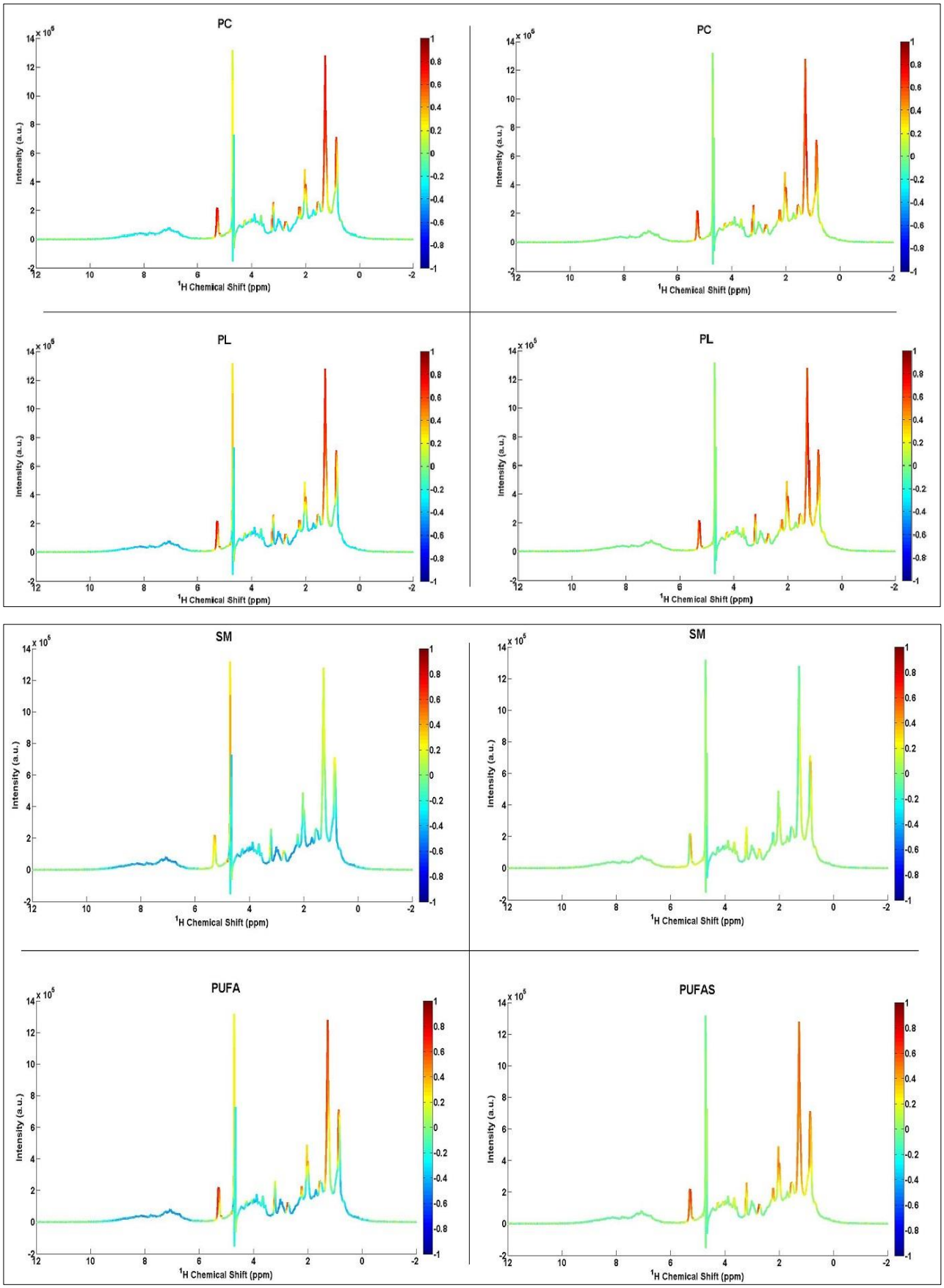

Figure 8. *STOCSY for all lipidic variables. Before z-score (left) and after z-score (right).* The X-axis shows the chemical shift while the Y-axis shows the intensity of the peaks. On the right, the intensity is shown as a colour gradient.

4.3. Partial least square regression models

Both for the creation of the predictive models with normal data and the models with standardised data, a total of 70% of samples were used for calibration and the other 30% for validation. Statistically, LV are variables that cannot be observed directly and try to project data from a higher dimensional space to a lower dimensional space, forming a condensed representation of the data. In this case, the LV were a linear combination of the original variables -the points that make up the spectrum-, achieving a reduction of dimensions. The models were selected considering the maximum variance captured in X, the best regression coefficient for each of them and always trying to minimise the number of LV as much as possible.

The validation performance of the PLS model was assessed by cross-validation splitting the data 10 times, which selects subsets of the training set data to estimate several similar submodels to test the validity of the overall model and to obtain the RMSECV values. RMSECV values should be the lower the better, as it means that the splitting of the data was done correctly, giving a good estimate of how the model built on the data set in question behaves for the unknown cases. As more information is fed to the model, the more likely it is that an effect called overfitting will occur. It is a concept that occurs when a statistical model fits exactly against its training data. When the model is trained with very complex data or a large amount of information, it may start to learn irrelevant information within the dataset, which in turn means that it cannot generalise new data well (validation set). This leads to an inaccurate performance of the algorithm against unseen data, thus defeating its predictive purpose.

4.3.1. PLS pre-standardization

Table 4 shows all relevant statistical parameters belonging to the PLS models with the best prediction for the lipid variables. These include the number of LVs, Pearson´s correlation coefficient (R), % of variance captured in X, the regions selected, and the RMSECV for each case. Pearson's correlation coefficients between the reference concentrations measured by the protocol described *in 3.1.3.* and the predicted concentrations were 0.85 and 0.925 for PC, 0.94 for PUFAs, 0.55 for SM, and 0.86 for PL. All these plots are shown in *Figure 9.*

Table 4. *All models selected for each target lipidic variable.* The top row shows the statistical parameters that were considered for the selection. **Include** is the name by which each different PLS model was designated.

VARIABLES	*INCLUDE*	*LV*	*SELECTED REGIONS*	*% CUM VAR*	*RMSECV*	*R*
PC	Include 5	3	1: (N-$(CH_3)_3$)	90	0.28	0.850
	Include 33	4	7: with highest correlation in STOCSY	95	0.215	0.925
PUFAs	Include 30	4	7: in correspondence with moieties	95	1.2	0.94
SM	Include 31	4	4: with highest correlation in STOCSY	92	0.07	0.55
PL	Include 37	4	7: with highest correlation in STOCSY	94	0.25	0.862

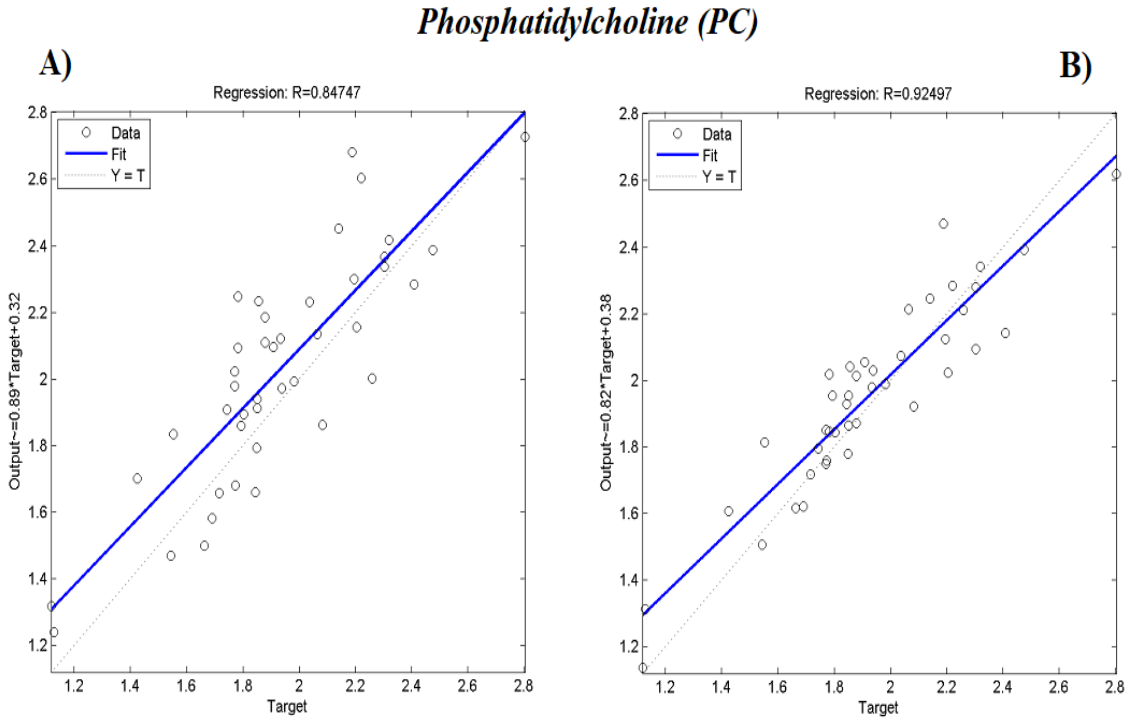

Figure 9. *Correlation plot of predicted versus experimental values for PC.* The dashed line marks the regression line where Y=T, i.e., when the experimental concentration matches the predicted concentration.

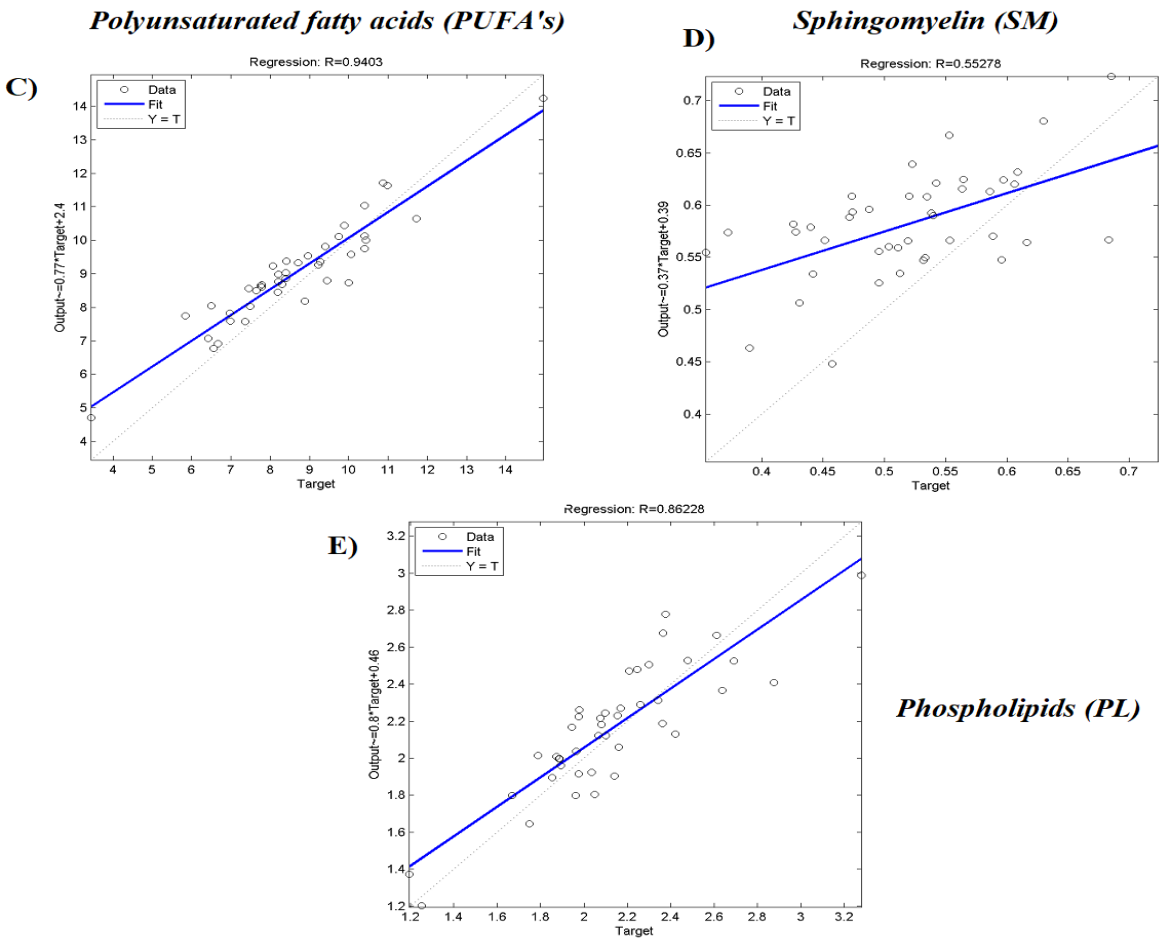

Figure 9 (continuation). *Correlation plot of predicted versus experimental values for PUFAs, SM, and PL.* The dashed line marks the regression line where Y=T, i.e., when the experimental concentration matches the predicted concentration.

4.3.2. PLS post-standardization

Table 5 shows the same statistical parameters as in the previous section. In the case of SM there was a very low correlation in the STOCSY, so a z-score was performed to improve it. Apparently, it was assumed that the low correlation was due to the high biological diversity between sets, as in some cases these were young individuals, adults with diabetes, and other adults with cancer, among others.

The same regions were chosen as inputs for the improved PLS models because they performed well in the originals. However, in general terms, an improvement of the regression coefficients was obtained despite reducing in some cases the number of LV, which is also an improvement, as better results are achieved by adding less information to the models. While it is true that we find an increase in the RMSECV and a slight

reduction in the variance captured in X, an improvement in the regression coefficients of all models can be seen as well as a reduction in the LV. Consequently, ensemble scaling by z-score allowed for more robust models, satisfying the two main requirements. Pearson's correlation coefficients between the reference concentrations and the predicted ones were 0.89 and 0.90 for PC, 0.91 and 0.953 for PUFAs, 0.82 for SM, and 0.91 for PL. All these plots are shown in *Figure 10*.

Table 5. *All models selected for each target lipidic variable after performing an improvement by a z-score scaling.* The top row shows the statistical parameters that were considered for the selection. **Include** is the name by which each different PLS model was designated.

VARIABLES	*INCLUDE*	*LV*	*SELECTED REGIONS*	*% CUM VAR*	*RMSECV*	*R*
PC	Include 5	3	1: (N-$(CH_3)_3$)	90	0.3	0.89
	Include 33	3	7: with highest correlation in STOCSY	94	0.2	0.9
PUFAs	Include 30	2	7: in correspondence with moieties	88	0.7	0.88
	Include 30	4	7: in correspondence with moieties	95	0.4	0.953
SM	Include 31	3	4: with highest correlation in STOCSY	92	0.7	0.82
PL	Include 37	2	7: with highest correlation in STOCSY	85	0.5	0.91

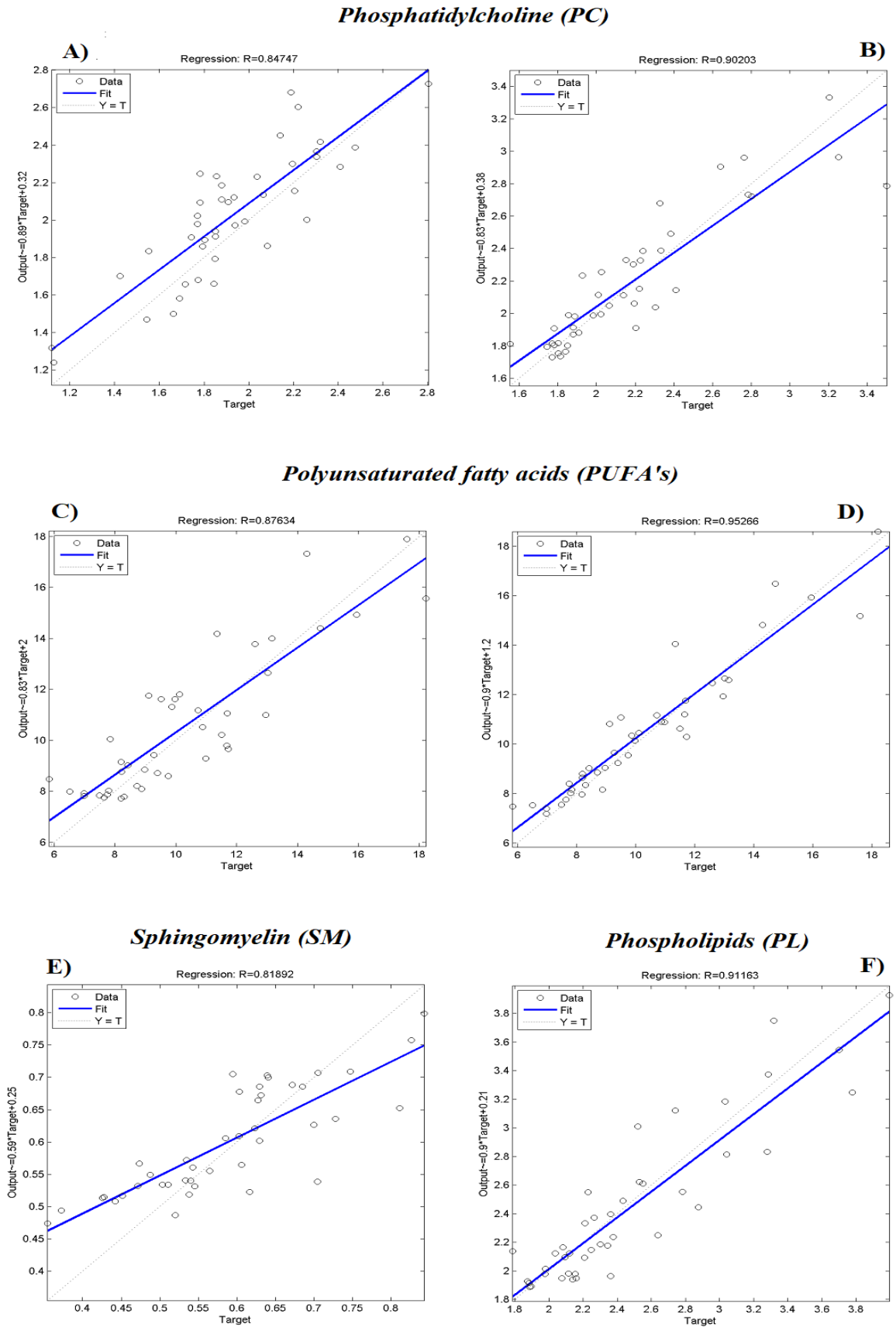

Figure 10. *Correlation plot of predicted versus experimental values for PUFAs, SM and PL.* The dashed line marks the regression line where Y=T, i.e., when the experimental concentration matches the predicted concentration.

5. Discussion

Previous studies have already shown that multivariate statistical methods can be very effective for the prediction of lipids in biological matrices such as serum or plasma (V. Mihaleva et al. 2014; Bathen et al. 2000; Sandra M. et al 2017). The results reported herein confirm the potential of PLS analysis to develop models with sufficient fortress to be applied, despite working with a reduced number of samples after the database cleaning process.

Through trial-and-error testing, many *Includes* were generated from which the results presented were selected based on the statistical parameters present in *Tables 4 and 5*. One of the main objectives was to achieve robust models, but with the minimum number of LV. As has been said, LV are a linear combination of the original variables, in this case, the points that make up the NMR spectra, so it was in the interest of achieving an optimal result with a reduced number of LV, to avoid a model saturation.

Another point to bear in mind is that it has been demonstrated that standardisation by z-score has served to optimally improve the initial models, making them fortress by reducing the biological variability of the research sets with different biological conditions. In agreement with (Petersen M. et al. 2005), two of the main criteria for the selection of the best predictive models were the regression coefficient and the RMSECV error. As can be seen in the PLS models post-standardisation, we ticked all the boxes obtaining a LV reduction, an improvement in the regression coefficients by adding less information and, finally, low RMSECV errors indicating that the data has been split correctly and the model has fitted the unknown information well to validate it. In addition, a considerable improvement in the results for the SM variable can be seen, so we can state that the high biological difference between the samples was causing the low correlation between the NMR spectral points and the SM concentration.

Furthermore, STOCSY can be said to agree with the results as SM showed the worst correlation compared to the other target variables, a factor that was reflected in the low regression coefficient shown in its *Include 31* model (r=0.55 pre-standardisation and r=0.82 post-standardisation). On the contrary, PUFAs, PC, and PL showed areas with a very strong correlation, and the selection of these regions as inputs was key to building models with an outstanding regression coefficient ($r > 0.90$). Despite the latter, a

reduction of the regression coefficient after standardisation from R=0.925 to R=0.90 is observed in the case of PC.

Finally, this study shows that prediction models of PC, PUFAs, SM, and PL perform well despite being trained on with samples from biologically diverse sets. Furthermore, the models prove to be robust with the participation of few samples, so training with more samples, as well as validation with more clinical cases could confirm that it is a generalisable method for the quantification of the target variables of this trial, following the same process that was performed for the evaluation of PLS prediction models for TG and other lipid variables in this study (Barrilero et al. 2015).

Considering the limitations of the study, which are the small number of samples used for the validation of the models and the lack of a factor that allows us to express the results in a range of concentrations, there are two proposals for further validation of PLS models. Firstly, despite having achieved models with great predictive capacity with few samples, models should be validated with a larger number of samples to train and validate them. The other proposal would be to select a general denormalization factor to be scalable to any sample. Finally, as the models have been shown to perform well for PC, PUFAS, SM, and PL, the development of models to predict other variables of pathological interest, such as LPC or MUFA should be tested.

6. Conclusion

This study has evaluated the performance of PLS prediction models based on ^{1}H-NMR spectroscopy. As demonstrated in this trial, RMN allows a robust and optimal prediction of lipids, being a technique that offers a few advantages over other quantification techniques already mentioned. Because of the above, it is of great importance to predict lipids to save time in the lipid quantification process and thus be able to provide a more comprehensive lipid profile in less time. This in turn allows a great improvement in preventing cardiovascular risk, because the link between lipids and cardiovascular disease has been demonstrated in several studies.

In summary, PLS regression models could be used as a general tool for quantifying the lipid profile in metabolomic studies in serum or plasma matrixes, which in turn could be used to better assess the cardiovascular risk.

References

- Allayee, Hooman, Nitzan Roth, and Howard N. Hodis. 2009. "Polyunsaturated Fatty Acids and Cardiovascular Disease: Implications for Nutrigenetics." *Journal of Nutrigenetics and Nutrigenomics* 2 (3): 140–48. https://doi.org/10.1159/000235562.

- Assmann, Gerd, and Antonio M. Gotto. 2004. "HDL Cholesterol and Protective Factors in Atherosclerosis." *Circulation* 109 (23 Suppl 1). https://doi.org/10.1161/01.CIR.0000131512.50667.46.

- Barrilero, Rubén, Miriam Gil, Núria Amigó, Cintia B. Dias, Lisa G. Wood, Manohar L. Garg, Josep Ribalta, Mercedes Heras, Maria Vinaixa, and Xavier Correig. 2018a. "LipSpin: A New Bioinformatics Tool for Quantitative 1H NMR Lipid Profiling." *Analytical Chemistry* 90 (3): 2031–40. https://doi.org/10.1021/acs.analchem.7b04148.

- 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, 271–288

- Cloarec, Olivier, Marc Emmanuel Dumas, Andrew Craig, Richard H. Barton, Johan Trygg, Jane Hudson, Christine Blancher, et al. 2005. "Statistical Total Correlation Spectroscopy: An Exploratory Approach for Latent Biomarker Identification from Metabolic 1H NMR Data Sets." *Analytical Chemistry* 77 (5): 1282–89. https://doi.org/10.1021/AC048630X.

- Ding, Ming, and Kathryn M. Rexrode. 2020. "A Review of Lipidomics of Cardiovascular Disease Highlights the Importance of Isolating Lipoproteins." *Metabolites*. MDPI AG. https://doi.org/10.3390/metabo10040163.

- Emwas, Abdul Hamid M. 2015. "The Strengths and Weaknesses of NMR Spectroscopy and Mass Spectrometry with Particular Focus on Metabolomics Research." *Methods in Molecular Biology* 1277: 161–93. https://doi.org/10.1007/978-1-4939-2377-9_13.

- Endo, Jin, and Makoto Arita. 2016. "Cardioprotective Mechanism of Omega-3 Polyunsaturated Fatty Acids." *Journal of Cardiology* 67 (1): 22–27. https://doi.org/10.1016/J.JJCC.2015.08.002.
- Engelking, Larry R. 2015. "Lipoprotein Complexes." *Textbook of Veterinary Physiological Chemistry*, January, 406–10. https://doi.org/10.1016/B978-0-12-391909-0.50063-3.

- Fahy, Eoin, Dawn Cotter, Manish Sud, and Shankar Subramaniam. 2011. "Lipid Classification, Structures, and Tools." *Biochimica et Biophysica Acta* 1811 (11): 637. https://doi.org/10.1016/J.BBALIP.2011.06.009.

- Fahy, Eoin, Shankar Subramaniam, Robert C. Murphy, Masahiro Nishijima, Christian R.H. Raetz, Takao Shimizu, Friedrich Spener, Gerrit van Meer, Michael J.O. Wakelam, and Edward A. Dennis. 2009. "Update of the LIPID MAPS Comprehensive Classification System

for Lipids." *Journal of Lipid Research* 50 (Suppl): S9. https://doi.org/10.1194/JLR.R800095-JLR200.

- Feingold, Kenneth R, and Carl Grunfeld. 2021. "Introduction to Lipids and Lipoproteins." *Endotext*, January. https://www.ncbi.nlm.nih.gov/books/NBK305896/.

- Frostegård, Johan. 2013. "Immunity, Atherosclerosis, and Cardiovascular Disease." *BMC Medicine* 11 (1): 117. https://doi.org/10.1186/1741-7015-11-117.

- Guasch-Ferré, Marta, Frank B. Hu, Miguel Ruiz-Canela, Mònica Bulló, Estefanía Toledo, Dong D. Wang, Dolores Corella, et al. 2017. "Plasma Metabolites from Choline Pathway and Risk of Cardiovascular Disease in the PREDIMED (Prevention with Mediterranean Diet) Study." *Journal of the American Heart Association* 6 (11). https://doi.org/10.1161/JAHA.117.006524.

- Ivanova, Ekaterina A., Veronika A. Myasoedova, Alexandra A. Melnichenko, Andrey v. Grechko, and Alexander N. Orekhov. 2017. "Small Dense Low-Density Lipoprotein as Biomarker for Atherosclerotic Diseases." *Oxidative Medicine and Cellular Longevity* 2017. https://doi.org/10.1155/2017/1273042.

- Jeyarajah, Elias J., William C. Cromwell, and James D. Otvos. 2006. "Lipoprotein Particle Analysis by Nuclear Magnetic Resonance Spectroscopy." *Clinics in Laboratory Medicine* 26 (4): 847–70. https://doi.org/10.1016/J.CLL.2006.07.006.

- Jiang, Xian Cheng, Furcy Paultre, Thomas A. Pearson, Roberta G. Reed, Charles K. Francis, Min Lin, Lars Berglund, and Alan R. Tall. 2000. "Plasma Sphingomyelin Level as a Risk Factor for Coronary Artery Disease." *Arteriosclerosis, Thrombosis, and Vascular Biology* 20 (12): 2614–18. https://doi.org/10.1161/01.ATV.20.12.2614.

- Khoury, Spiro, Cécile Canlet, Marlène Z. Lacroix, Olivier Berdeaux, Juliette Jouhet, and Justine Bertrand-Michel. 2018. "Quantification of Lipids: Model, Reality, and Compromise." *Biomolecules* 8 (4). https://doi.org/10.3390/BIOM8040174.

- Kikas, Petros, George Chalikias, and Dimitrios Tziakas. 2018. "Cardiovascular Implications of Sphingomyelin Presence in Biological Membranes." *European Cardiology Review* 13 (1): 42–45. https://doi.org/10.15420/ECR.2017:20:3.

- Lallemand, Tom, Myriam Rouahi, Audrey Swiader, Marie Hélène Grazide, Nancy Geoffre, Paul Alayrac, Emeline Recazens, et al. 2018. "NSMase2 (Type 2-Neutral Sphingomyelinase) Deficiency or Inhibition by GW4869 Reduces Inflammation and Atherosclerosis in Apoe -/- Mice." *Arteriosclerosis, Thrombosis, and Vascular Biology* 38 (7): 1479–92. https://doi.org/10.1161/ATVBAHA.118.311208.

- Law, Shi Hui, Mei Lin Chan, Gopal K. Marathe, Farzana Parveen, Chu Huang Chen, and Liang Yin Ke. 2019. "An Updated Review of Lysophosphatidylcholine Metabolism in Human Diseases." *International Journal of Molecular Sciences* 20 (5). https://doi.org/10.3390/IJMS20051149.

- Lee, Yujin, Ina Nemet, Zeneng Wang, Heidi T.M. Lai, Marcia C. de Oliveira Otto, Rozenn N. Lemaitre, Amanda M. Fretts, et al. 2021. "Longitudinal Plasma Measures of Trimethylamine N-Oxide and Risk of Atherosclerotic Cardiovascular Disease Events in Community-Based Older Adults." *Journal of the American Heart Association* 10 (17): 20646. https://doi.org/10.1161/JAHA.120.020646.

- Leiva, E., S. Wehinger, L. Guzmán, and R. Orrego. 2015. "Role of Oxidized LDL in Atherosclerosis." *Advances in Experimental Medicine and Biology* 285 (September): 353–65. https://doi.org/10.5772/59375.

- Li, Min, Li Yang, Yu Bai, and Huwei Liu. 2014. "Analytical Methods in Lipidomics and Their Applications." *Analytical Chemistry*. https://doi.org/10.1021/ac403554h.
- Löfgren, Lars, Gun Britt Forsberg, and Marcus Ståhlman. 2016. "The BUME Method: A New Rapid and Simple Chloroform-Free Method for Total Lipid Extraction of Animal Tissue." *Scientific Reports* 6 (June). https://doi.org/10.1038/srep27688.

- Mallol, Roger, Núria Amigó, Miguel A. Rodríguez, Mercedes Heras, Maria Vinaixa, Núria Plana, Edmond Rock, et al. 2015. "Liposcale: A Novel Advanced Lipoprotein Test Based on 2D Diffusion-Ordered 1H NMR Spectroscopy." *Journal of Lipid Research* 56 (3): 737–46. https://doi.org/10.1194/jlr.D050120.

- Marklund, Matti, Jason H.Y. Wu, Fumiaki Imamura, Liana C. del Gobbo, Amanda Fretts, Janette de Goede, Peilin Shi, et al. 2019. "Biomarkers of Dietary Omega-6 Fatty Acids and Incident Cardiovascular Disease and Mortality: An Individual-Level Pooled Analysis of 30 Cohort Studies." *Circulation* 139 (21): 2422–36. https://doi.org/10.1161/CIRCULATIONAHA.118.038908.

- Mihaleva, V. V., van Schalkwijk, D. B., de Graaf, A. A., et al. (2014). A systematic approach to obtain validated partial least square models for predicting lipoprotein subclasses from serum NMR spectra. Analytical Chemistry, 86, 543–550. doi:10.1021/ ac402571z.

- Mora, Samia, Nader Rifai, Julie E. Buring, and Paul M. Ridker. 2009. "Comparison of LDL Cholesterol Concentrations by Friedewald Calculation and Direct Measurement in Relation to Cardiovascular Events in 27 331 Women." *Clinical Chemistry* 55 (5): 888–94. https://doi.org/10.1373/clinchem.2008.117929.

- Mozaffarian, Dariush, Emelia J. Benjamin, Alan S. Go, Donna K. Arnett, Michael J. Blaha, Mary Cushman, Sandeep R. Das, et al. 2016. "Executive Summary: Heart Disease and Stroke Statistics--2016 Update: A Report From the American Heart Association." *Circulation* 133 (4): 447–54. https://doi.org/10.1161/CIR.0000000000000366.

- Ouimet, Mireille, Tessa J. Barrett, and Edward A. Fisher. 2019. "HDL and Reverse Cholesterol Transport: Basic Mechanisms and Their Roles in Vascular Health and Disease." *Circulation Research* 124 (10): 1505–18. https://doi.org/10.1161/CIRCRESAHA.119.312617.

- Petersen M, Dyrby M, Toubro S, Engelsen SB, Nørgaard L, Pedersen HT, Dyerberg J. Quantification of lipoprotein subclasses by proton nuclear magnetic resonance-based partial least-squares regression models. Clin Chem. 2005 Aug;51(8):1457-61. doi: 10.1373/clinchem.2004.046748. Epub 2005 Jun 16. PMID: 15961551.

- Posio, Pirjo, Mika Ala-Korpela, Matti Jauhiainen, and Petri T Kovanen. 2005. "Sphingomyelinase Induces Aggregation and Fusion of Small Very Low-Density Lipoprotein and Intermediate-Density Lipoprotein Particles and Increases Their Retention to Human Arterial Proteoglycans." https://doi.org/10.1161/01.ATV.0000168912.42941.60.

- Poznyak, Anastasia v., Nikita G. Nikiforov, Alexander M. Markin, Dmitry A. Kashirskikh, Veronika A. Myasoedova, Elena v. Gerasimova, and Alexander N. Orekhov. 2021. "Overview of OxLDL and Its Impact on Cardiovascular Health: Focus on Atherosclerosis." *Frontiers in Pharmacology* 11 (January): 2248. https://doi.org/10.3389/FPHAR.2020.613780/BIBTEX.

- Richard, Doriane, Pedro Bausero, Charlotte Schneider, and Francesco Visioli. 2009. "Polyunsaturated Fatty Acids and Cardiovascular Disease." *Cellular and Molecular Life Sciences: CMLS* 66 (20): 3277–88. https://doi.org/10.1007/S00018-009-0085-4.

- Sabzmakan, Leila, Mohammad Ali Morowatisharifabad, Eesa Mohammadi, Seid Saied Mazloomy-Mahmoodabad, Katayoun Rabiei, Mohammad Hassan Naseri, Elham Shakibazadeh, and Masoud Mirzaei. 2014. "Behavioral Determinants of Cardiovascular Diseases Risk Factors: A Qualitative Directed Content Analysis." *ARYA Atherosclerosis* 10 (2): 71. /pmc/articles/PMC4144369/.

- Schmitz, Gerd, and Katharina Ruebsaamen. 2010. "Metabolism and Atherogenic Disease Association of Lysophosphatidylcholine." *Atherosclerosis* 208 (1): 10–18. https://doi.org/10.1016/J.ATHEROSCLEROSIS.2009.05.029.

- Simopoulos, Artemis P. 2002. "Omega-3 Fatty Acids in Inflammation and Autoimmune Diseases." *Journal of the American College of Nutrition* 21 (6): 495–505. https://doi.org/10.1080/07315724.2002.10719248.

- Soininen, Pasi, Antti J. Kangas, Peter Würtz, Taru Tukiainen, Tuulia Tynkkynen, Reino Laatikainen, Marjo Riitta Järvelin, et al. 2009. "High-Throughput Serum NMR Metabonomics for Cost-Effective Holistic Studies on Systemic Metabolism." *The Analyst* 134 (9): 1781–85. https://doi.org/10.1039/B910205A.

- Stemmer, Kerstin, Brian Finan, Richard D. DiMarchi, Matthias H. Tschöp, and Timo D. Müller. 2020. "Insights into Incretin-Based Therapies for Treatment of Diabetic Dyslipidemia." *Advanced Drug Delivery Reviews* 159 (January): 34–53. https://doi.org/10.1016/J.ADDR.2020.05.008.

- Subbaiah, Papasani Venkata, Xian Cheng Jiang, Natalia A. Belikova, Buzulagu Aizezi, Zhi Hua Huang, and Catherine A. Reardon. 2012. "Regulation of Plasma Cholesterol Esterification by Sphingomyelin: Effect of Physiological Variations of Plasma

Sphingomyelin on Lecithin-Cholesterol Acyltransferase Activity." *Biochimica et Biophysica Acta - Molecular and Cell Biology of Lipids* 1821 (6): 908–13. https://doi.org/10.1016/J.BBALIP.2012.02.007.

- Tan, S. Z., P. Begley, G. Mullard, K. A. Hollywood, and P. N. Bishop. 2016a. "Introduction to Metabolomics and Its Applications in Ophthalmology." *Eye 2016 30:6* 30 (6): 773–83. https://doi.org/10.1038/eye.2016.37.

- Townsend, Nick, Melanie Nichols, Peter Scarborough, and Mike Rayner. 2015. "Cardiovascular Disease in Europe--Epidemiological Update 2015." *European Heart Journal* 36 (40): 2696–2705. https://doi.org/10.1093/EURHEARTJ/EHV428.

- Tymchuk, Christopher N., Jaana Hartiala, Pragna I. Patel, Margarete Mehrabian, and Hooman Allayee. 2006. "Nonconventional Genetic Risk Factors for Cardiovascular Disease." *Current Atherosclerosis Reports* 8 (3): 184–92. https://doi.org/10.1007/S11883-006-0072-2.

- Zhu, Weifei, Jill C. Gregory, Elin Org, Jennifer A. Buffa, Nilaksh Gupta, Zeneng Wang, Lin Li, et al. 2016. "Gut Microbial Metabolite TMAO Enhances Platelet Hyperreactivity and Thrombosis Risk." *Cell* 165 (1): 111–24. https://doi.org/10.1016/J.CELL.2016.02.011.