Pol Torné Charlez

Design and evaluation of linear prediction models for lipidic families based on ^{1}H-NMR LED spectra

Final Degree Project
directed by Dr. Xavier Correig
directed by Sr. Daniel Rodríguez

Degree on Biomedical Engineering

UNIVERSITAT ROVIRA I VIRGILI

Tarragona
2023

Index

Acknowledgments....1
Abstract....2
Abbreviations....3
1 Introduction....5
1.1 Importance of blood lipid characterization for the study and diagnosis of cardiovascular diseases....5
1.1.1 Lipid family description and classification....5
1.1.2 Role of lipids in Atherosclerosis....7
1.1.3 Assessment of cardio-vascular disease risk....9
1.2 Metabolomics and its relevance in healthcare....9
1.2.1 Lipidomics and its capability of providing new pathological markers for cardiovascular disease....10
1.3 Insights on nuclear magnetic resonance spectroscopy, ^{1}H-NMR....10
1.4 Serum compounds profiling through ^{1}H-NMR....11
1.4.1 The three molecular windows....12
1.5 Motivation and context of the project....13
2 Objectives and hypotheses....14
2.1 Objectives....14
2.2 Hypotheses....14
3 Materials and methods....15
3.1 Description of the blood samples sets used and of the lipidic families profiled....15
3.2 Baseline Data: Acquisition of the project's starting framework....16
3.2.1 1D diffusion-edited spectrum acquisition from native serum....16
3.2.2 Lipids quantification in serum extracts....16
3.3 Software tools....18
3.3.1 Matlab employment and starting point establishment....18
3.4 Statistical and multivariate analysis....18
3.4.1 Principal component analysis – PCA....19
3.4.2 Correlation heat-maps – STOCSY....19
3.4.3 Data normalization – Z-Score....20
3.4.4 Partial least squares linear regression technique and PLS toolbox employment....22
3.4.5 The SuperScript tool....24
3.5 Procedure for analysing normalization's impact in the predictive model building 28
3.6 PLS modelling procedure - training and validation processes execution....29
4 Results: development of PLS models for lipidic families quantification....34

4.1 PCA-based exploratory analysis of the sample sets 34
4.2 Study of the normalization's application 35
4.3 Sets selection for the study of each lipid 37
4.4 Selection of the spectral regions most associated with the concentration of each lipid and most efficient latent variables number 41
4.5 Distribution of samples in the training-validation split selection - final PLS models for lipidic families prediction 44
4.6 Validation of the PLS models against the model-building unemployed sets 47
5 Discussion 47
6 Concluding remarks 51
7 References 52
8 Annexes 57

Acknowledgments

This project is the outcome of several months of work, through the course of which many people have contributed to its success.

First, I would like to express my gratitude to Biosfer Teslab, not so much to the company, but to the people, who despite being a trainee student, have embraced and welcomed me as part of their family and granted me the opportunity to undertake and develop this project. Their constant support and assistance have been essential, during all the time spent with them I have grown professionally, knowledge-wise, and personally, I am very thankful to all of you, specially to Dani and Sara, who have closely helped me in the development of this this work.

Also, I would like to appreciate my academic tutor, Xavier, who has guided me with expertise and has provided continuous support through the course of the project.

Finally, I would like to thank my family, for their constant support and encouragement to thrive for success.

Abstract

New approaches for the assessment of cardiovascular disease risk are increasingly put to use and even more noticeably since the rise of metabolomics. This powerful science takes advantage of high-throughput analysis based on advanced analytical technologies such as mass spectrometry and nuclear magnetic resonance, tools that exceed the capabilities of classical methods like the enzymatical measurement of standard lipids for the evaluation of suffering a cardiovascular event. Lipid profiling through nuclear magnetic resonance is a cutting-edge technology that represents a shift in the cardiovascular disease field, providing revolutionary new approaches to lipid assessment that have set new standards, consequently creating a need for new sophisticate and effective tools that fit in these new advancements.

The present study focuses on the optimization of the NMR lipid profiling process, aiming at the suppression of the hitherto indispensable step of serum lipid extraction by designing linear regression models that can quantify lipid families directly from native serum's ^{1}H-NMR LED spectrum.

For a set of twelve lipidic families, being such: total cholesterol, esterified cholesterol, free cholesterol, triglycerides, lysophosphatidylcholine, linoleic acid, saturated fatty acids, ω-6 and ω-7, ω-9, ω-3, docosahexaenoic acid, and arachidonic acid together with eicosapentaenoic acid, a deep and exhaustive process for the development of a linear regression prediction model based on ^{1}H-NMR LED spectra has been successfully conducted.

Furthermore, an automatization of the entire linear regression predictive modelling process through the software MATLAB (MathWorks Inc.) and the Partial Least Squares (PLS) Toolbox has been performed and projected into a novel interactive software that through automatic database cleaning and processing, variables selection through custom-built iterative methods and a step-by-step presentation of the sequential outcomes through illustrative graphical representations, leads the user to the most refined predictive model possible.

Abbreviations

ARA: Arachidonic acid

ATP III: Adult Treatment Panel III

BUME: Butanol:methanol

CE: Cholesterol esters

CPMG: Carr-Purcell-Meiboom-Gill

CVD: Cardiovascular disease

DHA: Docosahexaenoic acid

DIPE: Diisopropyl ether

EC: Esterified cholesterol

EPA: Eicosapentaenoic acid

FC: Free cholesterol

FID: Free induction decay

HDL: High-density lipoproteins

HDL-C: HDL cholesterol

IDL: Intermediate density lipoprotein

LA: Linoleic acid

LDL: Low-density lipoproteins

LDL-C: LDL cholesterol

LED: Diffusion-editing pulse sequence with bipolar gradients and longitudinal eddy-current delay

LMWMs: Low-molecular-weight metabolites

LPC: Lysophosphatidylcholine

LV: Latent variables

MUFA: Monounsaturated fatty acids

NCEP: National Cholesterol Education Program

NMR: Nuclear magnetic resonance

NR: Quantifications normalization

NS: Spectra normalization

NSR: Spectra and quantifications normalization

PC1: Principal component 1

PC2: Principal component 2

PC: Phosphatidylcholine

PCA: Principal component analysis

PL: Phospholipids

PLS: Partial Least Squares

PPPM: Predictive, preventive, and personalized medicine

PUFA: Polyunsaturated fatty acids

R: Pearson correlation coefficient

RMSE: Root mean square error

rRMSE: Relative root mean square error

SFA: Saturated fatty acids

STOCSY: Statistical total correlation spectroscopy

TC: Total Cholesterol

TG: Triglycerides

TMS: Tetramethylsilane

VLDL: Very low-density lipoproteins

WHO: World Health Organization

1 Introduction

According to the findings of the World Health Organization (WHO), it has been determined that cardiovascular disease (CVD), which broadly speaking refers to the group of disorders that affect the heart and blood vessels, is responsible for being the most abundant cause of death on a global scale. In the year 2019 alone, these diseases claimed the lives of an estimated 17.9 million individuals, accounting for approximately 32% of all reported deaths worldwide. Of these deaths, the WHO has confirmed that an 85% were due to heart attack and stroke[1].

Moving to European scale and even though the European Heart Journal has reported an encouraging decrease in CVD in its epidemiological update articles, they have also stated that more than 4 million people still die each year across the continent, 1.4 million of whom die prematurely before the age of 75 [2]. This translated to a global level and out of the 17 million premature deaths reported in 2019, the 38% is found to be caused by CVDs [1]. Regarding the European Union and considering the Eurostat institution statistics latest updates on CVD, a number of 1.71 million of lives were lost to CVD in 2017 [3].

These statistics underscore the significant impact and prevalence of CVDs as a public health concern of utmost importance.

Most of these unfortunate cases, however, could be prevented by first, addressing behavioural risk factors such as unhealthy diet, sedentarism, physical inactivity, abuse of alcohol and tobacco among others [1], and also by the proper prognosis of the threat.

1.1 Importance of blood lipid characterization for the study and diagnosis of cardiovascular diseases

For us to understand the importance of blood lipid characterization when in need of a CVD study, a brief introduction to the lipidic macro biomolecules and their behaviour within the human body is required. This family of biological molecules, essential to any living organism, is principally characterized by the fact that its members are soluble in non-polar solvents, in other words, they cannot be dissolved in aqueous solutions. Lipids function as barriers, receptors, antigens, seconds messengers, sensors, electrical insulators, biological detergents, and membrane anchor points for proteins [4].

1.1.1 Lipid family description and classification

Among the members of the lipidic family biomolecules we find different chemical subclasses such as free fatty acids, triglycerides, phospholipids, glycolipids, and sterols [4]. Additionally, and even though not specifically belonging to this family but due to explanatory purposes, lipoproteins will be included to the presented group, whose members are of high concern for their relevance in this work's field of study.

To properly introduce these subgroups of the lipid family and the lipoproteins, a concise explanation of their role within our living bodies is mandatory.

Free fatty acids or broadly speaking, components of fats, are used as building blocks for many lipids and cell membranes, they provide a higher performance during growth and repair processes. Triglycerides are commonly stored in cytoplasmatic droplets, structure that will be discussed later, and represent a major source of potential cellular energy. Phospholipids are vital for cell membranes, this group is responsible for the formation of the bilayer structure for which cell membranes are characterised, providing the barrier and selective permeability necessary for cellular function. Glycolipids play an important role in the field of cell signalling, cell recognition, and cell adhesion. Sterol lipids, among which we find cholesterol, are generally

in charge of providing cell membrane structural stability, hormone synthesis and bile acid production. Last but not least, lipoproteins are complex structures composed of proteins and lipids that have the vital task of shipping various lipids through the bloodstream [4].

Digging a little deeper into the nature and functionality of lipoproteins, these are the vehicles for lipidic molecules in our body such as phospholipids, triglycerides, and cholesterol, all of which have already been intendedly described to have a certain relationship with membranes, in other words, they are directly related to lipoproteins. These three molecules bind to the structure of lipoproteins in order to be moved. Such structure ends up consisting of a core of triglycerides and cholesterol esters and an amphiphilic surface of free cholesterol and phospholipids. Besides, lipoproteins own a number of apolipoproteins that are found intercalated into surface lipids of the resulting lipid droplet, all along forming a mature plasma lipoprotein [5]. Apolipoproteins act as recognition molecules, facilitating the uptake and clearance of lipoproteins by specific receptors in target tissues, in other words, apolipoproteins determine the specificity and destination of lipoprotein particles [6].

Having described the general structure of a lipoprotein, a classification of the different subclasses of such molecule can be provided. The classes of lipoproteins vary in the major apolipoproteins present and the relative contents of all of the lipid components [7][8].

Chylomicrons are the most abundant lipoproteins found in the body after a meal and during the consequent lipid absorption process, they are produced in the intestine and primarily consist of triglyceride-bearing lipoproteins [7], they transport exogenous lipids.

Very low-density lipoproteins (VLDLs) are normally the main carriers of circulating triglycerides and are synthesised by the liver with a primary function of supplying fatty acids to tissues [7], they transport endogenous lipids from the liver to the cells [8].

Low-density lipoproteins (LDLs) appear as by-products of VLDLs metabolism and function as the main cholesterol carriers to cells in a normal-state scenario [7].

High-density lipoproteins (HDLs) are initially synthesised by the liver and intestine to then mature and become enriched with apolipoproteins acquired from exchanges with chylomicrons and VLDLs. These, as well as LDLs, are main carriers of cholesterol, but from cells back to the liver [7][8].

In-between the change of VLDL and LDL, another class named as intermediate density lipoprotein (IDL) was born, it is considered an intermediate stage between the mentioned lipoproteins [8].

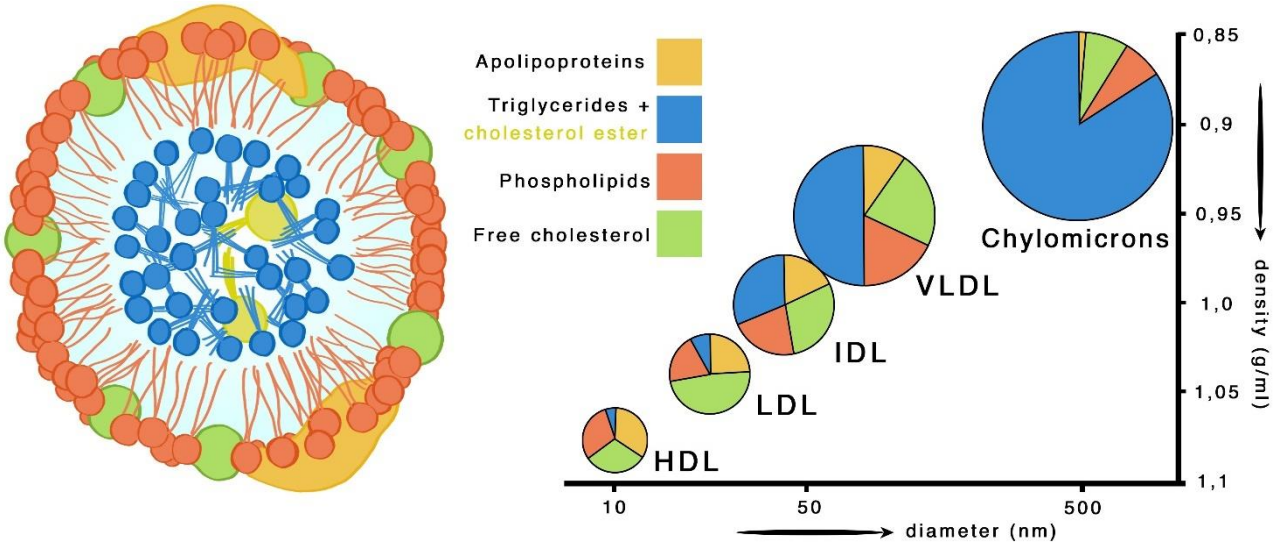


Figure 1. Relationship between size and density among plasma lipoprotein classes [57]

Regarding the apolipoprotein variance among subclasses, chylomicrons, VLDLs and LDLs all incorporate for the most part apoB whereas HDLs fundamentally incorporates apoAI and apoAII, specially in its nascent form. Although the mentioned apolipoproteins are the most abundant for each lipoprotein class, other apolipoproteins are also present [7].

The lipoprotein categories differ in size and density, ranging from the larger and less dense chylomicrons to the smaller and denser HDL. Within each category, a wide spectrum of particles that vary in size, density and relative proportions of lipid and protein is found [7].

A number of studies have focused on examining lipid components in lipoproteins, providing crucial information that opens the possibility of assessing whether a subject's lipoproteins composition is appropriate or not and studying what the different combinations of proportions may indicate [9]. In the following table, a representation of the variance in lipid composition is given for VLDL, LDL and HDL among healthy participants.

Lipid species*	**VLDL (%)**	**LDL (%)**	**HDL (%)**
Phospholipids (PL)	11.1	12	37.4
LPC	2.3	0.4	3
PC	8.5	11.6	31.5
PL – SFA	49	50	46
PL – MUFA	12	12	12
PL – PUFA	34	32	37
Triglycerides (TG)	59	10	6.3
TG – SFA	30	26	27
TG – MUFA	45	47	44
TG – PUFA	18	21	24
Cholesterol esters (CEs)	21.6	74.5	54.7
CE – SFA	13	12	12
CE – MUFA	26	22	22
CE – PUFA	58	67	61

Table 1. Lipid composition of VLDLs, LDLs, and HDLs [9]

*For lipid species, a percentage indicating the weight of each lipid species over weight of total lipids within each lipoprotein is given. Abbreviations: lysophosphatidylcholine (LPC), phosphatidylcholine (PC), saturated fatty acids (SFA), monounsaturated fatty acids (MUFA), and polyunsaturated fatty acids (PUFA).

1.1.2 Role of lipids in Atherosclerosis

Returning to the CVD field and for as to establish a link between it and lipid characterization, the atherosclerosis pathology must be introduced. Atherosclerosis refers to the chronic inflammation derived from the complex interactions between modified lipoproteins, monocyte-derived macrophages, components of innate and adaptive immunity and the normal cellular elements of the arterial wall [7]. The development of this disease ultimately leads to the synthesis of plaques infiltrated in the arterial lumen; process known as atherogenesis. Furthermore, the areas that develop an atherosclerotic plaque may be susceptible to injuries such as ruptures or erosions, provoking intravascular thrombosis resulting in the acute clinical complications of myocardial infarction and stroke [7].

Amidst the numerous factors contributing to atherogenesis, elevated cholesterol levels have been determined to have a significant role in both the initiation and progression of the disease, and even more, the clinical consequences of infarction, stroke, peripheral vascular disease, and heart failure [7].

Encouraging results from experiments with animal models show that, even though a series of genetic and environmental factors have been identified to have great influence on the formation of the lesion, atherogenesis will not occur in these models in the absence of greatly elevated plasma cholesterol levels (>800mg/dl) unless a direct arterial injury is executed [7].

For the development of atherosclerosis in human beings, hypercholesterolemia also appears to be obligatory, nevertheless, atherogenesis in humans is a slow process that occurs over many decades and consequently the threshold level of plasma total cholesterol that has to be exceeded to express clinically relevant disease is reported to be much lower than in animal models. Evidence shows that atherosclerotic clinical events are uncommon in humans with lifelong very low plasma cholesterol levels [10].

Consequently, cholesterol-lowering therapy suggests that only by reducing cholesterol levels, atherosclerotic disease could be avoided, halted, or reversed. This would be true if the reduction of cholesterol levels could be performed long before the usual time of development of the disease, nevertheless, conventional hypolipidemic therapy lacks the capacity of reducing cholesterol values enough for all kinds of patients and drug therapy is only initiated after disease episodes are noted, resulting in a non-effective mechanism to prevent atherosclerosis [7].

In spite of the disadvantages expressed regarding the treatment of the disease, an optimistic and extremely reliable attribute, total plasma cholesterol levels, has been established as a direct indicator of potential CVD development, and has been recognized worldwide.

Still, there is a clarification to be made, in reality it's the lipoproteins who interact with the arterial wall and set in motion the cascade of events that lead to atherosclerosis. Therefore, the quantification of total cholesterol is an indirect estimation of the lipoproteins, which transport the bulk of cholesterol in plasma and are indeed the most atherogenic. For the most part, and as previously described as the main cholesterol carriers, the lipoprotein classes aimed at studying are HDLs and LDLs [7].

For as to provide an overview of the atherogenesis process, LDL's abundance in plasma, among other factors, promote endothelial disfunction leading to an increased transport into the intima or medium layer of the artery, where LDL meets proteoglycans and binding happens, greatly prolonging its residence time. Under this scenario, LDLs are susceptible to a number of modifications that enhance macrophage uptake, eventually causing foam cell formation and initiation of the cascade events that ultimately result in the progression of an atherosclerotic lesion [11].

Going a little bit further and reminding ourselves of the variance that a lipoprotein's lipid composition proportions can present, some studies have been conducted aiming at addressing this variance's importance and relationship with the risk of experiencing a cardiovascular event. Since quantifying LDL-C has been proved to be an indirect measure of the lipoproteins that carry it [7], and its clinical application is recognized globally, it is logical to think that the capability of quantifying other lipidic components of lipoproteins could also be of great clinical relevance.

Indeed, such studies have concluded that direct association between lipoprotein's lipidic composition and cardiovascular events exists and is a promising research direction for developing and improving the assessment of CVD risk [9]. In the following section insights on this area are presented.

1.1.3 Assessment of cardio-vascular disease risk

Currently, as indicated by the National Cholesterol Education Program (NCEP) in the guidelines of the Adult Treatment Panel III (ATP III) [12], one of the main methods for assessing an individual's risk of developing CVD is the measurement of standard lipid concentrations in fasting blood. Standard lipids are a series of lipidic molecules that include total plasma cholesterol (TC), total plasma triglycerides (TG), HDL cholesterol (HDL-C) and LDL cholesterol (LDL-C) [12].

Conventionally, in routine biochemical assays, enzymatic methods are used to quantify TC, TG and HDL-C, this last one is possible only after a process of precipitation. On the other hand, LDL-C is calculated through the Friedwald equation [13], only valid when TG concentration is no more than 400 mg/dL. Non-HDL-C is the outcome of subtracting HDL-C from TC, and this measurement is not limited by TG concentration.

Improvements in this area are still to be done in view of the possibilities that determining the lipidic composition of lipoproteins may bring, as has been described in the previous section. Nevertheless, owing to the unquestionable relationship between lipidic composition and CVD, several studies in this field have been done, and understanding of the diversity of roles that lipids play in numerous metabolic pathways has increased. Aside from their role as organelles and membrane's building blocks and energy storing entities, lipids perform fundamental functions in signalling and metabolic regulation, and as has been reported in such studies, their capabilities to execute different functions is explained by their variance in structure [14]; however, conventionally available analytical procedures have limited the number of lipidic species that could be reliably assessed in clinical laboratories owing to restraints imposed by the nature of the techniques [14].

Again, conventional blood tests aiming to assess CVD risk are limited to providing standard lipids concentrations, but as has been reported in various studies [9][14], it is of great clinical interest and also technically feasible through emerging techniques to quantify human plasma lipids with greater depth and accuracy.

1.2 Metabolomics and its relevance in healthcare

To fulfil the need of evolving in the quantification of the lipidic composition of individuals, the metabolomics field springs into action.

Metabolomics comprises a series of important aspects to phenomics, which is the systematic measurement of the physical, physiological, and biochemical traits [15]. These aspects include the theory and methodology to study metabolome, including identification of biochemical and molecular characteristics of metabolome, characterization of interactions among various metabolites or between metabolites and genetic/environmental factors, and evaluation of biochemical mechanisms related to a given condition such as different pathophysiological processes [16].

To acknowledge the scope of this science, let us know that the metabolome includes all metabolites derived from nucleic acids, proteins, lipids and sugars in any given cell, tissue, biological system, or body-fluid [17]. The metabolites in a metabolome interact mutually in enzymatic reaction systems to form metabolic network systems [18]. Changes in metabolites are associated to a number of factors such as internal, external, genetic, environmental, drug, or dietary factors. Metabolomic variations can be translated to a reflection of the status of physiological and pathological processes, monitoring of the progression of a disease, and prediction and assessment of drug effects compared to the baseline of metabolic profiles, which benefits for disease stratification, and personalized/precise medicine in the context of predictive, preventive, and personalized medicine (PPPM) [19].

In light of all of this, it is no surprise that further insights into the study of diseases are provided by Metabolomics, and the CVD field is no exception.

In recent years, the reliability of only utilising standard lipids quantification and analysis for assessing all individuals CVD risk has been put in doubt, owing to the fact that a large proportion of individuals under CVD treatment and also undiagnosed ones, who express normal LDL-C levels, may end up suffering a cardiovascular event [20], confirming that there's still room for improvement in the assessment of CVD risk.

To fill this gap, several studies have shown that Metabolomics is of great use in the assessment of CVD risk owing to its capability to determine new pathological markers, which have already been identified and compared to the conventionally used standard lipids [21].

1.2.1 Lipidomics and its capability of providing new pathological markers for cardiovascular disease

Determining lipid class composition and the lipidic species pattern for lipoprotein fractions through metabolomic-derived techniques have been proved to be methods that provide new pathological markers for assessing CVD risk, on account of various research projects on the field [9][14][21][22].

To refer to this new study of the structure and function of the complete set of lipids, also known as lipidome, in a given cell or organism as well as their interactions with other cellular components that metabolomics has offered, the term Lipidomics was born [23]. This large-scale study of pathways and networks of cellular lipids presents new emerging techniques that allow for the high-throughput profiling of the lipidome, such as liquid chromatography – mass spectrometry, shotgun lipidomics and nuclear magnetic resonance (NMR) spectroscopy.

1.3 Insights on nuclear magnetic resonance spectroscopy, ^{1}H-NMR

In recent times, the use of NMR has increased in lipidomic assessment, but it is no recent breakthrough, it has been widely acknowledged in the field that lipid moieties of lipoproteins in plasma are highly visible in NMR fingerprints [24][25].

The technique itself relies on the inherent nuclear spin of atomic nuclei. Such phenomenon is produced when samples are placed under a strong and uniform magnetic field, aligning the nuclei of certain atoms, including ^{1}H, ^{13}C, ^{15}N, and ^{31}P. Determined by the atom nuclei aimed at studying, a radiofrequency pulse of a specific frequency, called Larmor frequency, is selected, and through the sample's exposure to it , NMR stimulates the nuclear spin of atoms and then records the electromagnetic radiation released following nuclei relaxation into the initial spin state induced by the magnetic field, this captured signal is known under the name of free induction decay (FID). The resonance frequency of the energy released by the atomic nuclei, recorded in the FID signal, reflects the microenvironment of adjacent nuclei [9][26].

The obtained FID signal is in the time-domain and doesn't provide direct valuable information. For as to obtain useful spectral information a conversion from time-domain to frequency-domain is needed, and the Fourier transformed is the indicated method. This shift allows for an analysis of the frequencies present in the signal and for the extraction of information about the chemical shifts, coupling patterns, and relaxation properties of the NMR-active nuclei [27].

To clarify, spins in a molecule experience slightly different magnetic environments, also meaning slightly different Larmor frequencies, depending on the surrounding nuclei. Such difference enables the detection of a set of signals dispersed along the frequency axis of an NMR spectrum, which are related to the surrounding functional groups of the studied nuclei (e.g. methyl, methylene, allyl, etc.) [27].

Thus, NMR has been widely used for structural elucidation. In biofluids such as saliva, urine, serum, and others, the application of NMR is most frequently through the analysis of proton ^{1}H, owing to its high sensitivity, fast relaxation, natural abundance, and its nearly ubiquitous presence in organic metabolites [27]. The success also arises from the fact that the sample is physically isolated from the NMR instrument, consequently being recognized as a non-destructive technique that in our field of study, can preserve lipoproteins in plasma unlike other techniques.

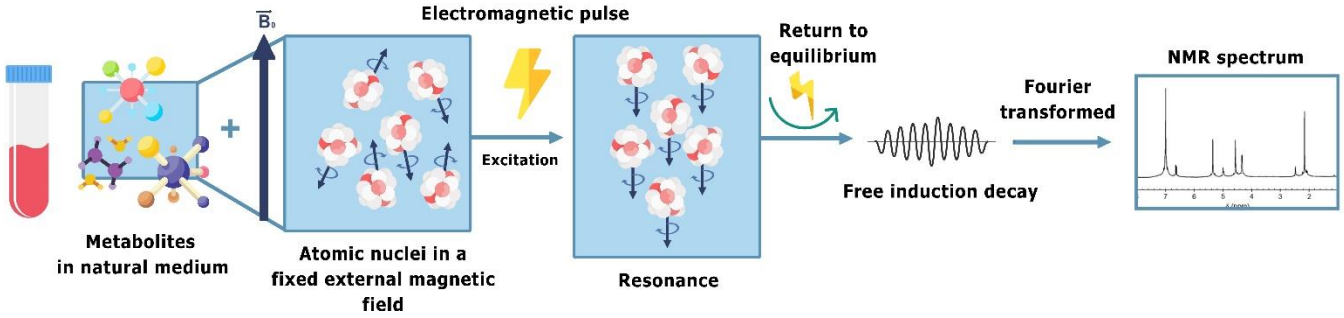


Figure 2. Operational outline of a nuclear magnetic resonance experiment

1.4 Serum compounds profiling through ^{1}H-NMR

Far from being only limited to the quantification of the lipidome, NMR technology provides a wide range of possibilities in the profiling of serum compounds with great depth and accuracy.

Serum compounds have different chemical natures and consequently identification procedures change among them. In order to manage the assessment of the different compounds found in human serum, NMR relies on the editing of the radiofrequency pulse that stimulates nuclei for as to obtain a signal that portrays the expected compounds, editing radiofrequency pulse sequences allows for the modification of the observable spectral information based on physiochemical properties. This technique is referred to as "NMR spectral editing" [27].

In spite of the upsides mentioned, ^{1}H-NMR serum compound profiling is not an easy task, the generally acquired spectrums consist of a conglomerate of several overlapping signals from a vast number of compounds at different concentrations, making it a challenge to provide reliable identifications and quantifications [27].

The identification of compounds relies on the support of existent public libraries that include lists of compound peaks, raw NMR files of standard compounds, and typical concentration ranges in common biofluids [28][29].

Regarding the metabolite quantification method, first, the quantification in area units of the signals assigned to known metabolites is performed. Integrating isolated signals is the classical approach and, unfortunately, has been proved to lead to various baseline issues with overlapping peaks [27]. Alternatively, the method of spectral deconvolution with line-shape fitting analysis turns out to be a more robust way of proceeding, it defines the ^{1}H-NMR spectrum into a finite number of Lorentzian/Gaussian line-shapes following quantum mechanical rules and some baseline functions [27][30].

The primary fluid substance connection to the metabolic system is undeniably blood, it reflects minimal changes in the whole metabolism making it the natural choice for vascular and systematic disease studies, as well as nutritional assays [31][32]. Blood's composition includes molecules of various sizes and mobilities: proteins, lipids, lipoproteins, cholesterols,

low-molecular weight metabolites and ions and their concentrations range from nano molar to micro molar.

The use of blood serum instead of blood plasma is not arbitrary, plasma consists of blood without blood cells while serum is blood plasma excluding blood clotting proteins or fibrinogens [33]. These clotting proteins lead to the appearance of interference signals in the ^{1}H-NMR spectrum; therefore, blood serum is preferred [27].

1.4.1 The three molecular windows

Despite the advantages presented by the use of blood serum over blood plasma, serum's ^{1}H-NMR spectra drawbacks need to be acknowledged, in fact, the complexity of the resulting spectrum disable the opportunity of characterizing the biochemical diversity of blood due to the overlapping of low-molecular-weight metabolites (LMWMs) sharp peaks with broad signals from macromolecules such as lipoproteins and albumin.

In response to the aforementioned problem arose the proposal of examining blood serum through the implementation of a three molecular windows model, such model relies on different preparations and analysis of the samples through ^{1}H-NMR, and it was proposed by Ala-Korpela and co-workers [34].

This model made possible the comprehensive high-throughput quantification of lipoprotein classes and constituent lipids, albumin, LMWMs such as amino acids, creatinine, glycolysis-related metabolites, and ketone bodies, with costs comparable to those of standard lipid measurements [27] The main difference among the three windows methods lies on the NMR spectral editing concept that has already been described in this section 1.4.

To start with, there is the lipoprotein window, which englobes any ^{1}H-NMR experiment of native serum where broad signals produced by macromolecules such as proteins and lipoproteins are visible. To acquire such spectrum, a pulse known as NOESY-presat is required, the second term of the name refers to the water presaturation process, which is actually mandatory for any serum ^{1}H-NMR spectrum since otherwise interferences generated by large residual signals from water protons would be visible [27].

The outcome of the described experiment is the spectral representation of proteins by a broad background signal, mainly describing albumin, and several broad picks assigned to lipid moieties of the lipoprotein subclasses VLDL, LDL and HDL, with contribution of LMWM peaks [27]. To avoid the interference of LMWM, diffusion-edited ^{1}H-NMR (LED) spectroscopy has been suggested and recently validated [35][36][37].

The second molecular window described is the LMWM window, encompassing those experiments that use a T2-edited ^{1}H-NMR, such as the Carr-Purcell-Meiboom-Gill (CPMG) pulse sequence, which enhances the detection of LMWM by suppressing the wide-ranging signals originated by fast-relaxing molecules such as proteins and lipoproteins [38]. Indeed, T2-edited NMR can be acknowledged as the reciprocal of diffusion-edited NMR.

Last but definitely not least, the lipid window arises, and given the whole elucidation thus far it is a foregone conclusion that this window is of great value for the characterization of the lipidome and consequentially the assessment of CVD risk. For the application of such window, the breakdown of protein and lipoprotein complexes enabling their lipid extraction is required and such process introduces several drawbacks.

The lipid extraction process, contrary to what might be expected, is still a manual and time-consuming procedure (the reader is referred to reference [39] for further information regarding lipid extraction, storage, and NMR sample preparation) that can become hazardous concerning the accurate obtention of the outcomes, that is due to possible errors originated by the human factor and the excessive manipulation of the samples.

Returning to the lipid window methodology, a standard ^{1}H-NMR 90º pulse is conventionally used, providing a spectrum that describes fatty acid families, free and esterified cholesterol, triglycerides, choline phospholipids and glycerophospholipids [40]. For the quantification of these metabolites there is a lack of software that employs automatic deconvolution and quantification methods, which is the reason why most studies rely on spectral integration and fingerprint analysis instead of the more proficient method [41]–[46].

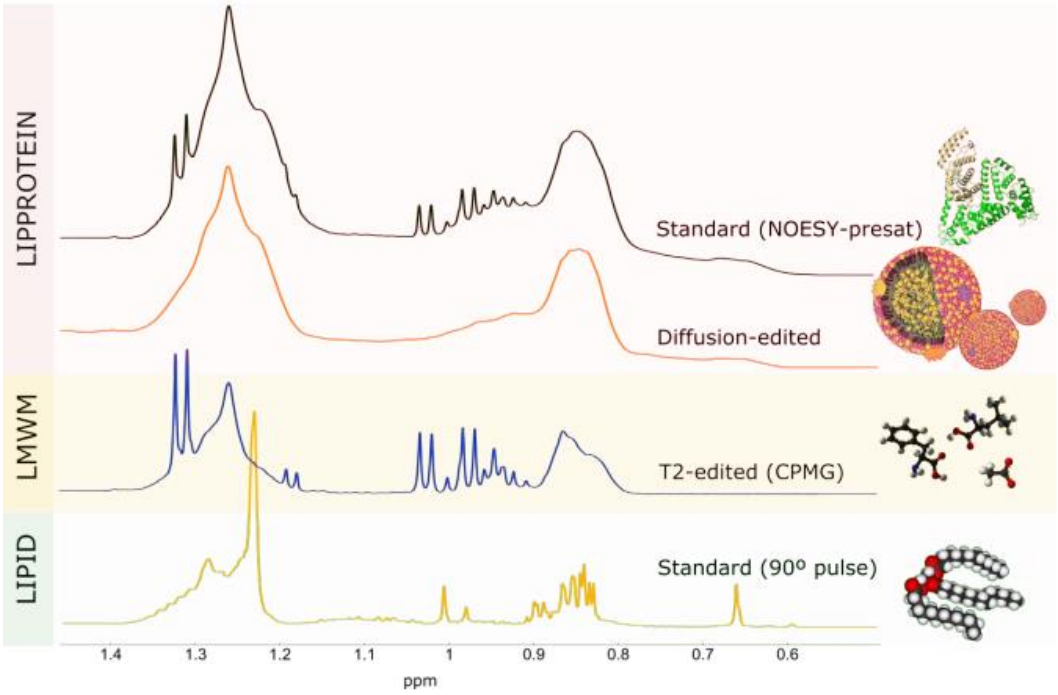


Figure 3. Methyl and methylene regions in the three molecular window model and examples of molecular species that are analysed with each window [27]

1.5 Motivation and context of the project

After acknowledging the profound value that NMR lipid profiling provides in the aforementioned CVD field and it's still to be exploited potential, the need for sophisticate and efficient tools for the development and evolution of this technology's application to healthcare arises.

Several research groups and companies have already ventured into this field and created tools based on the use of NMR technology for lipid profiling to be used in the healthcare sector. Fortunately, I have found myself within a marvellous group of people that have succeeded in such feat, and I was proposed to contribute to their project with the task of optimizing their lipoproteins' lipidome assessment process by enabling the quantification of lipids directly through native serum's ^{1}H-NMR spectrum, avoiding the laborious and almost hazardous process of serum lipid extraction.

Achieving such feat would be of great value for those who do research or work through the NMR technology for lipidic assessment, conventional quantification's issues such as overlapping signals would be avoided, and accurate and robust results would be obtained in a comfortable and faster way that will save time and money. The success of this project could provide more reliable quantifications than those of the actual methods, where various aspects of the procedure introduce variability to the final outcome.

2 Objectives and hypotheses

2.1 Objectives

This project aims on the optimization of the NMR lipid profiling process with the ultimate objective of suppressing the previously essential step of serum lipid extraction through the design of linear regression models that are capable of directly assessing the lipidome of lipoproteins from the ^{1}H-NMR LED spectrum of native serum.

For a set of 12 lipidic families, being such: total cholesterol (TC), esterified cholesterol (EC), free cholesterol (FC), triglycerides (TG), lysophosphatidylcholine (LPC), linoleic acid (LA), saturated fatty acids (SFA), ω-6 and ω-7 (W6W7), ω-9 (W9), ω-3 (W3), docosahexaenoic acid (DHA), and arachidonic acid together with eicosapentaenoic acid (ARA+EPA), a deep and exhaustive process for the development of a linear regression prediction model through the partial least squares regression technique and based on ^{1}H-NMR LED spectra, will be conducted.

Thus, the main goal of the project is:

- Development of 12 lipidic families linear regression prediction models through the partial least squares regression technique that enables lipid quantification directly from native serum's ^{1}H-NMR LED spectra, obtained through a diffusion-edited pulse.

Following, a segmented list of the main goal of the project is provided, where each point represents a task or step in the process of reaching it:

- PCA exploratory analysis of the sample sets employed in search of an initial comparison among them and a first selection criteria regarding the building of the models.
- Study of the possible benefits of normalizing the sample sets between them, aiming at improving the models development.
- Perform the best variable selection for building the predictive models through the STOCSY tool and other methods described below.
- Define a sophisticate modelling procedure for concluding with the best model possible.

For the proficient fulfilment of the listed objectives, another goal is set, the result of which will serve as a means to reach the main goal, and it consists of the automatization of the entire linear regression predictive modelling process through the software MATLAB (MathWorks Inc.) and the Partial Least Squares (PLS) Toolbox, to then project it into a novel interactive software that leads the user to the most refined predictive model possible. Therefore, additional to the main objective:

- Development of a novel interactive software based on MATLAB (MathWorks Inc.) that through automatic database cleaning and processing, variables selection through custom-built iterative methods and a step-by-step presentation of the sequential outcomes through illustrative graphical representations, guides the user to the best-possible linear regression predictive model.

2.2 Hypotheses

First, proficient prediction models will be obtained for each lipidic compound, although variances may exist among them due to the nature of each lipid's quantification, conditioned by the clearance of its spectral representation, the inherent attributes of the samples used, the influence of the serum lipidic extraction and the metabolite quantification process.

	Set 1 N = 134	Set 2 N = 398	Set 3 N = 127	Set 4 N = 569	Set 5 N = 101	Set 6 N = 199	Set 7 N = 117	Set 8 N = 150	Set 9 N = 83
Age (years)	NA	67.0 [56.0;76.0]	70.0 [55.0;84.0]	65.0 [53.0;76.0]	63.5 [54.8;76.0]	66.0 [54.2;80.0]	58.0 [52.0;66.0]	NA	NA
Sex: Male	NA	249 (63.0%)	74 (58.3%)	364 (64.3%)	56 (56.0%)	90 (54.5%)	27 (23.1%)	NA	NA
Type 2 diabetes	NA	218 (55.2%)	21 (16.5%)	128 (22.6%)	0 (0%)	32 (21.9%)	1 (0.85%)	NA	NA
Obesity	NA	189 (47.8%)	11 (10.5%)	167 (51.1%)	22 (22.0%)	42 (30.7%)	61 (52.1%)	NA	NA
Hypertension	NA	219 (55.4%)	43 (33.9%)	285 (50.4%)	0 (0%)	44 (58.7%)	25 (21.6%)	NA	NA

Table 2. Properties of the sets at disposal, consisting of the number of samples (N), age, male participants, type 2 diabetes, obesity, and hypertension. The age range and average value are provided, and for the rest of characteristics, the frequency in number and percentage is given.

Second, a software that gathers in an automatized manner all the steps required for the design of linear predictive models will be acquired through the merge of different codes that execute individually the different tasks required.

Third, this new software tool will be implemented for the development of the sought-after predictive models and will lead to the attainment of the best models possible, dodging all the bad influences and data patterns that could lead to the incompetence of the models.

Last, the optimization of the NMR lipid profiling process will be successful, and the outcome of this project will allow for the quantification of lipidic compounds on native serum ^{1}H-NMR spectra.

Furthermore, valuable knowledge will be obtained from the analysis of the used data sets that will provide insights on the actual lipid quantification procedure.

3 Materials and methods

3.1 Description of the blood samples sets used and of the lipidic families profiled

For the development of this project, a number of 9 sets of samples of human blood serum, constituting a sum of 1878 samples, were provided by Biosfer Teslab. From these sets of samples, the first 8 were obtained from various clinical centres across Spain during the Covid-19 pandemic and were collected right after the patient's admission to the centre, with set 7 corresponding to a control group. The ninth set does not belong to the group of previous sets, it corresponds to an external group that was obtained following the same procedures as the other 8 and will not be considered throughout the project's predictive models acquisition, it will serve as an external validation set to test the desired results.

Such samples were then analysed through different ^{1}H-NMR experiments, ultimately acquiring the quantification of a metabolic profile that included a total of 80 chemical compounds, ranging from low-molecular-weight compounds, lipidic families and lipoproteins to glycoproteins.

Within these samples, a great incidence of metabolic disorders such as diabetes mellitus type 2 were detected. Diabetes translated to the metabolomic field implies impaired glucose regulation and altered

lipid metabolism, resulting in elevated glucose, triglycerides, and free fatty acids in the blood. Among the sets, a wide range of metabolomic characteristics, such as the diabetes incidence, were achieved, making the whole database robust and representative, where the variance captured is provided by the gathering of profiles of all ranges and characteristics.

3.2 Baseline Data: Acquisition of the project's starting framework

For this work, the diffusion edited ^{1}H-NMR spectrum or LED spectrum of the samples' native serum and the lipidic families profile of the samples lipidic extraction, will be considered and serve as the baseline data, which stands to reasons with the project's intended objectives. To provide further clarity, for each blood serum sample a LED spectrum and a lipid families profile is at disposal, and the project focuses on the study of the relationship among the two of them for predictive modelling purposes. Following, a concise description of the acquisition procedure of both experiments is provided.

3.2.1 1D diffusion-edited spectrum acquisition from native serum

Frozen serum samples (250 µL) were shipped on dry ice to Biosfer Teslab (Reus, Spain) for the analysis of serum metabolomics profile by the NMR-based metabolomics. In brief, serum samples were first diluted with deuterated water and 50 mM pH 7.4 phosphate buffer solution before ^{1}H-NMR analysis. Then, ^{1}H-NMR spectra were recorded at 306 K on a Bruker Avance III 600 spectrometer operating at a proton frequency of 600.20 MHz according to the optimized experimental parameters for one-dimensional H-NMR pulse experiments described by Mallol et al., 2015 [20]. A 1D diffusion-edited spectrum, as previously said, also referred to as LED spectrum, with suppression of the low molecular weight compounds were recorded and a total of 32 scans were acquired.

3.2.2 Lipids quantification in serum extracts

Lipophilic extracts were obtained from a 200 µL aliquot of dry plasma using the BUME method [47] with slight modifications. BUME was optimized 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. This procedure was performed with a BRAVO liquid handling robot which has capacity to extract 96 samples at once. The upper lipophilic phase was completely dried in Speedvac until evaporation of organic solvents and frozen at -80 °C until NMR analysis.

Lipid extracts were reconstituted in a solution of CDCl3:CD3OD: D2O (16:7:1, v/v/v) containing Tetramethylsilane (TMS) 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. A 90° pulse with water presaturation sequence (ZPGR) was used. Quantification of lipid signals in ^{1}H-NMR spectra was carried out through an adaptation of LipSpin [48], a Biosfer Teslab in-house software. Briefly, spectra were phase-corrected, baseline-removed, and shift-referenced to TMS signal at 0 ppm. Finally, signals assigned to lipids [43] were quantified with line shape fitting analysis. After the quantification process, signal areas were converted into molar concentrations by using an external quantification.

Moving forward and to provide an overview of the distribution of lipid quantifications among the sets at disposal, graphical representations of each set's concentrations for each lipidic family aimed at studying were performed and are presented in Figure 4, enabling the comparison among sets through plotting one next to the other, and providing a graphical visualization of the starting point lipidic data.

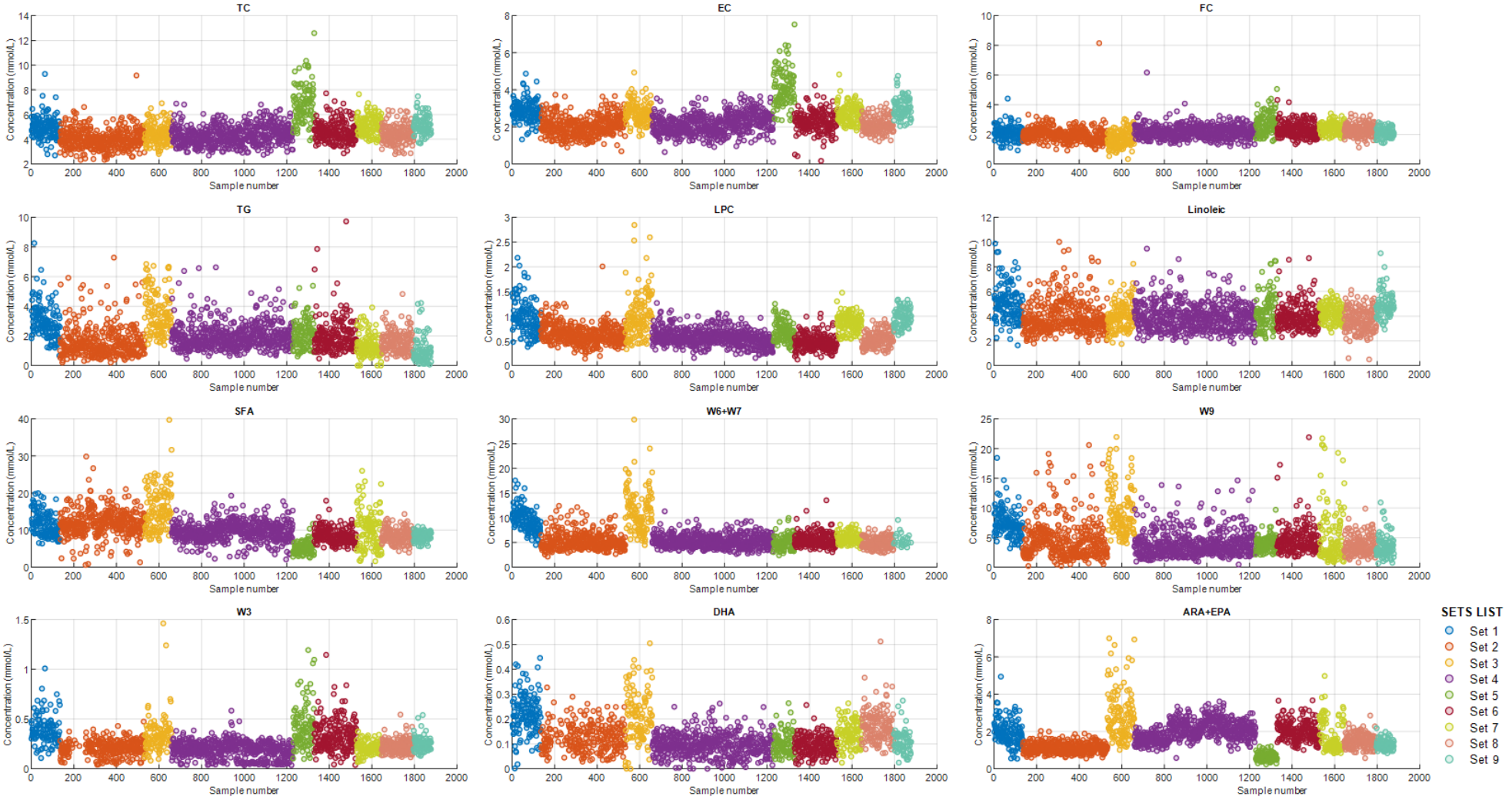


Figure 4. Lipidic families' concentration distribution among sets

12 different plots, one for each lipid family aimed at studying, representing each set's lipid concentration (mmol/L) distribution is provided. Through these graphs, an initial comparison of the concentration ranges each set's population presents can be noted, and will serve as an important decision criterion for the development of this project, mostly dealt with in the results section.

3.3 Software tools

For the proper fulfilment of this study's goals, mathematical software tools have been employed, exploiting the potential of statistical and multivariate analysis, large database management, and predictive model building. Through the use of these sophisticate tools, several advantages arise such as precise data evaluation, identification of patterns, and exploration of complex relationships, all of it combined with the proficient handling and storage of big databases.

3.3.1 Matlab employment and starting point establishment

Starring as field of work for the development of this project, the Matlab software has been harnessed, both v7.10 (R2010a) and v9.9 (R2020b). It is a powerful tool that encompasses a wide range of capabilities that align with the outlined objectives of the project.

By taking advantage of Matlab's robust features, the project benefits from enhanced analytical capabilities, data management processes, improved efficiency in constructing predictive models, and graphical representation tools among others, which greatly contribute to the project's overall efficiency.

One of the reasons for which Matlab is greatly suitable for this project is that it is the background for an extensive range of techniques and algorithms for constructing robust predictive models, such as the Partial Least Squares (PLS) toolbox from Eigenvector Research Inc., version 5.8.3 of which has been utilised for this project's development.

The Matlab starting point of this project is the importation of the acquired LED spectra from the 9 sets analysed (3.2.1), and their corresponding lipidic profiles which include the 12 lipidic families aimed at studying (3.2.2), both datasets serving as the starting point for the predictive modelling development. Therefore, an initial workspace of 18 datasets is settled.

The format of the LED spectra in Matlab consists of a n x m size dataset, with n representing the number of samples of the set of matter and m a discretization of the LED spectra with a number of 15274 discrete points. The values in the dataset correspond to the samples' LED spectrum signal intensity at every sampled point.

The lipid quantifications dataset format consists of a n x s size dataset, with n representing again the number of samples of the set of matter and s the number of lipids quantified. The values in the dataset correspond to each sample's lipid concentration value (mmol/L).

Be aware that the procedures employed for the acquisition of the predictive models are performed considering each lipidic family individually, which stands to reason with the fact that for each one of the lipidic families a prediction model is required, in other words, the predictive model building procedure has been executed 12 times, each one selecting the values of the lipid of interest from the lipid quantifications dataset.

3.4 Statistical and multivariate analysis

Statistical analysis involves collecting, organizing, analysing, and interpreting data to draw conclusions and make predictions [49]. Multivariate analysis extends this approach by considering multiple variables simultaneously, allowing researchers to understand complex relationships and interactions between variables [50]. Statistical analysis uses techniques such as descriptive statistics, hypothesis testing, and confidence intervals, while multivariate analysis employs methods like multiple regression, factor analysis, cluster analysis, discriminant analysis, and principal component analysis. These tools are essential for uncovering patterns, making predictions, and gaining insights from data in various fields.

3.4.1 Principal component analysis – PCA

Aiming at discovering tendencies and patterns among the sample sets used and to serve as a first approach for the model development, the principal component analysis (PCA) technique was employed.

The main task of PCA is data visualization and dimensionality reduction, it is a proficient method at transforming a high-dimensional dataset into a lower-dimensional space while retaining the most important information and minimizing the loss of variance in an unsupervised manner [51].

This technique was used against the LED spectral data and the whole lipidic quantifications data, considering the different spectrum points and the 12 lipid families as initial variables on which to run the dimensionality reduction procedure, while samples identifications are considered the observations.

Both analyses were performed in Matlab 2020 through the in-built function for PCA, where by default, a normalization process consisting of variable's mean subtracting is executed. This technique subtracts the mean of the data to adjust the scale and centre it, ultimately having a dataset with a mean of zero.

Underlying structures and patterns in the data can be identified through a PCA procedure, owing to the creation of new variables known as principal components, which are linear combinations of the original variables and are uncorrelated with each other. Each principal component is created based on the variables' covariance matrix, from which eigenvectors, representing a principal component's directions or axes in the original variable space, and eigenvalues, indicating the amount of variance explained by each eigenvector, are calculated, and these differ from each other depending on the impact that each of the original variables has had on it, also known as weights [51].

Therefore, higher eigenvalues correspond to more important and informative principal components and such measure serves as reference for ranking them from best to worst. After selecting the two most relevant principal components and verifying they were generated with a proper distribution of weights, a two-dimensional space where the original variables are projected onto was obtained. Each observation in the new dataset corresponds to a combination of the original variables, weighted by the corresponding eigenvectors.

3.4.2 Correlation heat-maps – STOCSY

Statistical total correlation spectroscopy (STOCSY) is a technique used in metabolomics and related fields to identify correlated signals in spectroscopic data [52]. This tool helps determine which regions of a spectrum are most correlated with a target variable of interest, in the present case consisting of a LED ^{1}H-NMR spectrum and a lipid of interest quantification values respectively.

This tool has been implemented in Matlab through a Biosfer in-house built script, that provides the user with the desired heatmap representation based over the plot of the LED spectrum.

$$r_{xy} = \frac{\Sigma(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\Sigma(x_i - \overline{x})^2 \Sigma\left(y_i - \overline{y}\right)^2}}$$

Equation 1. Correlation coefficient between

'x' and 'y' represent individual data points from two datasets and $\bar{x}$ and $\bar{y}$ are each dataset's mean. The summations are taken over all data.

Correlation coefficients, ranging from -1 to 1, are calculated between each sample's LED spectral points and their corresponding lipid of interest concentration, to further clarify and considering a single sample, for each one the 15274 spectral points, a correlation coefficient is calculated against this sample's lipid of interest concentration.

The graphical outcome is a merge of all the samples coefficients along the entire spectra, providing a representation of the most frequent regions where great relationships are found with the quantifications of the lipid of interest, which can also be seen as the areas of the LED spectrum where the lipid aiming at quantifying has a greater presence. Also, the areas with null relation or even negative relation with the lipid quantification can be noted. The correlation coefficients range is translated to a colour gradient that goes from dark blue for the most negative correlations or those close -1, to dark red for the most positive correlations or those close to 1, being the neutral coefficients or those close to 0 represented in green.

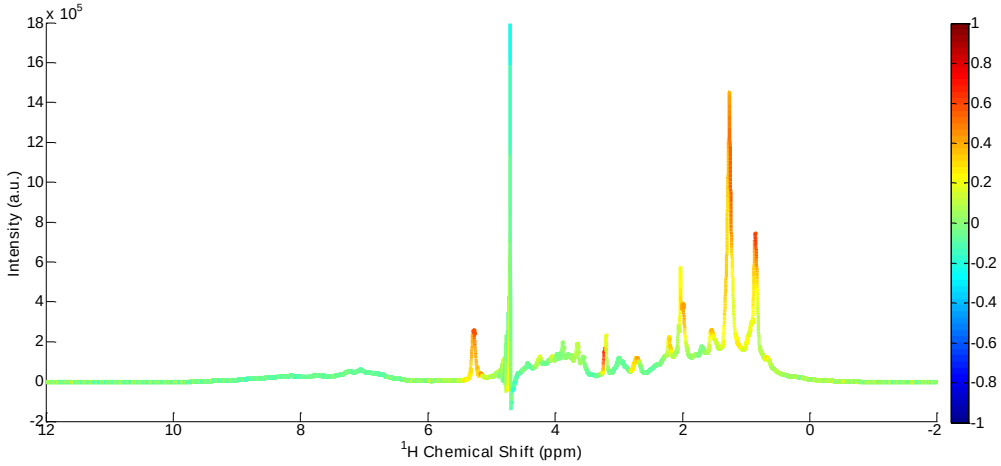


Figure 5. STOCSY heatmap

For this project, STOCSY heatmaps have been used as a tool to have a first set selection criteria and a spectral regions of interest selection criteria, both of which will be discussed in further sections of the project (3.4.4, 3.4.5, 3.6).

3.4.3 Data normalization – Z-Score

For several reasons, normalization is commonly used in data pre-processing and analysis, it converts different sets of variables into a standardized range or form that commonly results in a very helpful procedure [53].

Through data normalization comparability is assured, bringing variables with different scales or units to the same range to be analysed on an equal footing. This scaling allows for fair comparisons among data avoiding groups of variables with large values dominate over the others [53].

Nevertheless, this technique can be hazardous and lead to the devaluation of the initial data owing to the fact that this procedure brings all data to the same scale assuming that original differences among sets are not generated by intrinsic information of each set, instead, it assumes that it is negligible information that does not influence the value of the data, and that may not always be the case. When it is the case, and the bias among data arises from technical factors to the experimental data collection or other influential factors in the process, we are facing a batch effect in our data [53]. These effects are particularly relevant in high-throughput biological experiments such as metabolomics, where large-scale data is generated from multiple samples.

The presence of batch effect will most surely lead to incorrect conclusions, reduced reproducibility, and difficulties comparing or integrating data from different batches [53]. It is fundamental that batch effects are detected and solved to ensure accurate and reliable analysis and interpretation of the data.

In addition, normalization is also a means for outlier handling, mitigating their impact over the whole data, interpreting coefficients when performing regression analysis due to the ease of variables importance interpretation over the response variable, and visualizing data [54].

Being aware of the double-sided nature of the normalization process, a study of its application in this project has been performed and will be addressed in section 3.5.

Nevertheless, a normalization procedure through the Z-score technique has been established for our data. Z-score consists of a statistical technique that transforms sets of data by standardizing them relative to a selected set's mean and standard deviation, in other words, rescales a group of sets against another one so that the whole dataset presents the selected set's mean and standard deviation.

$$z = \left(\frac{x - \mu(x)}{\sigma(x)}\right) * \sigma(y) + \mu(y)$$

Equation 2. Z-score

The set wanted to normalize is represented by 'x', the reference set for the normalization is represented by 'y'.

The mathematical process consists of subtracting the mean of the whole data to each data point and then divide the result by the complete data's standard deviation. In this project, normalization has been proposed for both LED spectra datasets and lipid quantification datasets. Note that for the present study, normalization is only of relevance when coupling of various datasets happens, since normalizing individual sets serves no purpose. This coupling of datasets aspect of the project will be discussed in further sections (3.4.5, 3.5, 3.6).

The procedure for normalizing sets of LED spectra among them starts with the selection of one set as reference, which will be the one that the others will be normalized to, and within that set, a base-line spectral region where no metabolic information is meant to be acquired must be chosen. For this spectral region, mean and standard deviation are calculated and used to normalize the rest of sets in the group, proceeding as follows. For the remaining sets, the same base-line spectral region must be considered and mean, and standard deviation calculated to normalize each one of them, so that they individually present a mean of zero and standard deviation of one, when achieved, the reference set's mean and standard deviation is used to relocate the now normalized remaining sets to the same range as the reference one, this is performed by multiplying the reference set's standard deviation and then adding the mean to the remaining sets spectral points.

The normalization procedure of lipid quantification sets is more straight-forward, a set is selected as reference and its mean and standard deviation regarding the concentrations of the lipid aimed at studying are calculated. The acquired mean and standard deviation values are used to relocate the remaining sets within the group after z-score has been individually applied on each one of them.

It can be noticed that for both procedures, the sets of data are normalized based on a set selected as reference, which is the one against whom the others have been rearranged to,

in other words, remaining sets mean and standard deviation have been turned to that of the reference set. Reference sets for both procedures have been selected accordingly.

3.4.4 Partial least squares linear regression technique and PLS toolbox employment

The PLS method is a powerful statistical technique for modelling linear relationships between sets of variables, especially in high-dimensional datasets scenarios or complex multivariate relationships. It is often employed in situations where traditional regression techniques may face challenges, such as multicollinearity among predictor variables or when the number of predictors is large compared to the sample size [55].

The method itself focuses on maximizing the covariance between the predictor variables, which are variables from which a prediction is meant to be made, and the response variables, which are variables that are meant to be predicted and correspond to our predictor values expected results, by constructing a number of instances called components or latent variables in an iterative manner. A group composed of predictor variables and response variables used for model development is known as a training set, whereas its use for the validation of a model converts it to a validation set [56].

Each component is a linear combination of the original variables and is generated in a way that captures the maximum covariance between the two sets of variables. Ultimately, the process seeks to explain as much variance as possible in the response variables using a smaller number of components. The number of components to retain is determined by the analyst based on criteria such as the amount of variance explained per number of components or the prediction performance of the model. Generally, the more components the more variance explained, nevertheless, this is a double-sided factor since more components increases model complexity and overfitting happens, making the model too specific and unable to predict properly in unseen situations [56].

$$Cov_{XY} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Equation 3. Covariance

'x' and 'y' represent individual data points from the two datasets X and Y, $\bar{x}$ and $\bar{y}$ are each dataset's mean and n is the total number of paired observations. The summation is taken over all data.

Similar to the PCA, each component is developed through the linear combination of the original variables whose impact on the component is known by the term loading, in other words, loadings refer to the contribution of each original variable to the component. So again, we face another method of dimensionality reduction where latent variables preserve the relevant inherent information of the original data, nevertheless, the two techniques differ in a key aspect, PCA seeks for the unsupervised identification of principal components that explain the main patterns and sources of variation in the data, whereas PLS is a supervised method that seeks the identification of components that capture the maximum covariance between a given pair of variables, composed by predictor variables and response variables, ultimately to be used for modelling purposes.

The Matlab workspace enables the use of the PLS toolbox, already mentioned in 3.3.1, which is meant to be implemented to build linear regression predictive models and the user can do so via commands or an interactive display.

The development of the PLS predictive models in this project were carried out, again, through the PLS toolbox, and the procedure starts with the selection of our input, consisting of a number of chosen samples' LED spectrum and their corresponding quantifications of the lipid that wants to be studied. Then, follows the establishment of the training and validation sets from our data, data that to further clarify, is composed by the intensities of each sample's LED spectra sampling behaving as predictor variables and the corresponding lipid quantifications aimed at studying as response variables. The training and validation groups are both composed of a split selection of our two types of variables, the difference among them is the number of samples included, typically, to train the model, between the 70-80% of the samples are used, and the remaining 30-20% is used for the validation set. For this project, 70% of samples were assigned to the training set and the remaining 30% to the validation set.

This training-validation separation of the dataset is a crucial aspect in the building of the model, not so much due to the percentage of samples included in each group, but in fact, due to the range and quality of values acquired within the selected training percentage, since selecting a wide ranged training set from our data will make the model learn to predict properly in front of a bigger number of situations, making it more robust and efficient. Then, the idea of grouping different sets of data together for developing models through a large quantity of samples that provide a huge range of values seems the best choice, the application of this methodology is referred to in section 3.6.

Nevertheless, the PLS toolbox doesn't provide any resource for reaching the most proficient training-validation split possible. This matter is addressed in the following section (3.4.5).

In addition, and having a great impact on the final result, another aspect of the training set establishment procedure followed in this project, must be attended to. Following the process up to now, our predictor variables consist of a certain number of samples and for each sample, 15274 spectrum points are being considered. Proceeding this way would lead to inefficient and incoherent results since not all spectral points provide predictive information regarding the lipid of study, only some regions may be associated with it, and to tackle this issue, correlation heat maps are of vital importance. Through the generation of the STOCSY heatmaps of our LED spectra against our lipid of interest quantifications, the spectral points that express the highest correlations with the lipid quantifications out of the initial 15274, will become apparent, and by selecting these spectral points to be the ones that the PLS algorithm considers for the building of the model, more coherent, robust, and proficient results will be obtained.

When the training-validation split is complete and the spectral regions of interest are selected, the PLS toolbox provides data pre-processing for the two training set groups through the methods autoscaling and mean centring, the latter being the selected option for the model development in this project and the same methodology applied in the PCA pre-processing of data (3.4.1).

Following, the procedure of calculation and quantity assessment of latent variables is executed, and the user is provided with the cross-validation resampling method for assessing the performance of the generated model when using different numbers of components. Cross-validation can be employed through different techniques such as venetian blinds, which is the option of choice for this project. The training dataset is divided into multiple subsets or folds to train and evaluate the model iteratively, at the same time measurement of each fold's performance is done, ultimately aggregating them for an overall estimate of the model's proficiency. This procedure is automatically repeated with different numbers of latent variables, ranging from 1 to 20, and a representation of the root mean square error (RMSE)

with the increase of components regarding the cross-validation is provided through a red line plot, enabling the analyst to interpret the most efficient selection of components for the model.

Regarding the graphical representation of the RMSE against the increase of latent variables, the criteria for selecting the number of components consists of determining at which number of components the RMSE stabilizes, this can be determined in the graph through the detection of an elbow shape in plotted line, which means that the addition of further components has no impact on the improvement of the model, instead, it may be causing the undesired overfitting. Typical representations consist of high RMSE at low number of components and an exponential decrease when incrementing their number, which goes in accordance with what has already been explained about latent variables.

Finally, a model with a determined training set sample distribution, spectral regions selection and number of components is selected, enabling the test of its performance through validation, which proceed as follows. The obtained model is validated against the remaining validation samples' LED spectrum data to acquire a set of quantification predictions for the lipid of interest, these results are then compared to the original results, which are the remaining validation group's lipid quantifications, through the plot of a linear regression among both datasets, providing the Pearson correlation coefficient (R) and representing the success of the model through it. This described process is known as validation and it can also, and most importantly, be performed through external data instead of the remaining 30%, in other words, predictive models are put to the test by examining their performance at predicting data that does not belong to the dataset with which the model was trained, which is the ultimate purpose of a predictive model.

In terms of the whole PLS modelling procedure, several factors introduce variance in the building of the models and need to be acknowledged for the proper understanding of the project: the training-validation split, also referring to the possibility of merging different sets, the spectral regions or points of interest selected, and the number of components employed in the model.

3.4.5 The SuperScript tool

In the process of development of this project, the idea of building a software that automatizes the methodology for building linear regression predictive models, addresses the factors that introduce variance in the building of PLS models, and includes an interactive step-by-step workflow through which the user can monitor and lead the progression of the modelling, arose, and was carried out in pursuit of this project's successful accomplishment of the expected results.

In addition, the lack of efficiency in the execution of the initial procedure for developing predictive models through the already mentioned tools (3.4.4) contributed to venture with the development of this idea, the result of which will exploit the capacities of predictive modelling in the computational space we lie on, leading us to the most refined results possible.

The making of the software, named SuperScript, started with a merge of individual scripts initially programmed for automatizing different sections of the predictive modelling procedure, and seizing this opportunity, a sequential program that iterates through the different scripts, now considered steps of a whole procedure, was generated. The original scripts, later employed for the combined program, automatized the following tasks:

1. *Baseline data processing* (referring to the LED spectra and lipid quantifications datasets): provides the removal of missing values, ordering of the samples, and format-shaping of the datasets to a standardized format that is readable for the following procedures, which consists of the establishment of two cell arrays, one including the LED spectra of the different sets imported, and the other the

corresponding lipid quantifications for each set, both arrays following the same order.

2. *Data normalization:* executes normalization of the LED spectra and lipid quantification sets specified, providing three possible scenarios: LED spectra normalization, lipid quantifications normalization, and both spectra and quantifications normalization. All of this is performed following the procedures explained in section 3.4.3.

3. *Employment of the STOCSY tool:* applies the STOCSY tool on the selected data to be used as a starting point in the variable selection phase, concerning set selection and spectral regions of interest selection, for the later one, this task enables the selection of a baseline spectral regions, which consist of the STOCSY spectral points that surpass a certain user-specified threshold.

4. *Training-validation split:* performs random splitting of the selected data samples into 70-30% ratio, for training and validation respectively. Inputs can either be an individual set or various sets that are collapsed into a single coherent one.

5. *Custom-built spectral regions selection method:* considering the establishment of a training-validation split (task 4) and a baseline spectral regions selection (task 3), this code is meant to be used to assess through random iterations the best spectral regions selection for building a predictive model with the introduced training-validation split. The code takes as input both datasets, referring to the training and validation sets, and the baseline spectral regions, to automatically generate an up to the user number of models following the PLS toolbox procedure described in section 3.4.4. The models are generated through a randomized selection of the baseline spectral regions provided, with each randomization's model being portrayed using three different latent variables number in the following format: LV, LV+1, LV+2, with LV being a user-specified initial number of components. Then, this code validates the acquired models against their corresponding remaining validation 30%.

6. *Latent variables assessment:* employs cross-validation against a user-specified model to properly assess the best number of latent variables to be selected, the user is provided with a PLS-toolbox-generated graphical representation of the RMSE against the components increase. The outcoming plot's red line corresponds to the cross-validation's RMSE.

7. *Permutation of the samples' distribution in the training-validation split:* with selected spectral regions of interest and number of components, this code performs permutations in the training-validation datasets samples' distribution, to then have the models acquired validated against their corresponding remaining 30%.

8. *External validation of groups of models:* validates groups of models against a selected external dataset, providing a list of the 3 most proficient models in terms of the R acquired and the regression plot of the most successful model. Also, the STOCSY of the external validation set considering the lipid of study is provided through task 3.

Acknowledging the purpose of each task, it is noticed that the ones that are crucial for the successful development of models, in other words, the ones that deal with the factors that

introduce variance when building a model, are tasks 4, 5, 6 and 7. The impact of task 2 will be addressed in section 3.5.

Task 1 consists of a pre-processing step, prior and external to the actual development of the model, task 2 performs normalization when considering multiple sets of data seeking for a more efficient model development starting point, task 3 provides resources for having a selection criterion in the development of the model and finally task 8 informs the analyst of the performance of the models generated.

The coupling of tasks 2,3,4,5,6 and 8 resulted in the SuperScript tool, the workflow of which is presented below. Before, nevertheless, note that each of the tasks listed above can be executed both through the use of SuperScript, or through the execution of each individual task's script. Tasks 1 and 7 have not been implemented in the proper software but work in parallel with it, task 7's purpose in the procedure will be properly addressed in section 3.6.

The execution of SuperScript starts with the outcome of running task 1 through our input data, establishing our initial datasets. This initial datasets, consisting of a cell array of each set's LED spectrum data and a cell array of each set's lipid quantifications, are then put through task 4 to obtain the training-validation split in accordance to the user's asked specifications, being such the indication of which sets must be included for the development of the model, which if being more than one the program will collapse them into a single one, and which lipidic family are we aiming at studying, represented by the column number in the lipid quantifications dataset.

When the establishment of the training-validation split with a randomized sample distribution is complete, then the user is presented with the STOCSY heatmap of the data at issue. Acquired through task 3, this heatmap helps the analyst interpret the correlation between the LED spectra regions and the lipid of interest's quantifications, providing a baseline spectral regions selection, on the basis of which task 5 will work on, to further clarify, a baseline set of ranges that surpass a correlation threshold specified by the user is determined as a vector, and will serve as starting point for task 5. This step can be avoided if the user provides as input of the SuperScript software a vector that incorporates the baseline spectral regions that are meant to be considered for the development of the program. Nevertheless, generating the spectral regions of interest through the program itself provides a more accurate selection of spectral points, since correlation values are directly assessed and selected based on the specified threshold.

Straight after that, the user is presented with the application's menu, where the possibility of running any of the tasks incorporated in the software is provided. The options list goes as follows:

0. *Generate baseline spectral regions:* executes task 3 to provide a baseline spectral regions for the further development of the program. If a baseline spectral regions selection has already been made and the option is executed again, it provides the opportunity of rearranging the selected baseline spectral regions by running task 3 again. Also, if the user introduced a vector with spectral regions when executing the program, this option gives the opportunity of generating a second spectral regions of interest vector through the program's method.

1. *Selection of the baseline spectral regions:* it enables the user to swap between the two possible baseline spectral regions vector, consisting of the one generated by the program or the one introduced by the user.

2. *Latent variable assessment:* with the training-validation split sets introduced and the baseline spectral regions vector, task 6 is executed, providing a first approach to which number of components is more suitable for the development of the model at

issue. Then, the user is asked to specify which should be the number of components employed for the software's model development.

3. *Data normalization:* executes task 2 on the training-validation split datasets, and even more, it also normalizes the external validation dataset in accordance with the normalizations specified. Once this option has been executed, its second selection will enable the user to choose whether to continue working with the standardized sets or the non-standardized sets. Each one of the four normalization scenarios that can be generated through task 2 can be arranged through this option.

4. *Mode one-by-one:* this mode creates a single model with the training-validation split datasets under consideration and following the procedure of task 5. Providing as outcome the performance of a randomized spectral regions selection model portrayed with three consecutive numbers of latent variables that start on the number chosen in option 2.

5. *Mode ten-by-one:* this mode works exactly like the fourth mode except for the number of models being generated, this time, 10 models are generated at once through task 5, when generated, the user can continue with the generation of 10 more and so on, until satisfied. Note that in reality, for each one of the 10 models, 3 varieties of the model are created changing its number of components, which means that the total number of models acquired at each iteration of this mode is 30.

6. *Best models list:* this option provides a list of the three most successful models regarding their 30% validation, representing each one's performance with each three different components.

7. *External set validation:* option 7 performs task 8, validating the generated models through options 4 and 5 with an external set that has had to be initially introduced as input of the software. The outcome is a list that follows the format of option 6's outcome but expressing the performance against the external set.

8. *Lists emptying:* Results from options 4, 5 and 7, which are being saved in datasets to be provided as the software's outcome variables, can be erased through this option.

During all the software execution, results from options 4, 5 and 7 are recorded into two different cell arrays, the first one contains information regarding the building of the models: R corresponding to the remaining 30% of samples validation, number of latent variables utilised, randomised spectral regions selected, and the model's Matlab structure. Regarding the second cell array, each model's R corresponding to the external set validation is given, as well as the model generation number, 30% validation's R, number of components, and model's Matlab structure.

For the development of this project's main objective, the SuperScript tool has been employed in the acquisition of predictive models for each one of the lipidic families aimed at studying, and the procedure of application of such tool to acquire the most refined results possible is addressed in section 3.6.

Below in Figure 6, a block diagram of SuperScript's execution procedure is provided, with task 1's employment to obtain the input data in the upper part of the figure. The menu can be noted in the central part of the diagram since all options can be chosen from there. Inputs and outputs of the software are represented, as well as the required user specifications.

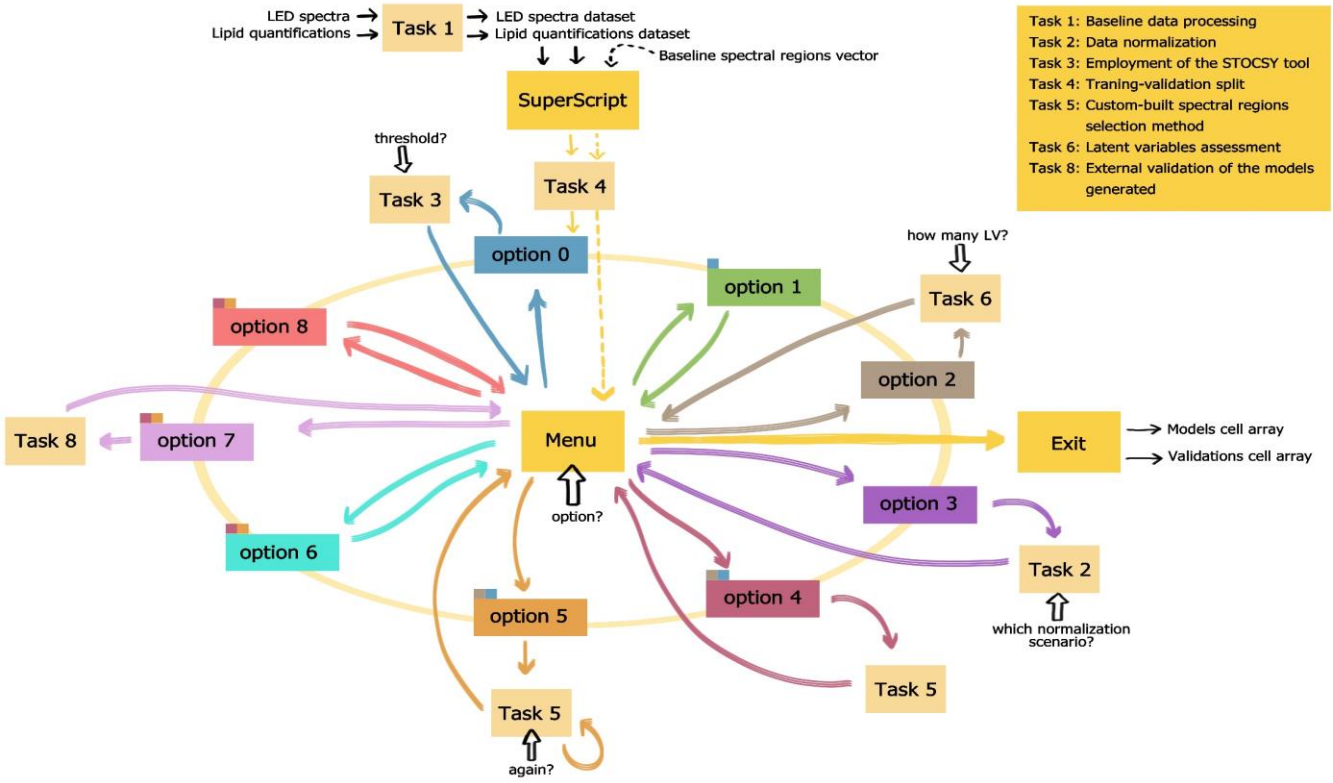


Figure 6. SuperScript's execution block diagram

Black arrows indicate inputs and outputs, the yellow arrows are SuperScript's menu entrance and exit, coloured arrows correspond to each option's execution path. Volumed arrows indicate user-specified inputs. Discontinuous arrows indicate alternative paths. Coloured squares above the options represent through their colour which other options must be previously executed for enabling its selection.

3.5 Procedure for analysing normalization's impact in the predictive model building

To address whether normalising our data leads to the better performance of predictive modelling, the SuperScript tool was employed.

Considering two sets for building a model, a third set for an external validation, and one of the 12 lipidic families, the process starts with the utilization of task 4 to merge into a single dataset both training sets, and generate the training-validation split, then this data was introduced into the SuperScript tool, as well as the external validation set.

The SuperScript employment consisted of building 900 models following the natural procedure of the software's model development, considered as the sequential execution of options 0, 2, and 5, but through four different normalization employments, achieved through option 3. Next, the selection of the best performing models in terms of the remaining 30% validation was performed, to then validate them externally against the previously selected set through option 7, seeking the observation of differences regarding the validation performance.

To elucidate more, the 900 models were generated in four different normalization scenarios. The first situation consisted of no data normalization, the second one of response variables' normalization or lipid quantifications normalization, the third one of predictor variables normalization or LED spectra normalization, and the fourth one of both predictor and response variables. For each of these scenarios, one of the two initially selected training sets was selected as the reference set for the normalizations, and the external validation set was normalized in accordance, all of this by running option 3 itself.

Again, out of the 900 models acquired in the four situations, the one with the highest remaining 30% validation R was selected, and then externally validated 4 times with the selected validation set normalized in accordance with the 4 scenarios listed. The regression plots of such validations, finally adding up to 16 plots, represent each normalization's employment performance. Bear in mind that the generation of 900 models is performed in search of convergence at reaching a model for each scenario and that comparability among them is possible.

3.6 PLS modelling procedure - training and validation processes execution

After introducing all the relevant tools employed for the accomplishment of the present study's ambitious goal, the whole modelling procedure employed needs to be addressed.

To start with, a strategy to tackle the different factors that condition the building of models was established and is explained below, also represented in Figure 7.

In general, variability in the models' development can be attributed to three different topics, which are represented by the following questions: which set, or groups of sets conforms the best data source for building each lipid family PLS model? Which spectral regions and number of components provide the most proficient model? Which distribution of samples in the training-validation split provides the most proficient model?

The three questions are written in the order they have been dealt with, and for each one of them, a procedure to reach the most proficient results has been designed and employed.

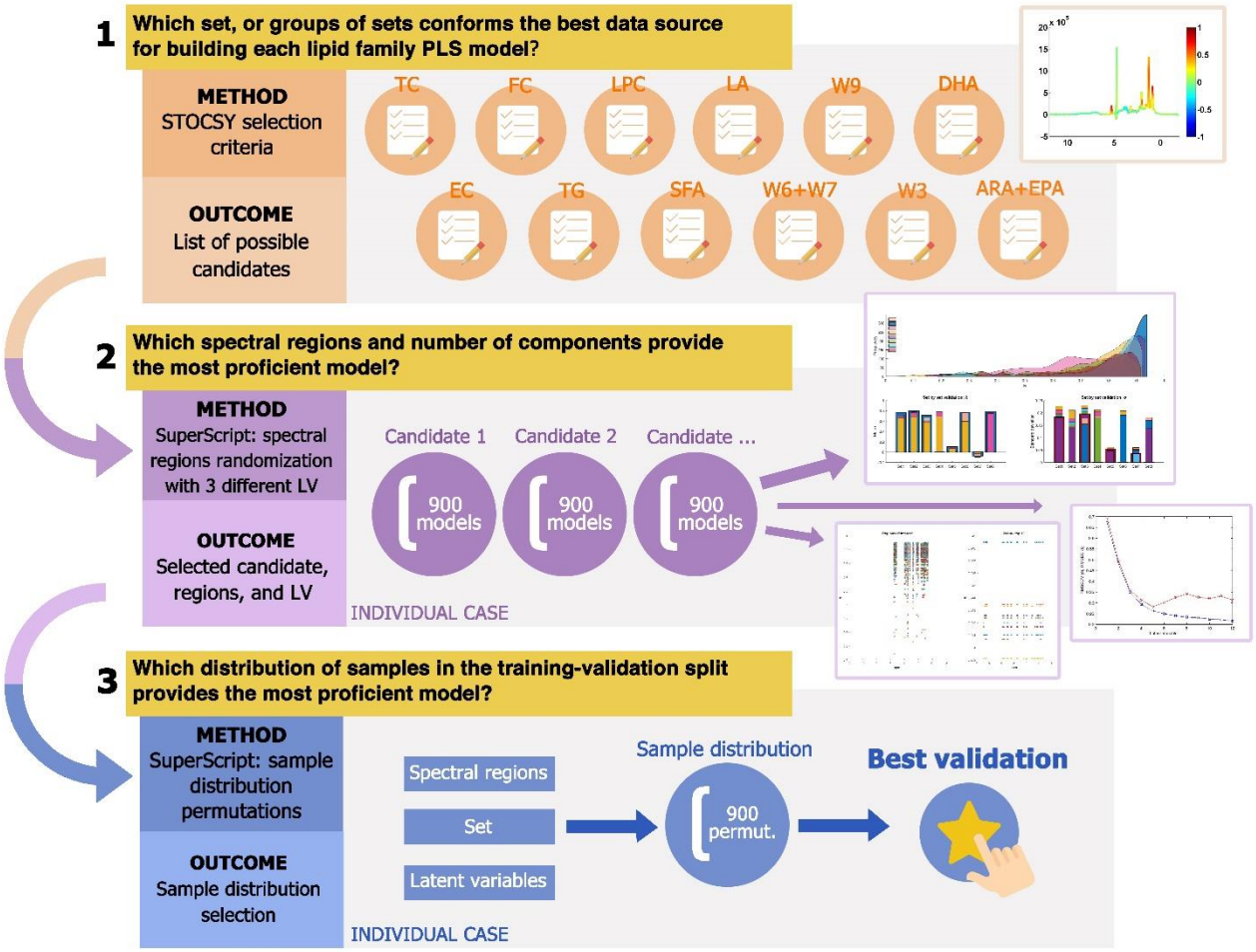


Figure 7. Outline of the procedures followed for addressing the three questions posed

For each question, the method and outcome are represented. The two last questions' procedures are represented for a single lipid case.

First, to obtain an initial selection criterion to rely on for the selection of sets employed in the development of each lipid's predictive model, STOCSY heatmaps were generated between each set, and each lipid family, resulting in a final number of 108 heatmaps. Then, the most strongly correlated pairings in each lipid family, referring to those maps where the most abundant correlations range from 0.3 to 1, indicated which sets were meant to be kept for the further development of each lipid's predictive modelling. With these sets, a series of combinations of candidates for the final lipid model building, were arranged, including individual sets, pairings, and even three sets merging depending on the lipid studied and the analyst's criteria. To refer to the group of candidates, the term *selected sets* will from now on be employed.

From now on and for explanatory purposes, the rest of the procedure will be explained regarding the development of a single lipidic family prediction model, nevertheless, be aware that the procedure was performed for each one of the twelve lipidic families aimed at studying.

Next, an iterative procedure through the SuperScript tool for reaching a conclusion in terms of which of the selected sets performs best for building the desired lipid predictive model and which spectral regions selection provide the most efficient representation of the lipid of interest was performed and evaluated by the analysis of the outcoming models themselves. The procedure consisted of the following, for each one of the selected sets or definitive model development candidates, 900 models were generated through the SuperScript tool procedure, starting from a randomized training-validation split (task 4), the program was used to generate a baseline spectral region selection (option 0), select an initial number of latent variables (option 2), and build 900 models randomizing the baseline spectral regions included (option 5), which is the natural procedure of the software. Note that, the generation of 900 models with option 5 translates to a number of 300 baseline spectral regions randomizations. Also, bear in mind that each model's performance is assessed by the R resulting from the validation of the model with the remaining 30%, which is automatically done in the execution of option 5.

When the aforementioned procedure was performed for each of the selected sets, a convergent and refined representation of each candidate's performance in the development of the lipid of interest predictive model was obtained and naturally saved in the outcoming datasets of the SuperScript tool, then these were exported to Matlab 2020 for the development of an illustrative graphical representation of each set's modelling behaviour for comparison purposes among them. The resulting plot, Figure 8-A, works as a histogram, it is an overlapped representation of each candidate's outcoming R's frequencies, with the horizontal axis representing the R's values and the vertical axis the number of models or frequency of obtention.

Furthermore, external validations of the models were performed through the following process. For each of the candidate's 900 models, external validations with the different model-building unemployed sets were executed, in other words, each model was validated against the sets that weren't employed for the development of the model itself. Again, this was executed for the 900 models generated with each candidate. Then, each set's 900 models performance in this external validation procedure was assessed through the average R of the external validations with each unemployed set, together with the standard deviation. To further clarify and considering only one of the selected sets for the lipid of interest model building, the 900 models generated through this candidate were individually validated against the remaining sets, and for each of the remaining sets, the R average and standard deviation of the 900 models' validations were calculated.

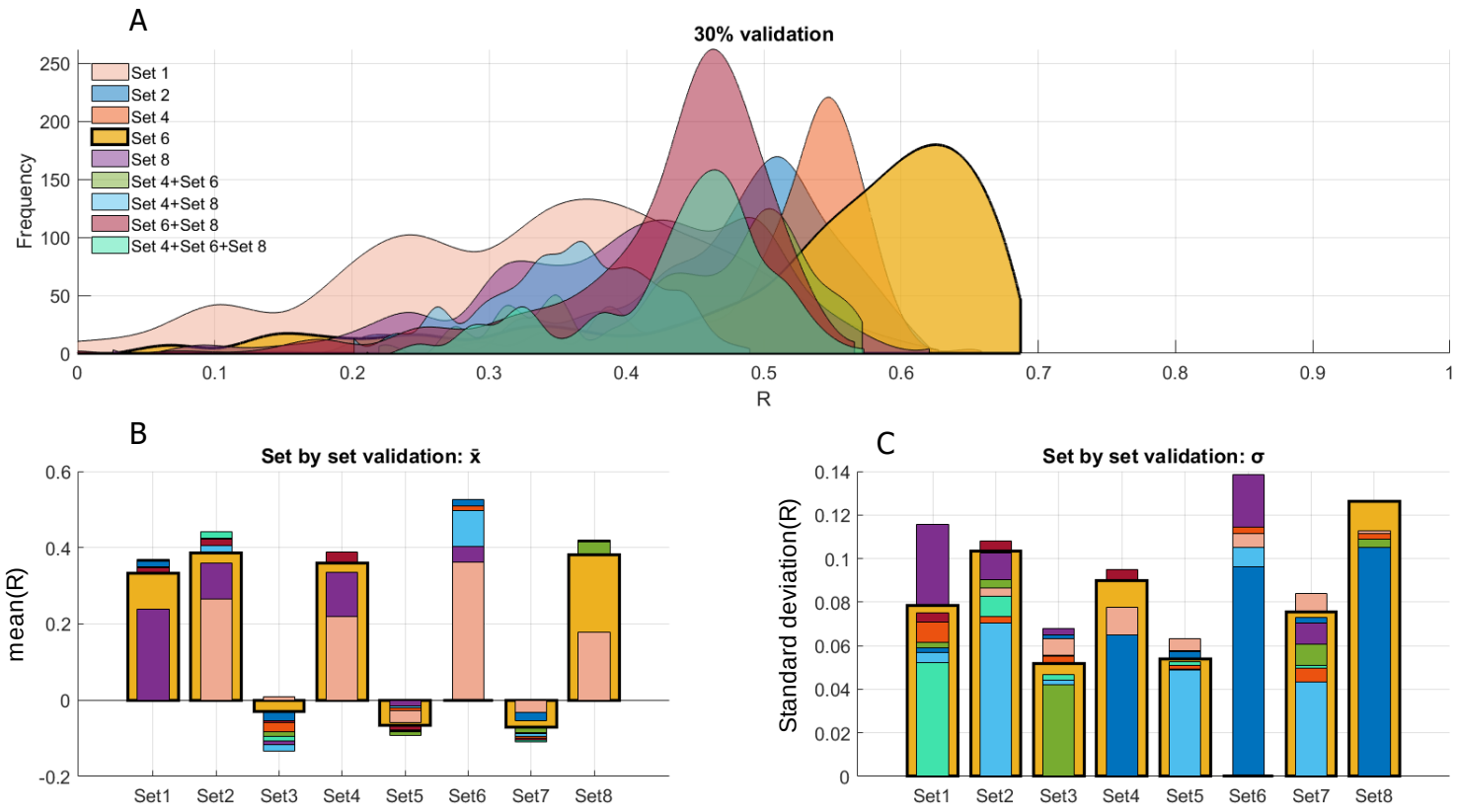


Figure 8. Set performance assessment plot

Figure A showcases each candidate's 900 models validation performance. Figure B expresses the average R of validating each candidate's 900 models against the remaining sets. Figure C shows each candidate's 900 models R validation standard deviation.

This information was also exported to Matlab 2020 to be represented through two bar plots, with the horizontal axis of both consisting of the lists of sets at disposal. For the R average plot, the vertical axis consists of R values, for the standard deviation plot, the vertical axis consists of standard deviation values, both represented in Figure 8-B and Figure 8-C respectively.

From this, the best performing set or what is the same, the one that reaches the highest R values more frequently when validated with the remaining 30%, presents the overall highest external validations average R, and smallest standard deviation among its external validations, was selected, and for that set, a representation of the randomly acquired spectral regions selection in the basis of the regions generated through option 0 was developed, aiming at determining which spectral regions selection works more efficiently for the portrayal of the studied lipid, based on the selected set. The outcoming graph, portrayed in Figure 9-A, is composed by the 15274 spectral points in the horizontal axis, enabling the linear representation of each model's selected spectral regions through dots, and R values in the vertical axis, so that each model's selected spectral regions plotting has the height of the R value it generated. Through this last plot, the most proficient spectral regions, corresponding to the ones at the highest position in the graph, were selected.

With the selection of sets and spectral regions, the two first questions regarding the variability in the development of the models are addressed. Next, the procedure followed for the assessment of the best sample distribution in our training-validation split, which addresses the third question, is explained.

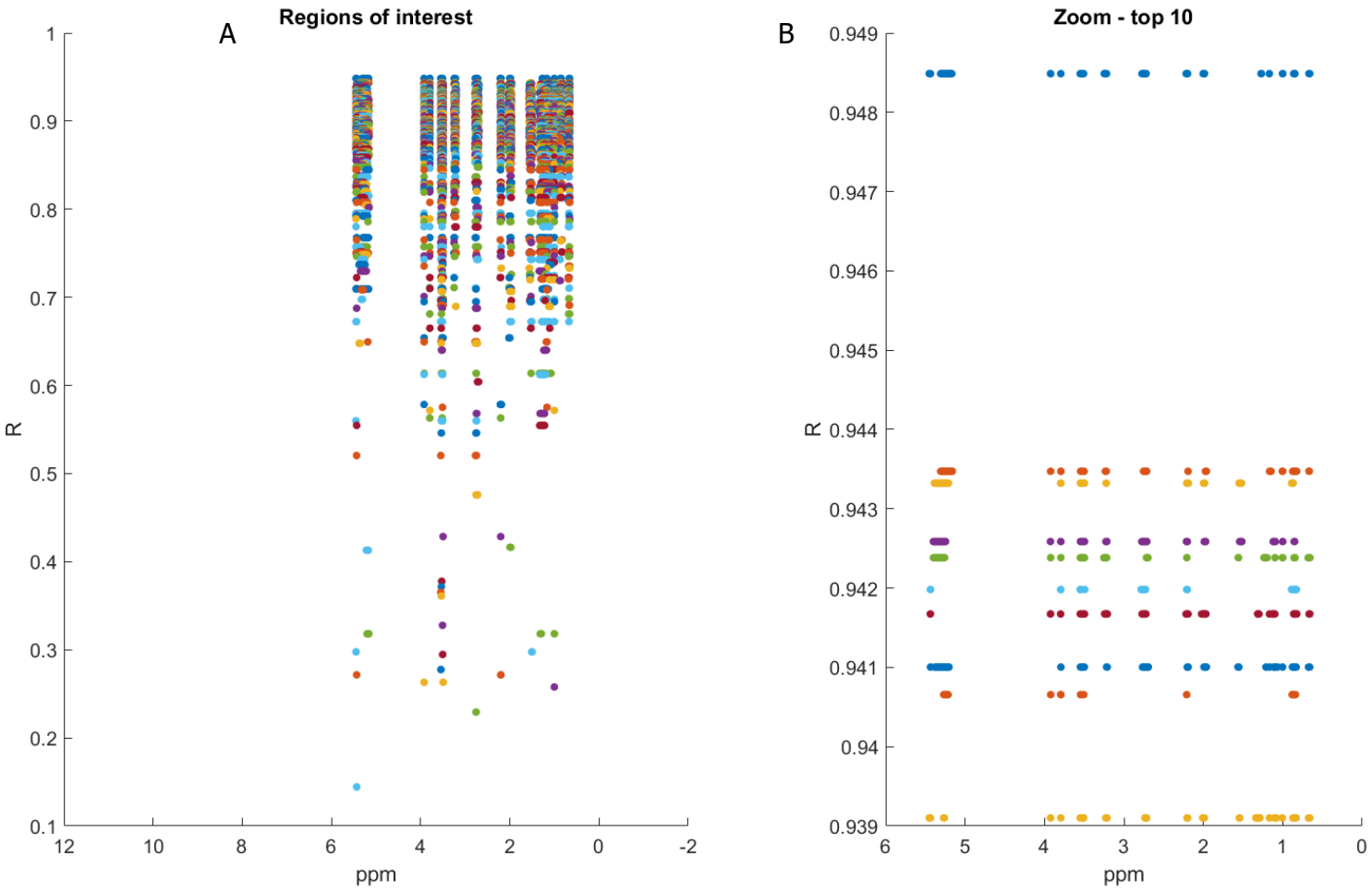


Figure 9. Spectral regions assessment plot

Figure A represents a selected candidate 900 models' spectral regions randomizations, with Figure B being a zoom on the top 10 best performing models in terms of the R they generated.

First, a firm decision regarding the number of components used for the final building of the models must be made. Up to this point, decisions taken for the selection of components number were based on option 2, which considers the randomly generated training-validation split performed through task 4 and the base-line spectral regions selected in option 0, nevertheless, changes in these two factors may result in subtle alterations of option 2's outcome. Since spectral regions have already been selected and will no longer introduce variance in the selection of components number, only the training-validation split sample distribution remains with influence. To address this issue, 10 different permutations of the training-validation split sample distribution were performed, and the most efficient number of latent variables was assessed through option 2's graphical outcome, expressed in Figure 10, selecting as the final number of components to be employed the one most frequently acquired.

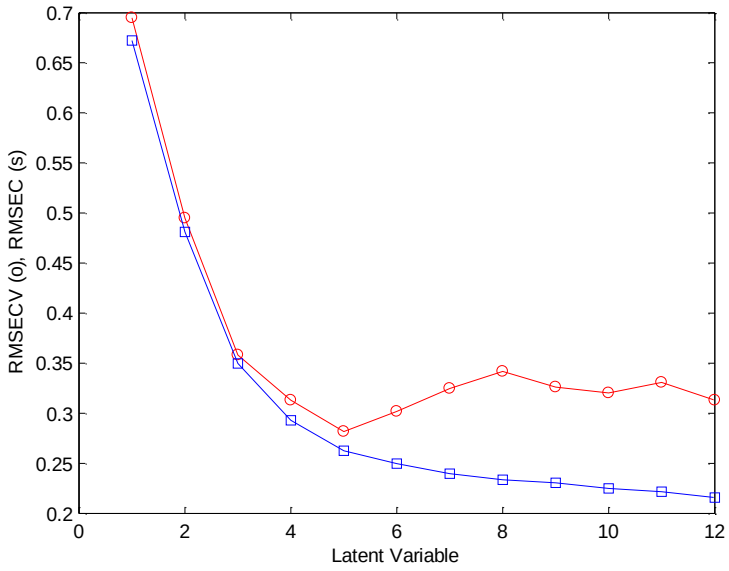


Figure 10. Latent variable number assessment through the RMSE

The red plotted line indicates the RMSE against latent variable increase through the cross-validation technique, and so does the blue plot, but without cross-validation employment.

Finally, with the selection of sets, spectral regions and latent variables performed, only the most proficient sample distribution in our training-validation split of the data for the building of the definitive model remains uncertain. To address this issue, 900 permutations of the training-validation split to acquire a wide range of sample distribution possibilities were performed, and together with the selected regions and latent variables number, predictive models for each one of the 900 permutations were built, all of it through the employment of task 7. Resulting from this procedure, a list with each permutation's predictive model and its remaining 30% validation R was acquired, selecting as the definitive model the one with the highest R. Regression plots the described validations were obtained for each model, also showcasing the relative root mean square error as a percentage (%rRMSE), calculated as follows:

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

Equation 4. Relative root mean square error expresses as a percentage

The predicted values are represented as 'y_{pred}' while experimental values as 'y_{exp}'.

After performing the entire procedure for each of the lipidic families studied, 12 linear regression predictive models were acquired and put to the test through their external validation against each one of the available sets for the project's development, excluding those used for the building of the model. Most importantly, in this final validation procedure set 9 was included, providing actual representation of a completely external set to those used for the models development. This final validation was conducted through task 8, and results were projected onto a heatmap, where the horizontal axis is a list of the 9 sets at disposal and the vertical axis a list of the 12 models acquired, the higher the R acquired from the models validation with each set, the brighter the colour representation.

4 Results: development of PLS models for lipidic families quantification

4.1 PCA-based exploratory analysis of the sample sets

PCA analysis regarding the LED spectra, Figure 11-A, results in a majorly overlapped representation of the 9 sets studied, confirming that homogeneity is found among spectra and discarding non-comparability. The first two principal components captured the most variance and showcase the values of 30% and 26%. The differential patterns of the sample sets found among the plot aren't very significant, in other words, few differences are found among LED spectra, instead, evidence of their similar behaviour has been acquired.

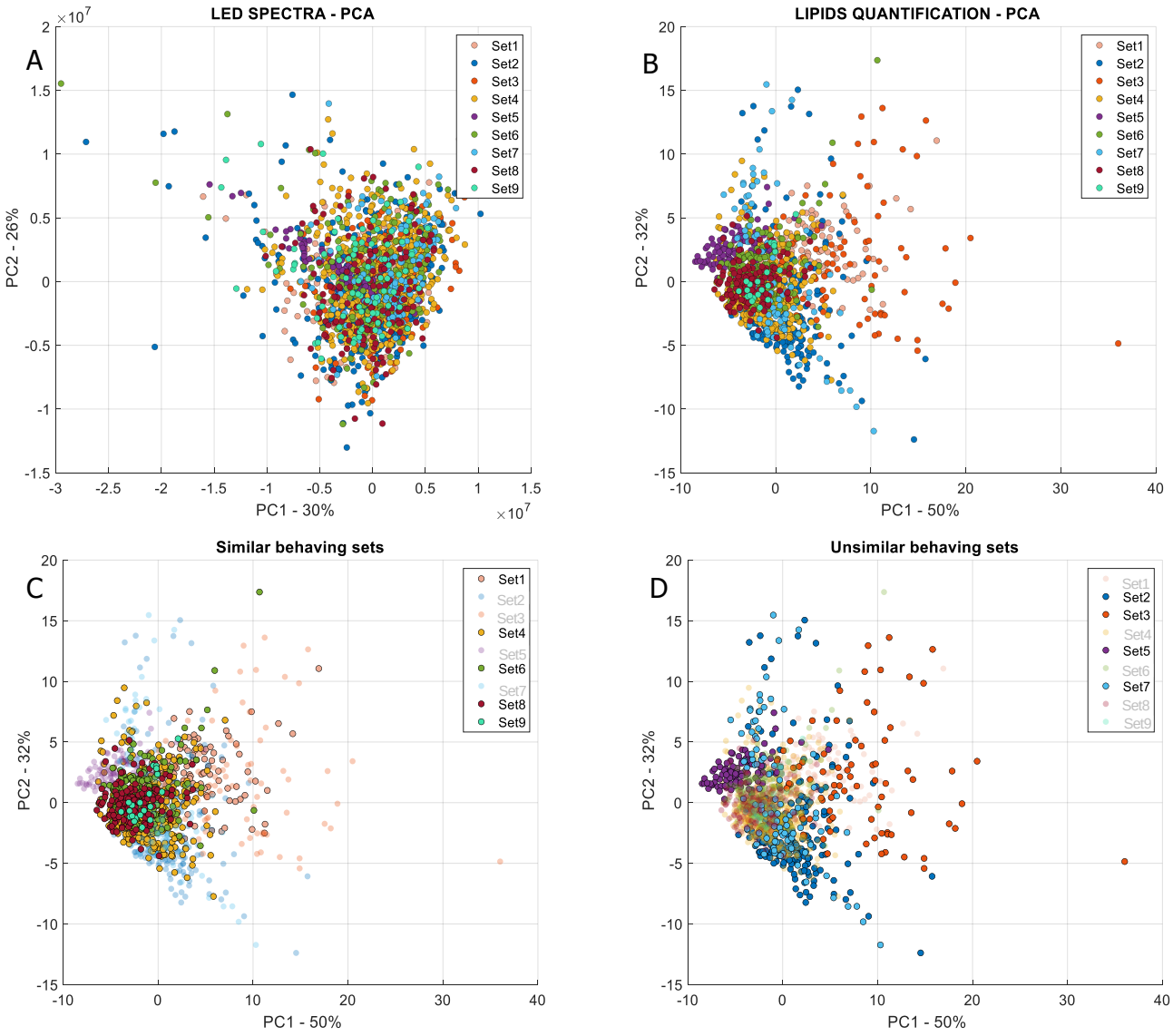


Figure 11. PCA of the LED spectra and the lipid quantifications

In the upper left corner (A), the PCA of the 9 sets available is provided, PC1 captures a variance of the 30% (PC1 – 30%) while PC2 a variance of the 26% (PC2 – 20%). In the upper right corner (B), the PCA of the lipid quantifications is provided, PC1-50% and PC2 – 32%. The left bottom corner graph (C) is the lipid quantifications PCA highlighting the similar behaving sets whereas the right bottom corner graph (D) highlights the unsimilar behaving sets.

Nevertheless, a small percentage of outliers can be appreciated in the PCA graph and seen dispersed across both principal components, although a little more significantly along the horizontal principal component (PC1).

Positively, these obtained results affirm that there is no batch effect in the 1H-NMR data acquisition nor differences in the sample collection and processing across the several data samples. Also, it confirms that merging of sets' LED spectra for the development of the models can be performed, since homogeneity among sets' spectra exists, all of this aiming at obtaining the best training set possible. Nevertheless, this does not mean that the merge of different sets for model development will lead to better results than not doing so, this issue will be addressed in section 4.3.

Concerning the lipid quantifications PCA, Figure 11-B, homogeneity among all data is not completely found, which is concordant to what can be expected after observing Figure 4. In this second PCA analysis, the two components that capture the maximum variance present values of 50% and 32%, which directly indicate that clearer differential patterns are found among the sets, in fact, two differently behaving groups can be appreciated and have been highlighted in Figure 11. A group of similarly behaving sets, Figure 11-C, is showcased composed by sets 1, 4, 6, 8, and 9, these are found overlapped in the centre position of the space, and although they have a small number of outlying samples, the main behaviour is homogeneous, ensuring comparability and sets coupling possibility. Set 9's similar behaviour to the other sets in this cluster provides evidence regarding its acceptance as validation set for the future procedures.

In terms of the unsimilar behaving sets, Figure 11-D, these include sets 2, 3, 5, and 7, which are seen scattered across both principal components, dodging the central space where the other groups are found and exhibiting major differences among them. Set 2 is greatly spread across the vertical principal component (PC2), as well as set 7. Set 5's data is homogeneous among itself but its location in the space is not consistent with the similarly behaving sets. Set 3 is broadly scattered across both principal components, proving that great variance is found within its samples.

Being aware that the lipid quantifications PCA's final PC1 and PC2 weights of original variables were checked to prevent having a single original variable be the source of major variance in the reduced space, the unsimilar behaving sets differences can be accepted and attributed to general variability among the lipidic extracts and quantifications, still, it is inevitable that some original variables have greater influence over others in each of the set's variability.

4.2 Study of the normalization's application

Several instances of this study were performed, mostly leading to uncoherent results, but out of all the perplexity, one of the performances is provided to represent the main tendency of the different tests executed.

The presented case was obtained under the following conditions:

Sets 1 and 6, and the EC lipidic family were selected owing to their intrinsic differences regarding the selected lipid's quantifications, which can be appreciated in Figure 4, making them the ideal scenario for the intended normalization's application. The external validation set chosen was set 3, which regarding the EC quantifications, resembles those of set 1. The reference set chosen was set 1, therefore set 6 and set 3 were relocated to set 1's scale and range.

Through the 4 normalization scenarios study, Table 3 was built, consisting of each scenario's selected model's validation with the 4 scenarios normalized version of set 3,

showcasing the impact of the 16 combinations possible with the four scenarios proposed on the model development.

First of all, it can be noted that the validation set's different normalizations are meaningless, owing to the fact that these do not alter the acquired R for a single row or model, nevertheless, changes are found in terms of the predicted values range when the validation set's quantifications are normalized, which is represented in the second column.

The only normalization influence in the model performance is found when altering the training set. As it can be seen in Table 3, normalizing the training set's quantifications, spectra, or both, generates different model performances, with the spectra normalization slightly but also insignificantly improving the model, and the other two situations worsening it.

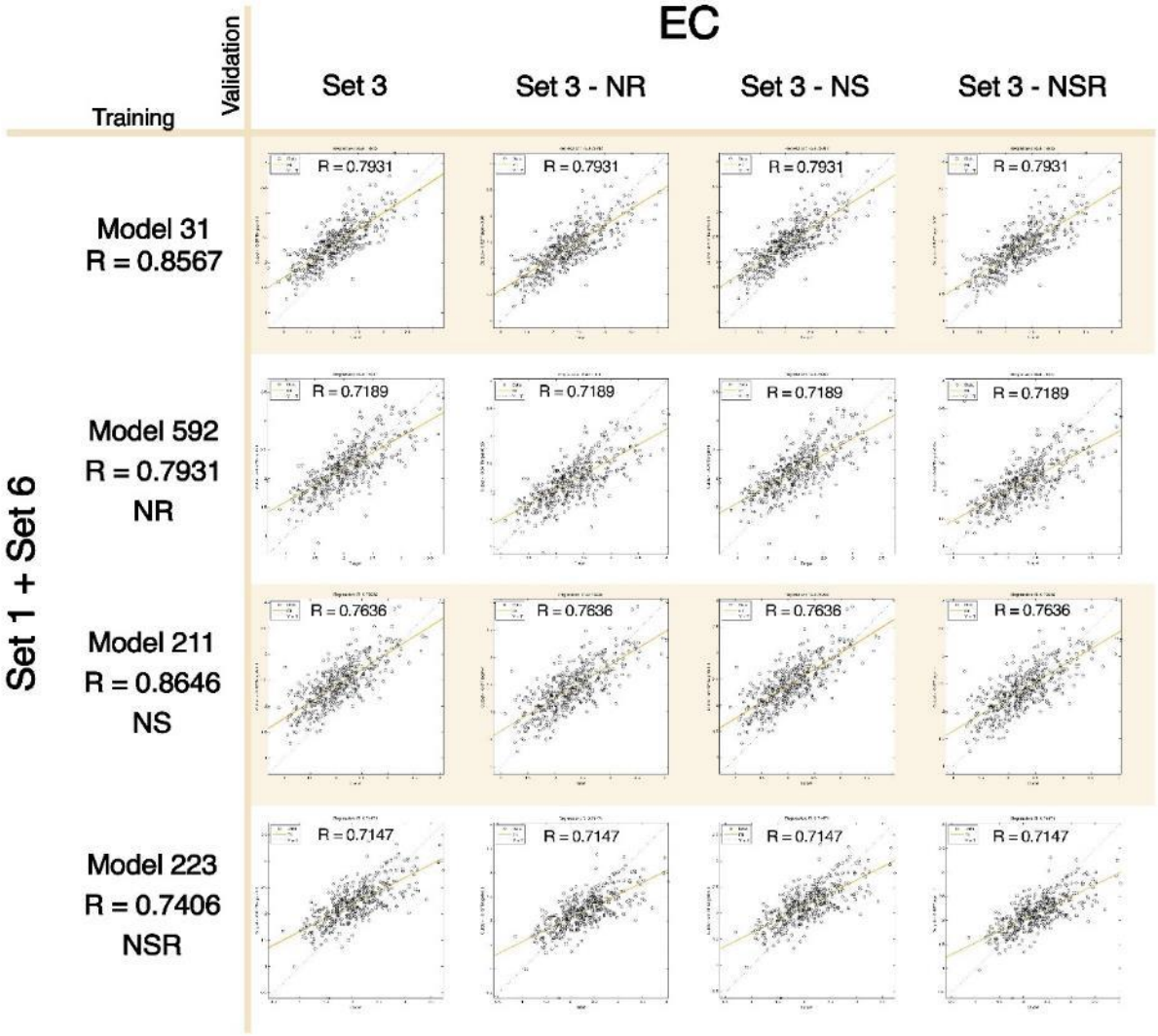


Table 3. Plot regression of each normalization scenario best model's validation, with the corresponding R

Plot regressions and Rs are given for each normalization scenario, ultimately obtaining 16 plot regressions or in other words, validations. Each row's model and column set represents a normalization scenario; no normalization, quantifications normalization (NR), spectra normalization (NS), and both spectra and quantifications normalization (NSR) respectively in descending order and left to right. The model number corresponds to the model generation number and the R provided under such name corresponds to the 30% validation R.

These results are meant to serve as an approximation of the tendency that using normalization generated when developing the predictive models. Ultimately, normalization employment was denied owing to the unsuccessful results acquired, which don't provide a clear explanation of the implementation of the method, therefore, following results did not incorporate any normalization procedure in their acquisition process.

4.3 Sets selection for the study of each lipid

Diving into the predictive model development results, the outcomes related to the answering of the first question posed in section 3.6, are provided below.

First, the 108 heatmaps (Annex 1) to undercover which sets provide more representative information about each lipidic family led to the following Table 4, where the initially selected sets for the study of each lipid are presented through coloured cells, and the coloration is not arbitrary, it is an approximate representation of the general correlation intensity interpreted through the STOCSY heatmap, informing about the strength of the relationship of each set's LED spectra with each lipid families corresponding quantifications. Uncoloured cells indicate that low or null relationships is found between the set and the lipidic family.

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8
TC								
EC								
FC								
TG								
LPC								
LA								
SFA								
W6+W7								
W9								
W3								
DHA								
ARA+EPA								

Table 4. Candidates selected for each lipidic family predictive model building

Each coloured cell represents the selection of the column's set for the development of the row's models, with colours indicating the set's LED spectra strength portraying the lipid of matter. The colour scale matches that of the STOCSY heatmap, with dark reds representing positive correlations close to 1 and dark blues negative correlations close to -1.

Generally, sets 1, 4, 6, and 8 present the most consistent relationship pattern with their lipidic quantifications, set 2 does not uniformly relate to its quantifications, and set 5 and 7 do not present strong enough correlations with the lipid quantifications, to be considered for the development of any lipid model.

Next, the combinations of the selected sets or candidates for each lipid family model development known under the term *selected sets* in section 3.6, is presented.

	1st	2nd	3rd	4th	5th	7th	8th	9th	10th	11th	12th	13th
TC	2	4	6	8	4+6	6+8	4+6+8	2+4+6+8				
EC	1	2	3	4	6	8	2+4	2+6	4+6	6+8	2+4+6	
FC	1	2	3	4	6	8	2+4	2+6	2+8	4+6	6+8	2+4+6
TG	1	2	3	4	6	8	2+4	2+6	2+8	2+6+8		
LPC	1	2	4	6	8	4+6	4+8	6+8	4+6+8			
LA	1	3	4	6	8	1+3	1+4	1+6	1+8	1+6+8		
SFA	4	5	6	8	4+8	5+8	6+8	5+6+8				
W6+W7	1	3	4	6	8	4+6	4+8	6+8				
W9	1	3	4	6	8	4+6	6+8					
W3	1	6	8	1+6	1+8	6+8	1+6+8					
DHA	1	6	8	1+6	1+8	6+8	1+6+8					
ARA+EPA	1	3	6	8	1+8	6+8						

Table 5. Candidate sets combinations for developing each lipidic family predictive model

Each row in the table presents the list of potential candidates for the development of the corresponding lipid family. The columns are labelled based on the position of each candidate within the corresponding lipid's candidate list, with the final column indicating the total number of proposed candidates for each set.

These presented combinations were employed for their corresponding lipidic family set selection procedure, and 900 models were generated through them. For the generation of the models, which again, was through the SuperScript tool, a threshold of 0.25 was selected for option 0 's spectral regions selection and a number of 5 latent variables was selected in option 2, ultimately having models with 5, 6 and 7 latent variables. These parameters selection was ended up being the same for all the candidates among all the lipidic families studied.

Next, the modelling behaviour representation of each of the candidates in each lipid family was represented through the procedure mentioned in section 3.6, resulting in the graphs of Annex 2, out of which the FC, LPC, and W6+W7 representations have been selected to have their results be commented upon, since the three provide a representative overview of the different outcomes obtained. Be aware that for the following results sections, exemplifications through the three lipid families mentioned may also be made for the same purposes.

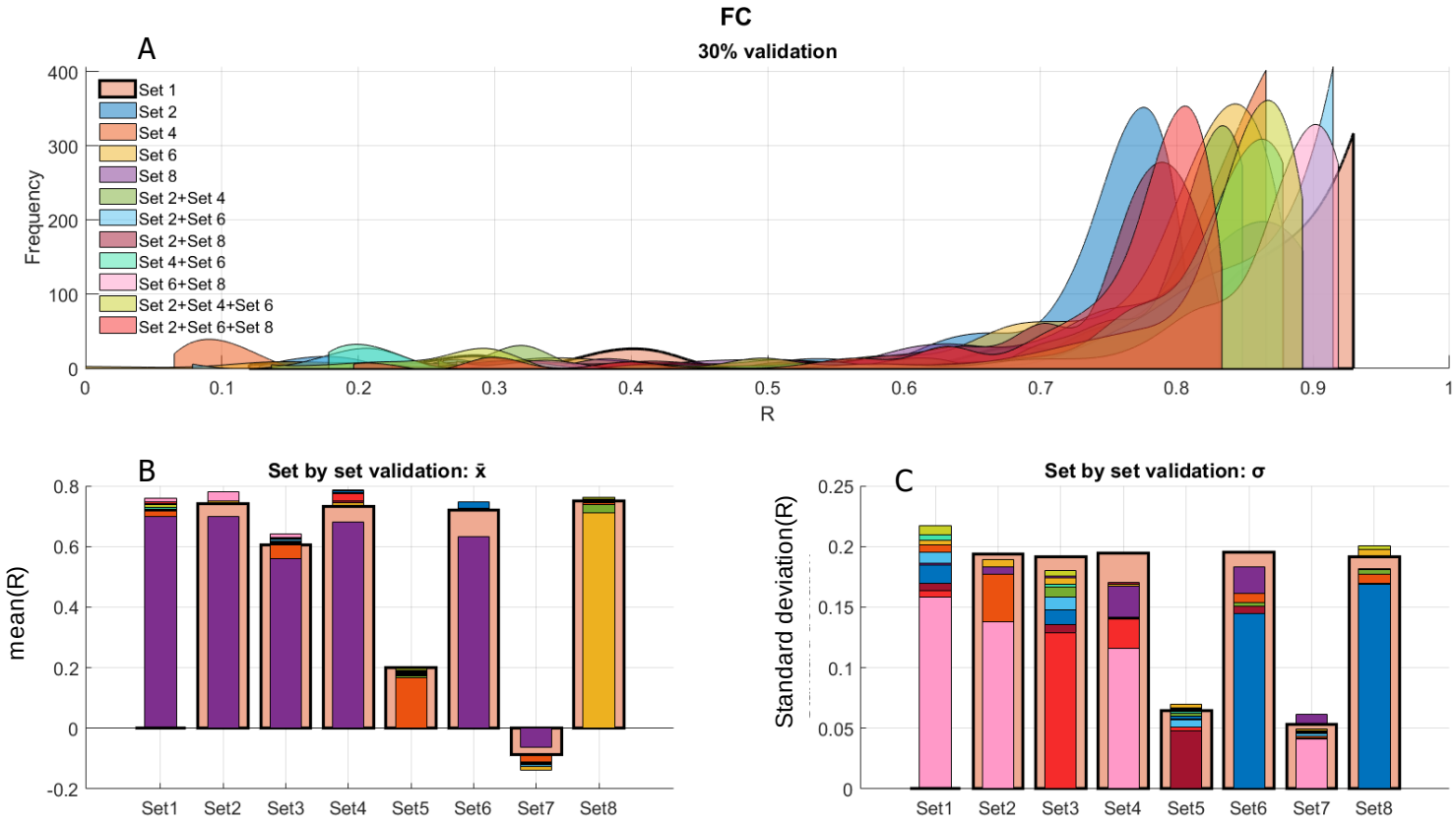


Figure 12. FC candidates' modelling behaviour representation

Wave-shaped signals (A) represent each candidate's 900 models performance based on the 30% validation. Below (B), average R of each candidate's models set by set external validation is provided, and in the same way, standard deviations (B).

With a proficient result, the FC lipidic family's candidates present high Rs with high frequencies, Figure 12-A, as well as high external validation Rs for a major part of the sets and low standard deviations, Figure 12-B and Figure 12-C respectively. Set 1 was chosen as the selected candidate for the final model development, being the one with highest R values, surpassing the 0.9, and providing remarkable set-by-set validations R average as well as standard deviations. The rest of candidates' performance vary within the Rs range of 0.7 to 0.9, falling behind set 1's performance.

Next, the LPC lipidic family results are presented. For this case, worse performance behaviours are obtained, with the best R acquired almost reaching 0.7, seen in Figure 13-A. Logically, the set that reached such R was chosen as the selected set for the development of the final LPC predictive model, consisting of set 6. The selected candidate does not stand out as one of the best in the set-by-set external validation but shows very low standard deviation values, especially if compared to those of the FC lipid selected candidate, Figure 13-C.

Last, Figure 14 showcases the results for the W6+W7 predictive modelling candidate assessment. In this last scenario and unlike the other two, bigger differences are appreciated among the candidates (Figure 14-A), the result can be approached by the identification of two clusters, one composed by set 1 and set 3, and the other one by the rest. The first cluster's performance is very weak compared to that of the second cluster, from which the selected candidate was chosen, also consisting of set 6. For the lipidic family W6+W7, set 6 provides R values that surpass the 0.9, very high Rs average regarding the set-by-set validations, and standard deviations in a similar range to those expressed in the LPC context, appreciated in Figure 14-B and Figure 14-C respectively.

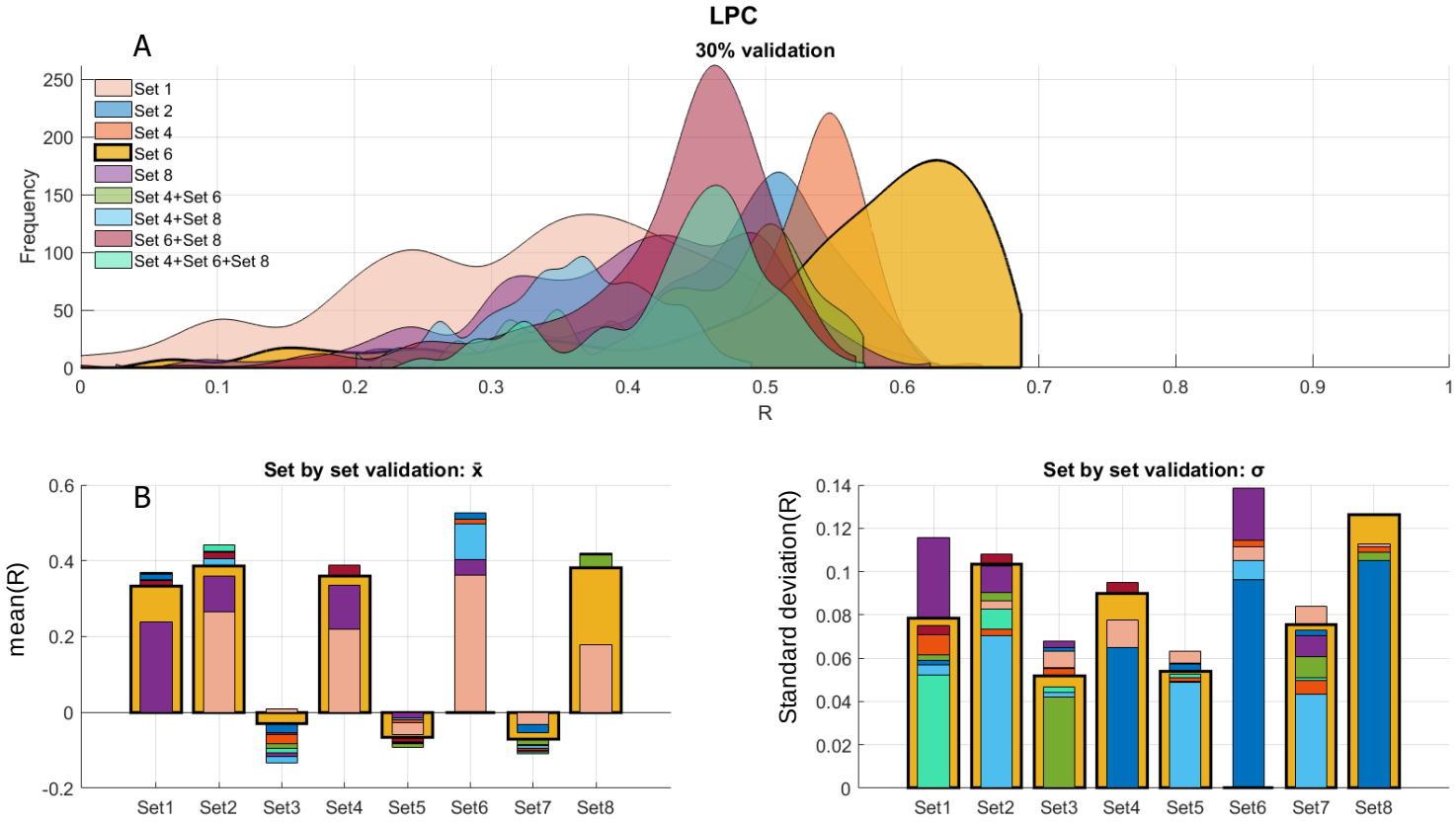


Figure 13. LPC candidates' modelling behaviour representation

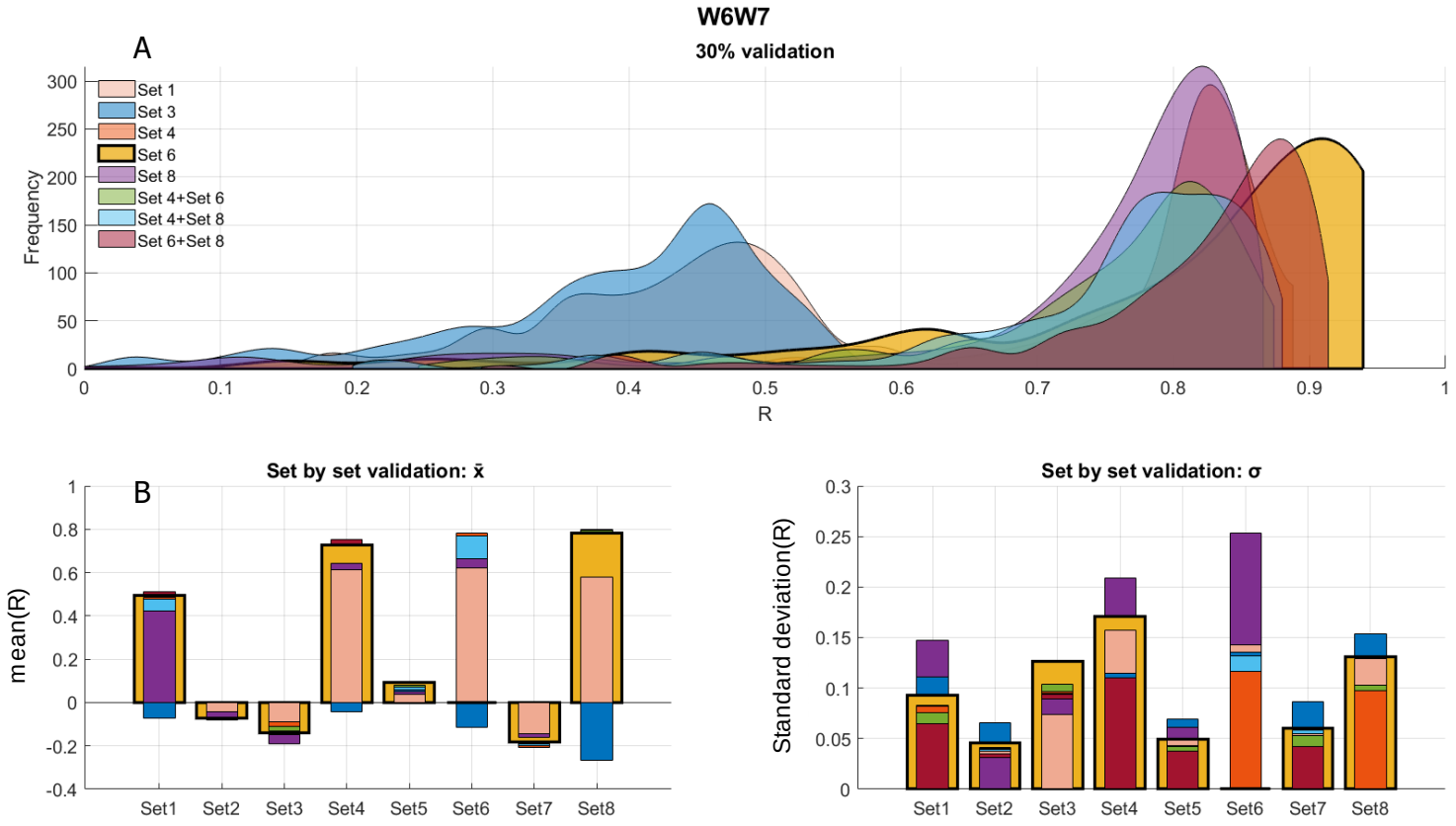


Figure 14. W6+W7 candidates' modelling behaviour representation

Regarding the rest of lipidic families chosen candidates, all of which were selected through the same reasoning as the one given for the three represented lipidic families, the following Table 6 enumerates them all, and again, graphical representations are presented in Annex 2.

	TC	EC	FC	TG	LPC	LA	SFA	W6+W7	W9	W3	DHA	ARA+EPA
Set1			X			X						
Set2				X								
Set3												
Set4	X											
Set5												
Set6		X			X	X		X	X	X		
Set7												
Set8							X				X	X

Table 6. Selected candidate for each lipid predictive model

4.4 Selection of the spectral regions most associated with the concentration of each lipid and most efficient latent variables number

Following the procedure of development for the three lipidic families approached in last section, each selected candidate was examined in search of the best performing model's spectral regions out of the 900 models generated. To represent the data dealt with, the following Figure 6-A and Figure 6-B representations were generated.

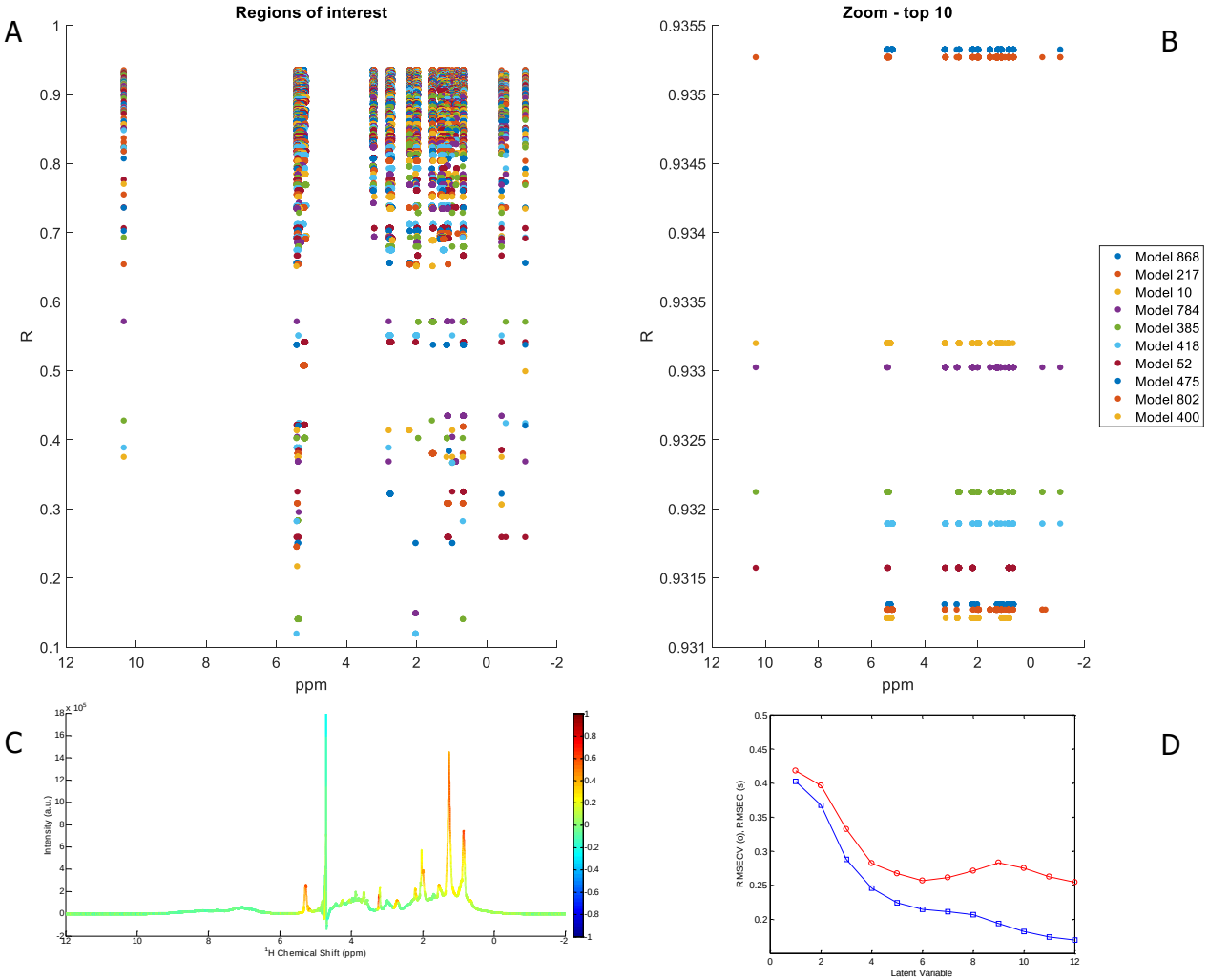


Figure 15. Baseline spectral regions randomizations of FC's selected set and RMSE vs LV increase prototype plot

To the left (A), the chosen set's, consisting on set 1, baseline spectral regions randomizations are represented against the hortizontal axis, which matches the axis of LED spectra, and for comparison purposes, the STOCSY is provided below (C). Randomizations' height is selected according the the R value they generated, ultimately acquring a dispered representation. In the upper right, the highest R acquiring regions or top 10, is specifically plotted (B). RMSE in cross validation vs latent variable increase protoype plot is provided through the red line on the bottom right corner (D).

The process of baseline spectra regions randomizations becomes apparent through Figure 15, baseline spectral regions can be noted framing the reach of the randomizations with approximately 8 initial regions, and through the aggregation of the STOCSY heatmap, matching regions between the heatmap and the plot can be appreciated. The model that acquired the highest R when validated with the remaning 30%, reacing the value almost 0.93, consists of model 868, which contrary to the other models shown in the zoom plot does not include spectral points in the range surrounding 10 ppm and the region before reaching 0 ppm. Model 868's spectral regions were selected for the further development of FC predictive model.

Furthermore in the development of the model, now that the final training set and spectral regions have been selected, the latent variable number selection remains. At the bottom right corner of Figure 15, a prototype plot of the RMSE against the increase of latent variables is provided, showcasing the most frequent outcome of the 10 permutations performed to asses this number. For this case, the elbow shape indicating the best number selection was determined to be at 4, which was selected to be the number employed in the final model acquisition.

Regarding the results for the LPC selected candidate, seen in Figure 16-A-B, the incorporation of a greater number of spectral points can immediately be observed, again, matching the STOCSY regions where major correlations are detected, specifically those regions that surpass the correlation threshold of 0.25, selected for option 0's execution in all the lipid models's development, as already detailed in the previous section.

The chosen spectral regions selection for LPC are those of model 97, which is the one that obtained the highest validation R, with a value of 0.7. With the selected set and spectral regions, permutations to assess the best number of latent variables for the further development of LPC's model were performed and through Figure 16-D, the selected number can be appreciated, which was 3.

Last, in the W6+W7 scenario represented in Figure 17-A-B, fewer and more concentrated baseline spectral regions can be noticed in the range of 0 to 6 ppm, again, matching the regions of the STOCSY where major portrayal of the lipid of intrest exists. With a validation R of 0.95, model 592's spectral regions were selected.

Regarding the latent variables number selection, the most abundant result of the permutations performed are presented through Figure 17-D, where an elbow shape can be appreciated at number 4, being the one chosen as the number of latent variables selected.

Regarding the rest of lipidic families' graphical results of the spectral regions selection with their corresponding STOCSY, and latent variables selection, these are all provided in Annex 3. In the following section, the listing of each lipid's selection is addressed through Table 7.

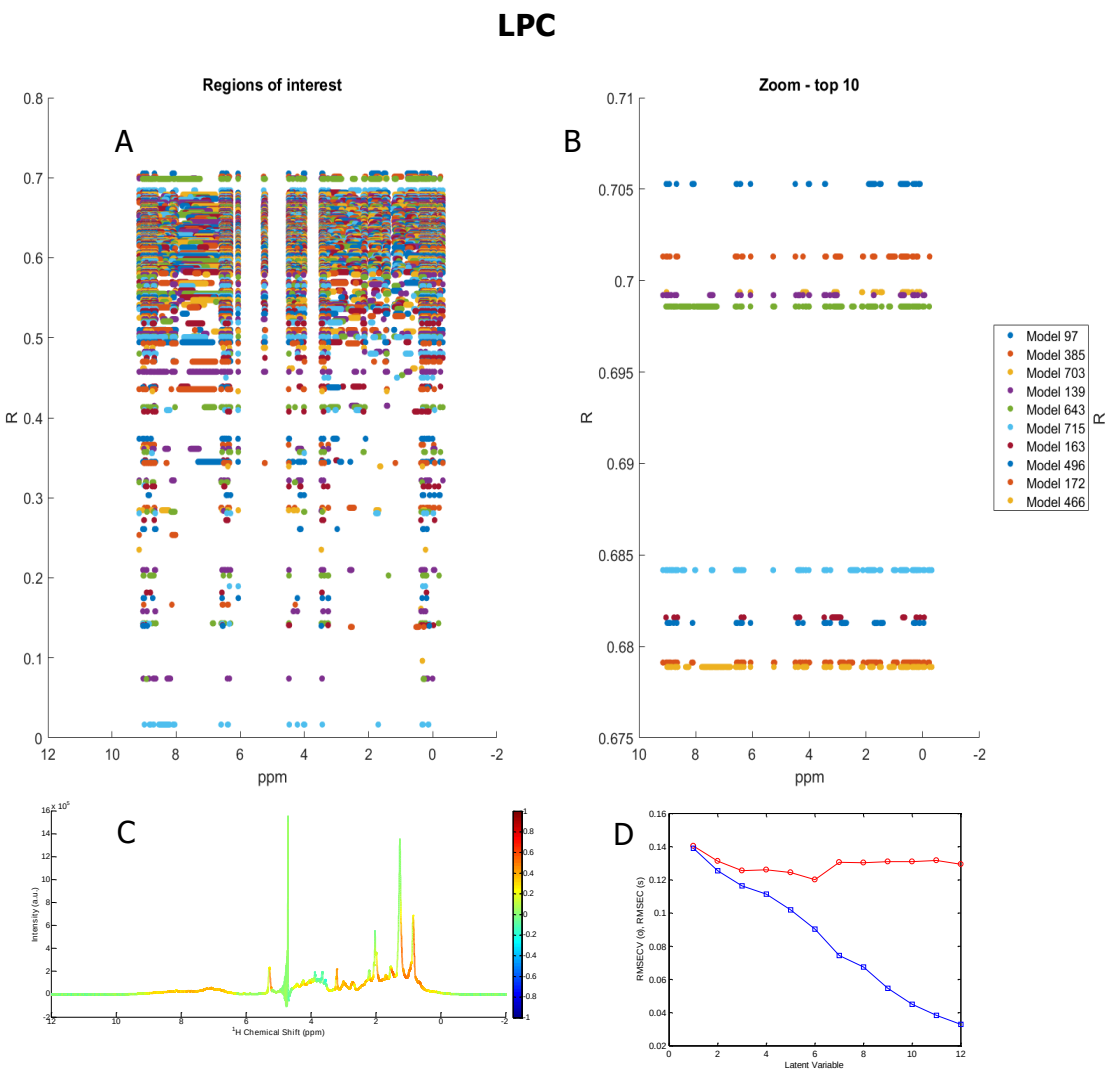


Figure 16. LPC's baseline spectral regions randomizations of the selected set and RMSE vs LV increase plot

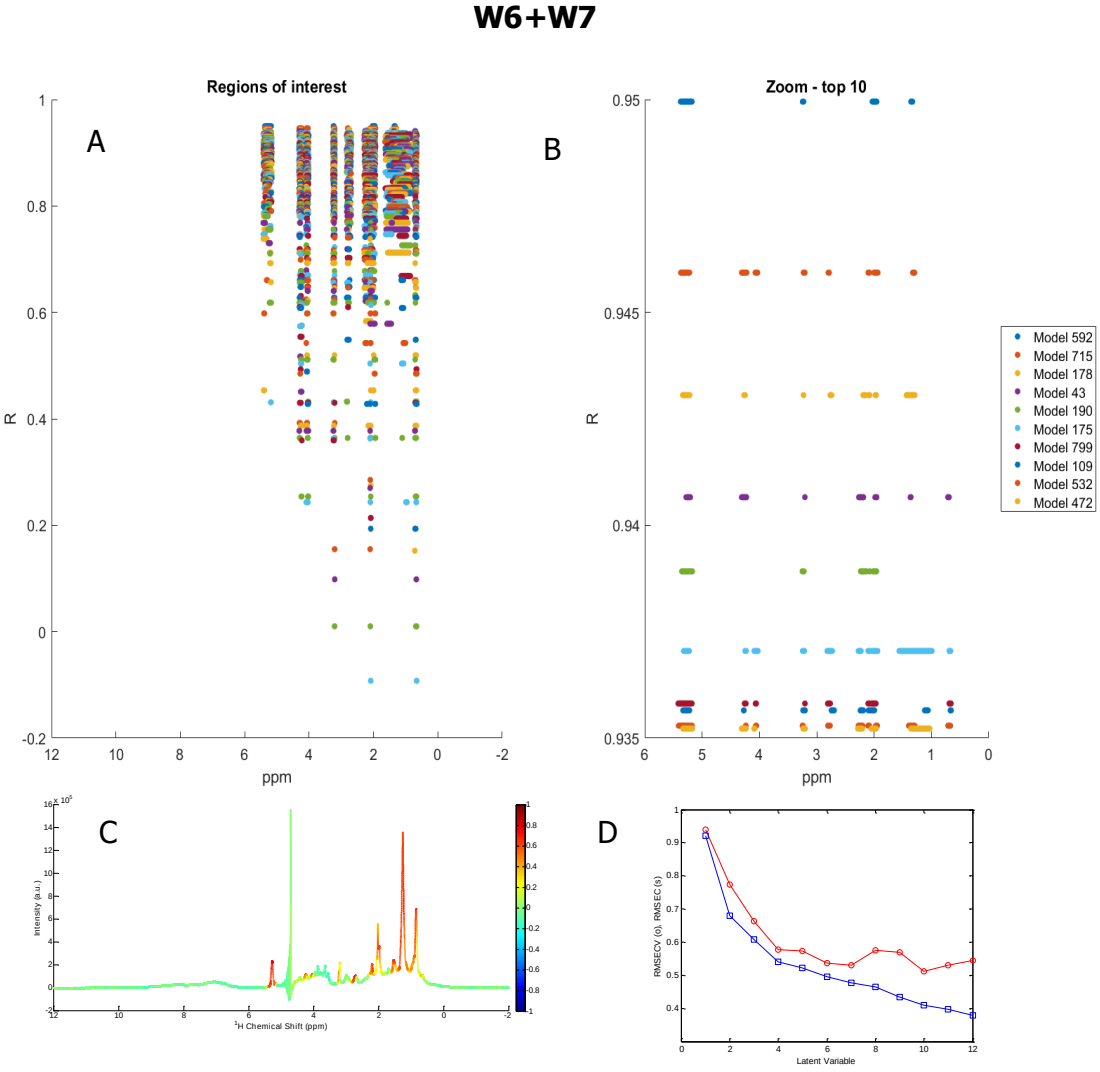


Figure 17. W6+W7's baseline spectral regions randomizations of the selected set and RMSE vs LV increase plot

4.5 Distribution of samples in the training-validation split selection - final PLS models for lipidic families prediction

For this final step in the development of the sough-after most refined models, results for each lipidic family are provided.

Considering each lipid family's selected training set, spectral regions, and number of latent variables, the best training-validation split sample distribution was selected out of the 900 permutations performed, proceeding as explained in section 3.6, ultimately achieving the final answers of the three questions initially posed. Such answers are portrayed in the following table.

	Training set	Spectral regions (model)	Latent variables	Permutation number	R-30%	σ (900 permutations)
TC	4	*526*	5	85	0.9637	0.0141
EC	6	*259*	5	893	0.9336	0.0413
FC	1	*868*	4	384	0.9468	0.0551
TG	2	*553*	2	146	0.9695	0.0154
LPC	6	*97*	3	716	0.8234	0.0716
Linoleic	1+6	*823*	5	520	0.9580	0.0143
SFA	8	*754*	5	684	0.8078	0.0917
W6+W7	6	*592*	4	315	0.9668	0.0394
W9	6	*493*	4	54	0.9822	0.0397
W3	6	*883*	5	813	0.8862	0.0943
DHA	8	*454*	5	670	0.9032	0.0952
ARA+ERPA	8	*409*	4	240	0.8624	0.0789

Table 7. Answering the three questions, each lipid family model's selected inputs and permutations procedure's results

Four first columns contents consist on: selected set (section 4.3), spectral regions (section 4.4), number of latent variables (section 4.4), and permutation or training-validation split sample distribution for the development of the final model. Fifth column includes the R validation with the 30% of the first four columns resulting model, and last column consists of the standard deviation of the 900 permutations.

Next, and referring to Table 7's fifth column, the series of regression plots obtained from that validation of the definitive model with its corresponding remaining 30% of samples, is provided, showcasing the acquired prediction performance for each lipidic family through the %rRMSE and the R, consisting in all cases of values over 0.8.

More specifically, most validations obtained are over 0.9, with some of them reaching values close to 1, which represents the perfect prediction. Nevertheless, instances of %rRMSE over 20% are found among the validations. In Figure 18, the plot regression for each acquired model and their regarding information are provided.

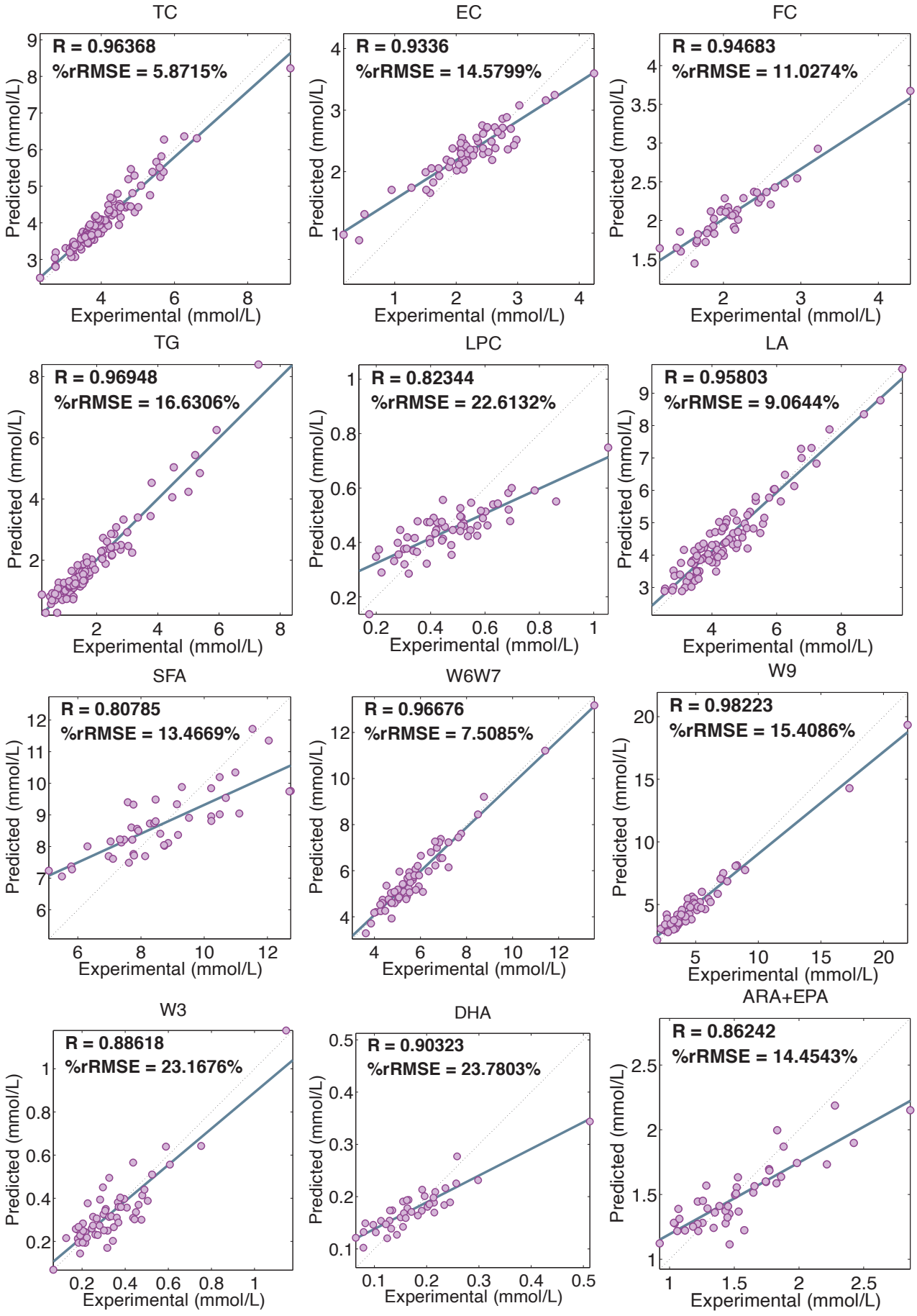


Figure 18. Regression plots of predicted vs experimental values employing each lipid's final predictive model

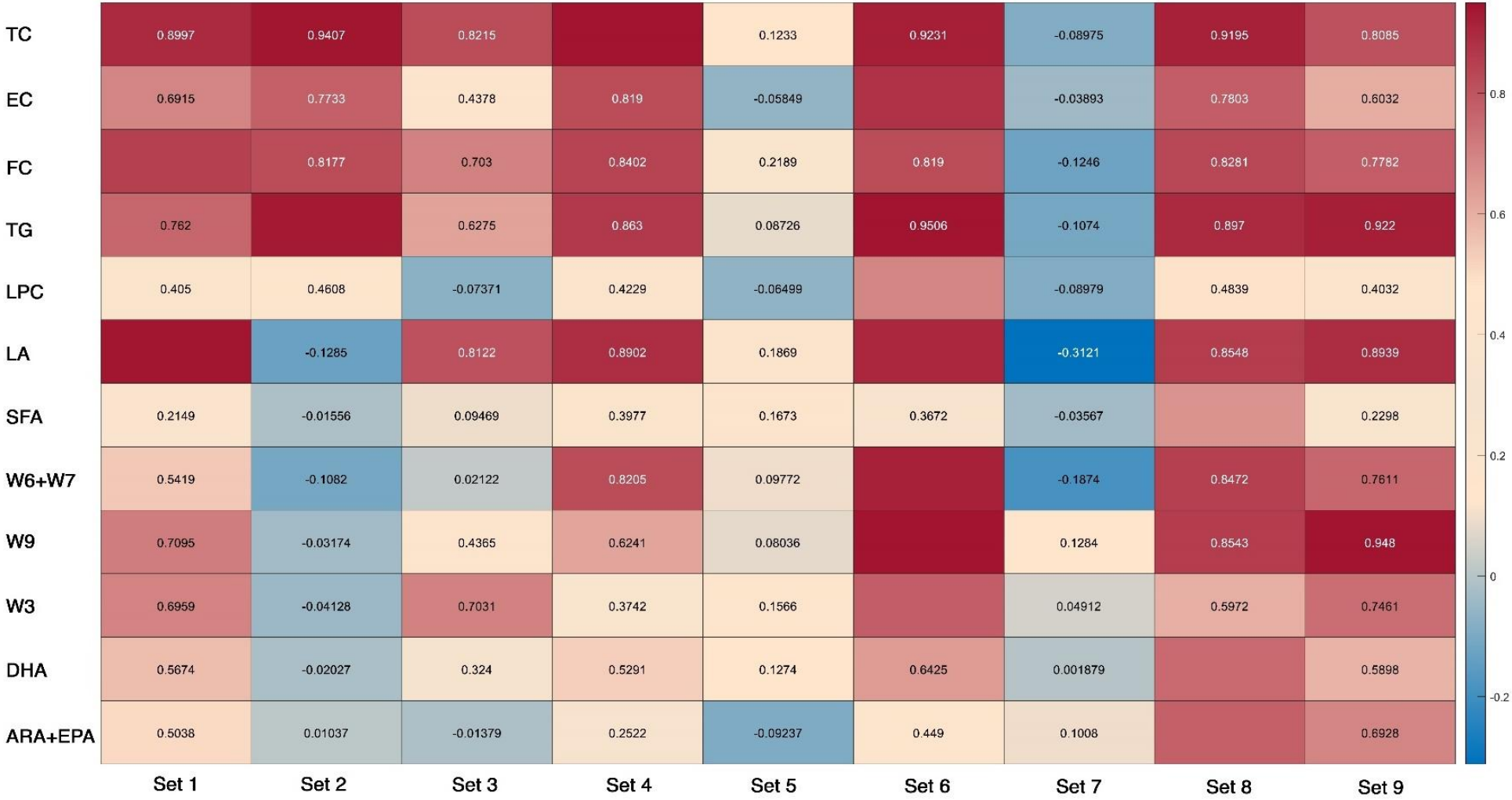


Figure 19. Pearson correlation coefficients heatmap of each model's employment against each set

Heatmap of the Rs acquired for each lipid model's validation against each set at disposal. Rows represent each lipidic family final predictive model and columns the sets at disposal. Dark reds indicate high positive correlations or values close to 1, dark blues indicate high negative correlations or values close to -1, and the neutral regions or the ones close to 0 are represented by colour beige.

4.6 Validation of the PLS models against the model-building unemployed sets

Last, the acquired models were tested performing external validations with each of the sets at disposal, now including set 9, and the heatmap from Figure 19 was acquired.

Same correlation patterns as the ones found in Table 4 and Annex 1 can be appreciated, mostly through the observation of set 6 and set 8. Such patterns consist of high positive correlations on the table's upper part and regarding sets 1, 2, 3, 4, 6, and 8, neutral correlations among the middle and lower left section, and negative correlations on set 2's bottom half, as well as set 7's upper half.

Sets 5 and 7's present discouraging validations, since for no lipid a coefficient above 0.22 is acquired, in the same situation is set's 2 bottom half. Generally, the rest of validations provide an acceptable range of Rs, ranging from values close to 0.6 up to values close to 0.95.

Regarding the validations against set 9, remarkable results are noticed, it is the set with the best overall lipid concentrations predictions, together with set 8. Its Rs range from 0.22 to 0.94, showcasing only two low results for LPC and SFA, three almost acceptable results for EC, DHA, and ARA+EPA, acceptable results for FC, W6+W7, and W3, and finally proficient results for TC, TG, LA, and W9 lipidic families.

5 Discussion

In light of all the acquired results, the following text provides a discussion on the implications and insights derived from this project's findings.

Concerning the PCA initial study, homogeneity among the different LED spectra was to be expected, since NMR is widely acknowledged to be a proficient and reproducible technique, in the most undesirable case, important differences among the LED spectra would have indicated that the acquisition process experienced some issues, and that the LED spectra at disposal is not reliable, which is definitely not the case. Through the principal components explained variance's low percentages, evidence about the uniformity of the LED data is provided, meaning that few differential patterns can be found among them, and that comparability is assured.

This point is crucial, since it enables the possibility of merging different sets together, aiming at obtaining a dataset with a really wide range of values that can be used to train predictive models with a representative amount of scenarios, which theoretically would result into the acquisition of a very robust model that can perform proficient predictions. Considering now the obtained results and being aware that merging of datasets was proposed in their process of acquisition, only one model out of the 12 collected was developed through a merged dataset, and even more, such dataset only consisted of two coupled sets.

Through the project's course, the inefficiency of merging datasets was detected, and the reason for such outcome still remains uncertain. Future research on this topic is necessary for as to assess whether the failure of merging sets is due to intrinsic characteristics of this project's sets, the model development procedure employed, or the fact that the supposed benefits of merging sets are a misconception, nevertheless, further into this section some explanatory ideas my arise, and comments on this topic will be provided. It could be argued, though, that through normalization, this activity may improve and provide the initially expected results. Regarding such idea and even though the normalization's study for this project presents some flaws, it still provides enough evidence to defend that it would not have had an important impact in improving the performance of merging sets for building models. Later, this topic will be addressed.

Moving to the lipidic quantifications PCA, this analysis provides more interesting information about the studied sets than that of the LED spectra, since homogeneity cannot be appreciated. In this case, the highest variance-explaining principal components provide a 50% and 32% of capturing. Even though these are not dramatically higher values than those of the LED analysis, clear differential patterns can be observed in the plot. Through appreciating Figure 4 at the beginning of the project, it can be noticed that for some lipids, quantifications are inconsistent, expressing different ranges between them. Among all the sets at disposal, integrands 3, 5 and 7 stand out for being the ones that have the most unsimilar concentrations ranges against the remaining sets for most of the lipidic families, this is very clear in the EC, LPC, and ARA+EPA quantifications plot in Figure 4.

Unsurprisingly, the outcome of this last PCA confirms major differences among the three sets mentioned together with set 2, and the rest of sets. Set 2 does not significantly stand out in the quantifications representations of Figure 4, but when examining Annex 1, the bottom half of this set's heatmaps provides uncertainty regarding its stability and coherence, which stands to reason with its incorporation to the unsimilar behaving sets group.

Now, examining the similar behaving groups, set 1 exhibits some dispersion along both principal components axis, which after observing Annex 1, Table 4, and Figure 19, can be attributed as the source of its worse correlating heatmaps and validations, such as the ones for EC, LPC, SFA, W3 and ARA+EPA, which are lower than the others. The other integrands of the similar behaving sets group, in comparison to set 2, are greatly overlapped in the centre region of the PCA space and provide high correlations in the STOCSY heatmaps. Within the similar behaving sets group, set 6 and set 8, have been used for the development of the majority of models. With this, it can be appreciated that a set's PCA observations' cohesion and coherence can be directly associated to that set providing high correlations in its STOCSY heatmaps, or in other words, having tight relationships between its LED spectra and lipid quantifications, which translates to it being a proficient candidate for predictive model building.

With all of this, a conclusion regarding the nature of the sets at disposal can be reached. Sets 1, 4, 6 and 8 are sets with strong relations between its LED spectra and lipid concentrations, whereas sets 2, 3, 5, and 7 have poor relations between, with each scenario showcasing the proficiency of the sets for model developing. Thus, regarding the just mentioned conclusions, the final conceiving of the models would be expected to be through one of the similarly behaving sets, and in fact, it wasn't. This indicates that, even though a general behaviour about each set's correlation between predictor and response variables can be detected, that does not mean that in the unsimilar behaving group's sets, poor connections exist for all the lipidic families, otherwise, the building of models through sets from such group would not have been possible, and indeed, for some of that group's integrands, stronger relationships between the LED spectra and some lipid quantifications, were found, and were then used for developing that lipid's predictive model, such as the use of set 2 for the TG model.

Up to this point, a question arises, why do these sets differ among each other in terms of the LED spectra connection to the lipid quantifications, when they have all been obtained through the same procedure? The answer to this question is still unknown, and further research is required to reveal the reason. For now, this project's development suggests that a priori, two of the methodologies employed for the lipid quantifications acquisition may be the cause, consisting of the lipidic extraction process, and the deconvolution of signals through the line-shape fitting technique. LED spectra is determined to have no influence on this issue owing to the already mentioned uniformity found in the PCA.

Moving forward to the normalization study and as previously indicated, normalization's benefits assessment concluded in unsuccessful results, which may indicate that the procedure

employed presents flaws, or that there was a misconception regarding this technique's actual improvements on the subject.

For as to address normalization's application on improving the merge of sets, the following scenario is presented. As explained in this project, when building linear predictive models, relationships between the predictor and response variables must be established, which enable the creation of a linear pattern that can be used for predictions against future predictor variables. Mentioned earlier, the utilization of a wide-ranged training set would correspond to the acquisition of a proficient prediction model, but when this range of training data is not so much a wide aggregation of different values but rather a big amount of samples that don't differ significantly from each other, and even more, the response variables do have a differentiated spectrum of values, then incoherence is found between the data, and outcoming prediction models perform poorly. In other words, for two slightly different predictor variables, two greatly differentiated corresponding response variables can be found, and this is thought to be the case of this project's data when merged, with LED spectra variations being very low, while lipid quantifications differing significantly, then, the situation where two similar predictor values have very different corresponding response variables is present, relationships patterns cannot be found, and incoherence in predictive model development occurs.

With the just introduced scenario, normalization is naturally brought to mind as the ideal solution, since standardization among these differing response variables would be achieved, resolving the whole problem. The methods employed in this project intended to perform as explained, and indeed, normalization of the response variables when building models was executed, but with disappointing results. The three actual normalization scenarios employed provided worse model acquisitions than those of the non-normalized one, with the response variables normalization scenario not solving the aforementioned problem, and nor did the predictor and response variables normalization one, as could be initially expected.

Thus, it can be concluded that other techniques for rearranging and normalizing the data to set a better starting point to develop predictive models than the one achieved in this project, must be explored, and may provide the results expected in this project's normalization application. Also, research on the normalization procedure employed in this project's application on other predictive modelling techniques such as random forests, support vector machines, neural networks, among others, could be performed and studied to assess its viability and functionality.

With this provided overview of the normalization study's outcome, it is no surprise that for the set selection phase, only one case of merged-datasets training set is present, all the other lipid family's models were chosen to be created through the use of a single set, as indicated by the results presented in section 4.3. Over such results, the different characteristics mentioned above and detected in each set through the analyses captured in Figure 4, Table 4, Figure 11, Annex 1, can be spotted, as for instance, set 5 and 7 providing extremely low validations when predictive models were used against them for almost all lipids, appreciated in any lipid, and set 6 together with set 8 performing competently among the major part of the lipids studied, both situations can be appreciated through any lipid's Figure 8. These graphs also provide illustrative information of how each lipid's candidate sets averagely perform for building the models, and Figure 8-C informs about that performance's stability against each set. General tendencies are found among all these plots, and can be appreciated in Annex 2.

With section 4.3's results, it can be affirmed that training set selection plays a crucial role in the process of developing predictive models, being the first factor to determine whether or not proficient results will be obtained, and the procedure performed in this project for such topic's assessment concluded in outstanding results, acquiring valuable information of the data studied through deep and robust representative plots of the sets modelling behaviour, that lead to a clear and decisive conclusion regarding which set to select.

Stepping forward to the spectral regions and latent variable selection, these two elements follow in the list of decisive data regarding predictive model building, and procedures employed in this project allow for a clear decision in the selection of both. Throughout all the graphs obtained, Annex 3, the influence of spectral regions selection can be appreciated to be very relevant, having great influence on the outcoming R, and giving evidence that across the LED spectra, lipids are portrayed in different regions, depending on the own nature of the compound. Thus, through the obtained results, the spectral regions that provide the most relevant information of the lipids studied in terms of the development of predictive model, are thoroughly assessed. The proficiency of this project's spectral regions selection procedure lies on the fact that not all highly correlating spectral points that one can detect through a STOCSY plot have to be selected in order to achieve the best predictive model, in fact, it has been proved that out of the baseline spectral regions selection, where the majority of properly correlating points are included, only a percentage remain for the final building of the model.

Latent variables assessment procedure relies on an established graphical representation acquired through the PLS toolbox employed, and following the same selection criteria, which adds to the validation of the success of the procedure. The 10 permutations performed provided very concordant results on amount of components to include, and did so for all lipidic families.

This project's proposed procedures for assessing the best selection of variables in terms of the factors that introduce variance to the development of predictive models, provided proficient results. The models obtained can be described to be the most refined version possible of themselves, and that is expressed by their R when validated against their remaining 30%.

Assessing the model's performance by validating them with the model-building unemployed sets is unfortunately concluded to be hazardous, owing to the detected differences among sets regarding the LED spectra and lipid quantifications connection, which is the main influential factor in the acquisition of a high R when validation is executed. This is logically deduced at this point, and sets 5 and 7 are clear examples of such behaviour.

Therefore, concluding validations out of the ones acquired are regarded to be those performed against the remaining 30%, set 1, set 4, set 6, set, 8 and set 9. Considering these, all models performance can be regarded to be proficient. Regarding set 9, its inclusion in the different analysis of the project, represented in Figure 4, Figure 11, and Annex 1, among others, enabled the assessment of its characteristics, which showcase great relationships between the LED spectra and all lipid quantifications, making it a great external validation set. In another case, this set could have been employed for the actual building of the models, since its outstanding lipid correlations would provide concluding results, but owing to the initial settling of this project, its contribution was determined to be that of an external validation set.

Now only considering the validation against the remaining 30%, the following reflection can be made regarding the development of predictive models in this project:

After selecting the best parameters possible for the building of a model through the project's employed procedures, if the outcoming model's validation with the remaining 30% is competent, such validation's weight on the assessment of the model's proficiency may be the most crucial one, since for the lipid of matter's model building, the set that had its best portrayal was used, with the remaining sets' portrayal of such lipid falling behind, owing to the fact that their quantifications are not as related to their LED spectrums, therefore, the values predicted by such model may provide more accurate results than those of the experimental quantifications, which explains possible low Rs when validation against the remaining sets is performed.

Thus, the practical applicability of the acquired lipid models can be partially assessed by externally validating them against sets 1, 4, 6, 8 and 9. However, the most crucial evaluation comes from the remaining 30% validation, where models with Rs above 0.8 can be considered competent performers, while models with Rs above 0.9 can be regarded as proficient performers.

6 Concluding remarks

This project aimed at the optimization of the NMR lipid profiling process by eliminating the need for serum lipid extraction, and successfully did so through the development of linear regression models using partial least squares regression that can be directly used against ^{1}H-NMR LED spectra of native serum.

In light of all of the achievements and findings, twelve lipidic models, whose validations against their remaining 30% are over 0.8 for four of them, consisting of LPC, SFA, W3 and ARA+EPA, and over 0.9 regarding the others, have been acquired, and their competency of prediction assessed, concluding with the following statement; the twelve lipidic models present competent prediction performances and can be used in real-case scenarios, assuring accurate outcomes.

The development of the predictive models involved tasks such as PCA exploratory analysis, normalization of sample sets, variable selection, and the implementation of a sophisticated modelling procedure. On one hand, LED spectra PCA analysis revealed comparability among them and discarded the possibility of batch effect, on the other hand and regarding the lipid PCA, differences within the lipid quantifications have been showcased, and since no batch effect can be associated to them, further research on this topic needs to be done to assess the cause.

Normalization's study was unsuccessful, it did not manage to correct the differences among the lipidic quantifications and provide benefits from sets merging, again, more research in this topic is needed to accurately determine its functionality.

In addition to the main objective, the project also aimed to automate the entire linear regression predictive modelling process using Matlab and the Partial Least Squares Toolbox. This automation led to the creation of an interactive software that guides users to the best possible linear regression predictive model, and owing to this software the acquisition of the sought-after best prediction models for the lipidic families studied was possible. This tool opens up possibilities for further research and practical applications in the predictive modelling field.

To conclude, outcoming of this project, the quantification of twelve lipidic families directly over native serum's ^{1}H-NMR LED spectra is now possible, which is a major breakthrough in the assessment of cardiovascular diseases, a powerful tool together with a refined procedure for developing predictive models have been settled, and new directions have been established for further research.

7 References

[1] World Health Organization (WHO), "Cardiovascular diseases (CVDs)," *Health topics*, Jun. 11, 2021. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed May 23, 2023).

[2] N. Townsend, L. Wilson, P. Bhatnagar, K. Wickramasinghe, M. Rayner, and M. Nichols, "Cardiovascular disease in Europe: epidemiological update 2016," *Eur Heart J*, vol. 37, no. 42, pp. 3232–3245, Nov. 2016, doi: 10.1093/eurheartj/ehw334.

[3] statistics explained Eurostat, "Cardiovascular diseases statistics," 2022. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Cardiovascular_diseases_statistics#Deaths_from_cardiovascular_diseases (accessed Jun. 01, 2023).

[4] D. M. and R. A. Z. R Small, "Lipids, structure and biochemistry of," *Encyclopedia of Human Biology*, vol. 4. Elsevier, pp. 442–462, 1991.

[5] Kostner Gerhard and Frank Sasa, *Lipoproteins Role in Health and Diseases*, 1st ed., vol. 1. Rijeka, Croatia: InTech, 2012. Accessed: May 24, 2023. [Online]. Available: https://www.google.es/books/edition/Lipoproteins/B5uUDwAAQBAJ?hl=es&gbpv=0&kptab=overview

[6] L. R. Engelking, "Lipoprotein Complexes," in *Textbook of Veterinary Physiological Chemistry*, Elsevier, 2015, pp. 406–410. doi: 10.1016/B978-0-12-391909-0.50063-3.

[7] J. D. Brunzell *et al.*, "Lipoprotein Management in Patients With Cardiometabolic Risk," *J Am Coll Cardiol*, vol. 51, no. 15, pp. 1512–1524, Apr. 2008, doi: 10.1016/j.jacc.2008.02.034.

[8] V. Aru *et al.*, "Quantification of lipoprotein profiles by nuclear magnetic resonance spectroscopy and multivariate data analysis," *TrAC Trends in Analytical Chemistry*, vol. 94, pp. 210–219, Sep. 2017, doi: 10.1016/j.trac.2017.07.009.

[9] M. Ding and K. M. Rexrode, "A Review of Lipidomics of Cardiovascular Disease Highlights the Importance of Isolating Lipoproteins," *Metabolites*, vol. 10, no. 4, p. 163, Apr. 2020, doi: 10.3390/metabo10040163.

[10] J. C. Cohen, E. Boerwinkle, T. H. Mosley, and H. H. Hobbs, "Sequence Variations in PCSK9, Low LDL, and Protection against Coronary Heart Disease," *New England Journal of Medicine*, vol. 354, no. 12, pp. 1264–1272, Mar. 2006, doi: 10.1056/NEJMoa054013.

[11] L. S. Athanasiou, D. I. Fotiadis, and L. K. Michalis, "Introduction," *Atherosclerotic Plaque Characterization Methods Based on Coronary Imaging*, pp. 1–21, 2017, doi: 10.1016/B978-0-12-804734-7.00001-4.

[12] E. and T. of H. B. C. in A. (Adult T. P. I. National Cholesterol Education Program (NCEP) Expert Panel on Detection, "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.," *Circulation*, vol. 106, no. 25, pp. 3143–421, Dec. 2002.

[13] W. T. Friedewald, R. I. Levy, and D. S. Fredrickson, "Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge.," *Clin Chem*, vol. 18, no. 6, pp. 499–502, Jun. 1972.

[14] O. Quehenberger *et al.*, "Lipidomics reveals a remarkable diversity of lipids in human plasma," *J Lipid Res*, vol. 51, no. 11, pp. 3299–3305, Nov. 2010, doi: 10.1194/jlr.M009449.

[15] M. Egea-Cortines and J. H. Doonan, "Editorial: Phenomics," *Front Plant Sci*, vol. 9, May 2018, doi: 10.3389/fpls.2018.00678.

[16] A. Tebani, L. Abily-Donval, C. Afonso, S. Marret, and S. Bekri, "Clinical Metabolomics: The New Metabolic Window for Inborn Errors of Metabolism Investigations in the Post-Genomic Era," *Int J Mol Sci*, vol. 17, no. 7, p. 1167, Jul. 2016, doi: 10.3390/ijms17071167.

[17] J. K. Nicholson, J. C. Lindon, and E. Holmes, "'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data," *Xenobiotica*, vol. 29, no. 11, pp. 1181–1189, Jan. 1999, doi: 10.1080/004982599238047.

[18] M. M. Khamis, D. J. Adamko, and A. El-Aneed, "Mass spectrometric based approaches in urine metabolomics and biomarker discovery," *Mass Spectrom Rev*, vol. 36, no. 2, pp. 115–134, Mar. 2017, doi: 10.1002/mas.21455.

[19] J. R. Everett, "Pharmacometabonomics in humans: a new tool for personalized medicine," *Pharmacogenomics*, vol. 16, no. 7, pp. 737–754, May 2015, doi: 10.2217/pgs.15.20.

[20] R. Mallol *et al.*, "Liposcale: a novel advanced lipoprotein test based on 2D diffusion-ordered 1H NMR spectroscopy," *J Lipid Res*, vol. 56, no. 3, pp. 737–746, Mar. 2015, doi: 10.1194/jlr.D050120.

[21] S. Mora, J. D. Otvos, N. Rifai, R. S. Rosenson, J. E. Buring, and P. M. Ridker, "Lipoprotein Particle Profiles by Nuclear Magnetic Resonance Compared With Standard Lipids and Apolipoproteins in Predicting Incident Cardiovascular Disease in Women," *Circulation*, vol. 119, no. 7, pp. 931–939, Feb. 2009, doi: 10.1161/CIRCULATIONAHA.108.816181.

[22] P. Wiesner, K. Leidl, A. Boettcher, G. Schmitz, and G. Liebisch, "Lipid profiling of FPLC-separated lipoprotein fractions by electrospray ionization tandem mass spectrometry," *J Lipid Res*, vol. 50, no. 3, pp. 574–585, Mar. 2009, doi: 10.1194/jlr.D800028-JLR200.

[23] K. Yang and X. Han, "Lipidomics: Techniques, Applications, and Outcomes Related to Biomedical Sciences," *Trends Biochem Sci*, vol. 41, no. 11, pp. 954–969, Nov. 2016, doi: 10.1016/j.tibs.2016.08.010.

[24] M. Ala-Korpela, "1H NMR spectroscopy of human blood plasma," *Prog Nucl Magn Reson Spectrosc*, vol. 27, no. 5–6, pp. 475–554, Nov. 1995, doi: 10.1016/0079-6565(95)01013-0.

[25] J. K. Nicholson, P. J. D. Foxall, Manfred. Spraul, R. Duncan. Farrant, and J. C. Lindon, "750 MHz 1H and 1H-13C NMR Spectroscopy of Human Blood Plasma," *Anal Chem*, vol. 67, no. 5, pp. 793–811, Mar. 1995, doi: 10.1021/ac00101a004.

[26] M. Balci, "Pulse NMR Spectroscopy," in *Basic 1H- and 13C-NMR Spectroscopy*, Elsevier, 2005, pp. 253–281. doi: 10.1016/B978-044451811-8.50011-5.

[27] R. Barrilero Regadera, "Development of 1H-NMR Serum Profiling Methods for High-Throughput Metabolomics," Universitat Rovira i Virgili, Tarragona, 2017. Accessed: May 30, 2023. [Online]. Available: https://www.tdx.cat/handle/10803/461603#page=8

[28] D. S. Wishart *et al.*, "HMDB 3.0—The Human Metabolome Database in 2013," *Nucleic Acids Res*, vol. 41, no. D1, pp. D801–D807, Nov. 2012, doi: 10.1093/nar/gks1065.

[29] E. L. Ulrich *et al.*, "BioMagResBank," *Nucleic Acids Res*, vol. 36, no. Database, pp. D402–D408, Dec. 2007, doi: 10.1093/nar/gkm957.

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

[31] R. Mallol, M. A. Rodriguez, J. Brezmes, L. Masana, and X. Correig, "Human serum/plasma lipoprotein analysis by NMR: Application to the study of diabetic dyslipidemia," *Prog Nucl Magn Reson Spectrosc*, vol. 70, pp. 1–24, Apr. 2013, doi: 10.1016/j.pnmrs.2012.09.001.

[32] P. Soininen *et al.*, "High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism," *Analyst*, vol. 134, no. 9, p. 1781, 2009, doi: 10.1039/b910205a.

[33] M. Jupin, P. J. Michiels, F. C. Girard, M. Spraul, and S. S. Wijmenga, "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*, vol. 228, pp. 81–94, Mar. 2013, doi: 10.1016/j.jmr.2012.12.010.

[34] P. Soininen, A. J. Kangas, P. Würtz, T. Suna, and M. Ala-Korpela, "Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Cardiovascular Epidemiology and Genetics," *Circ Cardiovasc Genet*, vol. 8, no. 1, pp. 192–206, Feb. 2015, doi: 10.1161/CIRCGENETICS.114.000216.

[35] E. O. Stejskal and J. E. Tanner, "Spin Diffusion Measurements: Spin Echoes in the Presence of a Time-Dependent Field Gradient," *J Chem Phys*, vol. 42, no. 1, pp. 288–292, Jan. 1965, doi: 10.1063/1.1695690.

[36] M. Dyrby *et al.*, "Analysis of lipoproteins using 2D diffusion-edited NMR spectroscopy and multi-way chemometrics," *Anal Chim Acta*, vol. 531, no. 2, pp. 209–216, Feb. 2005, doi: 10.1016/j.aca.2004.10.052.

[37] R. Mallol *et al.*, "Surface fitting of 2D diffusion-edited 1H NMR spectroscopy data for the characterisation of human plasma lipoproteins," *Metabolomics*, vol. 7, no. 4, pp. 572–582, Dec. 2011, doi: 10.1007/s11306-011-0273-8.

[38] H. Tang, Y. Wang, J. K. Nicholson, and J. C. Lindon, "Use of relaxation-edited one-dimensional and two dimensional nuclear magnetic resonance spectroscopy to improve detection of small metabolites in blood plasma," *Anal Biochem*, vol. 325, no. 2, pp. 260–272, Feb. 2004, doi: 10.1016/j.ab.2003.10.033.

[39] C. E. Kostara and E. T. Bairaktari, "Lipid Profiling in Health and Disease," in *Methodologies for Metabolomics*, Cambridge University Press, 2013, pp. 317–332. doi: 10.1017/CBO9780511996634.020.

[40] T. Tukiainen *et al.*, "A multi-metabolite analysis of serum by 1H NMR spectroscopy: Early systemic signs of Alzheimer's disease," *Biochem Biophys Res Commun*, vol. 375, no. 3, pp. 356–361, Oct. 2008, doi: 10.1016/j.bbrc.2008.08.007.

[41] R. Barrilero *et al.*, "LipSpin: A New Bioinformatics Tool for Quantitative ^{1} H NMR Lipid Profiling," *Anal Chem*, vol. 90, no. 3, pp. 2031–2040, Feb. 2018, doi: 10.1021/acs.analchem.7b04148.

[42] C. Jiang, K. Yang, L. Yang, Z. Miao, Y. Wang, and H. Zhu, "A 1H NMR-Based Metabonomic Investigation of Time-Related Metabolic Trajectories of the Plasma, Urine and Liver Extracts of Hyperlipidemic Hamsters," *PLoS One*, vol. 8, no. 6, p. e66786, Jun. 2013, doi: 10.1371/journal.pone.0066786.

[43] M. Vinaixa *et al.*, "Metabolomic Assessment of the Effect of Dietary Cholesterol in the Progressive Development of Fatty Liver Disease," *J Proteome Res*, vol. 9, no. 5, pp. 2527–2538, May 2010, doi: 10.1021/pr901203w.

[44] C. E. Kostara, A. Papathanasiou, M. T. Cung, M. S. Elisaf, J. Goudevenos, and E. T. Bairaktari, "Evaluation of Established Coronary Heart Disease on the Basis of HDL and Non-HDL NMR Lipid Profiling," *J Proteome Res*, vol. 9, no. 2, pp. 897–911, Feb. 2010, doi: 10.1021/pr900783x.

[45] O. Beckonert, J. Monnerjahn, U. Bonk, and D. Leibfritz, "Visualizing metabolic changes in breast-cancer tissue using1H-NMR spectroscopy and self-organizing maps," *NMR Biomed*, vol. 16, no. 1, pp. 1–11, Feb. 2003, doi: 10.1002/nbm.797.

[46] H. Fernando, K. K. Bhopale, S. Kondraganti, B. S. Kaphalia, and G. A. Shakeel Ansari, "Lipidomic changes in rat liver after long-term exposure to ethanol," *Toxicol Appl Pharmacol*, vol. 255, no. 2, pp. 127–137, Sep. 2011, doi: 10.1016/j.taap.2011.05.022.

[47] L. Löfgren, M. Ståhlman, G.-B. Forsberg, S. Saarinen, R. Nilsson, and G. I. Hansson, "The BUME method: a novel automated chloroform-free 96-well total lipid extraction method for blood plasma," *J Lipid Res*, vol. 53, no. 8, pp. 1690–1700, Aug. 2012, doi: 10.1194/jlr.D023036.

[48] R. Barrilero *et al.*, "LipSpin: A New Bioinformatics Tool for Quantitative ^{1} H NMR Lipid Profiling," *Anal Chem*, vol. 90, no. 3, pp. 2031–2040, Feb. 2018, doi: 10.1021/acs.analchem.7b04148.

[49] R. H. Bruhl, *Understanding statistical analysis and modeling*. Los Angeles, CA: SAGE Publications, Inc, 2018.

[50] W. Karl. Härdle, *Applied Multivariate Statistical Analysis*, 5th ed. 2019. Cham: Springer International Publishing, 2019. doi: 10.1007/978-3-030-26006-4.

[51] I. T. Jolliffe, *Principal component analysis*, 2nd ed. in Springer series in statistics. New York: Springer, 2002.

[52] O. Cloarec *et al.*, "Statistical Total Correlation Spectroscopy: An Exploratory Approach for Latent Biomarker Identification from Metabolic ^{1} H NMR Data Sets," *Anal Chem*, vol. 77, no. 5, pp. 1282–1289, Mar. 2005, doi: 10.1021/ac048630x.

[53] V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, "Group Normalization," in *Computer Vision - ECCV 2018*, in Lecture Notes in Computer Science, vol. 11217. Switzerland: Springer International Publishing AG, 2018, pp. 3–19. doi: 10.1007/978-3-030-01261-8_1.

[54] M. Chiesa, G. I. Colombo, and L. Piacentini, "DaMiRseq-an R/Bioconductor package for data mining of RNA-Seq data: normalization, feature selection and classification," *Bioinformatics*, vol. 34, no. 8, pp. 1416–1418, 2018, doi: 10.1093/bioinformatics/btx795.

[55] Herve. Abdi, W. W. Chin, Vincenzo. EspositoVinzi, Giorgio. Russolillo, and Laura. Trinchera, *New Perspectives in Partial Least Squares and Related Methods*, 1st ed. 2013. in Springer Proceedings in Mathematics & Statistics, 56. New York, NY: Springer New York, 2013. doi: 10.1007/978-1-4614-8283-3.

[56] P. H. Garthwaite, "An Interpretation of Partial Least Squares," *J Am Stat Assoc*, vol. 89, no. 425, pp. 122–127, 1994, doi: 10.1080/01621459.1994.10476452.

[57] E. M. van Leeuwen *et al.*, "A new perspective on lipid research in age-related macular degeneration," *Prog Retin Eye Res*, vol. 67, pp. 56–86, Nov. 2018, doi: 10.1016/j.preteyeres.2018.04.006.

LIST OF FIGURES

Figure 1. Relationship between size and density among plasma lipoprotein classes [57] 6
Figure 2. Operational outline of a nuclear magnetic resonance experiment 11
Figure 3. Methyl and methylene regions in the three molecular window model and examples of molecular species that are analysed with each window [27] 13
Figure 4. Lipidic families' concentration distribution among sets 17
Figure 5. STOCSY heatmap 20
Figure 6. SuperScript's execution block diagram 28
Figure 7. Outline of the procedures followed for addressing the three questions posed 29
Figure 8. Set performance assessment plot 31
Figure 9. Spectral regions assessment plot 32
Figure 10. Latent variable number assessment through the RMSE 33
Figure 11. PCA of the LED spectra and the lipid quantifications 34
Figure 12. FC candidates' modelling behaviour representation 39
Figure 13. LPC candidates' modelling behaviour representation 40
Figure 14. W6+W7 candidates' modelling behaviour representation 40
Figure 15. Baseline spectral regions randomizations of FC's selected set and RMSE vs LV increase prototype plot 41
Figure 16. LPC's baseline spectral regions randomizations of the selected set and RMSE vs LV increase plot 43
Figure 17. W6+W7's baseline spectral regions randomizations of the selected set and RMSE vs LV increase plot 43
Figure 18. Regression plots of predicted vs experimental values employing each lipid's final predictive model 45
Figure 19. Pearson correlation coefficients heatmap of each model's employment against each set 46

LIST OF TABLES

Table 1. Lipid composition of VLDLs, LDLs, and HDLs [9] 7
Table 2. Properties of the sets at disposal, consisting of the number of samples (N), age, male participants, type 2 diabetes, obesity, and hypertension. The age range and average value are provided, and for the rest of characteristics, the frequency in number and percentage is given. 15
Table 3. Plot regression of each normalization scenario best model's validation, with the corresponding R 36
Table 4. Candidates selected for each lipidic family predictive model building 37
Table 5. Candidate sets combinations for developing each lipidic family predictive model 38
Table 6. Selected candidate for each lipid predictive model 41
Table 7. Answering the three questions, each lipid family model's selected inputs and permutations procedure's results 44

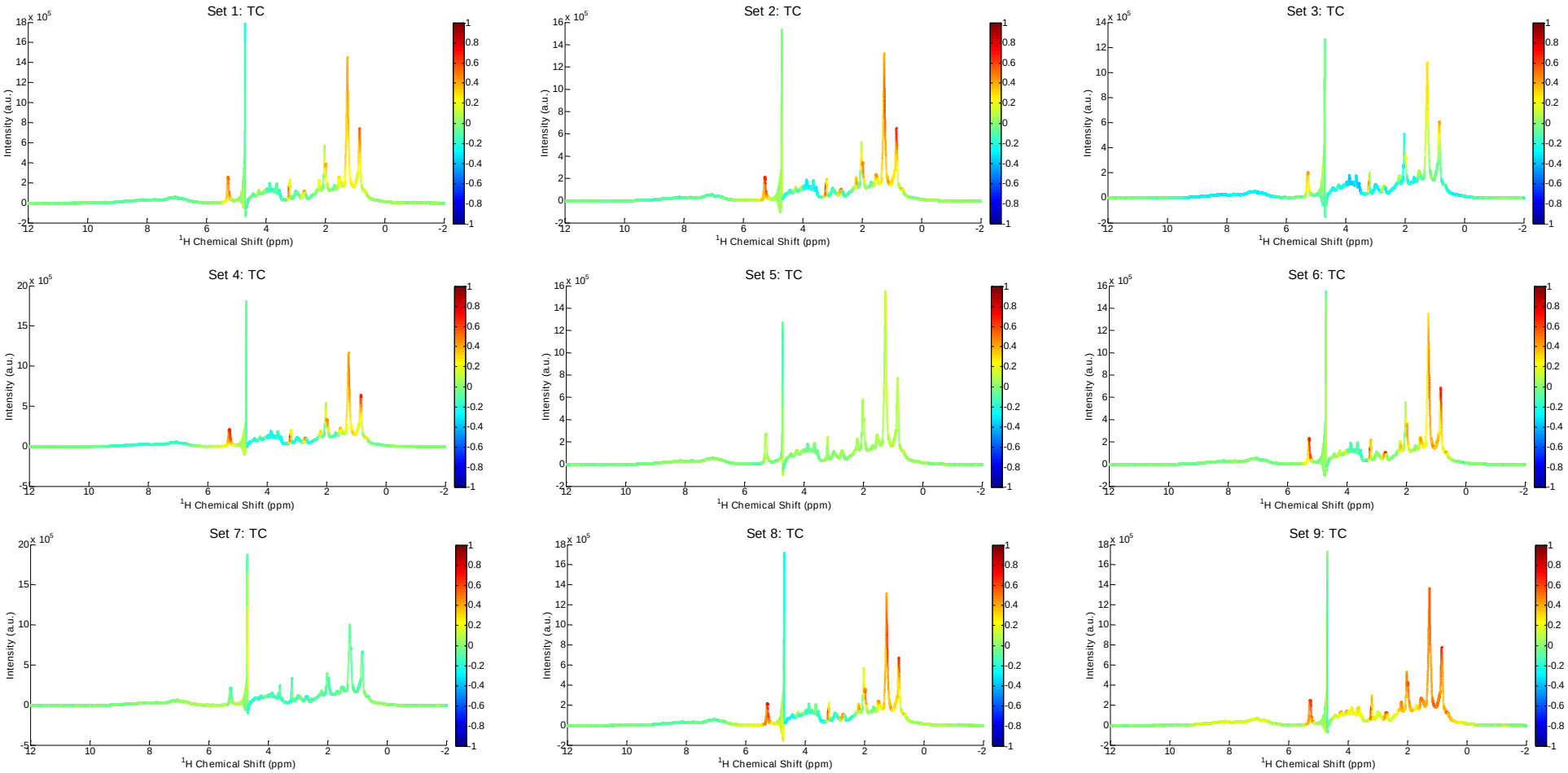
LIST OF EQUATIONS

Equation 1. Correlation coefficient between 19
Equation 2. Z-score 21
Equation 3. Covariance 22
Equation 4. Relative root mean square error expresses as a percentage 33

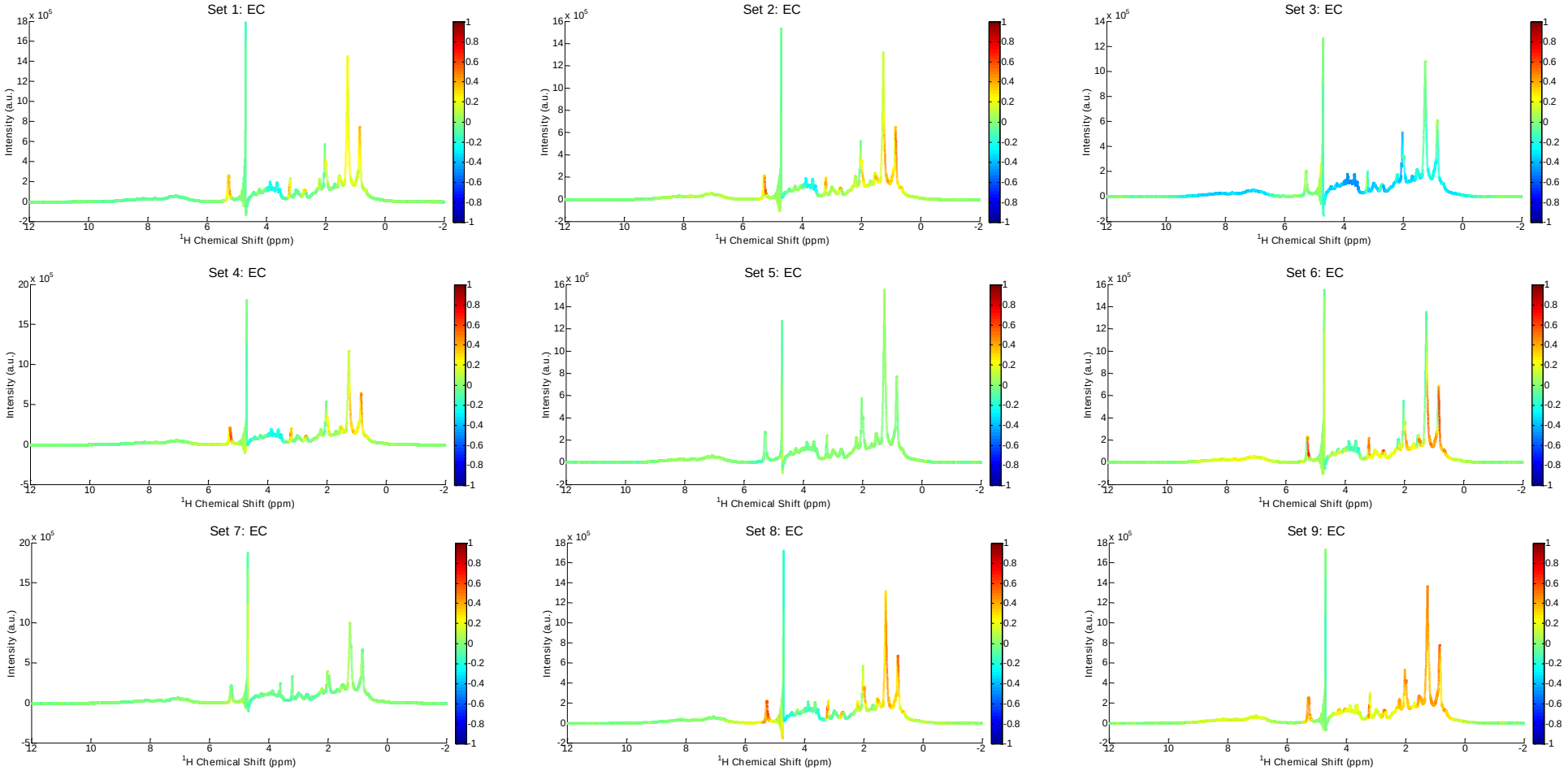
8 Annexes

Annex 1. STOCSY heatmaps for all the sets' 12 lipids

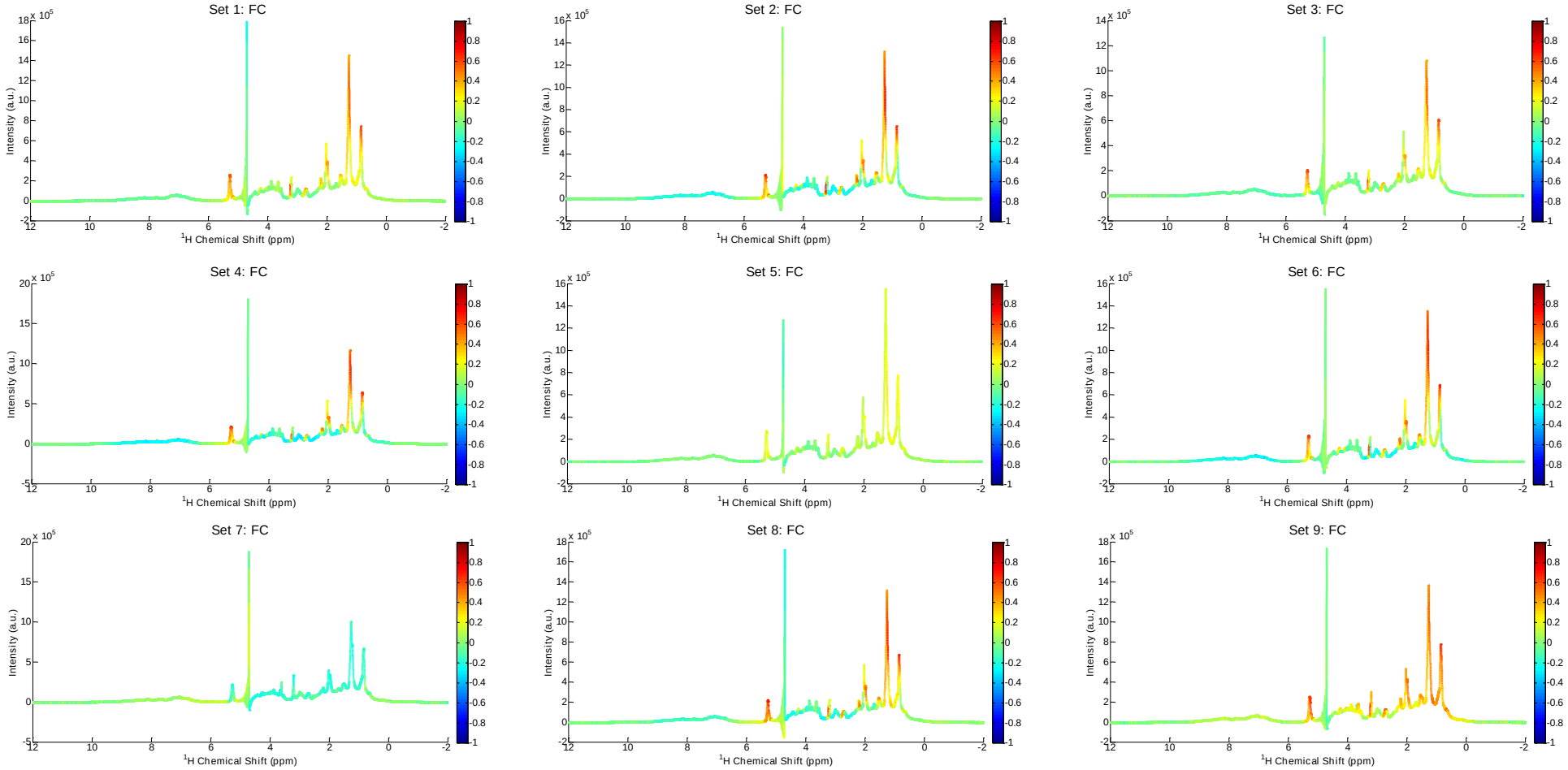
TC



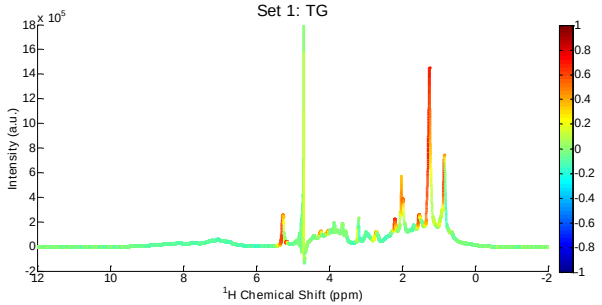
EC

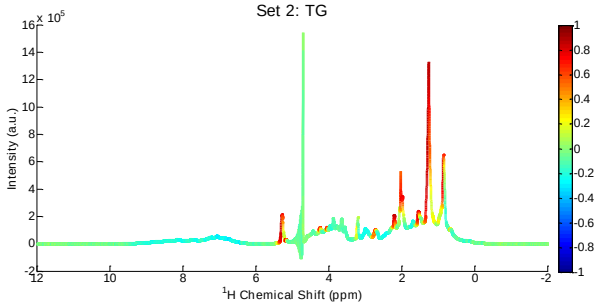

Set 1: EC
Set 2: EC
Set 3: EC
Set 4: EC
Set 5: EC
Set 6: EC
Set 7: EC
Set 8: EC
Set 9: EC
Intensity (a.u.)
x 10^5
1H Chemical Shift (ppm)

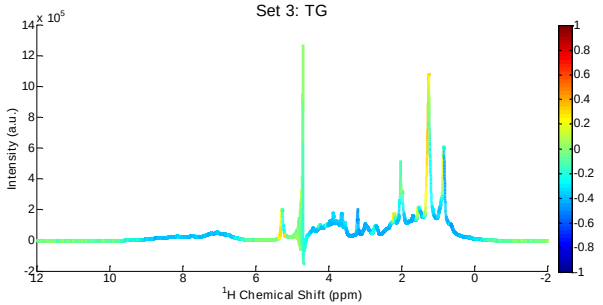
FC

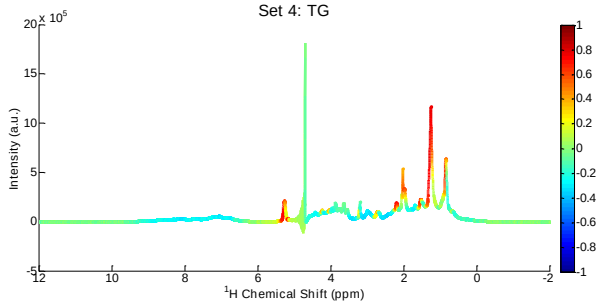

Set 1: FC
Set 2: FC
Set 3: FC
Set 4: FC
Set 5: FC
Set 6: FC
Set 7: FC
Set 8: FC
Set 9: FC
Intensity (a.u.)
x 10^5
1H Chemical Shift (ppm)

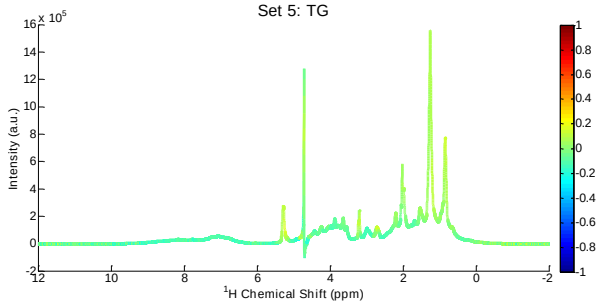
TG

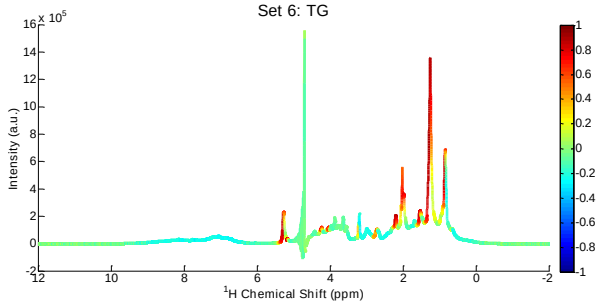

Set 1: TG
Intensity (a.u.)
x 10^5
^{1}H Chemical Shift (ppm)

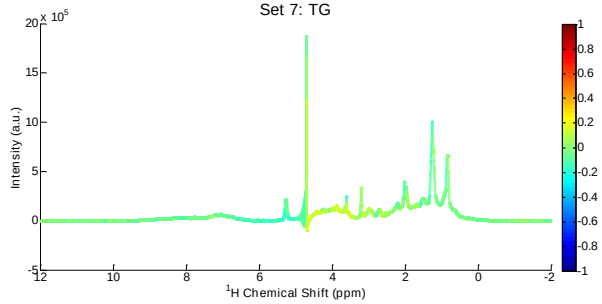

Set 2: TG
Intensity (a.u.)
x 10^5
^{1}H Chemical Shift (ppm)

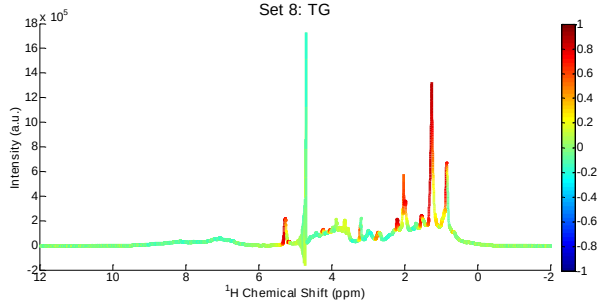

Set 3: TG
Intensity (a.u.)
x 10^5
^{1}H Chemical Shift (ppm)

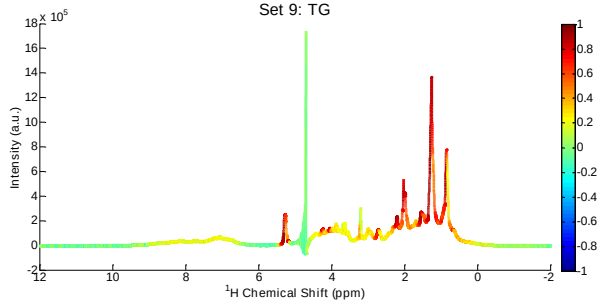

Set 4: TG
Intensity (a.u.)
x 10^5
^{1}H Chemical Shift (ppm)


Set 5: TG
Intensity (a.u.)
x 10^5
^{1}H Chemical Shift (ppm)

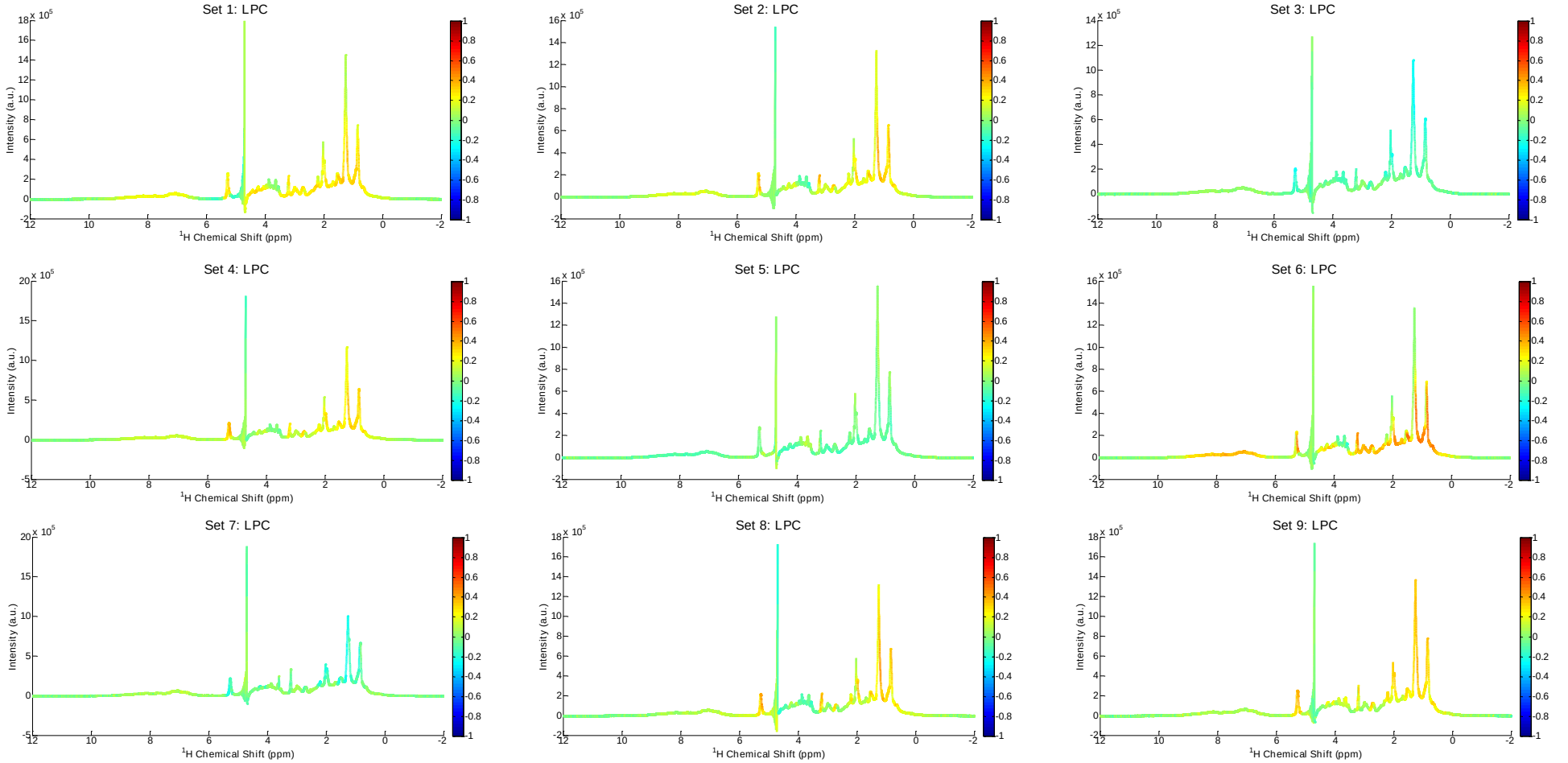

Set 6: TG
Intensity (a.u.)
x 10^5
^{1}H Chemical Shift (ppm)


Set 7: TG
Intensity (a.u.)
x 10^5
^{1}H Chemical Shift (ppm)

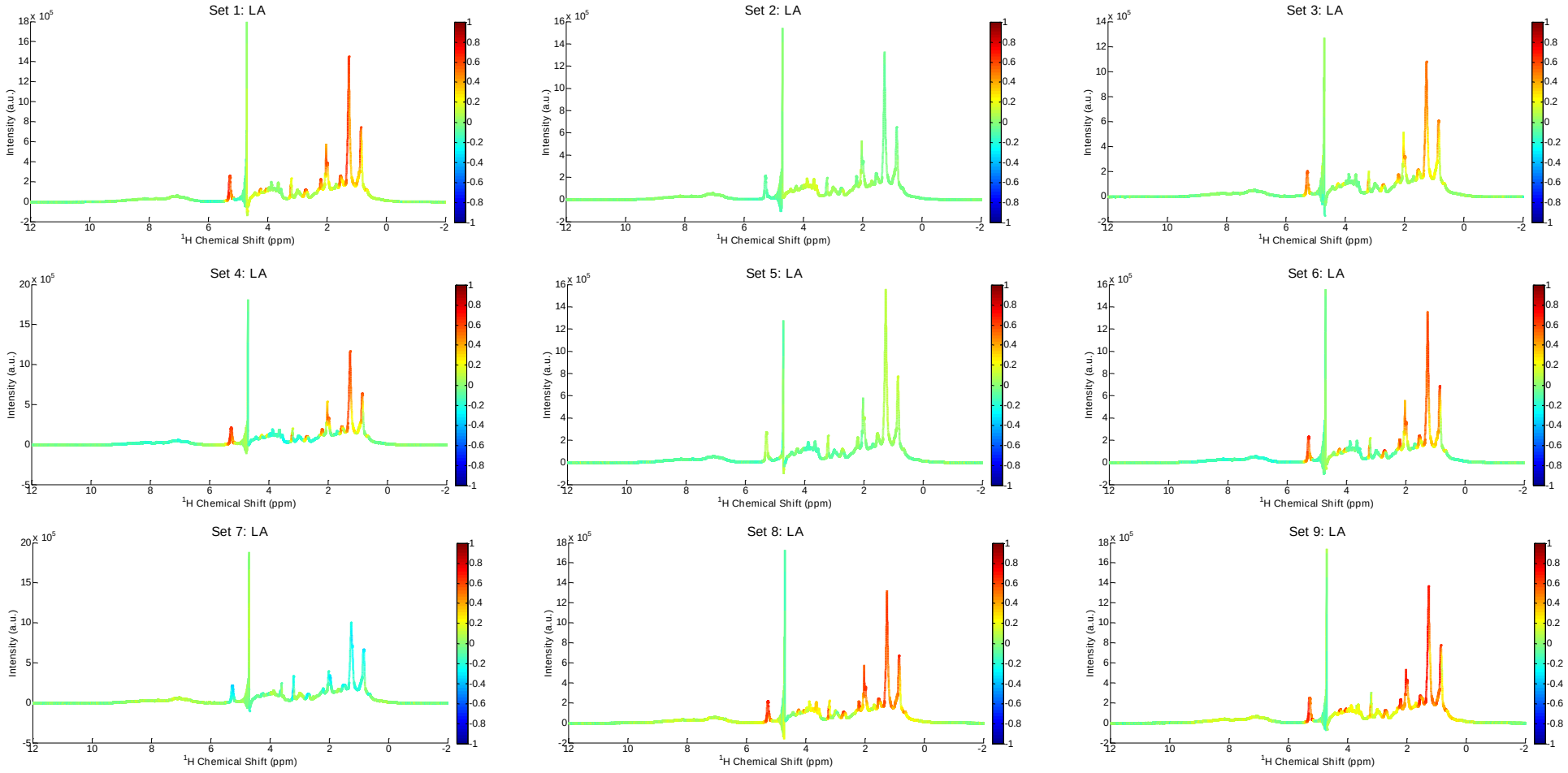

Set 8: TG
Intensity (a.u.)
x 10^5
^{1}H Chemical Shift (ppm)


Set 9: TG
Intensity (a.u.)
x 10^5
^{1}H Chemical Shift (ppm)

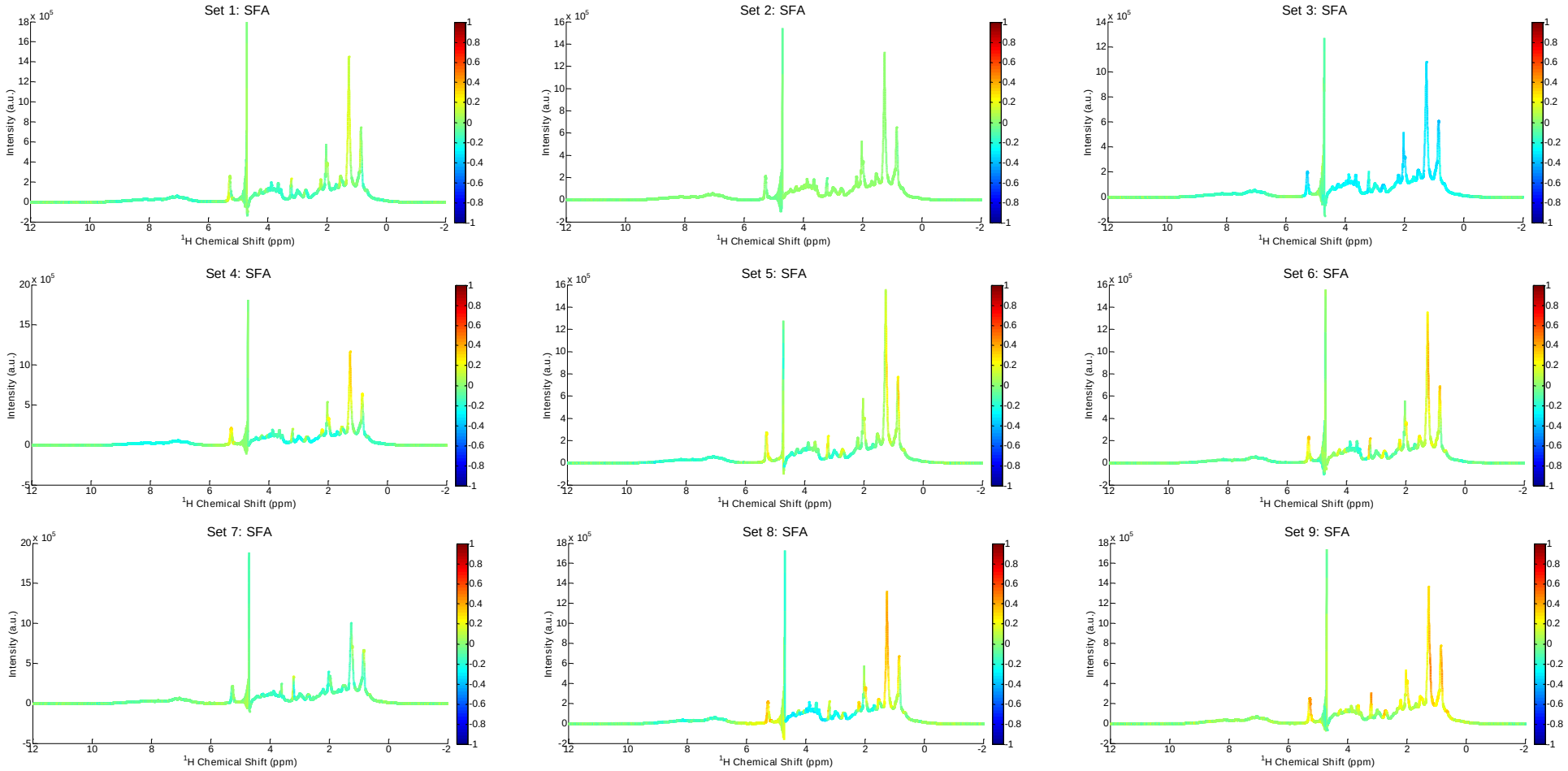
LPC


Set 1: LPC
Set 2: LPC
Set 3: LPC
Set 4: LPC
Set 5: LPC
Set 6: LPC
Set 7: LPC
Set 8: LPC
Set 9: LPC
Intensity (a.u.)
x 10^5
1H Chemical Shift (ppm)

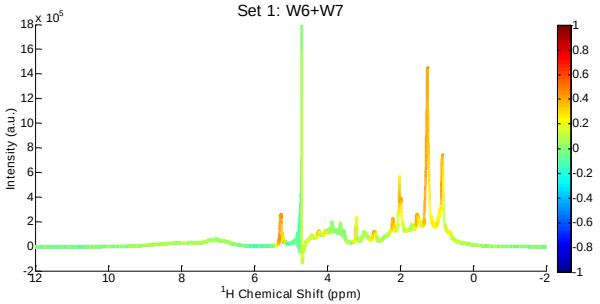
LA

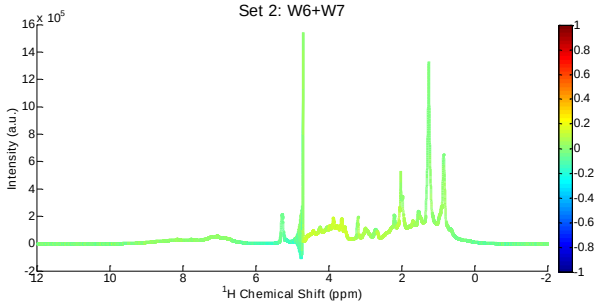

Set 1: LA
Set 2: LA
Set 3: LA
Set 4: LA
Set 5: LA
Set 6: LA
Set 7: LA
Set 8: LA
Set 9: LA
Intensity (a.u.)
x 10^5
1H Chemical Shift (ppm)

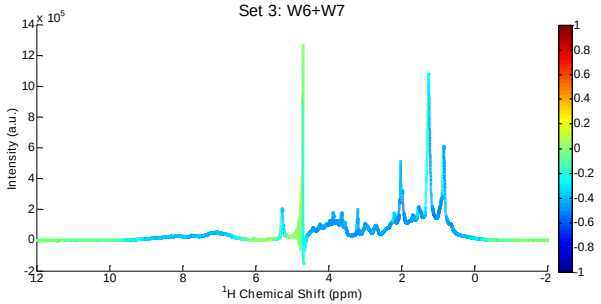
SFA

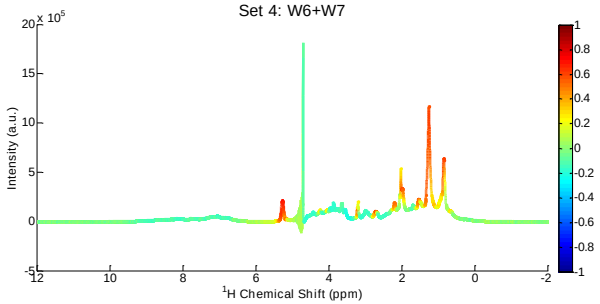

Set 1: SFA
Set 2: SFA
Set 3: SFA
Set 4: SFA
Set 5: SFA
Set 6: SFA
Set 7: SFA
Set 8: SFA
Set 9: SFA
Intensity (a.u.)
x 10^5
1H Chemical Shift (ppm)

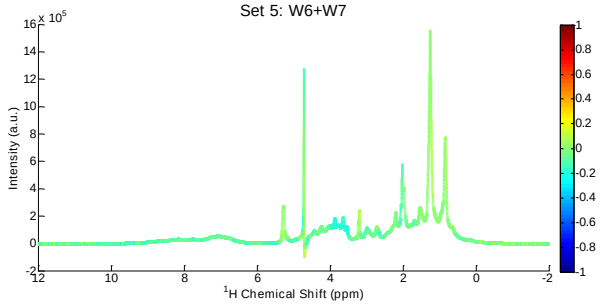
W6+W7

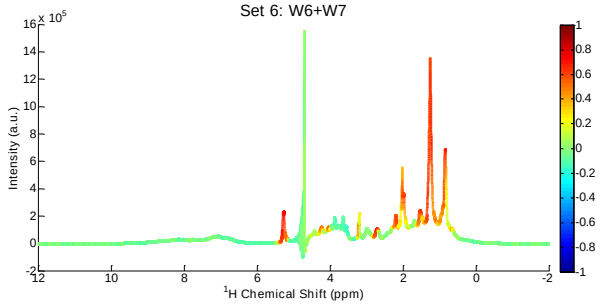

Set 1: W6+W7
Intensity (a.u.)
^{1}H Chemical Shift (ppm)

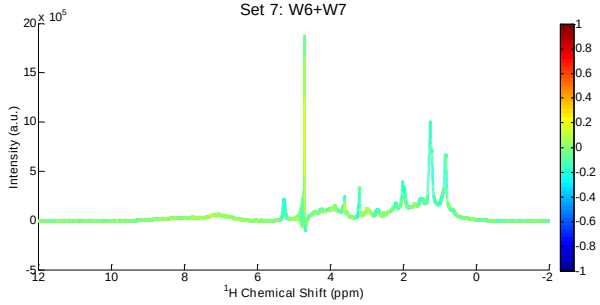

Set 2: W6+W7
Intensity (a.u.)
^{1}H Chemical Shift (ppm)

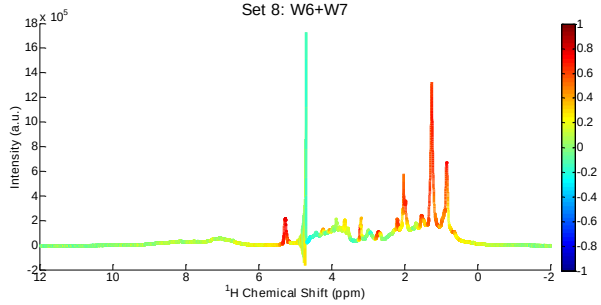

Set 3: W6+W7
Intensity (a.u.)
^{1}H Chemical Shift (ppm)

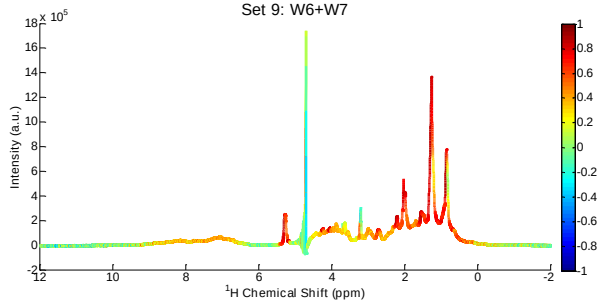

Set 4: W6+W7
Intensity (a.u.)
^{1}H Chemical Shift (ppm)


Set 5: W6+W7
Intensity (a.u.)
^{1}H Chemical Shift (ppm)

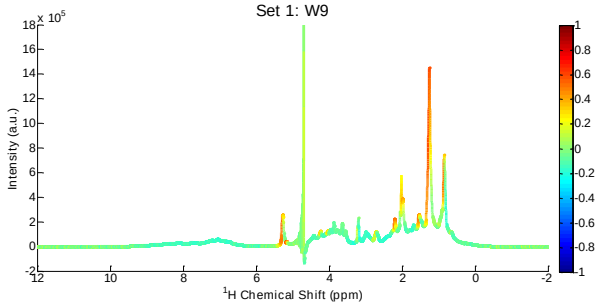

Set 6: W6+W7
Intensity (a.u.)
^{1}H Chemical Shift (ppm)

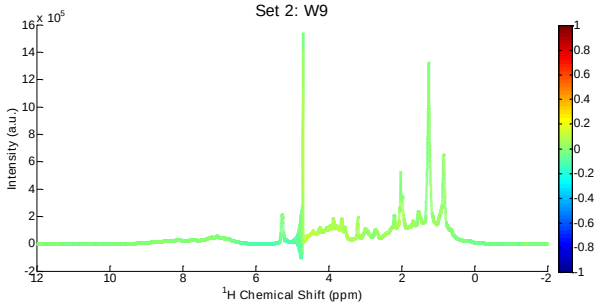

Set 7: W6+W7
Intensity (a.u.)
^{1}H Chemical Shift (ppm)

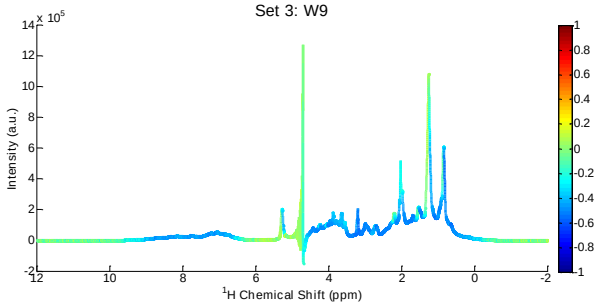

Set 8: W6+W7
Intensity (a.u.)
^{1}H Chemical Shift (ppm)

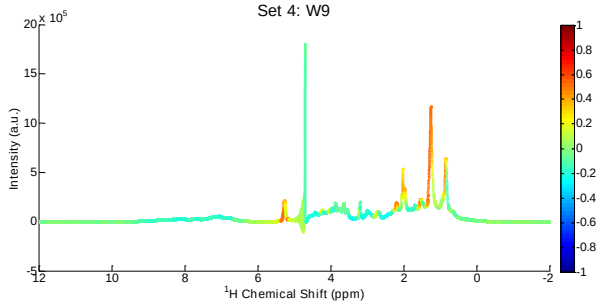

Set 9: W6+W7
Intensity (a.u.)
^{1}H Chemical Shift (ppm)

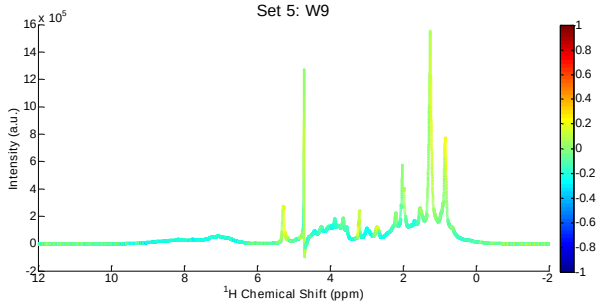
W9

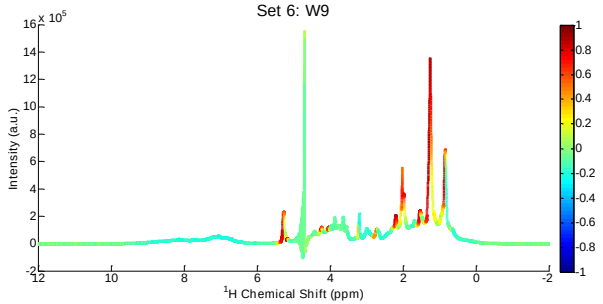

Set 1: W9
Intensity (a.u.)
x 10^5
^{1}H Chemical Shift (ppm)

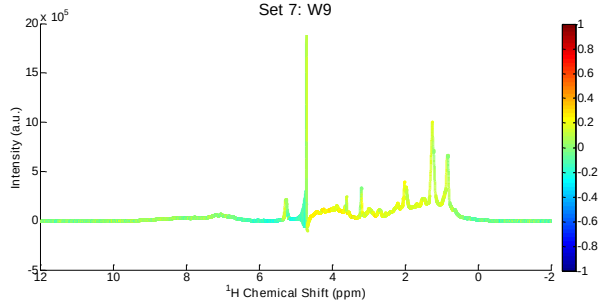

Set 2: W9
Intensity (a.u.)
x 10^5
^{1}H Chemical Shift (ppm)

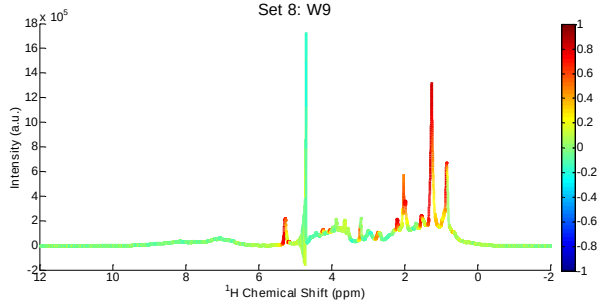

Set 3: W9
Intensity (a.u.)
x 10^5
^{1}H Chemical Shift (ppm)

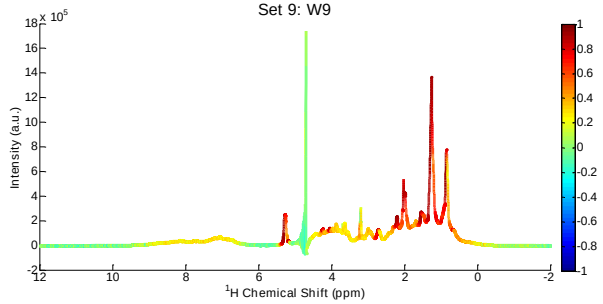

Set 4: W9
Intensity (a.u.)
x 10^5
^{1}H Chemical Shift (ppm)


Set 5: W9
Intensity (a.u.)
x 10^5
^{1}H Chemical Shift (ppm)

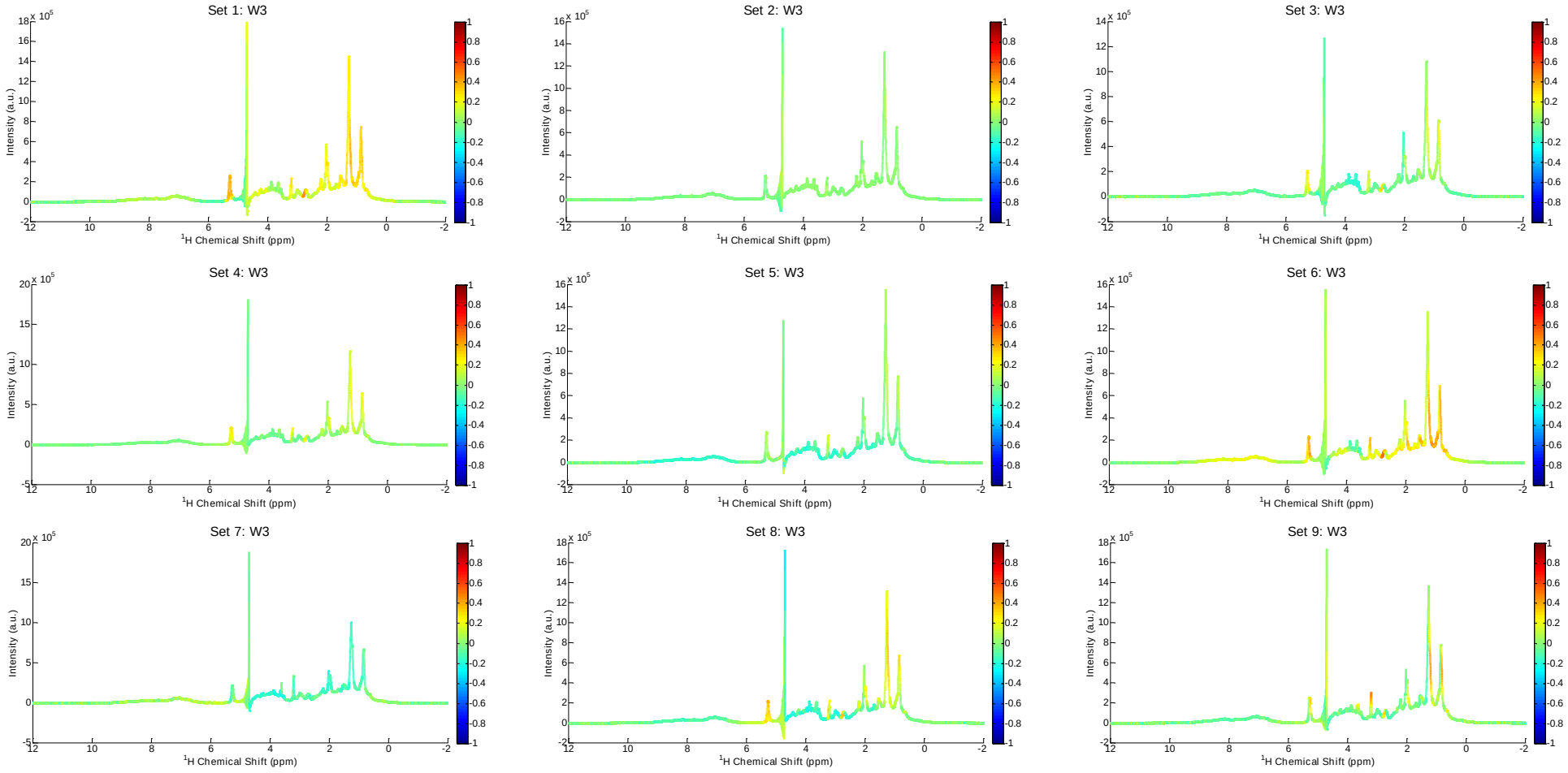

Set 6: W9
Intensity (a.u.)
x 10^5
^{1}H Chemical Shift (ppm)


Set 7: W9
Intensity (a.u.)
x 10^5
^{1}H Chemical Shift (ppm)

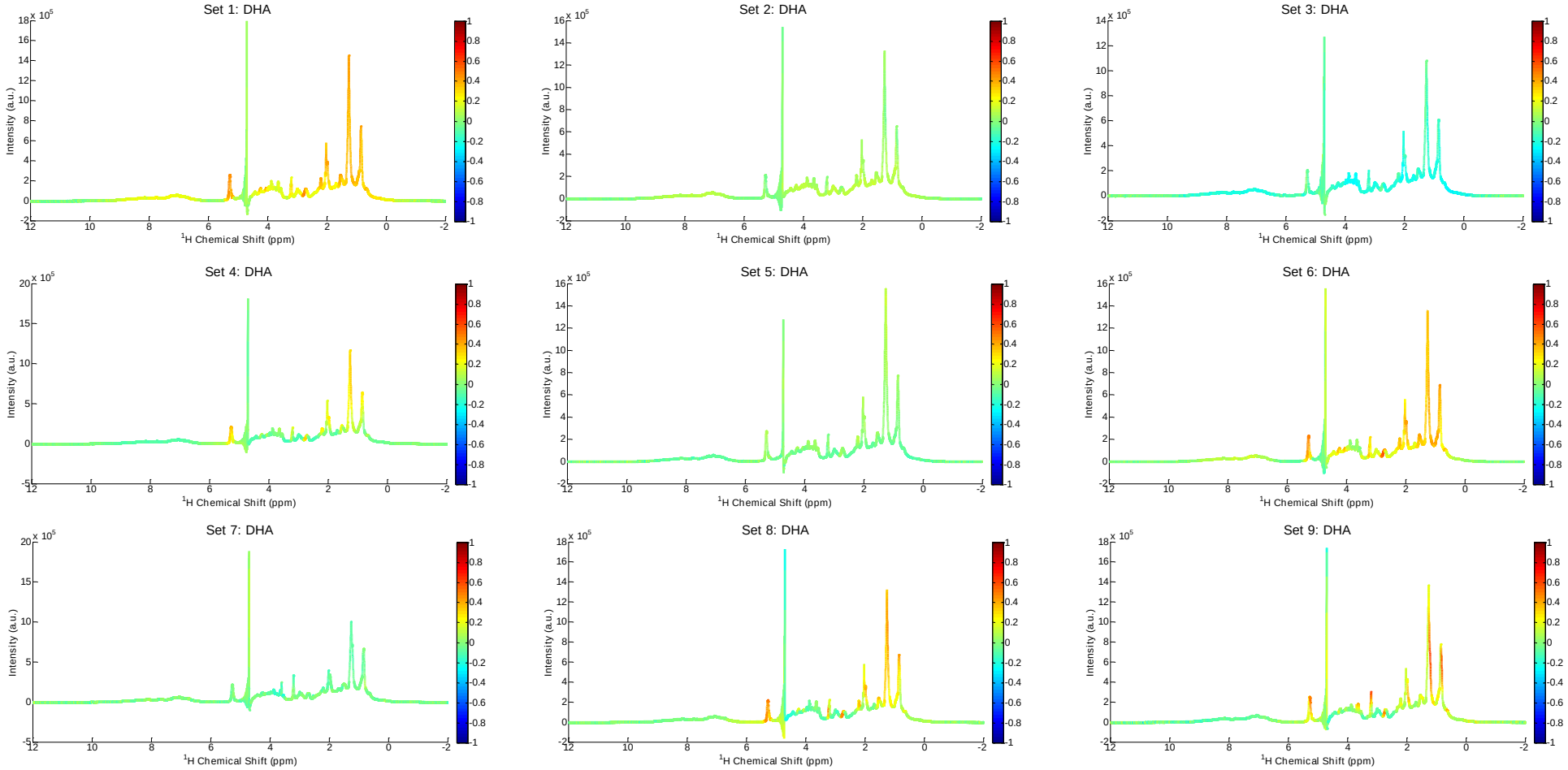

Set 8: W9
Intensity (a.u.)
x 10^5
^{1}H Chemical Shift (ppm)


Set 9: W9
Intensity (a.u.)
x 10^5
^{1}H Chemical Shift (ppm)

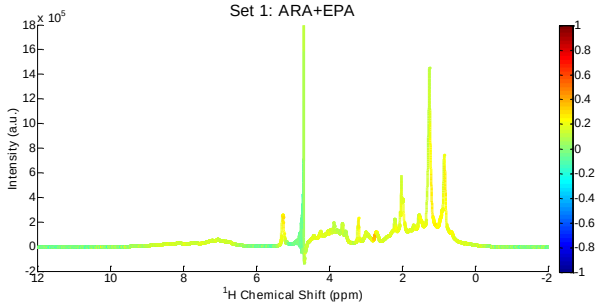
W3

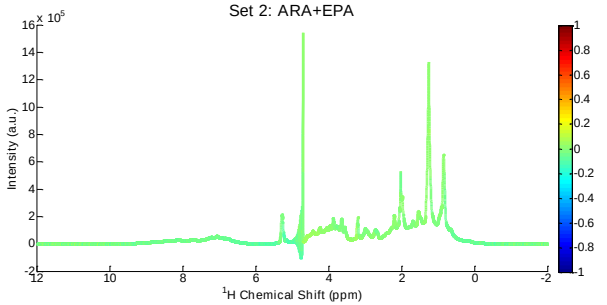

Set 1: W3
Set 2: W3
Set 3: W3
Set 4: W3
Set 5: W3
Set 6: W3
Set 7: W3
Set 8: W3
Set 9: W3
Intensity (a.u.)
x 10^5
1H Chemical Shift (ppm)

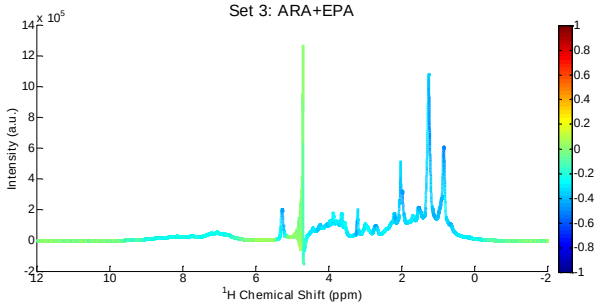
DHA

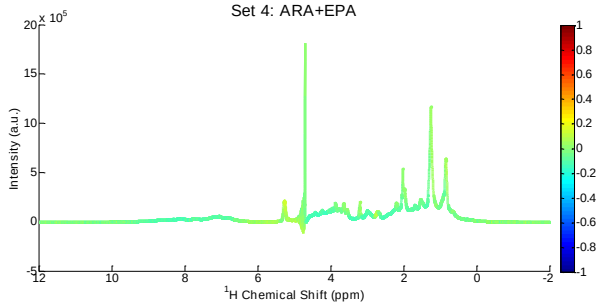

Set 1: DHA
Set 2: DHA
Set 3: DHA
Set 4: DHA
Set 5: DHA
Set 6: DHA
Set 7: DHA
Set 8: DHA
Set 9: DHA
Intensity (a.u.)
x 10^5
1H Chemical Shift (ppm)

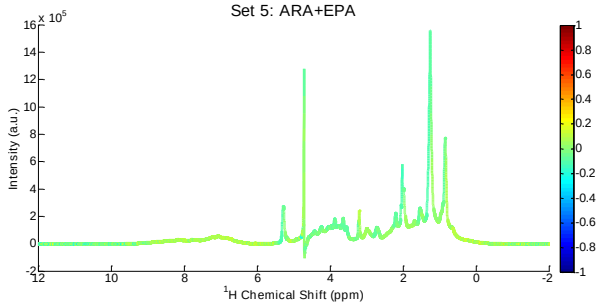
ARA+EPA

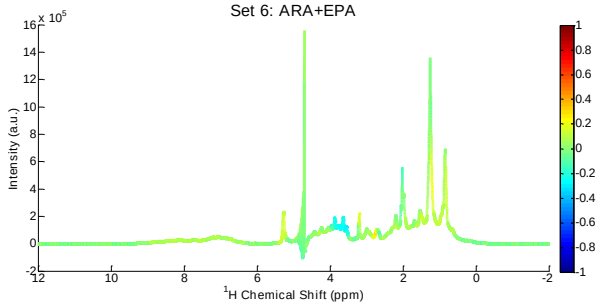

Set 1: ARA+EPA
Intensity (a.u.)
^{1}H Chemical Shift (ppm)

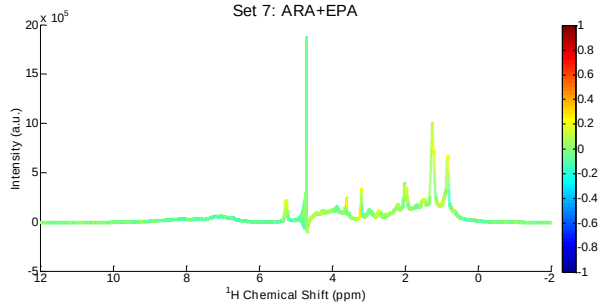

Set 2: ARA+EPA
Intensity (a.u.)
^{1}H Chemical Shift (ppm)

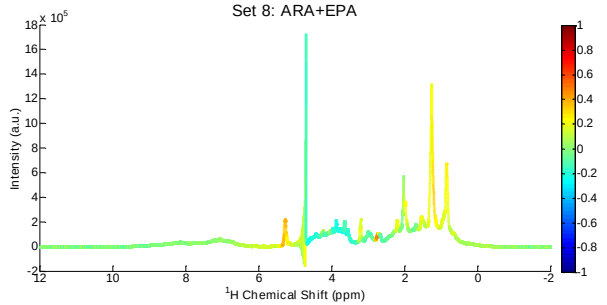

Set 3: ARA+EPA
Intensity (a.u.)
^{1}H Chemical Shift (ppm)

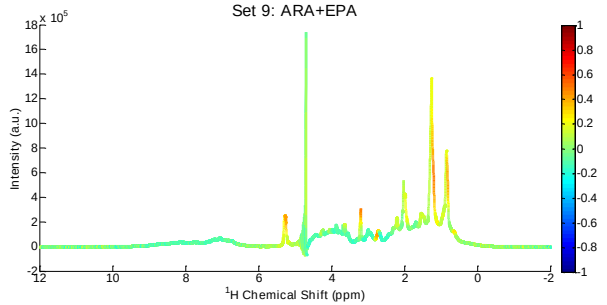

Set 4: ARA+EPA
Intensity (a.u.)
^{1}H Chemical Shift (ppm)


Set 5: ARA+EPA
Intensity (a.u.)
^{1}H Chemical Shift (ppm)

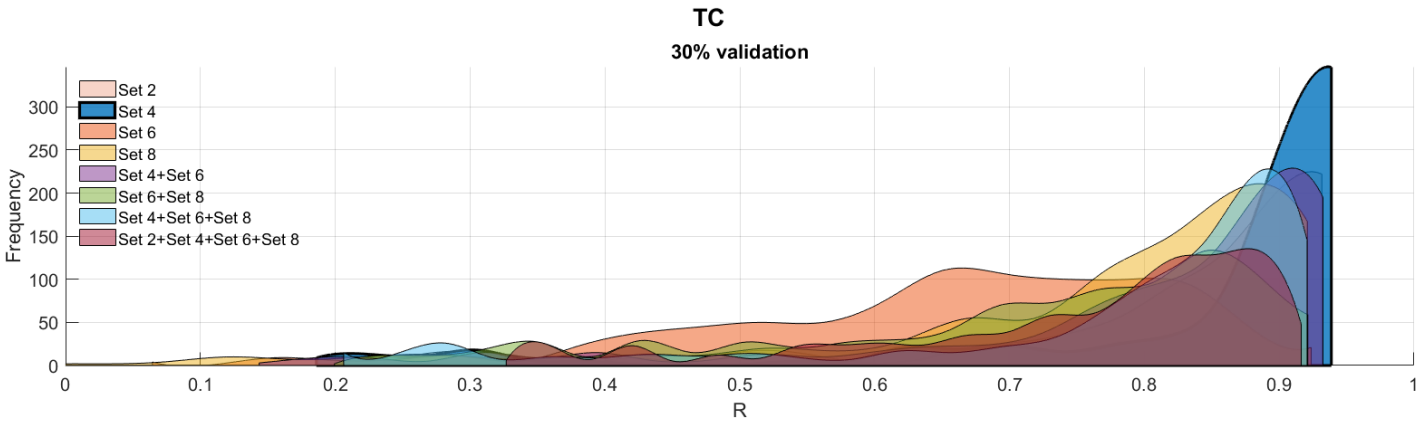

Set 6: ARA+EPA
Intensity (a.u.)
^{1}H Chemical Shift (ppm)

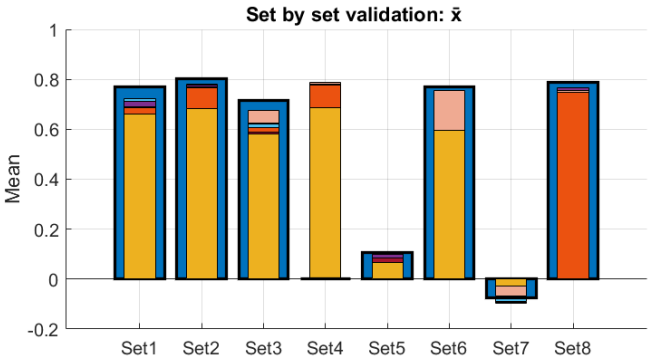

Set 7: ARA+EPA
Intensity (a.u.)
^{1}H Chemical Shift (ppm)

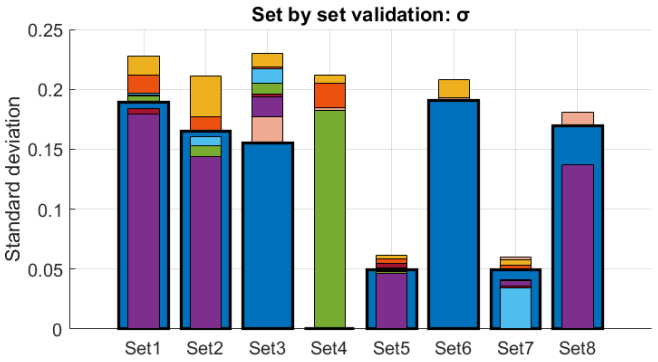

Set 8: ARA+EPA
Intensity (a.u.)
^{1}H Chemical Shift (ppm)

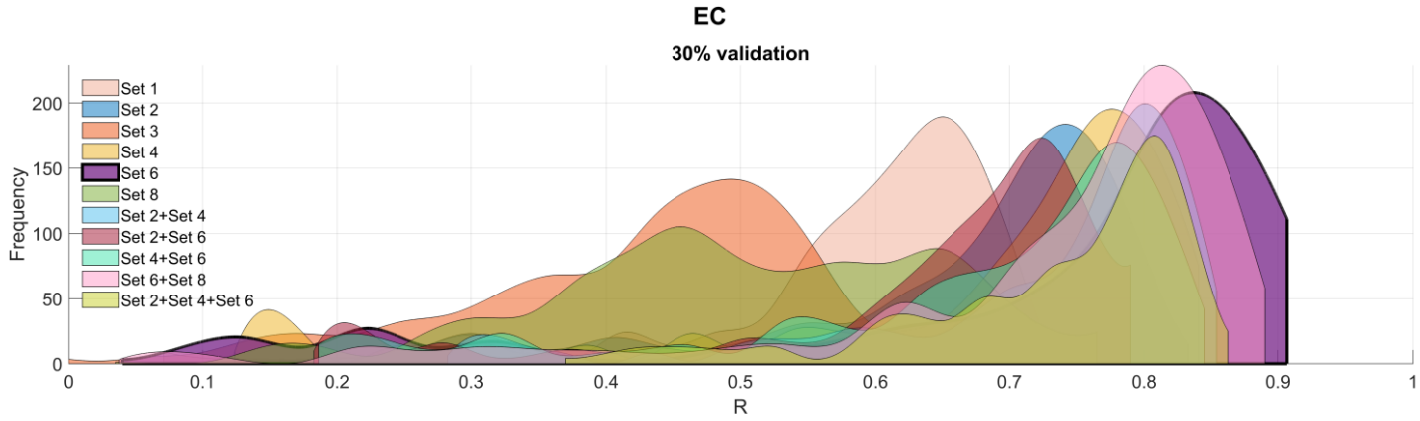

Set 9: ARA+EPA
Intensity (a.u.)
^{1}H Chemical Shift (ppm)

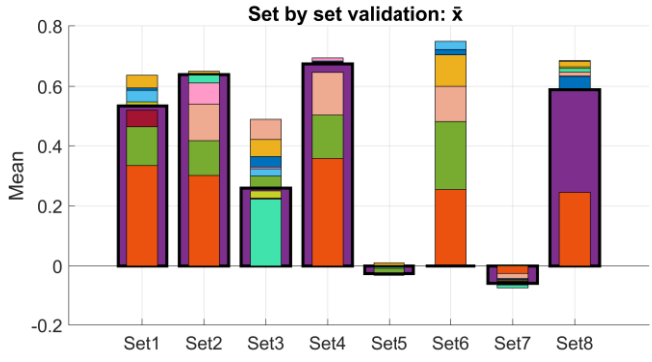
Annex 2. Set selection assessment plots for each lipid's study

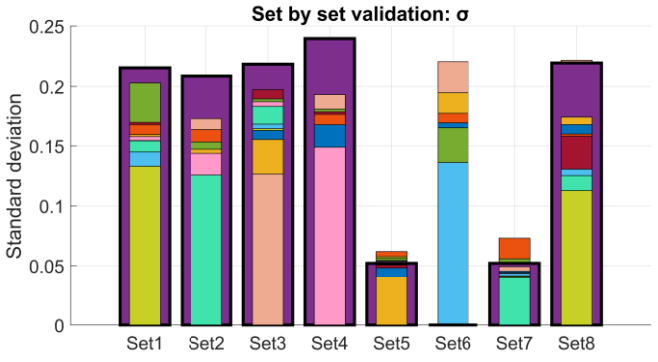


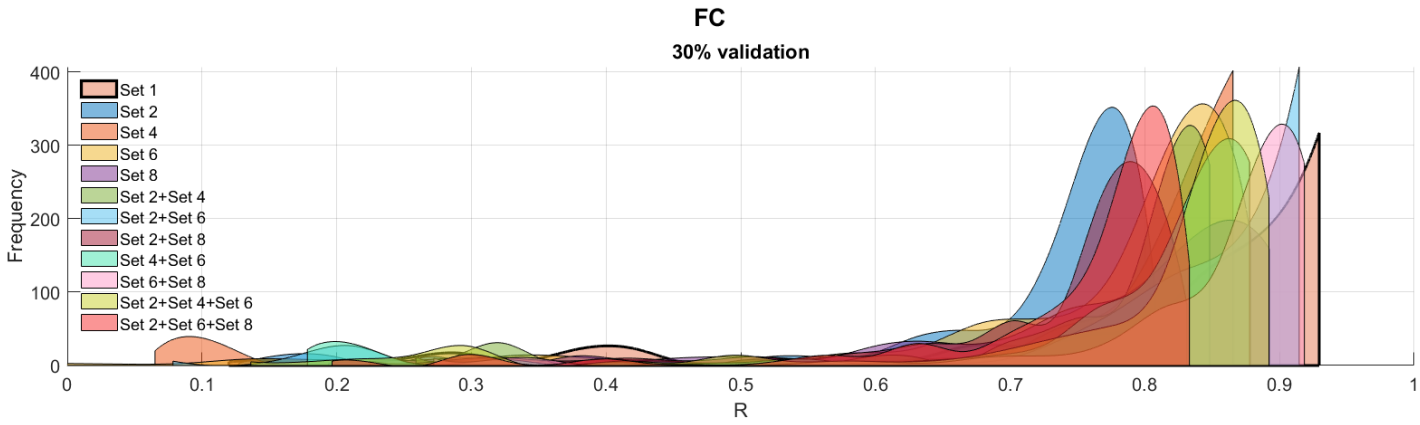


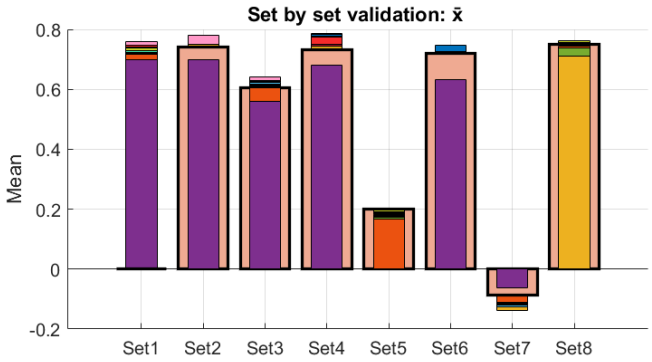


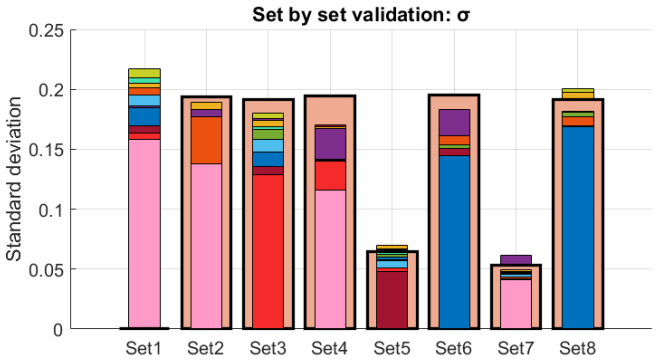


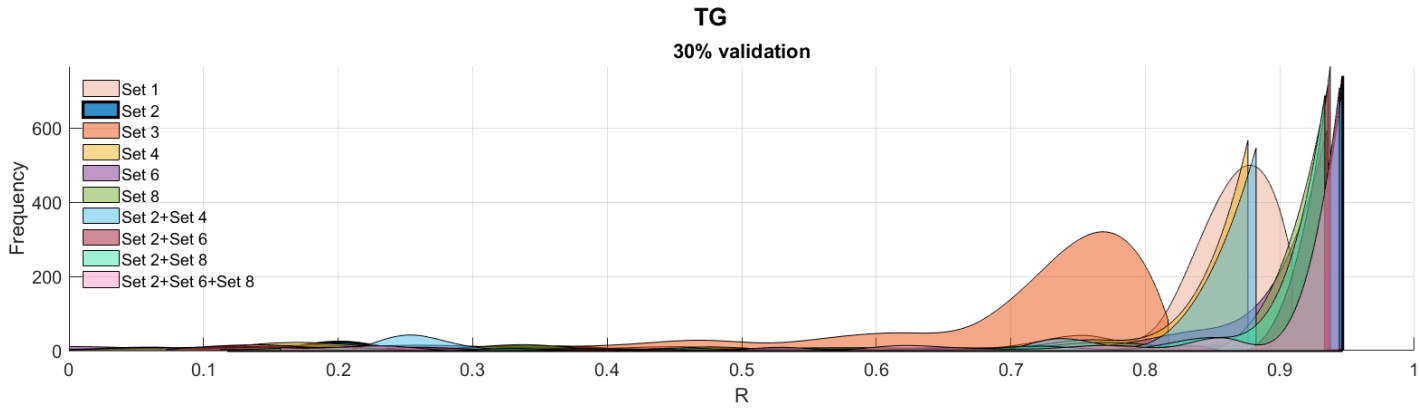


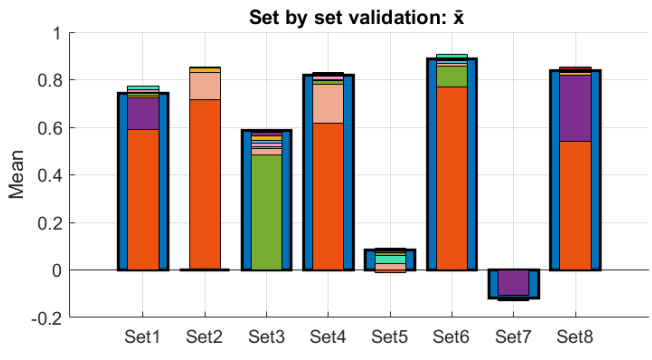


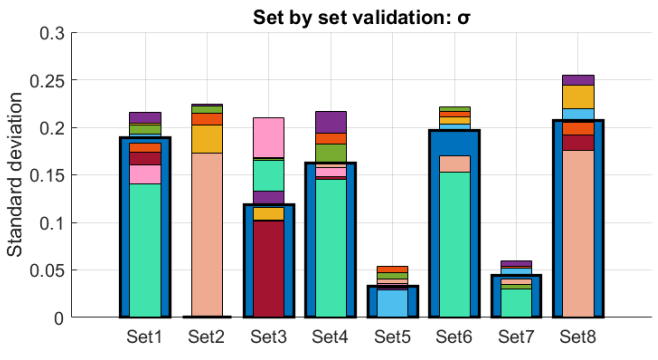

FC
30% validation
Set 1
Set 2
Set 4
Set 6
Set 8
Set 2+Set 4
Set 2+Set 6
Set 2+Set 8
Set 4+Set 6
Set 6+Set 8
Set 2+Set 4+Set 6
Set 2+Set 6+Set 8
Frequency
R

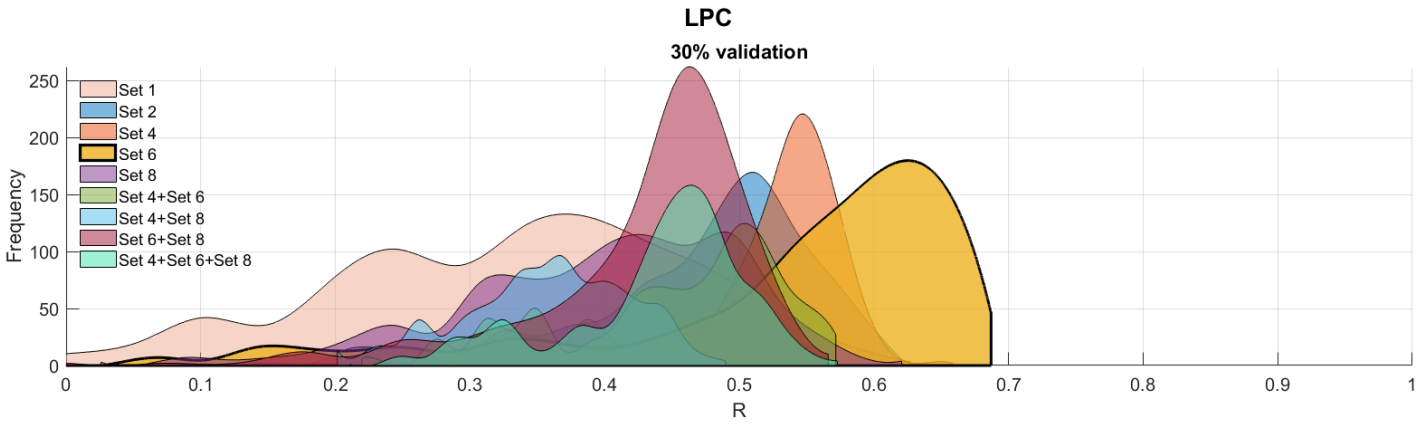

Set by set validation: x̄
Mean
Set1
Set2
Set3
Set4
Set5
Set6
Set7
Set8

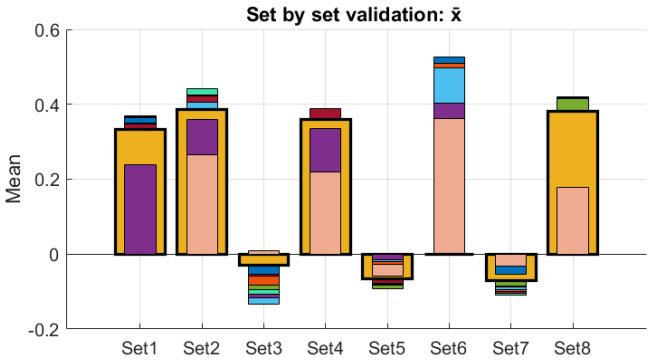

Set by set validation: σ
Standard deviation
Set1
Set2
Set3
Set4
Set5
Set6
Set7
Set8

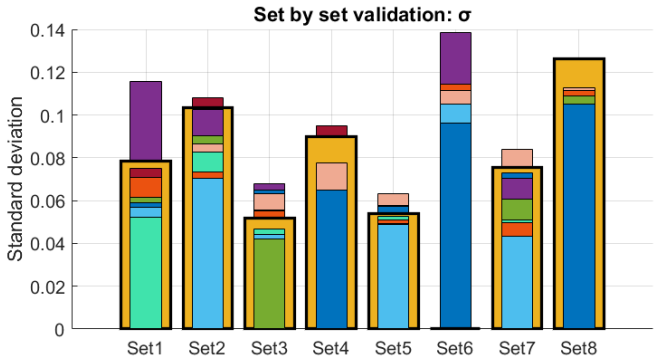

TG
30% validation
Set 1
Set 2
Set 3
Set 4
Set 6
Set 8
Set 2+Set 4
Set 2+Set 6
Set 2+Set 8
Set 2+Set 6+Set 8
Frequency
R

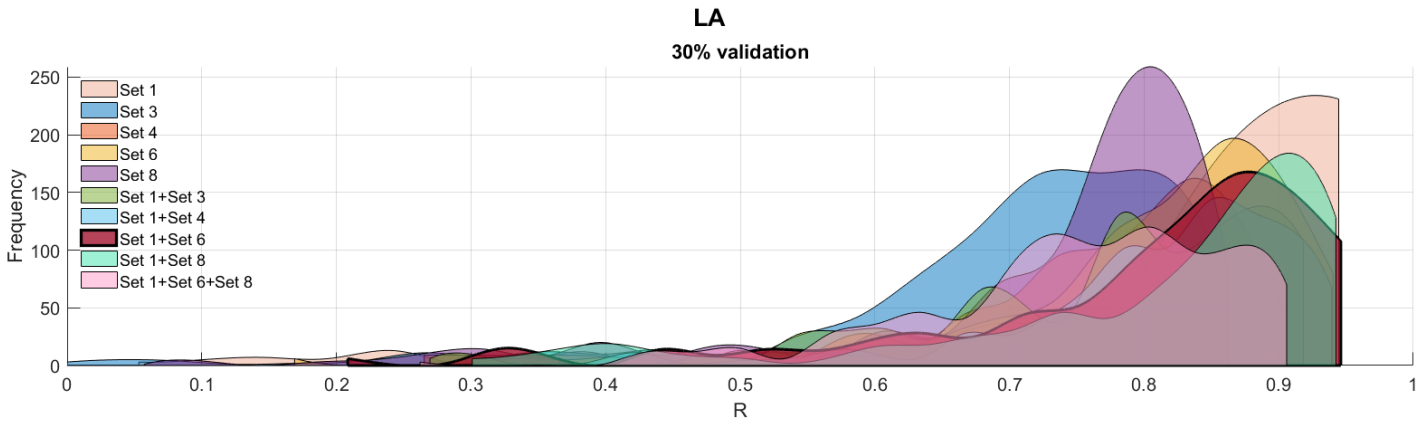

Set by set validation: x̄
Mean
Set1
Set2
Set3
Set4
Set5
Set6
Set7
Set8

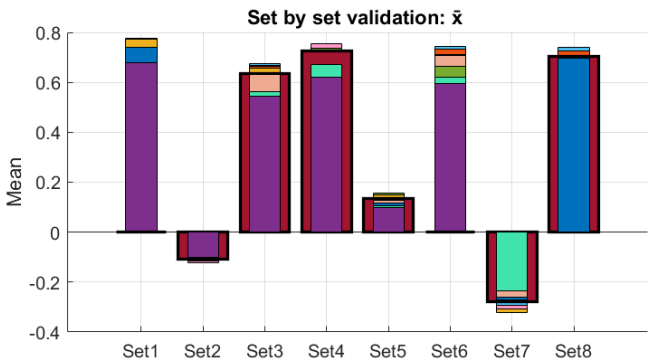

Set by set validation: σ
Standard deviation
Set1
Set2
Set3
Set4
Set5
Set6
Set7
Set8

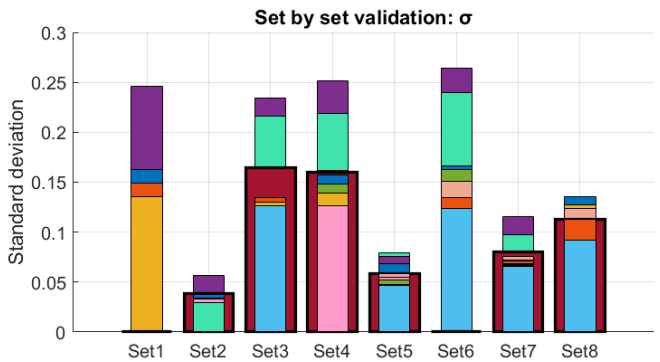

LPC
30% validation
Set 1
Set 2
Set 4
Set 6
Set 8
Set 4+Set 6
Set 4+Set 8
Set 6+Set 8
Set 4+Set 6+Set 8
Frequency
R

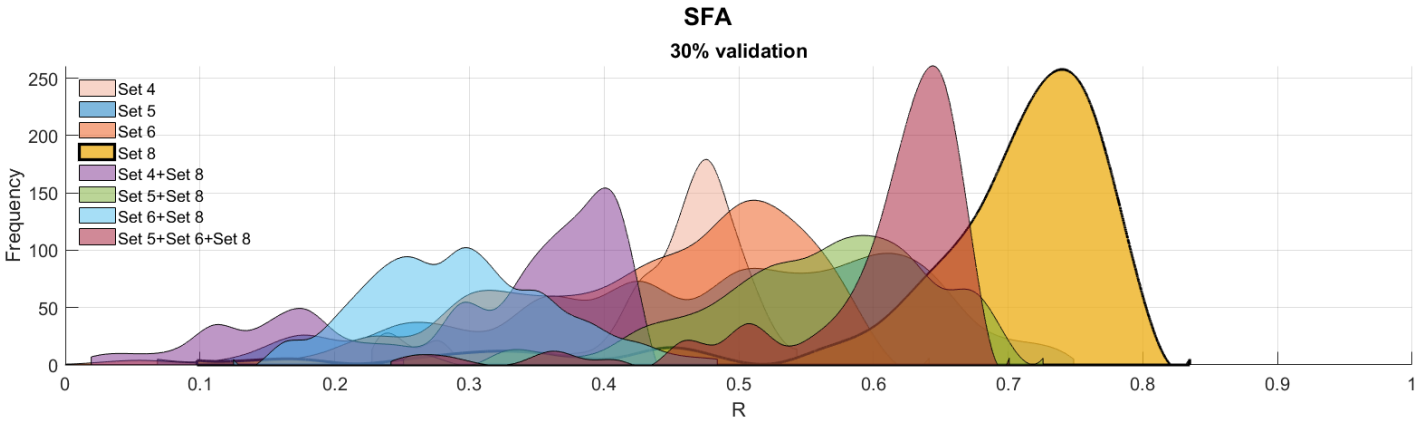

Set by set validation: x̄
Mean
Set1
Set2
Set3
Set4
Set5
Set6
Set7
Set8

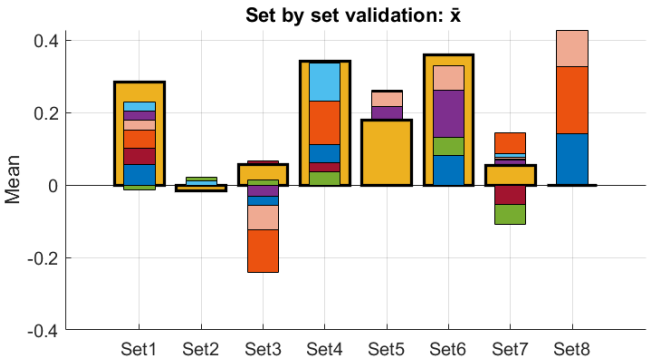

Set by set validation: σ
Standard deviation
Set1
Set2
Set3
Set4
Set5
Set6
Set7
Set8

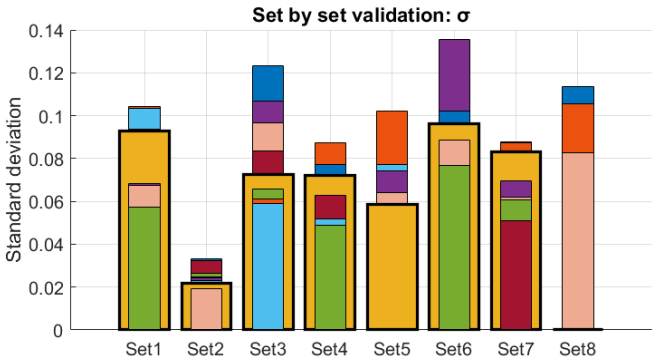

LA
30% validation
Set 1
Set 3
Set 4
Set 6
Set 8
Set 1+Set 3
Set 1+Set 4
Set 1+Set 6
Set 1+Set 8
Set 1+Set 6+Set 8
Frequency
R

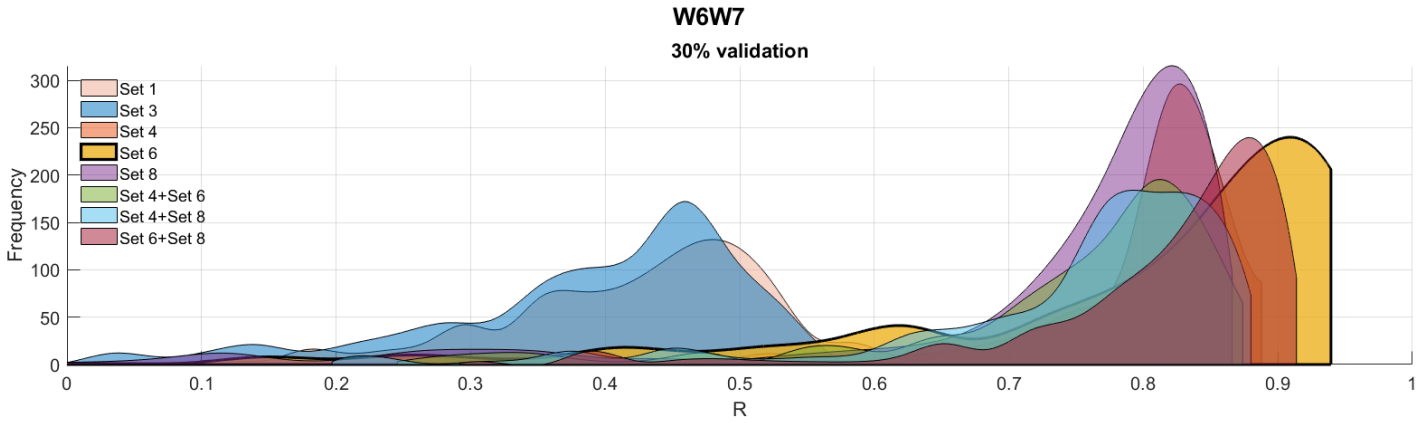

Set by set validation: x̄
Mean
Set1
Set2
Set3
Set4
Set5
Set6
Set7
Set8

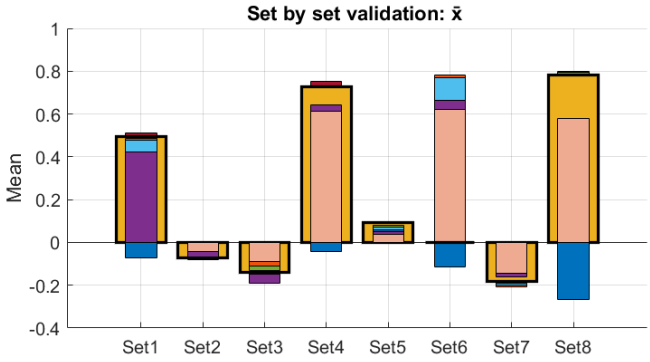

Set by set validation: σ
Standard deviation
Set1
Set2
Set3
Set4
Set5
Set6
Set7
Set8

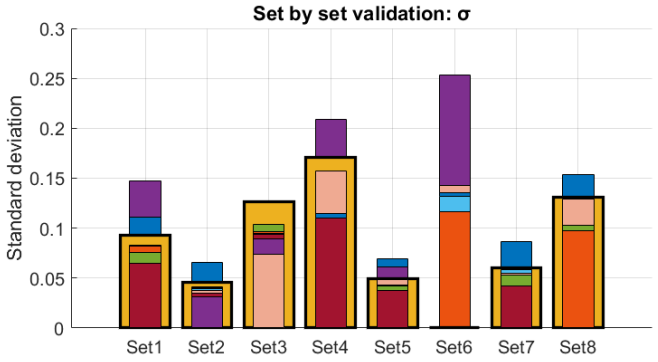

SFA
30% validation
Set 4
Set 5
Set 6
Set 8
Set 4+Set 8
Set 5+Set 8
Set 6+Set 8
Set 5+Set 6+Set 8
Frequency
R

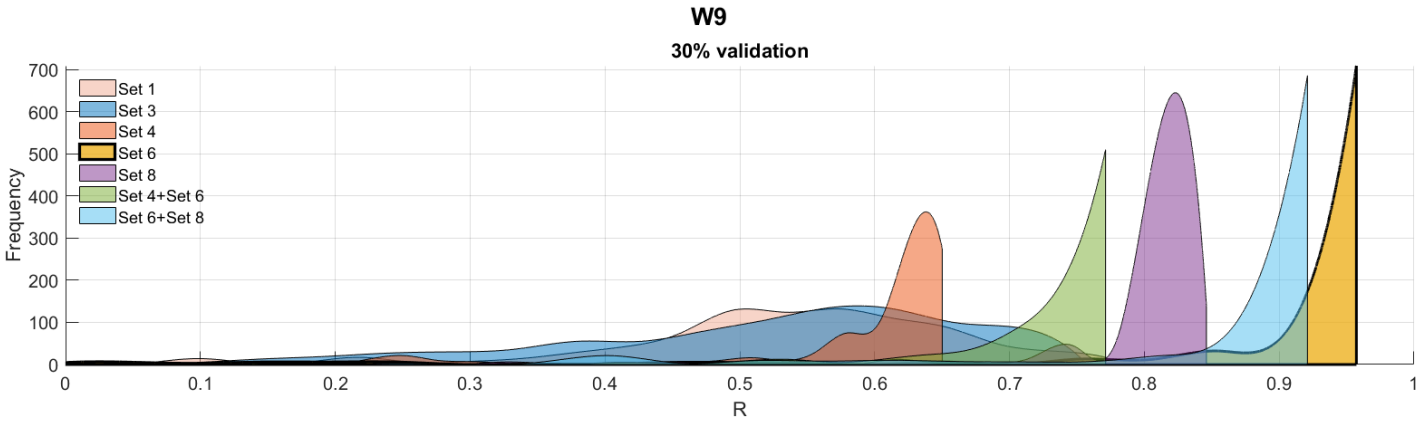

Set by set validation: x̄
Mean
Set1
Set2
Set3
Set4
Set5
Set6
Set7
Set8

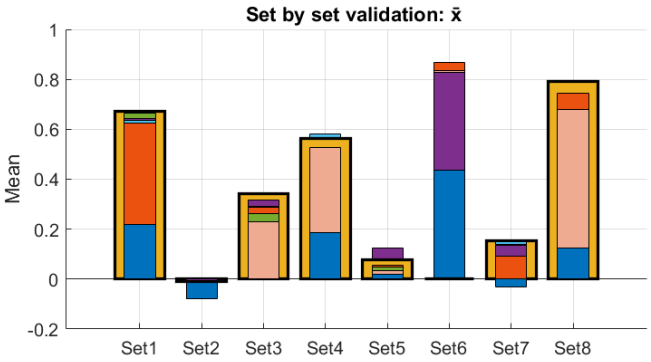

Set by set validation: σ
Standard deviation
Set1
Set2
Set3
Set4
Set5
Set6
Set7
Set8

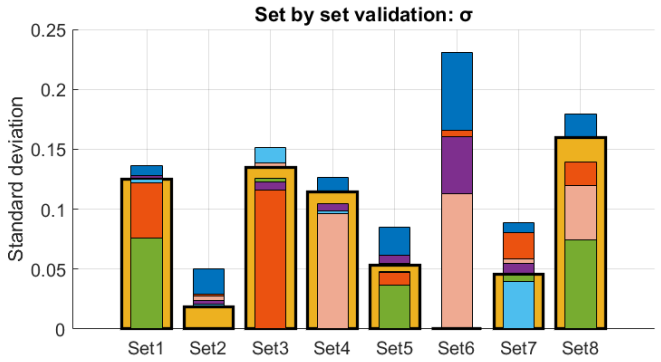

W6W7
30% validation
Set 1
Set 3
Set 4
Set 6
Set 8
Set 4+Set 6
Set 4+Set 8
Set 6+Set 8
Frequency
R

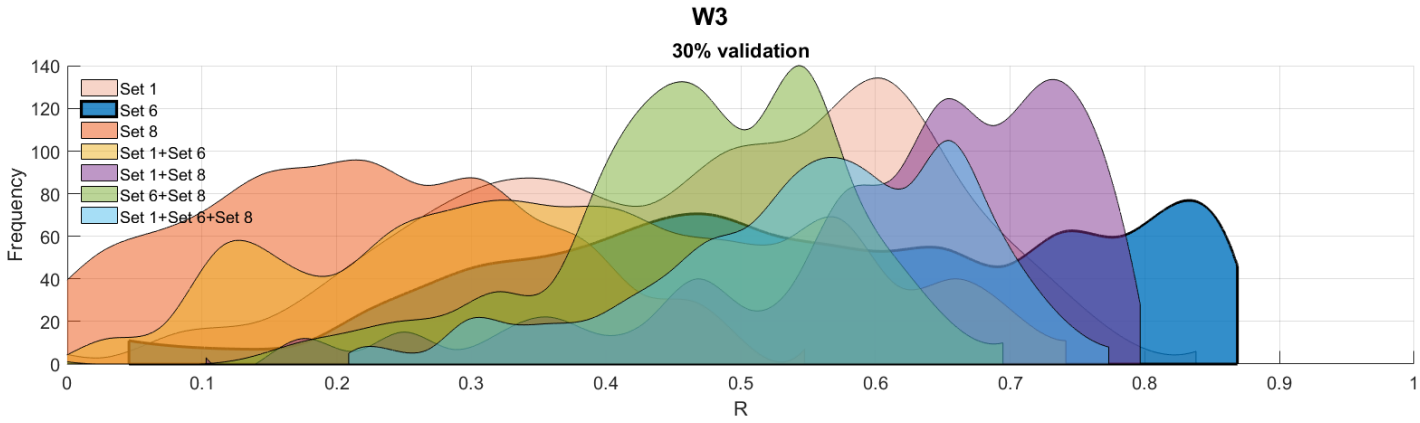

Set by set validation: x̄
Mean
Set1
Set2
Set3
Set4
Set5
Set6
Set7
Set8

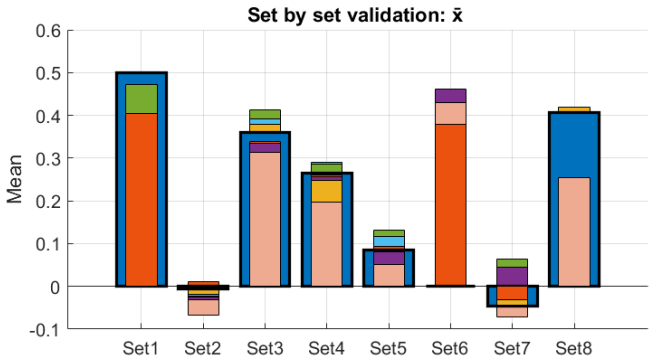

Set by set validation: σ
Standard deviation
Set1
Set2
Set3
Set4
Set5
Set6
Set7
Set8

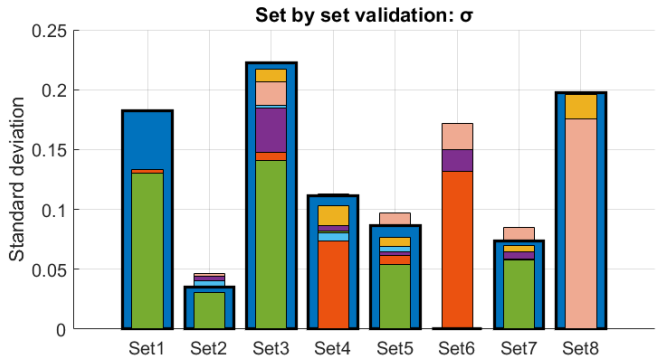

W9
30% validation
Set 1
Set 3
Set 4
Set 6
Set 8
Set 4+Set 6
Set 6+Set 8
Frequency
R

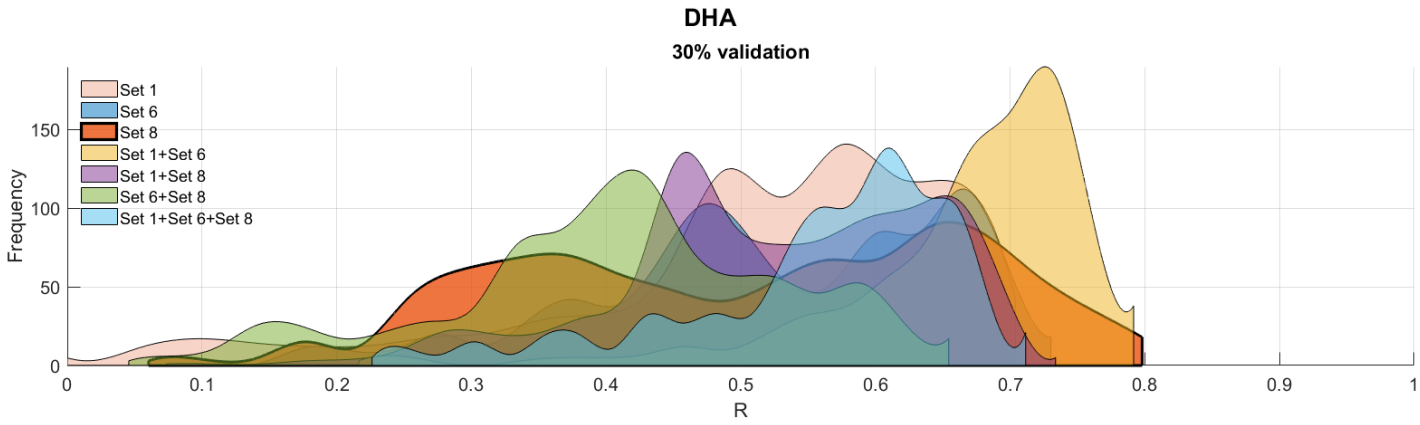

Set by set validation: x̄
Mean
Set1
Set2
Set3
Set4
Set5
Set6
Set7
Set8

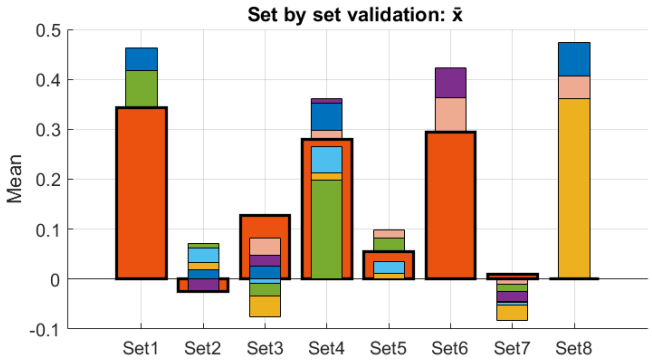

Set by set validation: σ
Standard deviation
Set1
Set2
Set3
Set4
Set5
Set6
Set7
Set8

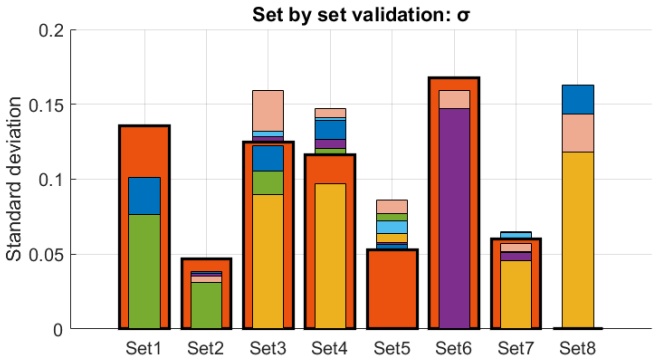

W3
30% validation
Set 1
Set 6
Set 8
Set 1+Set 6
Set 1+Set 8
Set 6+Set 8
Set 1+Set 6+Set 8
Frequency
R

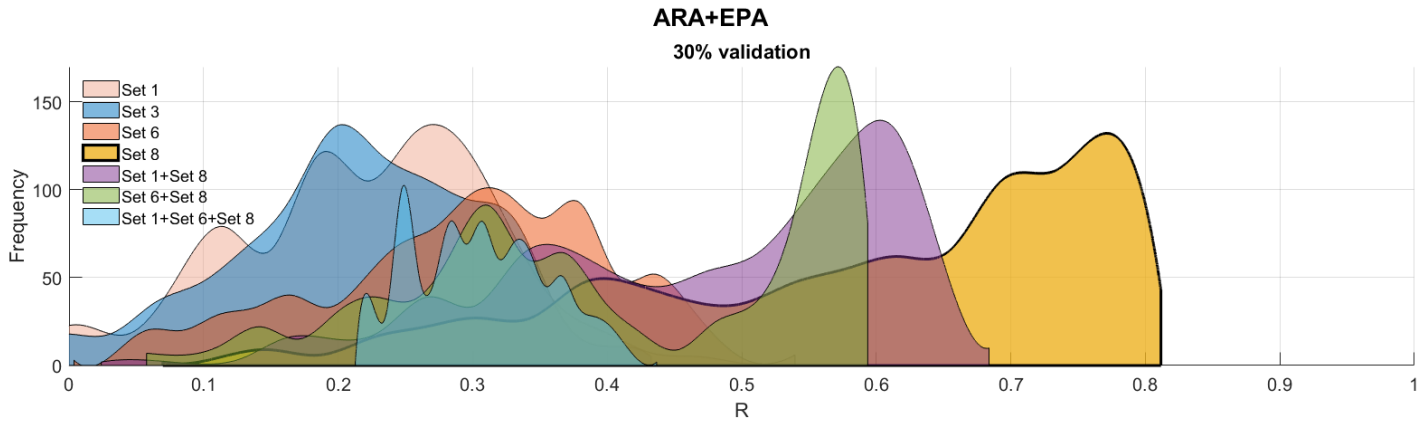

Set by set validation: x̄
Mean
Set1
Set2
Set3
Set4
Set5
Set6
Set7
Set8

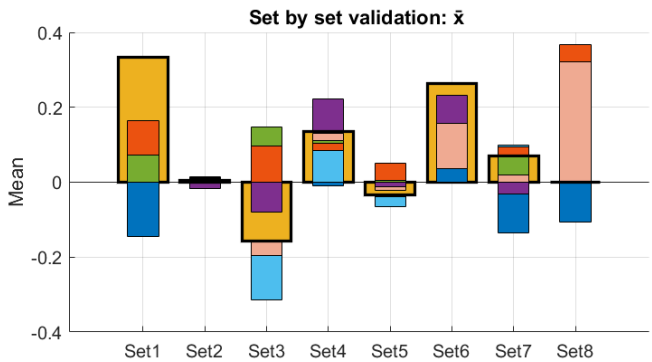

Set by set validation: σ
Standard deviation
Set1
Set2
Set3
Set4
Set5
Set6
Set7
Set8

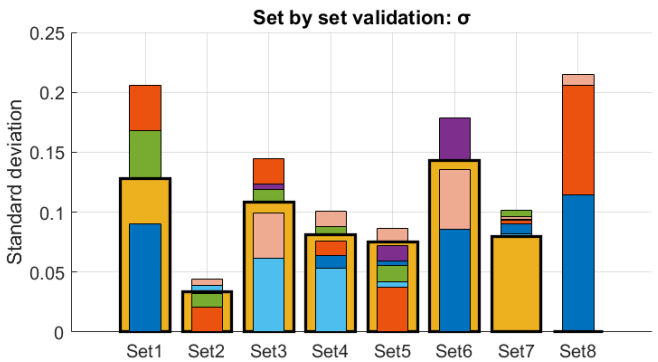

DHA
30% validation
Set 1
Set 6
Set 8
Set 1+Set 6
Set 1+Set 8
Set 6+Set 8
Set 1+Set 6+Set 8
Frequency
R


Set by set validation: x̄
Mean
Set1
Set2
Set3
Set4
Set5
Set6
Set7
Set8

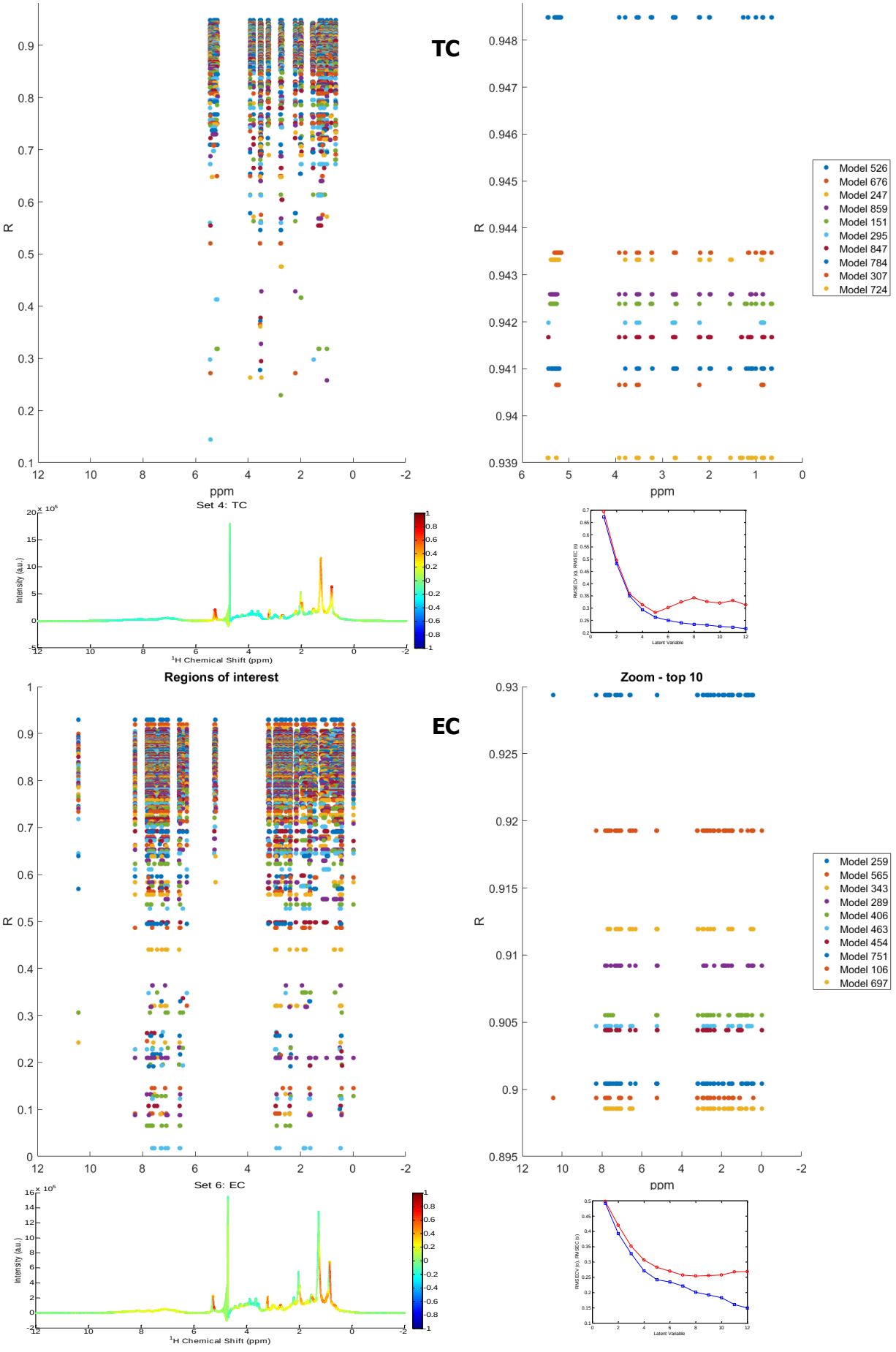

Set by set validation: σ
Standard deviation
Set1
Set2
Set3
Set4
Set5
Set6
Set7
Set8

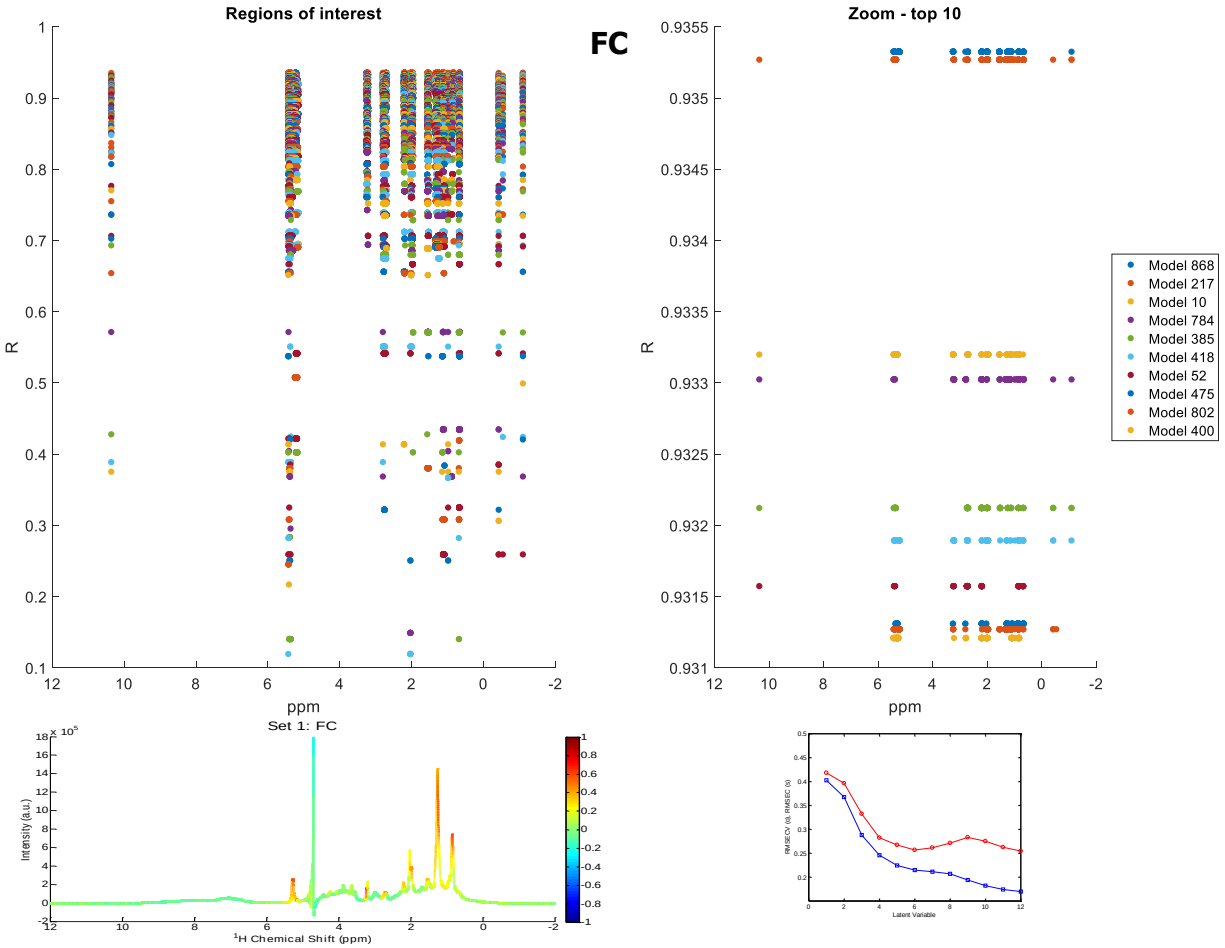

ARA+EPA
30% validation
Set 1
Set 3
Set 6
Set 8
Set 1+Set 8
Set 6+Set 8
Set 1+Set 6+Set 8
Frequency
R

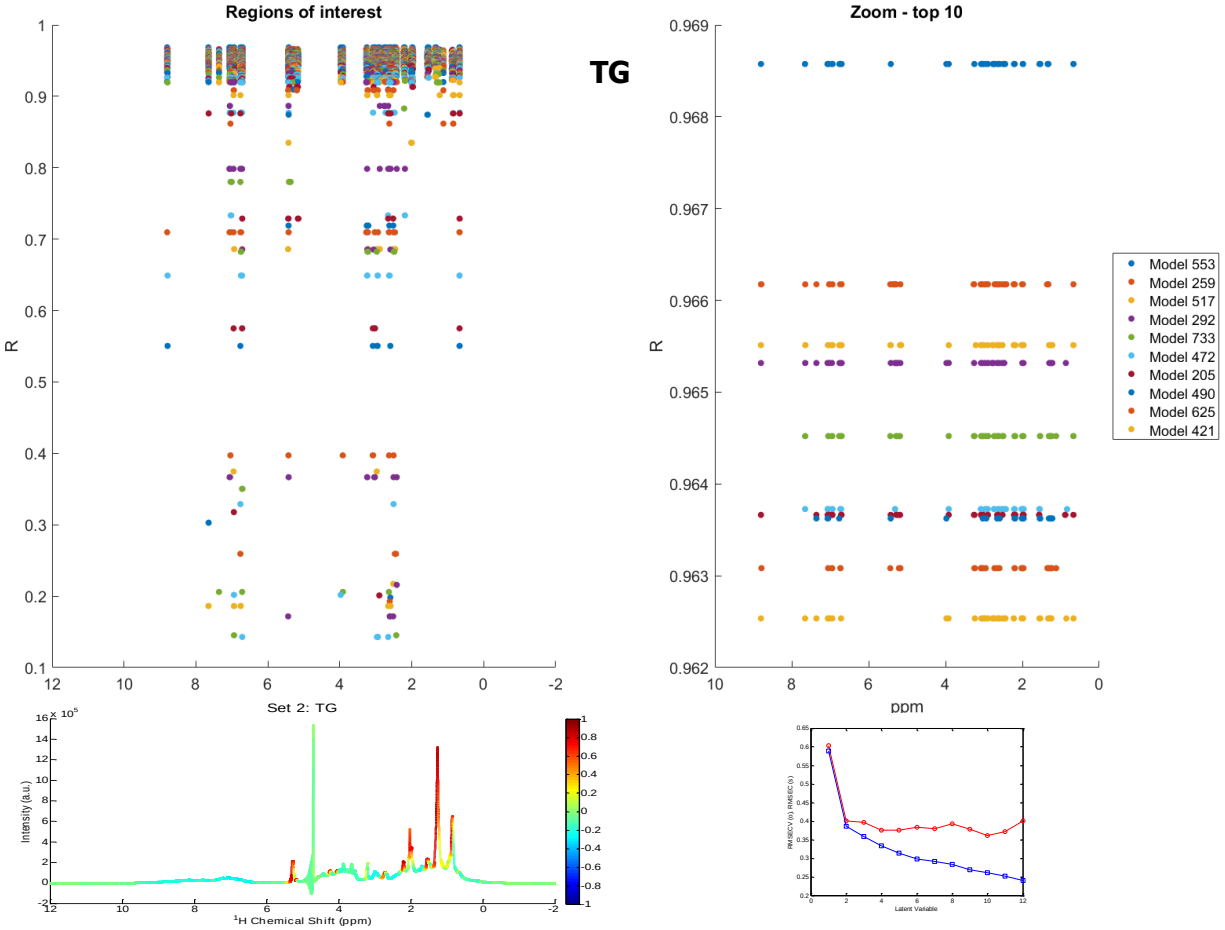

Set by set validation: x̄
Mean
Set1
Set2
Set3
Set4
Set5
Set6
Set7
Set8

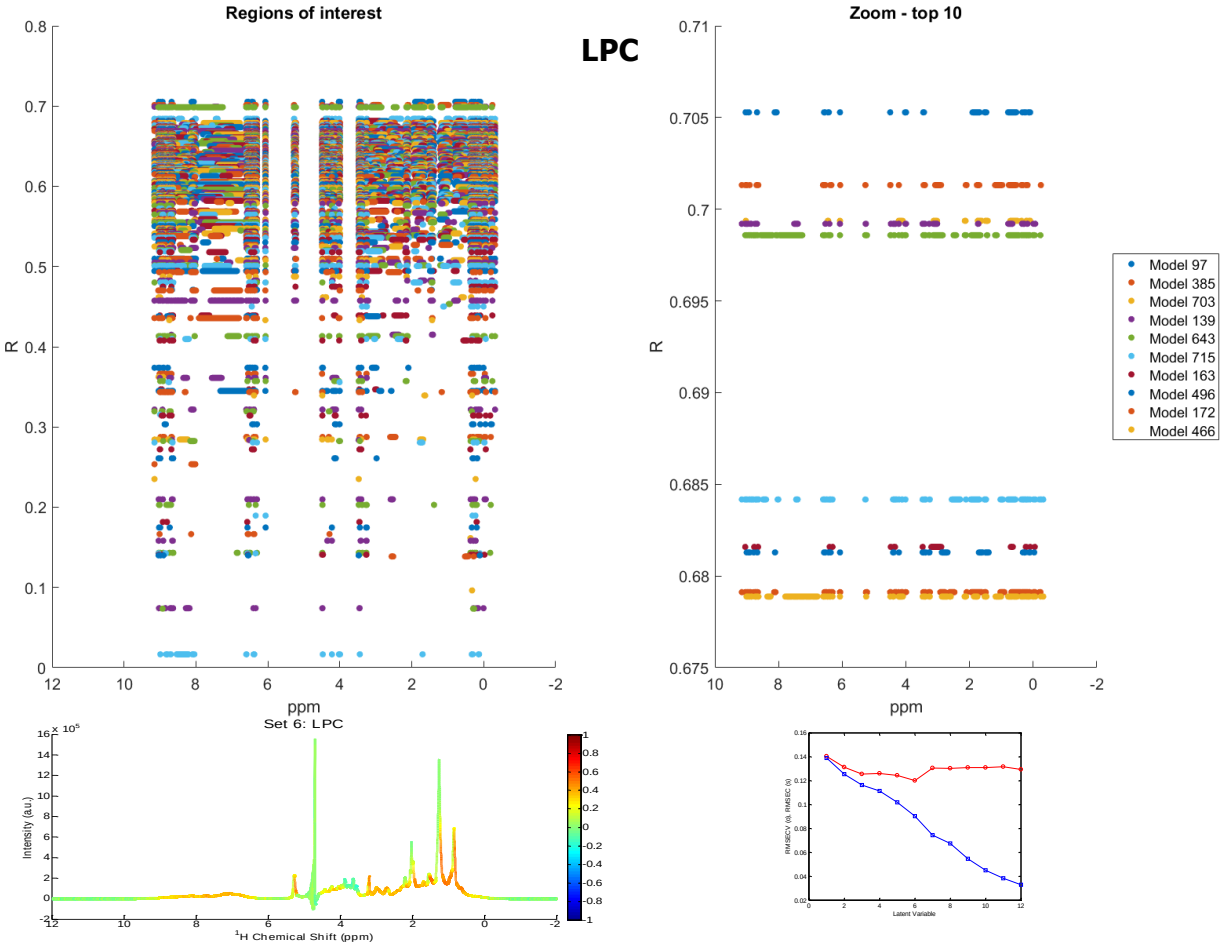

Set by set validation: σ
Standard deviation
Set1
Set2
Set3
Set4
Set5
Set6
Set7
Set8

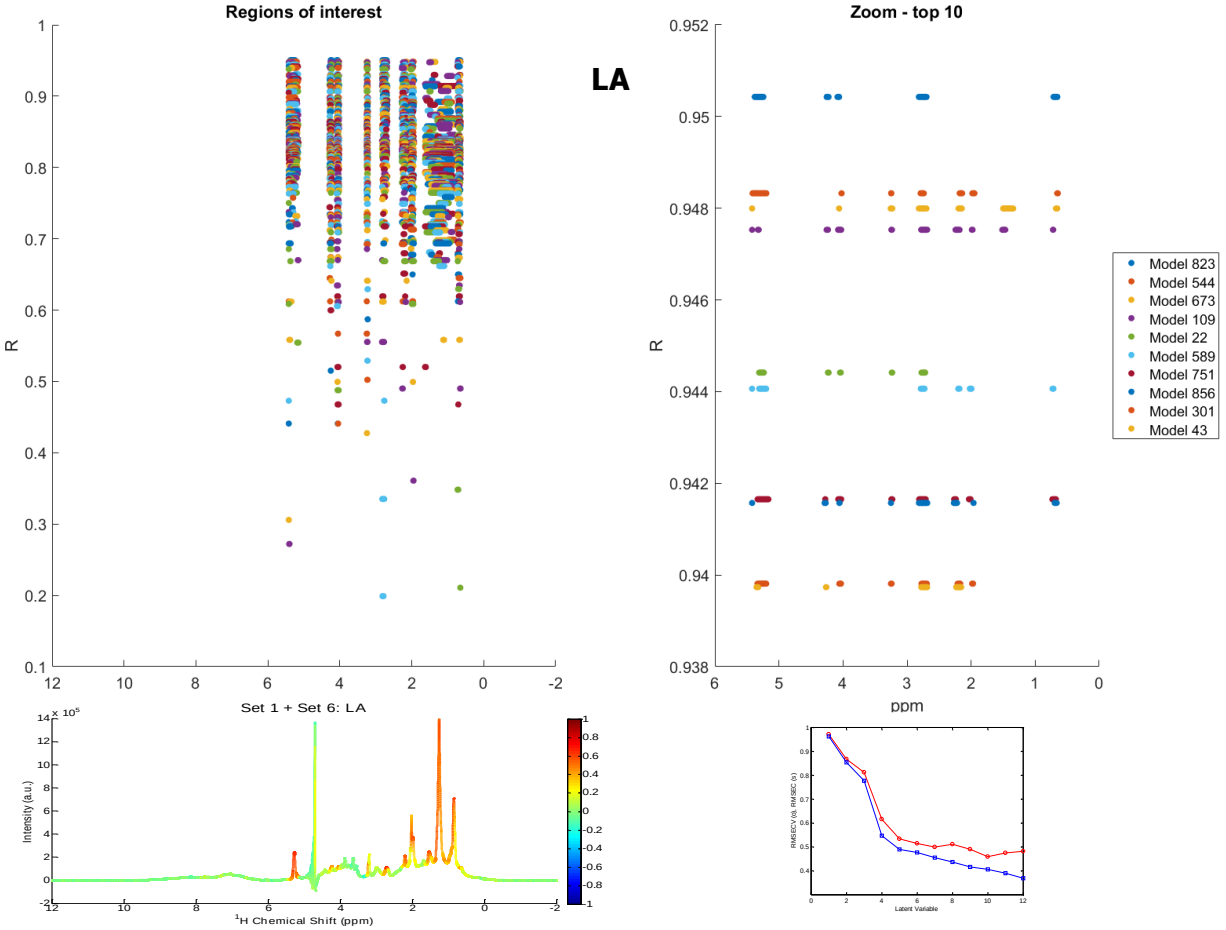
Annex 3. Each lipid selected set's baseline spectral regions randomizations and RMSE vs LV increase plot

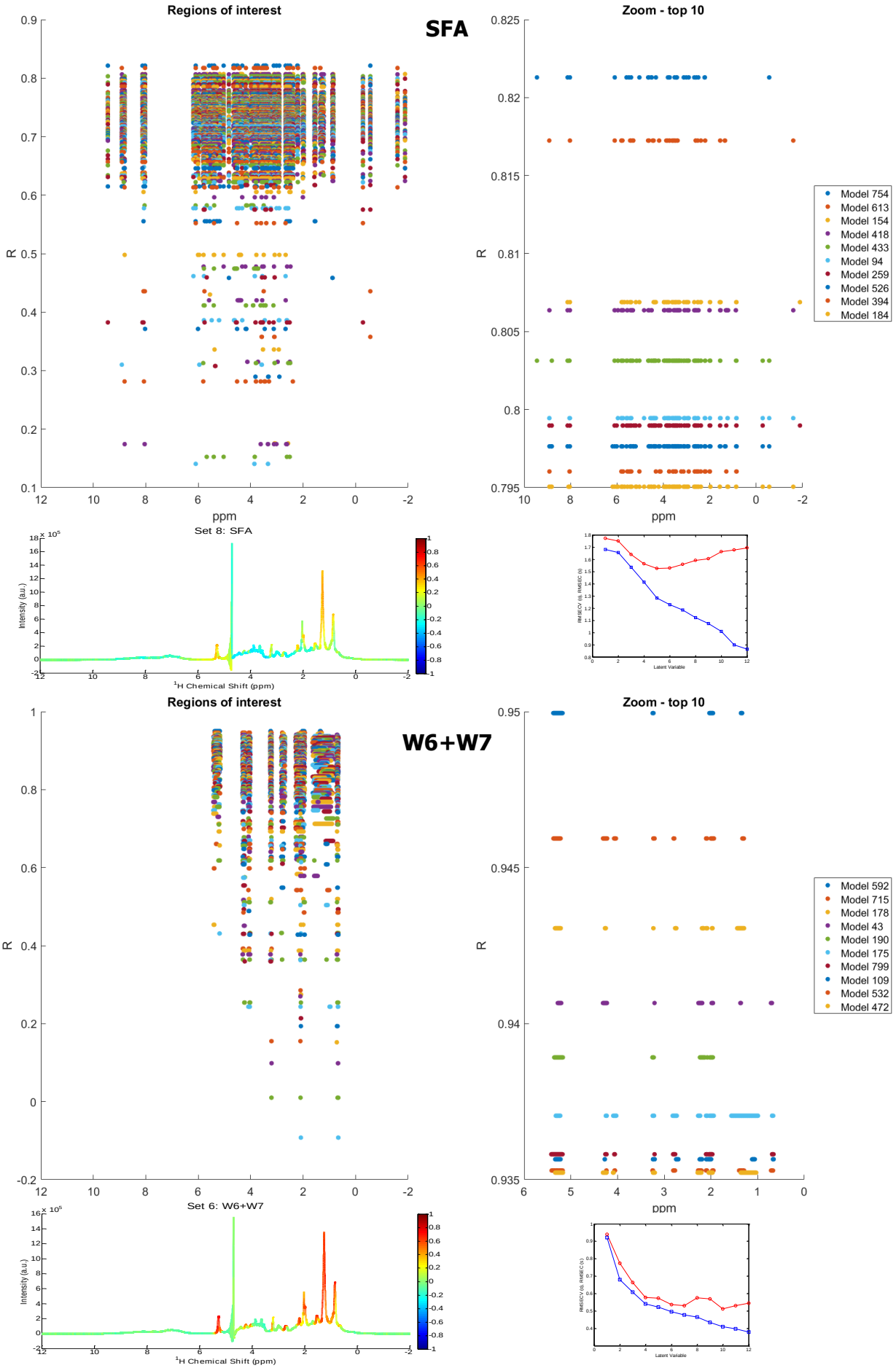


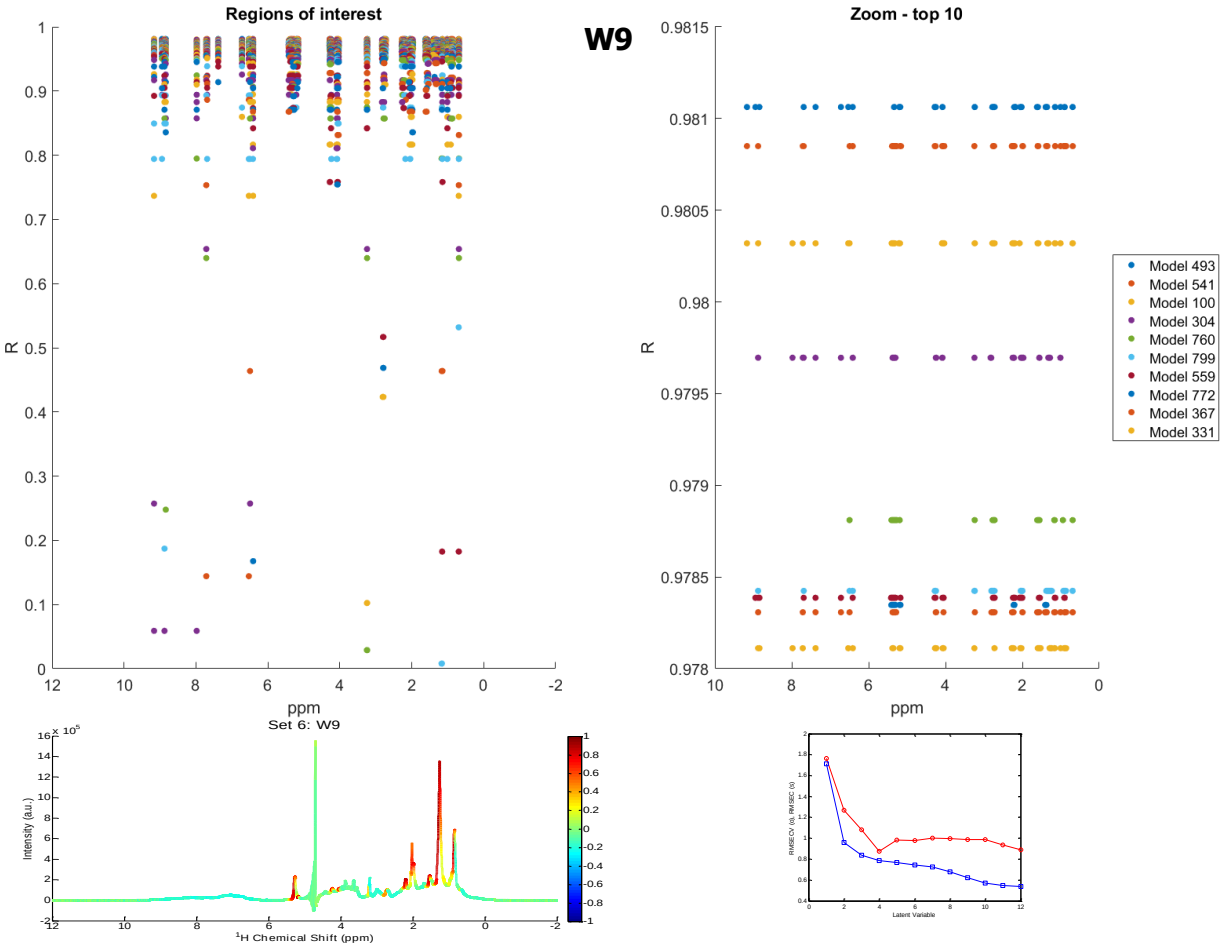

FC
Regions of interest
R
ppm
Zoom - top 10
Model 868
Model 217
Model 10
Model 784
Model 385
Model 418
Model 52
Model 475
Model 802
Model 400
Set 1: FC
Intensity (a.u.)
¹H Chemical Shift (ppm)
Latent Variable

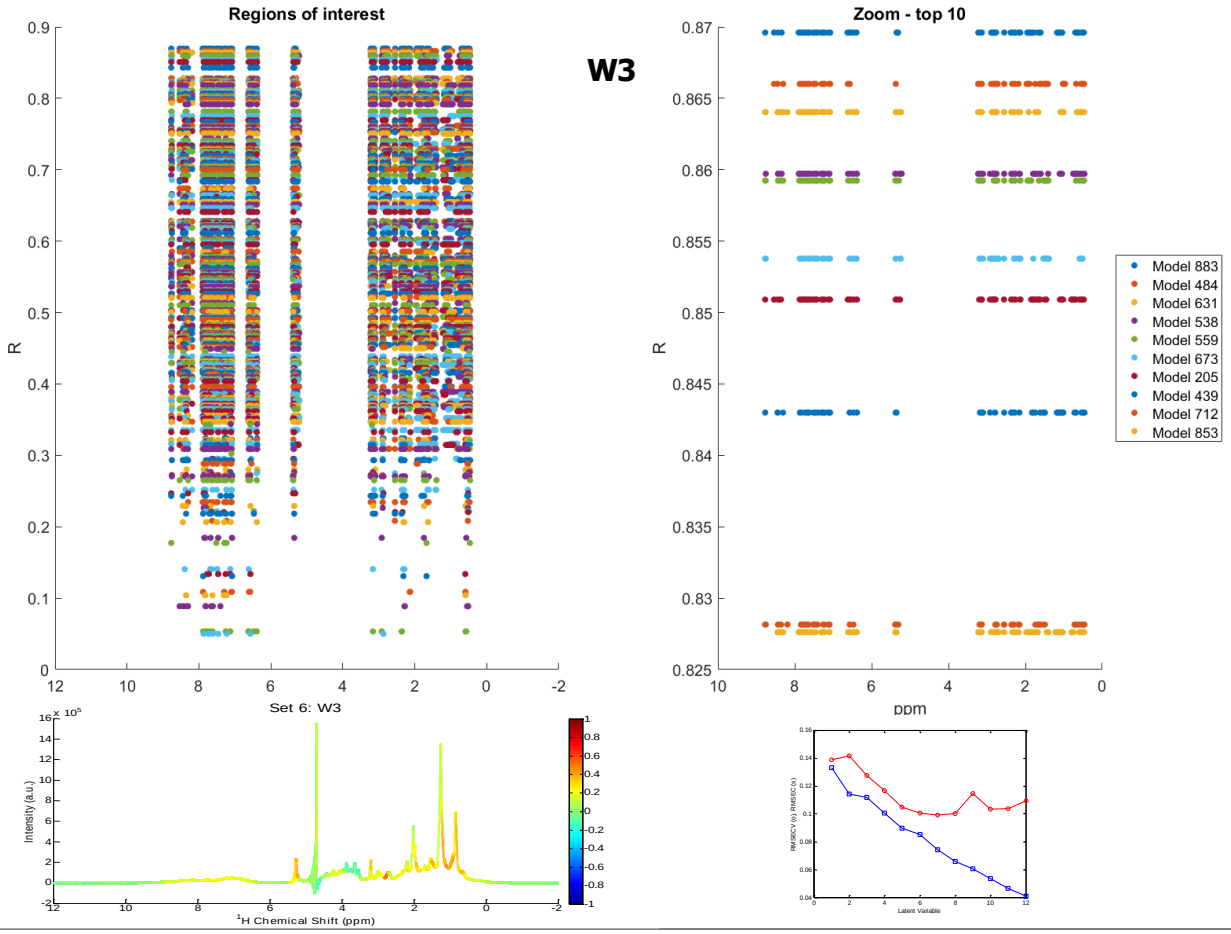

TG
Regions of interest
R
ppm
Zoom - top 10
Model 553
Model 259
Model 517
Model 292
Model 733
Model 472
Model 205
Model 490
Model 625
Model 421
Set 2: TG
Intensity (a.u.)
¹H Chemical Shift (ppm)
Latent Variable

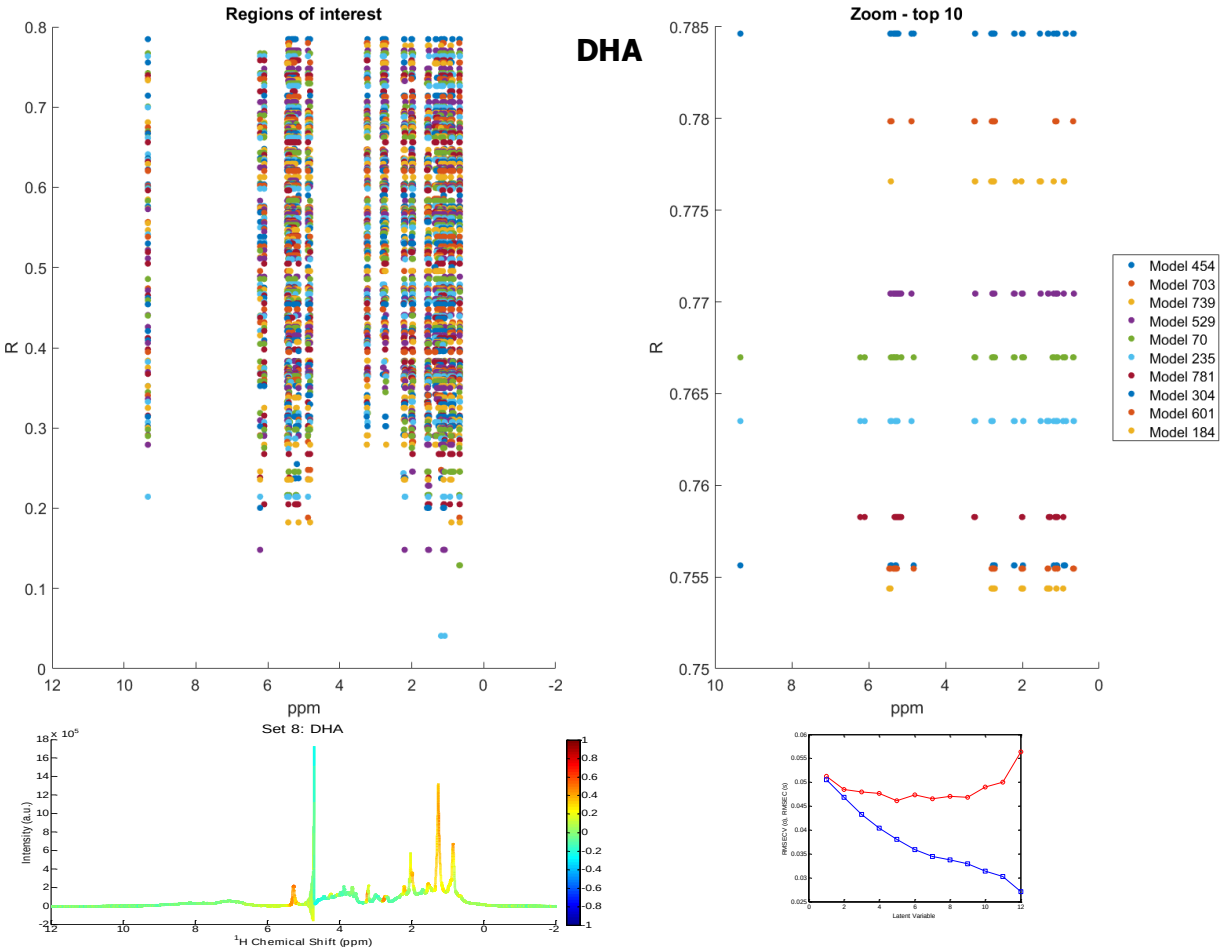

LPC
Regions of interest
R
ppm
Zoom - top 10
Model 97
Model 385
Model 703
Model 139
Model 643
Model 715
Model 163
Model 496
Model 172
Model 466
Set 6: LPC
Intensity (a.u.)
^{1}H Chemical Shift (ppm)
RMSECV (&, RMSEC (x)
Latent Variable

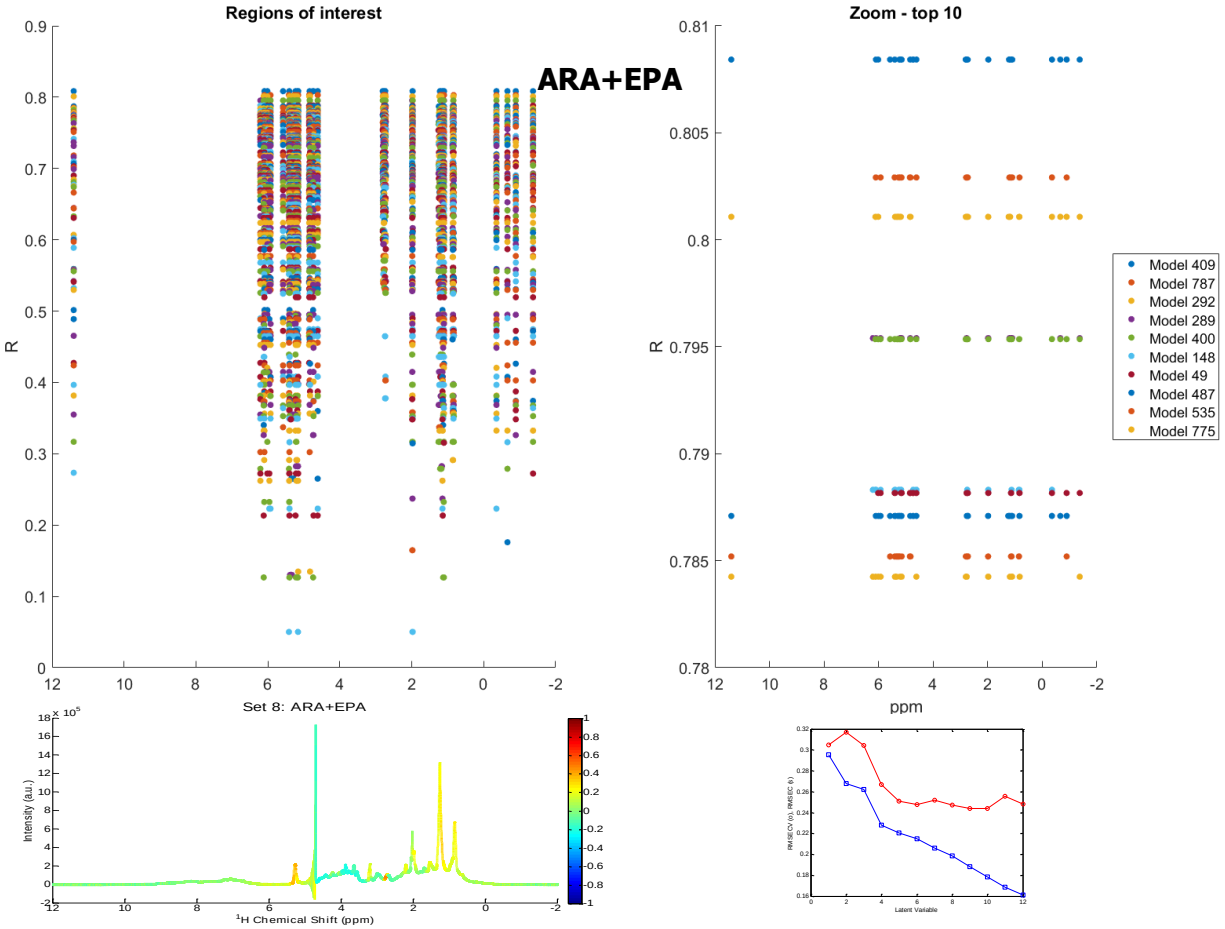

LA
Regions of interest
R
ppm
Zoom - top 10
Model 823
Model 544
Model 673
Model 109
Model 22
Model 589
Model 751
Model 856
Model 301
Model 43
Set 1 + Set 6: LA
Intensity (a.u.)
^{1}H Chemical Shift (ppm)
RMSECV (&, RMSEC (x)
Latent Variable


SFA
Regions of interest
R
ppm
Zoom - top 10
Model 754
Model 613
Model 154
Model 418
Model 433
Model 94
Model 259
Model 526
Model 394
Model 184
Set 8: SFA
Intensity (a.u.)
¹H Chemical Shift (ppm)
Latent Variable
W6+W7
Regions of interest
Zoom - top 10
Model 592
Model 715
Model 178
Model 43
Model 190
Model 175
Model 799
Model 109
Model 532
Model 472
Set 6: W6+W7
Intensity (a.u.)
¹H Chemical Shift (ppm)
Latent Variable


W9
Regions of interest
R
ppm
Zoom - top 10
Model 493
Model 541
Model 100
Model 304
Model 760
Model 799
Model 559
Model 772
Model 367
Model 331
Set 6: W9
Intensity (a.u.)
¹H Chemical Shift (ppm)
RMSECV (0), RMSEC (o)
Latent Variable


W3
Regions of interest
R
ppm
Zoom - top 10
Model 883
Model 484
Model 631
Model 538
Model 559
Model 673
Model 205
Model 439
Model 712
Model 853
Set 6: W3
Intensity (a.u.)
¹H Chemical Shift (ppm)
RMSECV (o), RMSEC (o)
Latent Variable


DHA
Regions of interest
R
ppm
Zoom - top 10
Model 454
Model 703
Model 739
Model 529
Model 70
Model 235
Model 781
Model 304
Model 601
Model 184
Set 8: DHA
Intensity (a.u.)
^{1}H Chemical Shift (ppm)
Latent Variable


ARA+EPA
Regions of interest
R
ppm
Zoom - top 10
Model 409
Model 787
Model 292
Model 289
Model 400
Model 148
Model 49
Model 487
Model 535
Model 775
Set 8: ARA+EPA
Intensity (a.u.)
^{1}H Chemical Shift (ppm)
Latent Variable