Universidade do Minho
Escola de Ciências

Vasco Rafael Rocha dos Santos

Lipidomic Profiler using NMR-phenotypic traits

outubro de 2022

Lipidomic Profiler using NMR-phenotypic traits

Vasco Santos

UMinho | 2022

Universidade do Minho
Escola de Ciências

Vasco Rafael Rocha dos Santos

Lipidomic Profiler using NMR-phenotypic traits

Dissertação de Mestrado
Biofísica e Bionanossistemas

Trabalho efetuado sob a orientação do(a)
Doutor Juan Gallo Paramo
Professor Doutor Pedro Miguel Amadeu Costa Santos

outubro de 2022

COPYRIGHT AND TERMS OF USE FOR THIRD PARTY WORK

Braga, October 31, 2022

Assinado por: **Vasco Rafael Rocha dos Santos**
Num. de Identificação: 15041470
Data: 2022.12.19 14:22:20+00'00'

(*Vasco Rafael Rocha dos Santos*)

ACKNOWLEDGEMENTS

Dear Weng Kung and Juan Gallo, if you are reading this, maybe is because you've cared for my words and you are willing to go even further with our sesquipedalian prose of scientific journals. Remember my naive questions and mistakes during this thesis, don't you? I used your answers for self-improvement and greatly finish this thesis, they will never be forgotten. Your high-bar standards, both at social and scientific level, made this work and future ones possible. To both, my sincere thanks.

In a more personal tone, Ana Gonçalves (wife), Diogo Félix, Naney (aka. Ney Setas), Tiago Pozo, Pedro Moleiro, Nuno David and João Tiago (the guitar man), and obviously my sponsors (my parents, not so much my sister), without you i wouldn't have the motivation to keep moving forward. Although achievements don't taste more than obligations, somehow you make me tirelessly always want more, and more. Kisses to all of you.

Including International Iberian Nanotechnology Laboratory, to everybody who felt useful during and for this work, may 'God' (no preferences here) be with you.

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity.

I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

Braga, October 31, 2022

Assinado por: **Vasco Rafael Rocha dos Santos**
Num. de Identificação: 15041470
Data: 2022.12.19 14:26:41+00'00'

(*Vasco Rafael Rocha dos Santos*)

RESUMO

As diferenças no ambiente molecular dos óleos vegetais (e.g., insaturação, ácidos gordos livres) induzem alterações nas medições de NMR. Manipulando um equipamento mais antiquado (i.e., NMR de domínio temporal) com ferramentas atuais (p.e., algoritmos de *machine learning)*, foi concebido o *Lipidomic Profiler*. Esta metodologia permite rastrear desvios fenotípicos de perfis lipídicos. Com base neste conceito, podemos identificar (em minutos) e classificar líquidos (i.e., óleos de palma, amendoim, azeitona (azeite), abacate, sésamo, girassol, milho) de forma não destrutiva. Mais pormenorizadamente, é demonstrado que, o *Lipidomic Profiler* proposto tem potencial na caracterização do perfil lipídico (p.e., quantidade de ácidos mono- e poli-insaturados) e na classificação do azeite pelo grau de acidez (p.e., extra-virgem, virgem ou refinado) e pela região de origem. A caracterização do perfil lipídico alcançou um bom nível de previsão na caracterização do teor de ácidos gordos mono- (R^2=0,86) e poli-insaturados (R^2=0,89). Além disso, na classificação do azeite por grau de acidez, o conceito prosposto (AUC=0,95) revelou-se mais sensível e preciso do que as metodologias atuais, como, a espectroscopia de infravermelho próximo (AUC=0,84) e a espetroscopia do UV-Visível (AUC=0,73), respectivamente. Devido às ferramentas utilizadas, tais metodologias podem fornecer futuras avaliações e classificações de amostras *in situ* (devido ao reduzido tamanho do equipamento de NMR no domínio temporal) não rotuladas num curto espaço de tempo.

ABSTRACT

The differences in molecular environment of the vegetable oils (e.g., unsaturation, free fatty acids) induce substantial changes in the time-domain NMR-phenotypic traits. Using an old-fashioned equipment (i.e., time-domain NMR) augmented with modern tools (i.e., machine learning models), was conceptualize a Lipidomic Profiler, a scientific tool for tracing down phenotypic deviation in lipid profiles. Using this concept, we can rapidly (in minutes) identify and classify (e.g., palm, peanut, olive, avocado, sesame, sunflower, corn) in label-free and non-destructive manner. In more detail, is demonstrated that the proposed Lipidomic Profiler, has potential in characterizing the lipid profile (i.e., amount of monounsaturated and polyunsaturated fatty acids), and classifying olive oil by its grading (e.g., extra-virgin, virgin or refined) and region of origin. Characterization of the lipid profile achieved an prediction level in the fatty acid content of monounsaturated (R^2=0.86) and polyunsaturated (R^2=0.89) species. In addition, in classifying olive oil by grade, the proposed Lipidomic Profiler (AUC=0.95) proved higher sensitive and specificity than the current gold-standards, i.e., near infrared spectroscopy (AUC=0.84) and ultraviolet-visible spectroscopy (AUC=0.73), respectively. Due to the tools used, such conceptual methodologies may provide future rapid assessments and object classification *in situ* (NMR point-of-use) of unlabelled samples with a short delay.

KEY WORDS fatty acids, lipid profile, NMR-based traits, time-domain NMR.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

Fraudulent food industry is one of the major public health concerns. Vegetable oils, for example, are one of the main targets due to their indispensable nutritional values (e.g., bioactive compounds) and attractive organoleptic properties. Therefore, the high-value authentic products (e.g., extra-virgin olive oils) are often blended with the counterfeit low-value oils (e.g., sunflower oils) [1]. For example, in Spain (1981), there was a report of an oil (e.g., contaminated rapeseed) fraudulently sold with olive oil who affected 20,000 individuals, killing around 1,000 (i.e., toxic oil syndrome [2]). Although regulation became tighter, olive oil is still the one of the most notified product in the EU [3]. Thus, in short-terms, adulteration serves to capitalise on consumer and boost profits. In part, this is due to the raising support for 'natural' products (i.e., unprocessed food), which uses only the highest quality materials, otherwise it may deteriorate during storage [4]. For instance, conventional food processing techniques (e.g., refining) have lost popularity to cold-pressed oils since there is no contact between cold press oil and chemicals [5].

The complexity of adulteration detection increases with blended oils mixtures, whose biochemical properties may mimic the one of an authentic product [6]. As consequence, some of the fraudulent products are untraceable. For example, as much as 82% of avocado oils distributed in the US is reported to be either adulterated or expired [7]. Thus, the search for newer adulteration 'detectors' with similar or better properties (i.e., time to results, user-friendly, price per assay) than the current gold-standard techniques (spectroscopy [8–12]), or gel- or high-performance liquid-chromatographies [13, 14], is fundamental in a growing and technological-adapted society. For this purpose, time-domain NMR measures augmented with machine learning models (i.e., Lipidomic Profiler) are tested in vegetable oils characterization (i.e., detailed content of the lipid profile) and classification (i.e., olive oil grade).

1.1 Objectives

Time-domain Nuclear Magnetic Resonance (NMR) (i.e., low-field NMR or relaxometry) is a powerful method to study interactions occurring in biological systems. The primary purpose is tracking down relaxation times (T_1, T_2). It provides spectral, one- or two-dimensional information (e.g., exponential, bi-exponential or *Laplace* raw data fitting) regarding the relaxation or diffusion properties of a sample. The measured signal 'contains' generalizable knowledge from molecular environment differences, [8, 15–18], being a highly unique molecular signature. With an increase in the dimensionality of information (i.e., various raw data fittings), the coupled machine learning algorithms were proven to improve accuracy and precision [15]. Based on this, the concept of a Lipidomic Profiler (Figure 1) was built to trace down phenotypic deviation in several vegetable oils (i.e., characterization, classification).

Vegetable plant oils are majorly constituted by fatty acids. Due to fatty acid species (e.g., saturated, unsaturated) combinations, there is a plethora lipid profiles. As examination, the Lipidomic Profiler is employed in characterizing (i.e., detailed content of saturated and unsaturated), and classifying olive oil by grade and region of origin. Using just a single droplet on a bench-sized time-domain NMR, the subtle differences in lipid profile and specific molecular environment of oils, are expected to induce subtle changes in the relaxation mechanism (i.e., NMR-phenotypic traits). Similar or higher prediction levels (with machine learning) than current gold-standard may develop this scientific tool (i.e., new adulteration 'detector') to futurely assist the food science community.

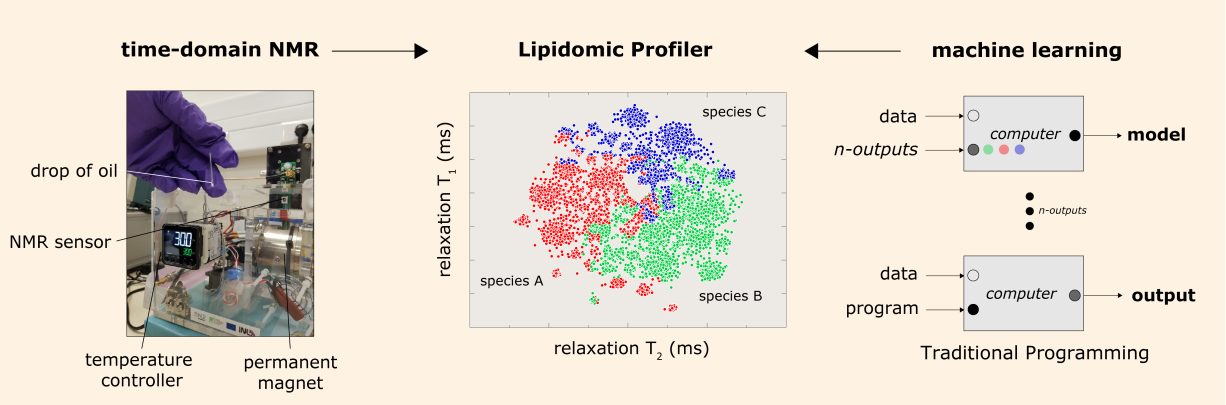

Figure 1: **Conceptual Lipidomic Profiler using NMR-phenotypic traits.** The concept of using Lipidomic Profiler as proposed in this work. The time-domain NMR device consists of a portable commercial console, a circuit coil and a palm-sized permanent magnet (magnetic field of 0.5T). For the analysis, a microcapillary tube with the sample (e.g., a drop of oil) is slotted into the NMR detection coil. The entire assay completes in less than 10 minutes. The measured NMR-based traits (e.g., T_1, T_2 relaxation times), augmented with machine learning models (to improve clear-cut classification and precise labelling) may provide scientific advances in the food science community [15, 19].

1.2 Work structure

This work is divided in order to approach less-knowledgeable readers (of NMR-based techniques) in a gradual manner. The following part ('State of the art'), encloses how the composition (i.e., lipid profile) of vegetable oils is synthesised and regulated in biological systems, with a few insights on it may change with storage and cooking. Subsequently, a chapter on NMR (more specific in time-domain NMR) wherein is described how the phenomenon is detected, and how it changes the relaxation times (T_1, T_2) in lipid profiles. The part is finished with a small chapter in machine learning.

The second part ('Experimental results') starts with a framework of analyses (i.e., resumed protocol of the work done) to combine both vegetable oils and NMR-based information (i.e., abbreviations, expressions). It is followed by the methods (e.g., sample preparation, NMR acquisition parameters, statistical analyses) and the whole experimental results, for both characterization and classification of vegetable oils. These results are evaluated and compared to the gold-standard techniques in the field (i.e., near infrared spectroscopy and ultraviolet-visible spectroscopy). In last, the part ('Discussion'), is used to resume the overall scientific discoveries and how to improve them futurely using the Lipidomic Profiler concept.

Part I

STATE OF THE ART

2

VEGETABLE OILS

Vegetable oils are mostly constituted by fatty acids (FA) (higher than 90%) and they represent one of the main sources of essential FA in human beings [20]. The term essential refers to the polyunsaturated FA's (PUFA) that are present in our diet yet they are not able to be synthesized in our body. PUFA are vital for the overall health being [21] (e.g., diminishing risks of cardiovascular diseases [22, 23], maintaining health of ageing individuals [24]). In addition to PUFA, vegetable oils are often characterized by the amount of saturated FA (SAFA), monounsaturated FA (MUFA), and 'trace' compounds (e.g., tocopherol, phenolic compounds, chlorophyll) [25]. Vegetable oils of different vegetable origin, mainly differ as consequence of their lipid profile (e.g., saturation levels, hydrocarbon chain properties).

In order to unveil the main causes of phenotypic variations (i.e., NMR-based traits) and where they may arise a description of FA biophysical properties, how they are synthesized (i.e., metabolic pathway and regulator factors) and their 'shelf-stability' (i.e., factors in oil oxidation) are presented below. The scope includes the main synthetic pathway (fatty acid synthase type II) for most of the vegetable oils in study. Classification of vegetable oils will be based on their dominant FA such as SAFA (palm), MUFA (peanut, olive and avocado) and PUFA (sesame, sunflower, corn). The lipid profile of the most used vegetable oils as inferred by the manufacturer displayed in Table 1.

Table 1: **Quantitative traits (lipid profile) of the main vegetable oil samples** (e.g., palm, peanut, olive, avocado, sesame, sunflower and corn). Grey color represent the predominant specie of FA on this organism, while 'bracketed' values, show the ratio of a FA versus the total lipid profile (FA) (all samples in the characterization study are in Appendix, Table A1).

Oil type	**Palm**	**Peanut**	**Olive**	**Avocado**	**Sesame**	**Sunflower**	**Corn**
Manufacturer brand	Guineas™	Fula™	Herdade do Esporão™	Graduva™	Emile Noël™	Fula™	Fula™
Energy (kJ/kcal)	3700 / 900	3374 / 821	3375 / 821	3397 / 826	3700 / 900	3397 / 826	3397 / 826
Fatty acids (g/mL)	100	90.7	92	92	100	92	91
SAFA	48	16	13.2	11	16	13	10
	(0.48)	(0.17)	(0.15)	(0.12)	(0.14)	(0.14)	(0.11)
MUFA	37	61	71.2	67	42	28	28
	(0.37)	(0.66)	(0.79)	(0.73)	(0.42)	(0.31)	(0.30)
PUFA	15	15	6.3	14	42	50	53
	(0.15)	(0.16)	(0.07)	(0.15)	(0.42)	(0.54)	(0.58)
Unsaturated	52	76	77.5	81	84	81	78
	(0.52)	(0.83)	(0.85)	(0.88)	(0.84)	(0.89)	(0.85)

2.1 Biophysical properties and effects of the lipid profile

The lipid profile traits of each vegetable oil type has an abundance of combinations due to the interconnected biophysical, and geometrical differences of each FA (and other trace compounds). In a bottom-up type approach, the main key points are the triacylglycerols species (e.g., glycerol bonded with three fatty acids), and FA variability (e.g., hydrocarbon chain length, number, position, and stereochemistry of *cis*-bonds (double bonds), see Table 2). This accounts for the free FAs content (FFA, FA decoupled from glycerol), a valuable parameter for accessing vegetable oil quality with known adverse health effects [26, 27]. However, their poor solubility in water in their undissociated form, due to reactiveness to potassium or sodium salts (i.e., agents of neutralization) makes their removal slightly easier.

In oil-phase, FA do not form ideal fluids but small domains of hydrogen-bonded layers somewhat similar to the molecular organization seen in their X-ray crystal structures [28, 29]. Generally, Van der Waals attractive forces are predominant in well-packed hydrocarbon chains. The degree of presence of these forces is proportional to viscosity (i.e., liquid friction). Consequently, an increase in average hydrocarbon chain length, leads to a increase of viscosity (η) and density of the oil [30]. In the other hand, disrupting the packing 'efficiency' (i.e., weakening of Van der Walls forces) will disrupt in the viscosity.

Thus, unsaturation level (i.e., number of *cis*-bonds), or small hydrocarbon chain length [30, 31], will both be proportional to fluidity ($1/\eta$) (see Figure 2). In a similar event, the decrease of the melting point, is somewhat correlated with the presence of branching or unsaturation. In addition, the melting point of a FA depends on whether the chain is even- or odd (generally higher melting point) numbered [32]. The effect of glycerol (in triacylglycerols), is the addition of density and rigidity (i.e., packing) to the lipid profile. For example, triolein (glycerol + 3 oleic acids) has higher viscosity than individual oleic acids [33].

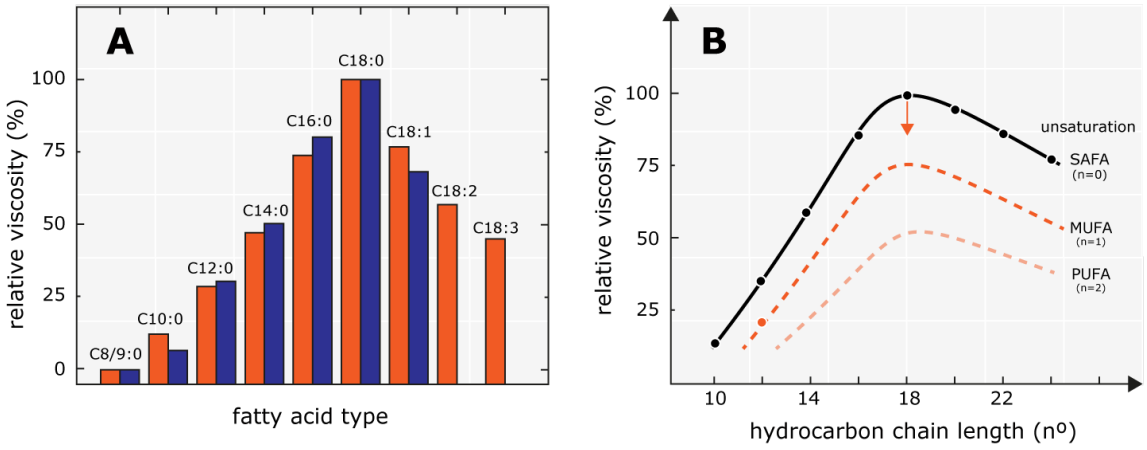

Figure 2: **Physical properties of individual fatty acids in the lipid profile.** (A) normalized relative kinematic viscosity from, orange [34, 35], blue [33] works; (B) the conceptual change in the viscosity of a lipid profile due to FAs intrinsic variability (e.g., hydrocarbon chain length (n°), double bond effect). Note that further unsaturation (n) will decrease the viscosity (orange arrow representing phenotypic variation to unsaturation). Maximum point, or highest viscosity, is conferred to stearic acid (18:0).

Beside the physical effects on the lipid profile, individual or complexed FA differ in their biological impact. Normally they are associated with storage and transport of energy, however they [32, 36] exhibit a hindered connection with signaling pathways and consequently, with cell metabolism. Complexed FA (e.g., phospholipids, triacylglycerols) are integral members of a cell system (e.g., cell lipid bilayer, storage of fat). Wherein, their effect is (generally) only adverse in surplus, for example, in hypertriglyceridaemia enhancing risk of pancreatitis [37].

In contrast, individual FA (in this case the FFA content), and especially long-chain SAFA, has been linked with unhealthy effects such as the mediation of insulin resistance, impaired glucose tolerance, and β-cell toxicity [26, 38, 39]. This worsens with *trans* isomers displaying high risks of ischemic heart disease [40], inhibition of the enzymatic desaturation of PUFA species and increasing low-density lipoproteins plasma concentration [41, 42]. In counterpart from SAFA, PUFA are vital (some essential) in reducing the risk of cardiovascular diseases [22, 23], and on diminishing ageing in individuals (ω-3,6 PUFAs) [24]. However, one of the concerns with PUFA is their ease of oxidation (due to high level of unsaturation). This facilitates the peroxidation of low-density lipoproteins that might be endocytosed by macrophages and initiate the development of atherosclerosis [43, 44]. Further adverse physiological or health effects are directly correlated with FA oxidation pathways products (i.e., effect of cooking, process of extraction).

Table 2: **Fatty acid variability in vegetable oils:** individual FA may be characterized by their unsaturation levels, lipid number (e.g., C:D, C being the number of carbons in the chain, D the number of double bonds), or (specially unsaturated FA) by the omega denotation (ω-x, double bond located on the x^{th} carbon–carbon bond, starting from the methyl end). The table presents the name of some of the most commons FA species in vegetable oils. The *cis*-bond position starts counting from the carboxylic group.

FA variability	Common name	Lipid number	cis-bond position
SAFA	lauric acid	12:0	-
	myristic acid	14:0	-
	palmitic acid	16:0	-
	stearic acid	18:0	-
	arachidic acid	20:0	-
MUFA	palmitoleic acid	16:1 ω-7	9
	oleic acid	18:1 ω-9	9
	gondoic acid	20:1 ω-9	11
	erucic acid	22:1 ω-9	13
PUFA	linoleic acid	18:2 ω-6	9,12
	eicosadienoic acid	20:2 ω-6	11,14
	linolenic acid	18:3 ω-3	9,12,15
	α-linolenic acid	20:3 ω-6	8,11,14

2.2 Fatty acids metabolic pathways

A pathway is a series of biochemical interactions via specific enzymes to produce a certain type of metabolite. In prokaryotic, plants and algae plastid, the synthases (FAS) is denominated as type II and consists of a multi-protein complex (Figure 3) [45]. On the other hand, FAS type I is found on the cytosol of eukaryotes being based on a large single multifunctional protein capable of catalyzing every step of the biosynthesis (these FAS's may work together) [46]. Focusing on FAS type II, our attention goes towards the knowledge of the predominant enzymes in the final percentage of unsaturated (e.g., MUFA and PUFA) and saturated FA (e.g., SAFA), in a lipid profile. In the main vegetable oil samples (Table 1), a variable percentage of FA are present. The major percentage of PUFA is found in corn oil (58%), MUFA in olive oil (79%) and SAFA in palm oil (48%). This approach is essential in order to constrain the length of the phenotype dimensions (protein-genes in focus).

It becomes relevant when defining a phenotype-genotype mapping. However, biological and environmental factors (e.g., up- and down-regulations of proteins) need to be further elucidated to correctly understand the key-changers in the final FA pool within a phenotypic change.

The first step of *de novo* FA biosynthesis starts with the same substrate, pyruvate from three different routes [47]. This compound is then converted to acetyl-CoA by pyruvate dehydrogenase (PD) and, from here, two different routes can be taken. In first route, acetyl-CoA carboxylase (ACC) catalyzes the irreversible carboxylation to form malonyl-CoA [48, 49]. Then malonyl-CoA:ACP transacylase (MCAT) transfer the malonyl group from previous molecule to holo-ACP (acyl carrier protein, a complete protein constituted only by an amino acid chain plus a prosthetic group) and malonyl-ACP is obtained [50]. The second route (coloured in black in Figure 3) consist in an association of acetyl-CoA and malonyl-ACP forming acetoacetyl-ACP, catalysed by KAS III [48]. It is important to refer that route two is dependent on the malonyl-ACP produced in route one. Acetoacetyl-ACP is the first element of the elongation cycle (i.e., addition of carbons to the chain) and is reduced to β-hydroxyacyl derivate by β-ketoacyl-ACP reductase (KAR) in NADH presence [48]. Afterwards, the final compound is dehydrated by β-hydroxyacyl-ACP dehydratase (DH) and reduced by enoyl-ACP reductase (ENR), through NADPH oxidation [51]. These two steps result in a four-carbon (butyryl) acyl-ACP [48]. This substrate is condensated with malonyl-ACP by KAS I and a new chain elongation cycle begin. The principal difference between KAS I and KAS III is that, the protein uses acetyl-CoA as substrate to react with malonyl-ACP, instead of using acyl-ACPs (e.g., butyryl). Later on (after five elongations) the final product (palmitoyl-ACP) is directly formed, by having KAS I a determinant impact in the condensation necessary to produce acyl-chains [48]. Palmitic acid is one of the most common SAFA, palmitoyl-ACP (C16) is one form of this acid found in cells [52]. This compound can be further elongated in a cycle, and a last condensation catalyzed by KAS II, produces stearoyl-ACP (C18). KAS II can use palmitoyl-ACP for condensation with malonyl-ACP [51].

Similar to palmitic acid, stearic acid (e.g., stearoyl-ACP) is one of most common long chain saturated FA founded in vegetable cells [53]. At this point, activity of desaturases introduce, if needed, double bonds into the FA (forming unsaturated FA). In plants, desaturases work via an aerobic mechanism giving rise to two water molecules (oxygen being reduced by four hydrogen) [54]. More concretely, stearoyl-ACP desaturase (SAD) in plastid stroma is responsible for the oleaoyl-ACP formation [55]. Further reactions occur by action of FA desaturase (FAD2) present in the endoplasmic reticulum [56]. For seed oils, FAD 2 is the main pathway to form PUFA's outside the plastid [54]. PUFAs, especially linoleic acid, are the most abundant FA in plants [51].

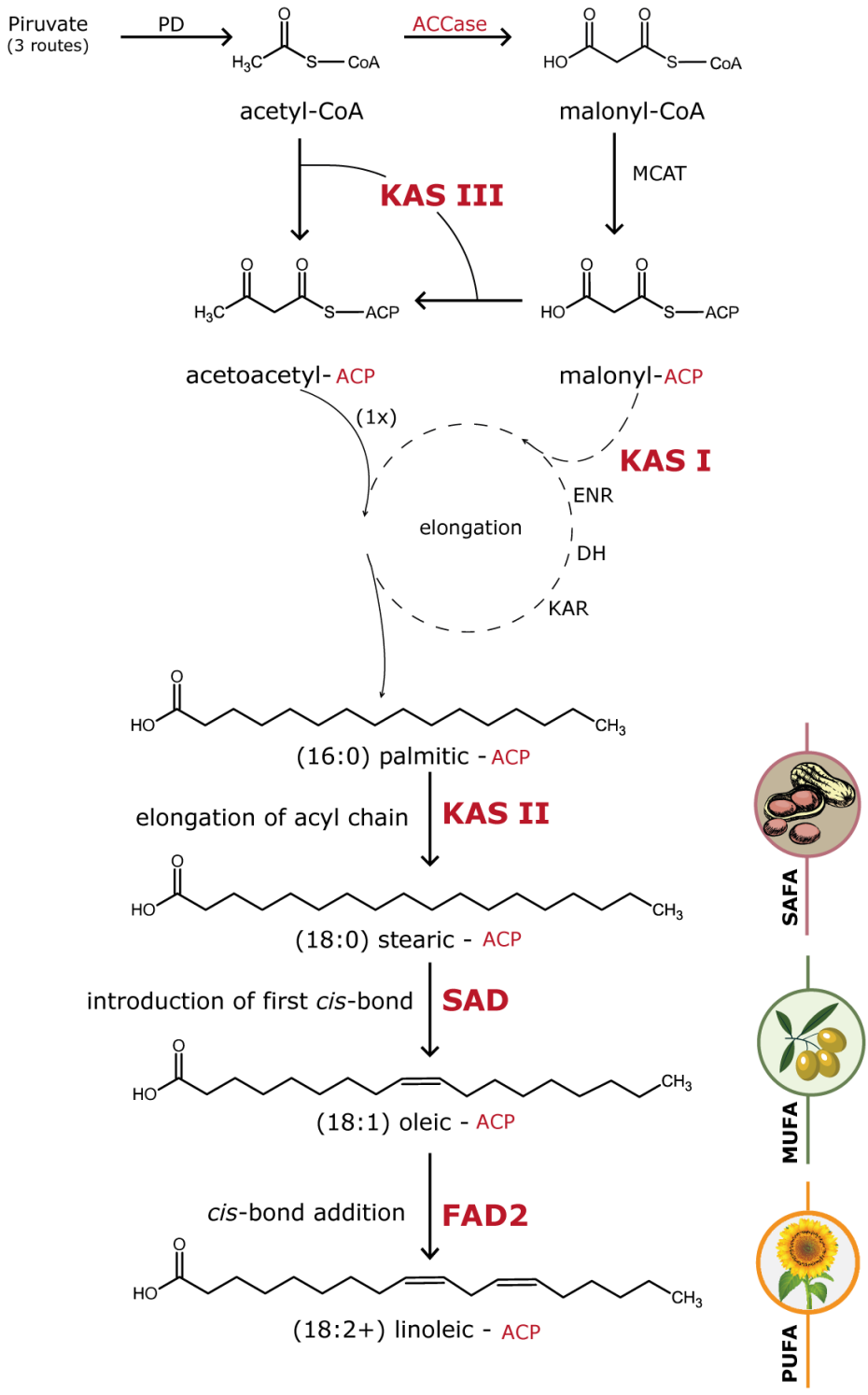

Figure 3: **Fatty acid biosynthesis pathway - Type 2.** Detailed description in the chapter 'Fatty acids metabolic pathways'.

2.3 Metabolic pathway regulatory factors

Although the crucial precursor for the *de novo* FA biosynthesis is acetyl-CoA, its synthesis is not directly correlated with changes in concentrations of FA. It becomes relevant for the formation of triacylglycerols (specificity of acyltransferases) and synthesis of malonyl-CoA [48, 57]. Also, proteins related with the elongation process, directly correlated with SAFA concentration and sixteen/eighteen-hydrocarbon chain length (C16/C18) ratio of FA, are totally sequenced but without high liability in these plants. Additionally, genome-wide surveys [58] and metabolic pathways were studied for the proper choice of predominant FA-regulatory networks. Within this, proteins such as SAD, ACC, FAD2 and KAS family, are distinctively described as potential main regulators wherein, some are described as fundamental to the up or down-regulation in the biosynthesis of FA [59].

ACC is a multifunctional biotin-dependent enzyme that catalyzes and regulates the obligatory the first step in FA biosynthesis in both bacteria and plants [60, 61]. It is the only known plant FA metabolism protein encoded by a plastid genome. In plants, the majority binding ion for ACC is manganese, but, in the other organisms this protein shows zinc as cofactor [57, 62].

Next in the pathway, the protein family KAS comes into play. It can be subdivided into three proteins with different functions (KAS I, KAS II, KAS III). Firstly appears KAS III, that catalyzes the formation of acetoacetyl-ACP. Previous articles suggest that this protein may have a rate-limiting role in the pool of FA [63, 64]. When KAS III is overexpressed it shows an increase in C16 chain FA's in spinach seeds [63], and in PUFA with the consequence of lowering the amount of MUFA in *B. napus* [64]. Moving onto the KAS I, not much information about his effects when up or down regulated exists. However, when this protein is mutated (i.e., presents deficiency) a significant change in the polar FA composition appears (due to agglomeration, and posterior degradation, to the chloroplast) [65]. This protein presents sensibility to cerulenin. Unlike the KAS III, butyrate is a good substrate for KAS I [66]. Lastly, KAS II is used for chain lengthening (final condensation) and controls the final C16/C18 ratio in a FA pool. Is distinguished from KAS I due to its sensitivity to arsenite [51].

Steps ahead in the FA metabolism we find SAD, a protein responsible for the formation of the first unsaturarion (i.e., SAFA to MUFA) in FAs metabolism [48]. Its down-regulation enhanced stearic acid content by 40% in *B. napus* (due to the lack of conversion of saturated to unsaturated FA) [58]. Finally, FAD2 is responsible for the addition of more *cis* bonds in FA (MUFA to PUFA). This protein is regulated by temperature, light, environmental and mechanical stresses. When temperature increase to 35°C the expression of gene decrease. FAD2 genes indicate a light-dependent transcriptional regulation, and PUFAs are signaling precursors molecules for the defense system and wound-healing [67].

Two conformations have been described, FAD2-1 and FAD2-2. FAD2-1 is found in a young seed, is involved in desaturation and, FA storage. One the other side, FAD2-2 proteins are present at the mesocarp and mature seeds being mostly responsible for reserve FA desaturation [48].

2.4 Factors in oil oxidation

Vegetable oils tend to present higher unsaturation levels than animal fats. Although unsaturated FA enhance a healthy dietary, they are more susceptible to spontaneous oxidation (i.e., deterioration) than their counterparts (i.e., SAFA), thereby, increasing the degree of formation of primary (e.g., hydroperoxides) and secondary oxidation (e.g., low molecular weight volatile compounds) products. For example, MUFA rich oils (e.g., olive) have greater oxidative stability than PUFA rich oils (e.g., corn) [68]. Besides oxidation, oils are deteriorated by hydrolysis and polymerization [69]. Hydrolysis increases the amount of FFAs, monoacylglycerols and diacylglycerols (i.e., by breaking FA bond to the glycerol). Polymerization occurs at higher temperatures (e.g., cooking) creating dimers and polymers.

Oxidation stability is the resistance to oxidation during processing and storage [70]. It can be expressed as the period of time necessary to attain a critical point of oxidation, establishing one important indicator for oil quality and shelf-life. Oil oxidizes via enzymatic, autoxidation or photosynthetic pathways (enhanced in the presence of metal ions and energy from heat/light) [70]. FA oxidation products from non-enzymatic sources may pose as health threat, reportedly being associated with an increasing risk of chronic diseases [71]. Non-enzymatic pathways have in common their trigger, oxygen. Both atmospheric triplet oxygen ($^3O_2$) and singlet oxygen ($^1O_2$) react with vegetable oils, starting autoxidation and photosensitized oxidation pathways, respectively (see Box 1, below). Note that autoxidation has a higher rate in the final products. Although the saturation level plays an major role in oxidation stability, vegetable oil processing, concentration of oxygen and other trace compounds of vegetable oils (e.g., antioxidants, chlorophyll's), exposure to energy of light and temperature, do play a minor role in the final stability.

In addition, common oil quality is monitored and compared with parameters such as: color and odor, FFA content (i.e., acidity or acid value), polar compounds (polymerization), peroxide (primary oxidation), p-anisidine (secondary oxidation) and iodine (unsaturation level) values. During this work, comparison in the detection classification of NMR-based traits with the current gold-standard techniques (e.g., NIR and UV-Vis) will be presented.

Box 1. **Autoxidation and Photosensitized oxidation mechanisms.**

Autoxidation. Also known as the free radical chain reaction, it is a three-phased event: a initiation, propagation and termination steps. (Initiation) FA or acylglycerols need to be in radical forms (i.e., by hydrogen removal) [70] due to the effect metal catalysis, ultraviolet, visible light, or processing effects. Double bonds are the most energetically favoured regions where it can happen. For example, to remove an hydrogen from C11 vs C8 and C14 is about 209 kj/mol vs 314 kj/mol, respectively. While the homolytic dissociation energy between hydrogen and C17 or C18 is about 418 kj/mol [72]. After removal, the double bond adjacent to the carbon radical shifts to a more stable next carbon and from the *cis* to the *trans* form. (Propagation and termination) The remaining FA alkyl radical easily reacts with $^3O_2$, resulting in FA peroxide radicals. This reaction occurs quickly at normal pressure, hence increasing the concentration of FA peroxide radicals over FA alkyl radicals. Continuously, FA peroxide radicals remove hydrogen from other FA alkyls to form hydroperoxides (i.e., propagation), completing the primary oxidation products. Finally radicals react with each other and nonradical species are produced (i.e., termination), to end the cycle. Additionally, the degree of formation of primary oxidation products is only dependent on the temperature and oxygen availability.

Photosensitized oxidation. Oil oxidation is accelerated by light in the presence of sensitizers such chlorophylls. Excited sensitizers (due to absorption of light energy) react with $^3O_2$ and produce superoxide anion. Superoxide produces hydrogen peroxide by spontaneous dismutation (redox reaction). Then, the reaction of hydrogen peroxide with superoxides results in singlet oxygen formation ($^1O_2$), which either reacts chemically or by transferring its energy to them. Due to its higher energy, it directly reacts with high-electron density regions (i.e., double bonds) without the formation of alkyl radicals, and forms hydroperoxides in these regions. When hydroperoxide is formed, *cis*-bond shift and *trans* FA occur, producing both conjugated and nonconjugated hydroperoxides. Production of nonconjugated hydroperoxides is not observed in the autoxidation.

3

NMR-BASED THEORY

Nuclear magnetic resonance (NMR) is a physical phenomenon based on the intrinsic properties of a nuclei. However, these intrinsic magnetic properties do not appear when the nuclei is spinless (i.e., zero nuclear spin, $I = 0$), therefore, NMR silent. 'By a quirk of fate' this includes ^{12}C and ^{16}O, some of the most common and abundant isotopes of organic substances. For all the other atoms, with one-half spins or quadrupolar nuclei (i.e., $I > 1/2$), detection is possible. In this thesis the focus is towards ^{1}H protons ($I = 1/2$) which, unlike other nuclei, presents an spherical shape with convenient magnetic properties [73].

A simple NMR measure starts by applying a magnetic field on a sample. Because nuclei spins behave like a compass needle, they will rotate and align with this external magnetic field in the same or in the opposite direction (minimizing the magnetic energy). At this stage, the energy levels and their population are no longer equal due to nuclear *Zeeman splitting*, wherein, a nuclear state is $(2I + 1)$-fold degenerated. Note that this splitting is far smaller than the thermal energy (i.e., ground/excited nuclear states) [73]. Afterwards, a second magnetic field, radio frequency pulse (tuned to the precessional frequency of target nuclei, *Larmor* frequency, ω_0) allows the nuclei to absorb and then emit electromagnetic energy (polarization of nuclei spins) [74]. The magnetic field experienced by each nuclei differs slightly from the applied field since it's *Larmor* frequency is affected by the chemical environment (i.e., electron shielding).

In frequency-domain NMR (i.e., high-field NMR or NMR spectroscopy), the spectrum is obtained as a function of the chemical shift (ppm) which is independent of specific experimental conditions yet dependent on the resonant and applied frequency. Therefore, the NMR signal is a peak with a defined amplitude and placed in a characteristic region of the spectrum. Wherein, the chemical structure of a molecule is elucidated.

On other hand, time-domain NMR (i.e., low-field NMR or relaxometry) lacks on molecular 'resolution'. Although this may appear as a problem in certain types of analyses, time-domain NMR detects generalized information about the mobility of the nuclei (e.g., longitudinal and transversal relaxation times) in a given environment (e.g., vegetable oils). The produced molecular signature is highly unique and rapidly obtained (in minutes) compared to high-field NMR which is much more expensive (i.e., magnetic fields required are much stronger) and time-consuming [15].

The acquisition of these relaxations times is mostly done by data manipulation (i.e., fitting data to a function). Relaxometry studies often process data by single-exponential fittings. However, bi-exponential or even inverse *Laplace* transformation algorithms may be applied to dimensionally up-scale the analysis.

3.1 Relaxation mechanisms and detection of NMR-based traits

In NMR, a signal is observed when the sample of interest is exposed to a magnetic field and the resonance condition is satisfied:

$$\omega_0 = \gamma B_0 \tag{1}$$

where γ is the gyromagnetic ratio, and B_0 the external applied magnetic field. The detected signal, free induction decay (FID) is measured as a decay in the time domain, which is then mathematically operated with *n*-exponential fittings to simplify the information obtained, *Laplace* transforms to unveil the relaxometry spectra, or with *Fourier* transforms to move back to the frequency-domain.

The interaction between the applied field, matter, and the radio frequency pulse are exploited through a sequence of pulses in order to maximize the signal-to-noise ratio of FID. The ensemble spin network (after applying B_0) has an average magnetization vector (i.e., bulk magnetization) due to stem degeneracy of the energy levels through Nuclear *Zeeman splitting*. In ^1H, this leads to two observable energy levels ($I = 1/2$), wherein, the probability of spin alignment is unequal but occupation is favourable for the lower energy state (as *Boltzman* equilibrium describes). The corresponding difference in the occupation of the two energy levels forms a finite state *z*-magnetization where the nuclei precess around this axis. Note that the equilibrium value and axis are arbitrarily chosen in the *z*-axis, thus precessing is oscillating in the *x-y* plane. However, to be rendered by NMR a transverse radio frequency pulse (i.e., tilting from equilibrium state) is irradiated in repeated pulse patterns based on the phenomena in study. The resonant phenomenon is therefore, caused by the external B_0 and the frequency matching irradiation leads to the rotation of the bulk magnetization by an angle (β), see Figure 4. Hence, the information retained from both spectroscopic or/and relaxometry studies is remarkably rich as the nucleus can be effectively used as extremely sensitive probes to their surrounding environment (e.g., detailing structural and dynamical information dependent on spin).

The bulk magnetization vectors in NMR can be described using the *Bloch* equations with the relaxation phenomenon (i.e., the process that drives the spins back to equilibrium).

There are two kinds of relaxation processes in low-field NMR: longitudinal (spin-lattice, T_1), occurring in the *z*-axis, and transverse (spin-spin, T_2) relaxation times occurring in the *x-y* plane. Both measure how much time (average) the nuclei takes to return to the equilibrium state (after irradiation). Developing *Bloch* equations with the relaxation phenomena terms (i.e., exponential decay) and before any irradiation, gives the following relationship:

$$M_x(t) = M_\perp \cos(\omega_0 t + \phi) \exp\left(-\frac{t}{T_2}\right) \tag{2}$$

$$M_y(t) = M_\perp \sin(\omega_0 t + \phi) \exp\left(-\frac{t}{T_2}\right) \tag{3}$$

$$M_z(t) = M_{eq} + (M_\parallel - M_{eq}) \exp\left(-\frac{t}{T_1}\right) \tag{4}$$

where, M_i is the magnetization component over the *i*-axis, $M_\parallel$ or $M_\perp$ are representative of the applied magnetization provided by B_0 under a relative direction, ϕ is the phase term, M_{eq} is the mean magnetization at equilibrium and *t* time. From an initial interpretation of these equations, is clear that transverse relaxation will be modulated (i.e., cosine under *x*, sine under *y*), oscillating at a characteristic *Larmor* frequency while decaying over the time. Consequently, this becomes the voltage oscillation detected (exponentially fitted) as the FID in a standard NMR experiment.

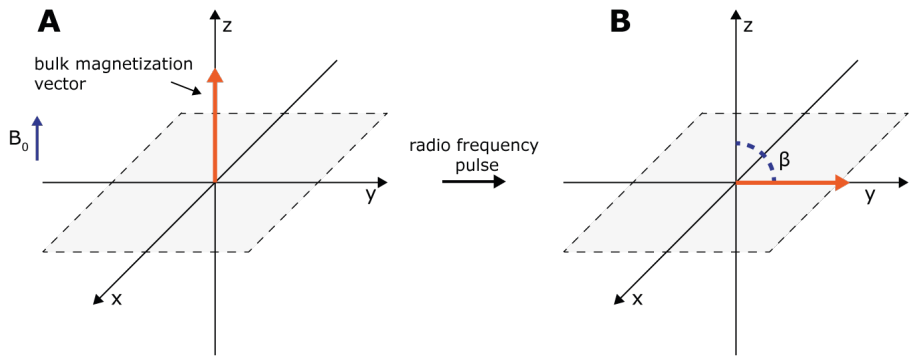

Figure 4: **Interaction between external magnetic field, irradiation and matter.** (A) Average bulk magnetization vector after applying B_0, thus representing the finite z-magnetization in the equilibrium state; (B) Bulk magnetization is flipped ($\beta = \pi/2$) with irradiation (i.e., radio frequency pulse). At the instant when the pulse is completed, the bulk magnetization vectors are: $M_z = 0$ and $M_y = M_{eq}$. After τ time, the M_z and M_{xy}, which, respectively, are parallel and perpendicular to the field B_0, progressively return towards their equilibrium states, (A) configuration: $M_{xy} = 0$, and $M_z = M_{eq}$.

Therefore, the study of the relaxation phenomena in NMR (i.e., low-field NMR or relaxometry) contemplates the behaviour of the T_1 and T_2 constants, which we denominate of NMR-based traits.

3.1.1 *Longitudinal relaxation - Inversion Recovery*

T_1 (i.e., longitudinal or spin-lattice relaxation) defines the process that describes how the spins lose their intrinsic energy (i.e., absorbed by the lattice) and restores the *Boltzmann* and z-magnetization equilibrium (after pulse is applied). Two routinely employed pulse sequences for estimating the behaviour of T_1 are: Inversion Recovery (IR) and Saturation Recovery. Both observe how the system evolves under a perturbation offset of the z-magnetization. The IR pulse sequence (details in Figure 5) was the standard procedure in this work.

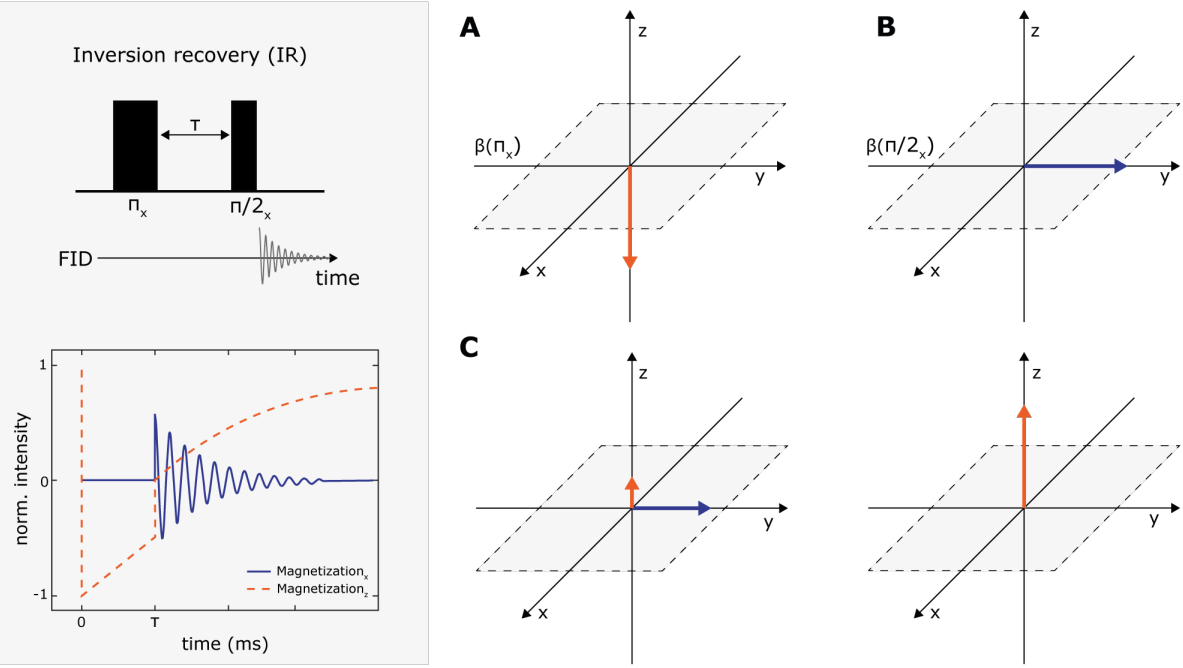

Figure 5: **Inversion recovery pulse sequence.** Detailed pulse sequence, FID signal, and inversion recovery fitting curve (left). (A) Pulse initializes by flipping ($\beta = \pi$) the bulk magnetization (from M_{eq} to $-M_{eq}$), which is allowed to evolve during τ time. (B) A second pulse ($\beta = \pi/2$) that produces a decaying oscillating signal from the relaxation phenomenon. (C) System return to equilibrium *z*-magnetization. Varying the τ, with estimation of how the magnetization recovers based within a time interval, ergo estimating the exponential behaved T_1. The standard protocol is to wait $5T_1$, since after oscillation decay the M_z component will continue to evolve (due to $T_1 \gg T_2$). Orange (M_z) and blue (M_{xy}) represent the bulk magnetization components.

3.1.2 *Transverse relaxation - CPMG*

The decay of the M_{xy} components to equilibrium is denominated as T_2 (transverse or spin-spin relaxation). It is the responsible for modulating the FID signal and may be measured by the Carr Purcell Meiboom and Gill (CPMG) pulse sequence. In summary, the pulse was originated from two works, Carr-Purcel [75] and Meiboom-Gil [76], which improved the pulse sequence by introducing a phase shift between the two first pulses. For estimation of the T_2 relaxation, its needed the acquisition of the echoes (e.g., FID repetitions) intensity points. In the expected decay signal (see Figure 6), and after the refocusing pulse, we observe the subsequent coherence leading to the formation of an echo.

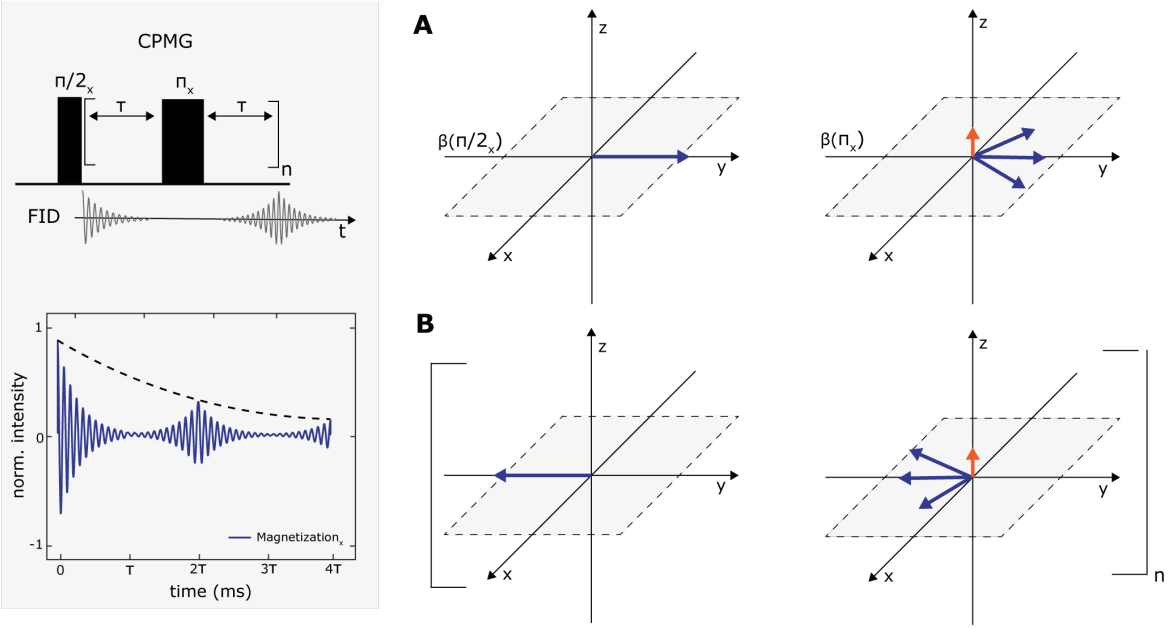

Figure 6: **CPMG pulse sequence.** Detailed pulse sequence, FID signal, and CPMG fitting curve (left). (A) A pulse ($\beta = \pi/2$) with a FID decay for τ times. (B) It is followed by a 'train' of $\beta = \pi$ pulses, which recovers magnetization up to the amount lost in the relaxation phenomena. Nonetheless, the spins of each region still retain precession and magnetization. Refocusing pulse is needed to reverse nuclei relative motion in a way that, and acummulation of phase differences, leads to a coherent state after τ time. Orange (M_z) and blue (M_{xy}) represent the bulk magnetization components.

3.2 Time-domain NMR phenotypic mechanisms in lipid profiles

As above-mentioned, lipid profiles display a remarkable variability in hydrocarbon chain composition, particularly in chain length and in the number, position, and stereochemistry (i.e., *trans* or *cis* conformations) of double bonds. These differences directly impact the geometrical properties and consequent molecular packing (i.e., Van der Walls interactions) of FA. [28, 31]. In overview, it's described how unsaturation level (i.e., level of *cis*-bonds), and hydrocarbon chain length vary at low-field NMR lens, and further weaken these interactions creating more degrees of freedom for molecular movements in the liquid (i.e., less friction, more mobility).

Due to the dense network of $1/2$ spins in the FA hydrocarbon, the dominant relaxation mechanism is through dipole–dipole (D-D) interactions [73]; it may be written as the following relationships [77]:

$$\frac{1}{T_1}^{D-D} = 2C\left[\left(\frac{\tau_c}{1+\omega^2\tau_c^2}\right)+\left(\frac{4\tau_c}{1+4\omega^2\tau_c^2}\right)\right] \tag{5}$$

$$\frac{1}{T_2}^{D-D} = C\left[3\tau_c+\left(\frac{2\tau_c}{1+\omega^2\tau_c^2}\right)+\left(\frac{2\tau_c}{1+4\omega^2\tau_c^2}\right)\right] \tag{6}$$

where C is a constant related to the rigid lattice second moment [78], τ_c is the rotational correlation time of a molecule in a environment, and ω the resonant frequency ($\omega/2\pi = \nu_0$). The constant C is unique for each FA and is acquired using the inter proton distances and angles as the main source of information [79]. It is important to mention that if protons are free to rotate the value of the constant will be small, or, if the intrinsic lattice (i.e., hydrocarbon chain) is mainly rigid, C will be larger. Studies show that with an increasing length hydrocarbon chain the lattice rigidity increases up to 18 carbons, decreasing afterwards [80]. For example, for palmitic acid (16:0) the value of C is $2.0 \times 10^{10} s^2$ [81], but for stearic acid its slightly higher (18:0) and then it slightly decreases at 20 carbons [79, 82] (note that: $C \propto 1/\mathrm{T}_i$).

Adding to this, τ_c may be generalized using the Stokes–Einstein–Debye relationship for rotational diffusion (D_r):

$$D_r = \frac{1}{\tau_c^i} = \frac{k_B T}{\eta V f_i} \tag{7}$$

in which, k_B is Boltzmann's constant, T is the absolute temperature, V the molecular volume, η is the absolute viscosity (i.e., fluid's internal flow resistance or friction), and f_i is the dimensionless constant related to the geometrical properties of the molecule.

Robinson et al. [31] discriminate that, in the extreme case where $T_2 = T_1$ and $(\omega\tau_c)^2 \gg 1$, considering the equations 6 and 7, leads to the following proportionality:

$$T_i \propto 1/\tau_c \propto D_r \propto 1/\eta \tag{8}$$

wherein, we hypothesize that our NMR-based traits (e.g., T_1 and T_2) should exhibit a close to linear relationship with fluidity (e.g., $1/\eta$). Moreover, and since vegetable oils are a mixture of triacylglycerols, in the narrow case where glycerol impact on the viscosity and individual FA molecular dynamics is 'constant', the final sample viscosity may be segmented into the summation of n viscosities mixtures from n-FA's types [33].

Interestingly, SAFA's kinematic viscosity, ν, (η/ρ, where ρ is density) and rigid lattice second moment display the same pattern (see Figure 2); ρ pattern is inverted [83]. An increasing length of the hydrocarbon chain (until 18 carbons) is proportional to a rise in viscosity, decreasing afterwards (number of carbons > 18). The addition, of unsaturation or branching, will lead to a disruption of the molecular packing of FA, thus to a decrease in viscosity (see Section 2.1).

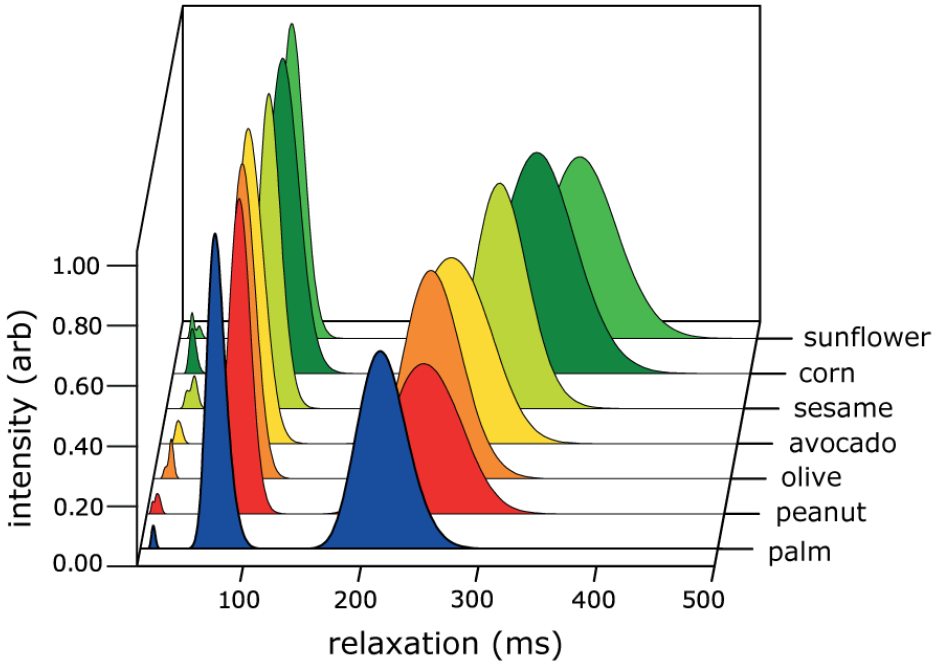

Figure 7: **Vegetable oil relaxometry spectra as resolved by inverse Laplace transform.** The main vegetable oil samples (see Table 1) evaluated under the T_2 relaxation time spectra. Raw data was fitted with a inverse *Laplace* transform.

In time-domain NMR, working under just one- or two-dimensional information may lead to equal phenotypes with different lipid profiles (i.e., vegetable oils viscosity phenotypic landscape is narrower than for individual FAs). Therefore, the generalized mobility information from nuclei (e.g., T_1, T_2 relaxation times and A-ratio (T_1/T_2), can be up-scaled by adding bi-exponential decay expression (e.g., T_{xa},T_{xb}, where x is either longitudinal or transversal relaxation) or even inverse Laplace transformation algorithms (e.g., peak intensity and position). Further prediction of the lipid landscape is proven to improve based on the dimensionality of NMR-based traits [16].

Reports employing low-field NMR-based traits for adulteration detection between vegetable oils [8, 17, 18] use time-domain traits, or personalized fingerprints varying with the pulse sequence parameters. In order to overcome this reductionist approach, spectral relaxometry (i.e., data fitted with inverse *Laplace* transformations) is already reported in assessing FA variability [31] (e.g., double bond position and stereochemistry) and the phenotypic impact. On the region of milliseconds, in pure FA samples, 2 or 3 peaks are resolved based on the clustering of hydrogen atoms in the FA molecule. Each peak varies (i.e., horizontal shift) proportionally with the lipid profile viscosity, having 2 domains displayed for SAFAs, or MUFAs when the double bond position is close to the carboxyl atoms, otherwise displaying three. Double bond *trans* conformation, comparatively with *cis*, displayed much faster relaxation mechanisms (i.e., higher viscosity, less mobility). For illustrative purposes, and seeing vegetable oils as a whole mixture of different FA, three peaks are well resolved using our time-domain NMR equipment (see Figure 7). This is in agreement with Robinson et al. [31] work from our NMR.

4

MACHINE LEARNING ALGORITHMS

The effectiveness of a ML augmented-work is directly linked to the nature, source and characteristics of the data used. However, we often overlook how the type (mostly) and the quality of the data influences the final results. Most of the time, ML studies follow a bottom-top approach, where, some criticism to the source of the data is frequently forgot. Not only should we have a sceptical approach from the start, but also adapt our ML algorithms to the type of data (increase the likelihood to extract insights or useful knowledge from the analysis). During this thesis, ML models will be evaluated using the Area Under the curve of the Receiver Operating Characteristics (AUC), a valuable metric in checking any classification model's performance. Wherein, the Receiver Operating Characteristics (ROC) is a graphical plot of true positive rate (i.e., sensitivity) versus the false positive rate (i.e., specificity) of the classifications.

Based on this, it is important to know and discuss the various types of data that currently exist [84, 85]: *unstructured* can be understood as the data that is randomly distributed or with no predefined format or organization. This type of information is much more harder to capture, process and analyze comparatively to the others due to the difficulty in data segmentation; *structured* has a data model which follows a standard order with most of the information labeled, being highly organized and easily accessed. These factors enhance processing and analysis efficiency/quickness [84]; *semi-structured* fits in structured data but does not conform with the formal structure of data models associated with data tables (e.g., relational databases). However, contain tags or other markers for separate semantic elements and inflict hierarchies of records and fields within the data [84]; *metadata* is a "data about data" and describes the relevant information being very useful for users. The main difference between "data " and "metadata" is that data can simply classify, measure, or even document something relative to an organization's data properties [85].

While the term taxonomy is often used in biology, it also may be applied for ML algorithms based on the desired outcome. We can divide them in four distinctive groups: supervised learning, unsupervised learning, semi-supervised and reinforcement learning [84, 86]. A detailed explanation of the supervised and unsupervised learnings (used in this work), the associated models/algorithms and applications is continued below.

4.1 Supervised learning

May be defined as a learning with a task-driven approach [84]. Here the output is directly connected to the input information with a mapping function constructed upon structured data. Thus, this data consists on a set of training examples. A supervised algorithm is often used in situations where we pretend to learn some kind of pattern from the training data set [87]. This teaching can provide predictions/regressions or classifications from test data sets [86].

To thoroughly understand this, three models, decision trees, Naïve Bayes classifier, and support vector machines will be presented:

- **Decision trees:** this algorithm displays an output, based on conditional statements, in a flowchart like model. Every node represents an attribute that needs to be classified, where each branch represents a value that the node can take. Applications of this type of classifier are common into random forest algorithms [88]. In this method, decision trees are ensembled in parallel (i.e., forming sub-trees), and uses the averaging voting from each tree for the final output. Thus minimizing the problem of over-fitting but increasing the prediction accuracy [89].

- **Naïve Bayes' classifier:** as explicit in the name, this model is based on the bayes theorem. It relies on two assumptions: it assumes that the output is conditionally independent given the class; it posits that latent or hidden attributes influence the prediction [90, 91]. Naïve Bayes is mostly applied to group and classify information. Comparatively to others more sophisticated algorithms, it needs lower amount of data to train effectively. Some variants to this classifier are the Gaussian and Bernoulli statistical distributions or even non parametric for calibration of data [84].

- **Support vector machines:** is a linear classification model that separates classes with the largest gaps (denominated as optimal margin) between a support vector (i.e., median border line between aggregates) [92]. However, this model can perform non-linear classifications using the kernel trick (mapping inputs into higher dimensions) and post-creation of hyperplanes. The margins between data are drawn to maximize distances, thus minimizing errors in classifications [91].

4.2 Unsupervised learning

As the name unsupervised suggest, this method of learning does not require a specific training set. These algorithms are specialized in dealing with unstructured data with the objective of exploring meaningful trends and correlations, being defined as an data-driven process [84, 93]. This is achieved because algorithms are left on their own devises to discover patterns with taskings such as: clustering, feature learning, dimensionality reduction and anomaly detectors [94].

Some well known groups of algorithms, based on their function, for unsupervised learning are described below:

- **Clustering:** the main function of a clustering method is to group and identify data. Having x number of defined clusters provides a collection of objects with similarity between each other. Examples of algorithms for this task are: K-means [95], Mean-shift [96] and Density-based spatial clustering of applications with noise (DBSCAN) [97]. They differ from each other on the cluster method. For example, K-means uses a partitioning method where the centroids of a group of data are calculated to be as far away from each other. This same method is applied on Mean-shift analysis, however, DBSCAN utilizes an density-based method where a cluster is defined as a contiguous region with high point density [84]. Points not belonging to this high density region are considered as noise. Consequently, partitioning methods are less sensible to outliers (affecting directly the median) comparatively to density approach's.

- **Dimensionality reduction:** for better interpretation of high-dimensional data, like which feature has a highest impact on the final results, two type of reduction algorithms come up: feature extraction (creation of new subsets) and feature selection (keeps the initial set, choosing the most relevant features). An example of feature extraction algorithm is the Principal component analysis (PCA). On the other hand, methods such as Analysis of variance (ANOVA), t-distributed stochastic neighbor embedding (T-SNE) and Chi-squared tests are samples of feature selection algorithms [84].

Part II

EXPERIMENTAL WORK

5

METHODOLOGY

5.1 Framework of the analyses

As aforesaid, vegetable oils lipid profiles variability induce detectable changes in the relaxation mechanisms (T_1 relaxation, T_2 relaxation). The concept of a Lipidomic Profiler, time-domain NMR-phenotypic traits augmented with machine learning models (see Figure 1), was tested in the characterization of an lipid profile of several vegetable oils (i.e., amount of saturated, monounsaturated and polyunsaturated FA), and classification of olive oils by their grading and region of origin. For both analyses, the NMR-based traits measured from vegetable oils were saved in a database format. According with the experiment, data was processed (e.g., addition of nutritional information) and trained by ML models. Due to this tools, we traced down the dominant NMR-phenotypic variation mechanisms in both characterization (interspecies) and classification (intraspecies) of oils. Further, and with the aid of machine learning, the sensitivity and specificity of our models were compared with current gold-standards (e.g., ultraviolet-visible (UV-Vis), near-infrared spectroscopy (NIRS) for oil quality.

In the characterization of vegetable oils, we trained the models to predict and map changes in the NMR-based traits (T_2, T_1) in various lipid profiles (e.g., palm, olive, avocado, peanut, sesame, corn, grapeseed, sunflower, linseed). Nutritional information (i.e., lipid profile) as inferred by manufacturer and average measurements displayed in Appendix, Table A1-A2, respectively. With it, database was created in pair with combinations of NMR-based traits. Wherein, we tested the precision and accuracy of single-phase systems (T_1, T_2, A-ratio), and/or, biphasic systems (addition of bi-exponential fittings, T_{xa}, T_{xb}). Then, the prediction ability of NMR-based traits was studied to predict monounsatured (MUFA) and polyunsaturated FA (PUFA) content (i.e., characterization).

For the classification of the grading and region of origin of olive oils (OO), measures were performed blindly (without manufacturer (e.g., region of origin or type disclosed). Our sample pool was composed of Portugal, Spain, Italy, Greece based OO, with different gradings (e.g., extra-virgin OO (EVOO), virgin OO (VOO), and refined OO). Manufacturer information and averaged measurements are in Appendix, Table A3-A4, respectively. Augmented with the ML models, we traced down the effect of free FA (i.e., acid value) in the lipid profile, being a key-changer in NMR-phenotypic deviation.

Specifically in identifying the region of origin, we employed, using AUC comparison matrix, a phylogenic tree to ensure cluster proximity with geographical distance. All of this work was replicated and compared with the gold-standards techniques (e.g., UV-VIS, NIRS).

5.2 Methods

Details and vegetable oils sample preparation. The characterization of the lipid profile (i.e., amount of MUFA and PUFAs) of vegetable oils (interspecies), and classification of olive oil by their grading and region of origin (intraspecies) had samples bought locally in Braga, Portugal or purchased online (e.g., international brands). The commercial (i.e., manufacturer) brands names for of all vegetable oil are disclosed in Appendix, Table A1, and specifically for olive oils, in Appendix, Table A3. No sample processing was made before the NMR measurements, and all other measurements.

UV-VIS and NIR measurements and detection. UV-Vis measurements were performed in a SHIMADZU UV-2550 spectrophotometer (Kyoto, Kyoto, Japan), while for NIR measurements a PerkinElmer LAMBDA 950 instrument was used. All samples were measured in matched 1 cm path length quartz, with a empty cell as a reference. UV-Vis spectra were measured within 200 to 800 nm spectral range at 1 nm spectral resolution, while NIR, spec-tra were obtained within 500 to 2200 nm with 5 nm steps. NIR spectra spike removal algorithms were applied (cut-off=6, threshold=10) [98]. Every sample was measured three times and the mean values were taken as representation.

Acid value measurements. The acid value, or the free FA content, was determined under the EN ISO 660:200940 [99] protocol for oleic acid quantification. Simply, 10 mL of vegetable oil were weighted and diluted in 20 mL of ethanol (ϕ = 99 %) with small amounts of phenolphthalein. Titrations with 0.1 mol/L of potassium hydroxide (KOH) were done under magnetic stirring until slight color changes appear (and persisted for +10s). Measures were executed twice per sample. The acid value was extrapolated from the amount of KOH required for each sample, defined as the amount of KOH required to neutralize one gram of chemical substance, with the following formula:

$$W_{AV} = \frac{56.1 \times cV}{m} \tag{9}$$

where, c is the exact concentration of the standard KOH solution (mol/L), V the volume of KOH added (mL), and m the mass (g) of the test portion. Acidity, or the free FA content, can be estimated by:

$$W_{FFA} = \frac{VcM}{10m} \approx 0.5 \times W_{AV} \tag{10}$$

wherein, M is the molar mass (g/mol) of the predominant FA in the vegetable oil, in this case oleic acid (282.47 g/mol).

NMR-based traits acquisition and parameters. The ^{1}H magnetic resonance measurements of olive oils were acquired at the average resonance frequency of 21.7579 MHz polarized using a portable permanent magnet (Metrolab Instruments, Switzerland), B_0=0.5T, using a benchtop-type console (Kea Magritek, New Zealand). A temperature controller was set to maintain the measurement chamber at 33°C. The T_1 relaxation and T_2 relaxation times were acquired using standard IR and CPMG train pulse sequences, respectively. The experimental parameters used were echo time (200), number of echoes (8,000) and signal averaging (64). A recycle delay of two seconds was set to provide sufficiently long time to allow all molecular spins to return to thermal equilibrium. (T_2 relaxation, T_1 relaxation) measurements were carried out on a wide variety of vegetable oils (e.g., palm, olive, avocado, peanut, sesame, corn, grapeseed, sunflower, linseed) and commercial EVOOs, VOOs and refined OOs. Averaged values for characterization of vegetable oils and identification of olive oil presented in Appendix, Table A2-A4, respectively. Clustering based-NMR methodology [16] uses a pair of relaxation times (T_2, T_1) for each object (each oil in this case).

Machine learning algorithm and workflows. Using statistical programming languages (e.g., Orange 3.1.2 or R), the raw datasets were processed using supervised and unsupervised learning techniques. The machine learning algorithms were written and run on a personal laptop (Intel Core Pentium i7 CPU @ 3.20 GHz, 16GB RAM). Once the model in machine learning was built, all the tasks run simultaneously and completed typically in less than 1 min.
Using unsupervised learning, the relationship between each object was rapidly constructed (e.g., hierarchical clustering) and its quantitative linkages shown on heat map and dendogram, or through evaluation metrics (e.g., Receiver Operating Characteristics (ROC) curve, Area Under the Curve (AUC) of ROC). Supervised learning models (AdaBoost, k-Nearest Neighbors (kNN), Linear Regression, Logistic Regression, Naïve Bayes, Neural Network, Stochastic Gradient Descent (SGD) and Random Forest) were used to train the datasets and the best model with the highest accuracy was chosen to predict the object classification and predictions (e.g., oil classification, or characterization) using pre-trained datasets.

Statistical analysis. A separation between two different vegetable oils was considered statistically significant when this criterion (P<0.5) is achieved or otherwise denote as non-significant (n.s). The Student's unpaired t-test was used throughout this study. One tailed were used, or otherwise mentioned in the figure captions. OriginLab - Pro 8 was used to handle all the graphs plotting. Pearson correlational coefficient (r) was calculated with the same approach.

Receiving Operating Characteristic. Analyses were used to evaluate the specificity and sensitivity of the diagnostic techniques. Various supervised models were used for the ROC tests. These were namely the (AdaBoost, k-Nearest Neighbors (kNN), Logistic Regression, Naïve Bayes, Neural Network, and Random Forest) models. A fitting of all ROC tests, was performed with a power function $y = ax^b$ for the classification of olive oils by its grading and region of origin. Iterations were run with the Levenberg–Marquardt algorithm until a chi-squared tolerance of 10^{-9} was achieved. Final function AUC was compared to the real averaged AUC from all assistive models (Appendix, Figure A3).

6

EXPERIMENTAL RESULTS

6.1 Characterization of the lipid profile

The NMR-measurements were carried out on nine different vegetable oil types (i.e., palm, olive, avocado, peanut, sesame, corn, grapeseed, sunflower, linseed) at 33°C. The vegetable oils can be classified based on their dominant FA content such as saturated (SAFA), monounsaturated (MUFA) and polyunsaturated (PUFA), details of the lipid profile as inferred by manufacturer in Figure 8A.

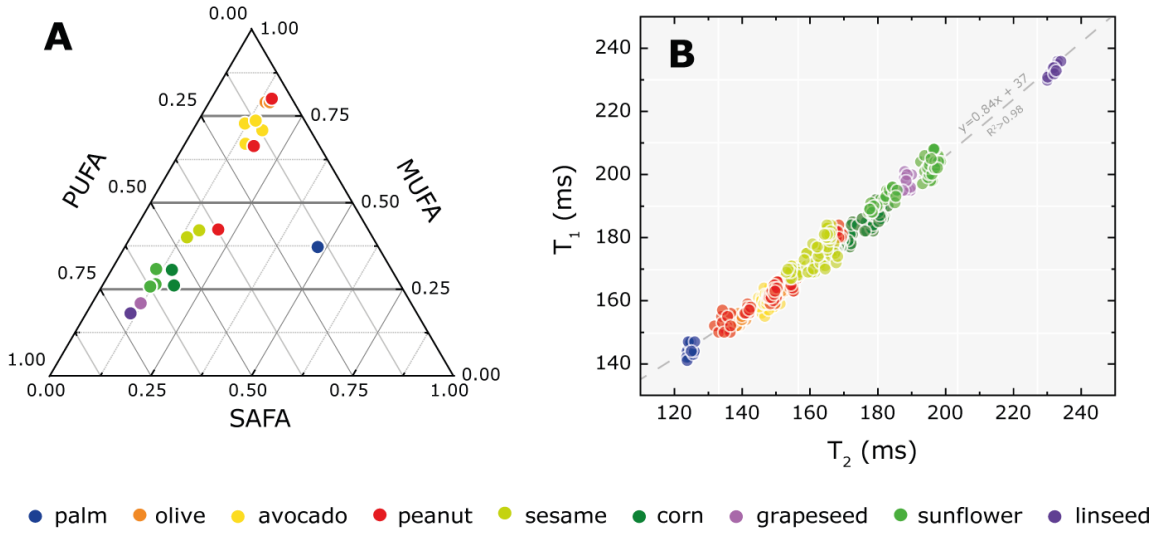

Figure 8: **Single-phase system for identification of vegetables oils.** Time-domain NMR measures were performed in 36 vegetable oils (palm, olive, avocado, peanut, sesame, corn, grapeseed, sunflower, linseed). In total, 504 pairs of relaxation (T_2, T_1) were collected. Single-phase refers to using single exponential decay fit. (A) Ternary plot of the lipid profile (e.g., SAFA, MUFA, and PUFA content) of the labelled (23 out of 36) vegetable oil samples (i.e., nutritional information). Note that their sum equals unity. Full nutritional information in Appendix, Table A1. (B) Two-dimensional plot of the relaxation pairs. Points can be linearly fitted ($y=0.84x+37$, $R^2>0.98$). One-dimensional plot of all the NMR-based traits in Appendix, Figure A1.

The mean values as evaluated by the single-phase relaxation pair (T_2, T_1), were for palm (124.7, 144.3) ms, olive (147.5, 161.5) ms, avocado (150.2ms, 163.2ms), peanut (151.7, 165.5) ms, sesame (161.8, 174.1) ms, corn (174.4, 182.5) ms, grapeseed (189.4, 196.5) ms, sunflower (198.4, 189.9) ms, and linseed (232.0, 232.2) ms (Figure 8B). With the Pearson correlation coefficients (r), we observed that with an increase in the PUFA content, the single-phase relaxation pair is enhanced towards slower relaxations (r = 0.81, 0.80), resulting in a decrease on A-ratio (r = -0.79). For example palm oil has an A-ratio of (1.16), while linseed (1.01). In overall, the clustering was efficient (i.e., good separation between interspecies) both on T_2, and T_1 dimensions (P<0.005, in Appendix, Figure A1A-C).

In time-domain NMR, working under just one- or two-dimensional (e.g., single-phase) leads to overlapping (i.e., equal phenotypes with different lipid profiles). Therefore, the generalized mobility information from nuclei was up-scaled by adding bi-exponential decay expression (e.g., T_{xa},T_{xb}). Thus, when the system is treated as a biphasic (only in T_2 dimension), one fast (T_{2a}) and one slow relaxation component (T_{2b}) are unfolded.

The mean values as evaluated by bi-exponential fitting, for the fast component were (69.0, 78.8, 79.9, 79.9, 82.5, 86.6, 89.6, 91.7 108.6) ms, while for the slow component were (220.0, 245.8, 258.6, 261.8, 278.7, 299.9, 339.2, 332.2, 437.3) ms for (palm, olive, avocado, peanut, sesame, corn, grapeseed, sunflower, linseed), respectively (Figure 9A-B). Similar patterns as the single-phase system were obtained with a correlational study (r>0.6). Both systems delineate a positive relation with unsaturation level (higher PUFA's content) and negative with MUFA.

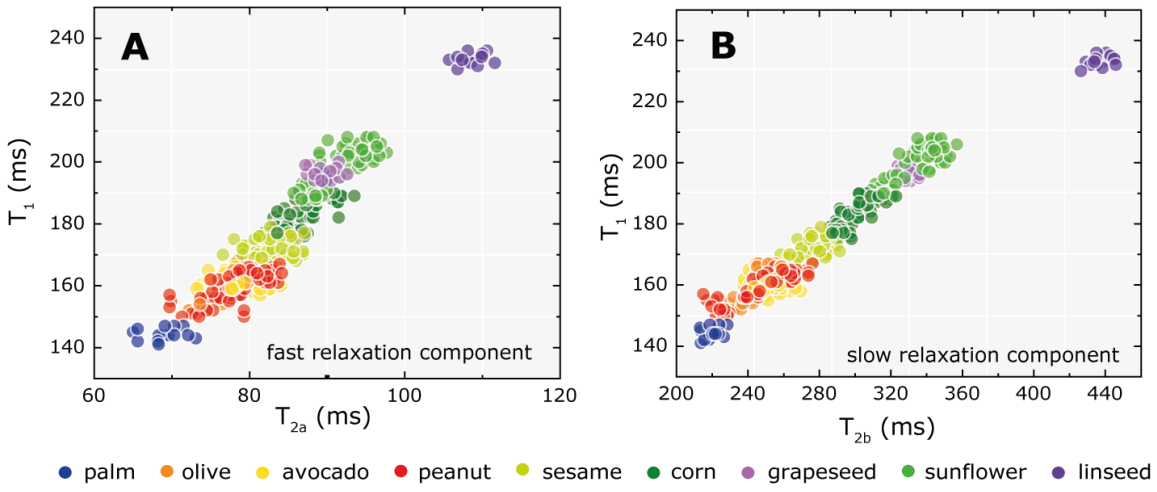

Figure 9: **Biphasic system for identification of vegetables oils.** Time-domain NMR measures in vegetable oils evaluated using bi-exponential decay fits (i.e., biphasic system). (A) Two-dimensional plot of the fast relaxation pair (T_{2a}, T_1), and (B) of the slow relaxation pair (T_{2b}, T_1). One-dimensional plot of all the NMR-based traits in Appendix, Figure A1.

The NMR-based traits from both systems (i.e., single-phase, biphasic), were further tested in vegetable oil identification (i.e., classification by specie) using Receiver Operating Characteristics (ROC) analysis (Figure 10). The single-phase system relaxation pair presented an AUC of (0.95) while the biphasic system (0.93). When combined (Figure 10B), the NMR-based traits achieve an ability to identify the vegetable oil of (0.96). Its observed that an increase in NMR-based traits dimensionality, enhances the precision and accuracy of the detection.

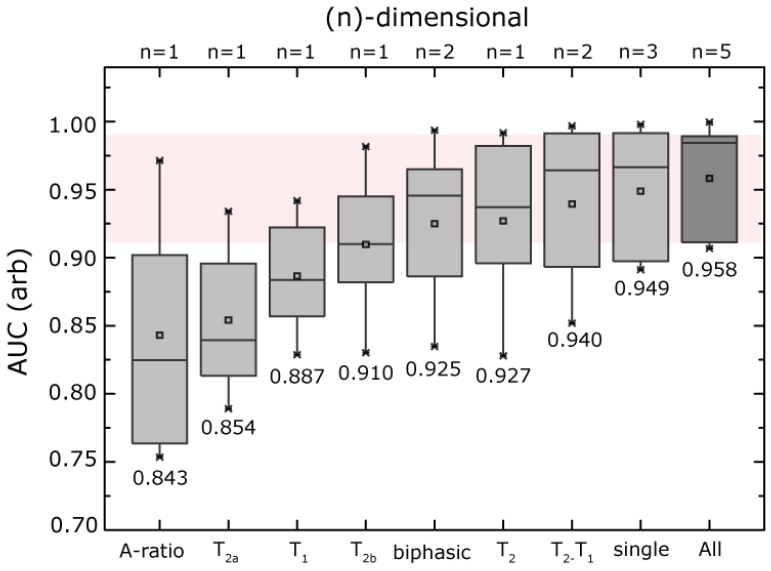

Figure 10: **Identification of vegetable oils using Receiver Operating Characteristics.** AUC plots evaluated by the ROC of various supervised models (i.e., kNN, Logistic Regression, Naïve Bayes, Neural Network, Random Forest) using various combination of *n*-dimensional phase parameters. One-dimensional plot of single-phase traits (T_1, T_2, A-ratio), biphasic traits (T_{2a}, T_{2b}) and their combination (All, in darker grey). The box plots represent 25% and 75% quantile of the measurements. ROC-AUC results for single-phased and biphasic systems in Appendix, Table A5-A6, respectively.

Due to the feasibility of the NMR-based traits (both from single-phase and biphasic systems), we moved forward to use them in the characterization of the lipid profile (i.e., MUFA and PUFA content). Predictions, as averaged by the supervised models (Appendix, Table A7), achieved an R^2(0.86, 0.89) for MUFA and PUFA, respectively. ML models were, for visual help, trained to predict the T_1 and T_2 phenotypic landscape of various lipid profiles (Figure 11). Thus, for fast characterization of blind samples, characterization of MUFA and PUFA content can be accessed by the relaxation pair (T_2, T_1) directly into the machine learning model, or searched within the prediction landscape (i.e., search algorithm). SAFA content is not as viable as its counterparts R^2(0.65), due to database bias towards higher unsaturation level samples (i.e., lack of mechanistic behaviour in process of model learning).

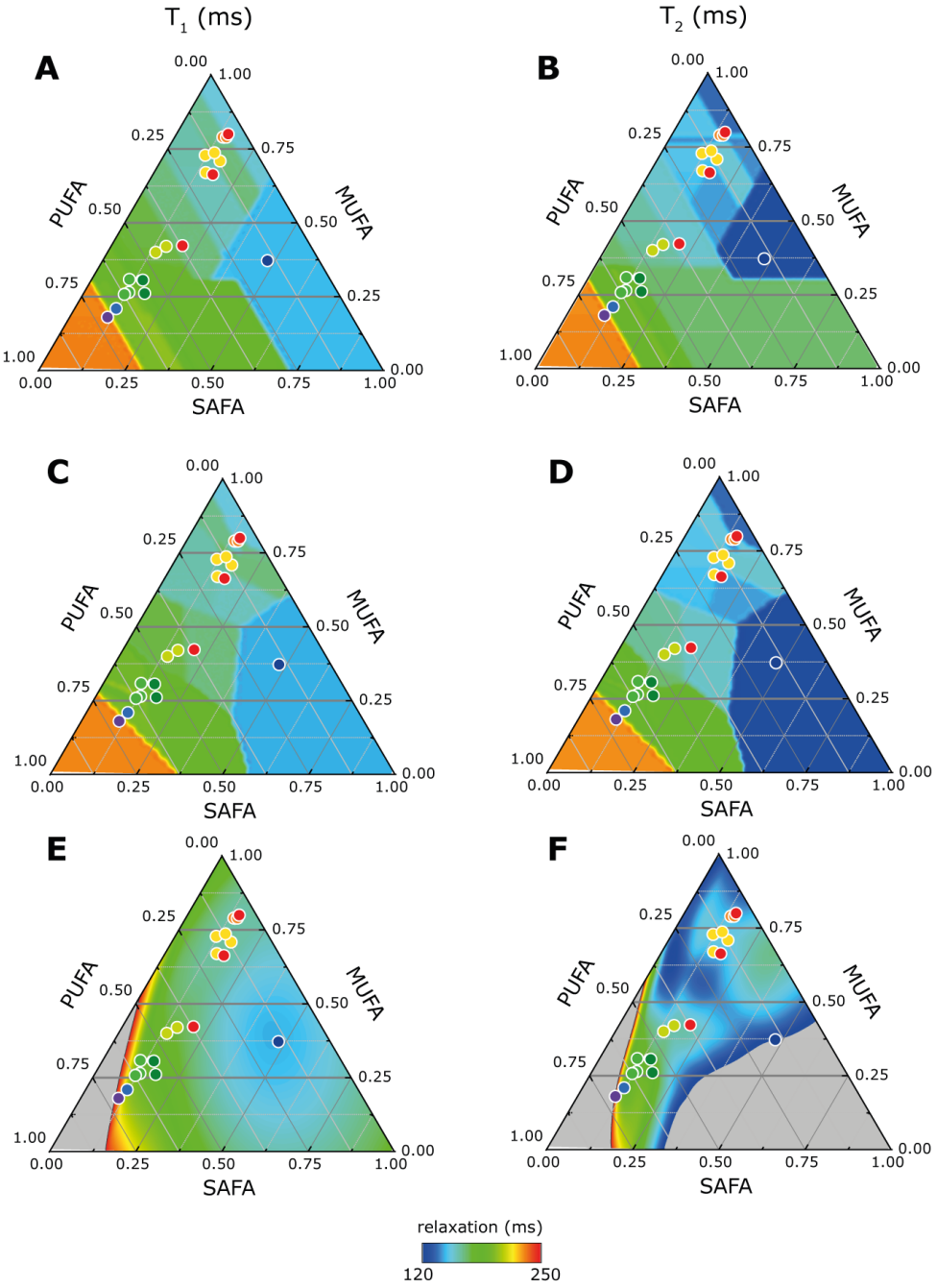

Figure 11: **Phenotypic landscape of the single-phase system in vegetables oils.** The NMR-phenotypic landscape in lipid profiles represented by ternary plots. Each point corresponds to exact ratios between SAFA, MUFA and PUFA that sum to unity. Color intensity is based on T_1 and T_2 predicted relaxation values from supervised models. (A, B) T_1, T_2 as predicted by AdaBoost, (C, D) kNN, (E, F) Neural Network. Evaluation metrics of the validated model in Appendix, Table A8. Each dot corresponds to the lipid profile of the labelled (23 out of 36) vegetable oil samples. Missing values (out of 120 to 250ms range) coloured in grey.

6.2 Classification of olive oils by grade and region of origin

In order to demonstrate the industrial applications of time-domain NMR, we use the proposed technique (i.e., Lipidomic Profiler) to validate the authenticity of EVOO from VOOs and refined OO (Figure 12). A wide variety of olive oils (i.e., 21 EVOOs, 8 VOOs, and 7 refined OOs) were purchased from different manufacturers off-the-shelf in Braga, Portugal, or through online platforms. The relaxometry measurements and acid value (i.e., free FA content) determination were performed on 36 types of OO without disclosing the manufacturers label and country of origin. For each sample, the relaxation measurements were carried out in double using 5 different samplings (i.e., 10 relaxation pairs per sample).

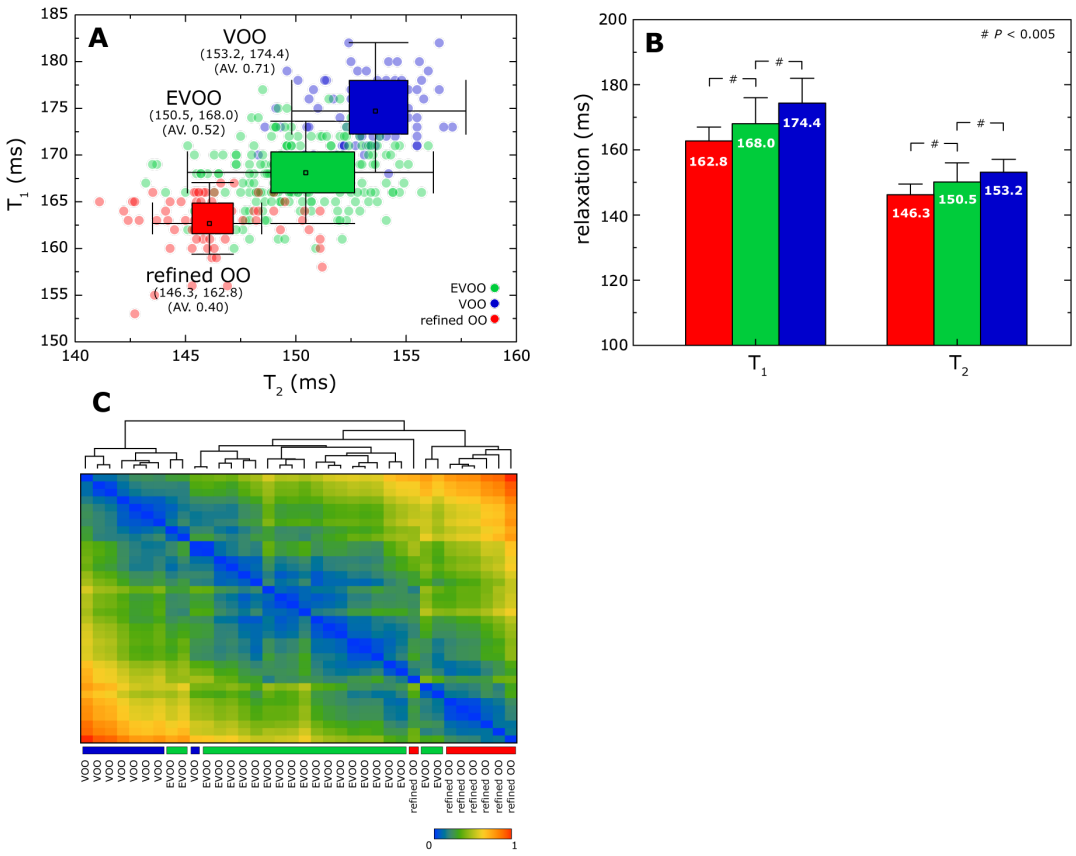

Figure 12: **Classification of olive oil grade using single-phase system.** Time-domain NMR measures were performed in 36 olive oils (360 relaxation pairs). (A) Two-dimensional plot of (21) EVOOs (green), (8) VOOs (blue), and (7) refined OO (red) by their relaxation pairs measures (T_2, T_1). The averaged relaxation pair and acid values (AV) were denoted below. (B) Average T_1 and T_2 for the different types of OOs. The statistical analysis of the data was calculated using unpaired two-tailed Student *t*-tests (P<0.005). (C) Classification of OOs using the single-phase traits in the form of clustering analysis. This hierarchical clustering was constructed based on the Euclidean distance between the averaged measures per sample. The color code (vertical axis) of each OO type is illustrated for eye-ball purposes. One-dimensional plot of NMR-based traits and AV measured in Appendix, Figure A2.

The mean single-phase relaxation pairs (T_2, T_1) generalize the composite intrinsic properties of the vegetable oils, thereof, forming a calibration standard for OO grading (EVOOs, VOOs, refined OOs), averaging (150.5, 168.0) ms, (153.2, 174.4) ms, and (146.3, 162.8) ms, respectively (Figure12A). There is a well-clustered effect (P<0.005) in NMR-based traits and in AV (Appendix, Figure A2), implying that the intra-variation were much smaller than the inter-variation of the OOs (Figure 12B). The details breakdown for each commercial brand is shown in heatmap (Figure 12C). The classification of olive oil by ROC analysis indicated that relaxometry measures have excellent detection sensitivity and specificity with AUC of (0.95) (Appendix, Table A9).

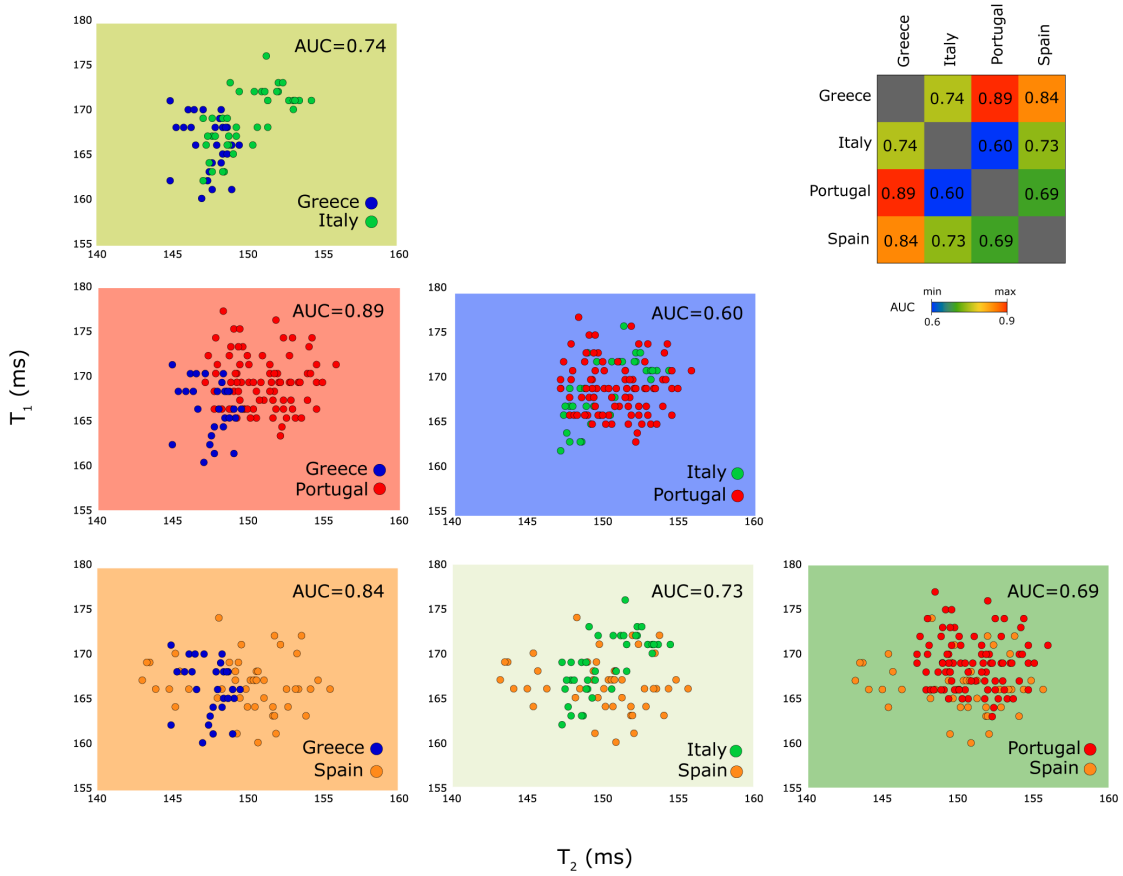

Figure 13: **Identification of olive oil region of origin using single-phase system.** EVOO samples were studied based on their regions of origin. Off-the-shelf EVOO samples from different European regions (i.e., Portugal (red), Spain (orange), Italy (green), Greece (blue)) according to their labelling. Pair-wise two-dimensional mapping of EVOOs origin as inferred by relaxation pairs (T_2, T_1). The sensitivity and specificity of each pair of regions were calculated using ROC analysis. The substantially high AUC, ranging from 0.6 to 0.9 of each pair-wise region were evaluated (AUC correlation matrix at top right). The models were validated using the Leave-one-out method using single-phase system (Appendix, Table A10)

We further proposed NMR-analysis in classification based on their regions of origin. The variation in phenotypic traits is now governed by number of factors, such as migration drift (e.g., diversification and domestication events [49]), and abiotic factors (e.g., local climate, soil factors [100, 101]) which have a direct effect of the lipid profile. For the identification of the regions of origin for OO, a matrix of data subsets (i.e., only EVOOs), encompasses four different regions (i.e., 3 Greece, 4 Italy, 9 Portugal, 5 Spain) that were plotted in pair-wise form using the single-phase system (Figure 13).

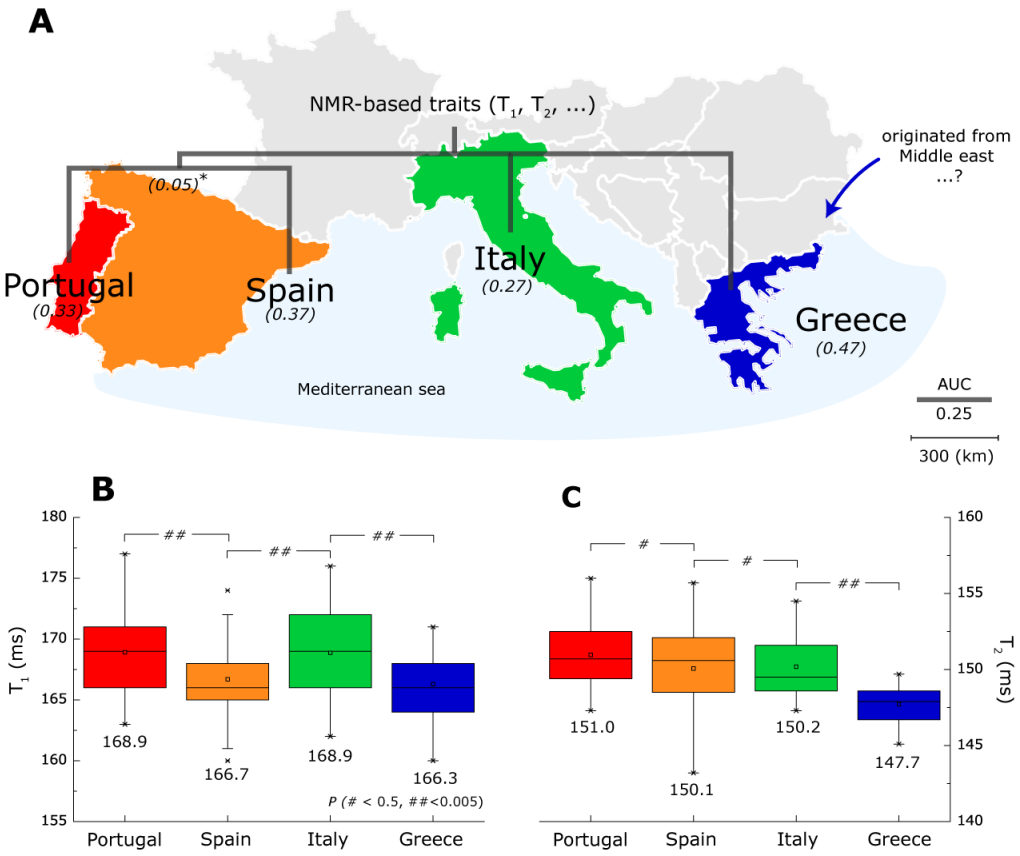

Figure 14: **Single-phase system in identification of the regions of origin.** (A) The NMR-based phylogenetic tree was built using the AUC comparison matrix (Figure 13) using neighbour joining algorithm which splits the NMR-based traits into three main regions (i.e., Iberian, Italy, Greece). The proposed NMR-based traits (legends of AUC is 0.25 in vertical) in agreement with their geographical orientation (shown in legend of 300 km per bar). Higher similarities are expected to be found species that are closely related. Neighbouring countries are expected to have higher species exchange and genes flow due to their geographical proximity. Similarities fade away with, for example, geographical distance. (B, C) One-dimensional plot of measured T_1, and T_2 for the different regions of origin of EVOOs. The statistical analysis of the data was calculated using unpaired two-tailed Student *t*-tests (P<0.5). Note that higher AUC means higher separability

At a pair-wise point of view, higher separability is achieve between (Portugal, Spain with Greece EVOOs), achieving AUC of (0.89, 0.84), respectively. This is to be expected as neighbouring countries are expected to have much higher of species exchange due to its proximity in geographical location. We confirmed this by applying the AUC comparison matrix (Figure 13 at top right) and employing an algorithm to construct a phylogenetic tree (Figure 14A). In addition, the mean relaxation pairs (T_2, T_1) were, for Portugal (151.0, 168.9) ms, Spain (150.1, 166.7) ms, Italy (150.2, 168.9) ms, Greece (147.7, 166.3) ms (Figure 14B-C). The overall regional-based identification for time-domain NMR is AUC of (0.71) (Appendix, Table A10).

6.3 Comparison of NMR-based traits with current gold-standards

In order to compare NMR-based traits precision and accuracy with gold-standard techniques (e.g., UV-VIS, NIRS), the OO analysis (i.e., see 'Classification of olive oils by grade and region of origin') was replicated and the Limit-of-detection (LOD) of all techniques was evaluated. ROC analysis (i.e., spectra peak with higher deviation versus single-phase system traits) was used as evaluation metric between techniques.

As previously done for NMR-based traits, a calibration standard for OO grading and region of origin was repeated with 12 random samples (3 readings) with UV-VIS and NIRS techniques (details in 'Methods'). In OO classification of grading (e.g., EVOO, VOO or refined OO) ROC analysis between the techniques indicated that time-domain NMR measures have higher accuracy and precision than gold-standards (Figure 15A). Averaging an AUC of (0.95), while NIRS (0.84) and UV-Vis (0.73) as evaluated by supervised models (Figure 15B-C). Full results in Appendix, Table A9. Equivalent prediction was obtained in ROC analysis for OO identification or origin (Figure 15D-E). Time-domain NMR AUC of (0.71), while NIRS and UV-Vis (0.69). Results in Appendix, Table A10. Nonetheless, clustering effect (P<0.5) was still well defined in gold-standard calibrations.

As final experiment, we evaluated the limit-of-detection (LOD) of NMR-based traits by mixing sunflower oil into a selected EVOO to mimic the cases of adulteration. For each sample, the relaxation measurements were conducted in double using five different samplings, covering from 0% (sunflower oil) to 100% of OO (control) in the mixed edible oil (Figure 16). As clearly indicated by the relaxation pairs, a linear relation (R^2=0.93) between NMR-based traits and the amount of sunflower oil (PUFA-rich oil) reduced into EVOOs (MUFA-rich) relaxation effect becomes clearer (due to a decrease in saturation level as seen in characterization of vegetable oils). Therefore, the averaged (T_2, T_1) relaxation pairs were (188.3, 202.9) ms and (155.3, 174.6) ms for sunflower oil and EVOO (control), respectively (Figure 16A-B). The LOD for single-phased traits were approximately (1%), which are comparable to NIRS (1%, Figure 16C-D) or much better than UV-Vis (5%, Figure 16E-F).

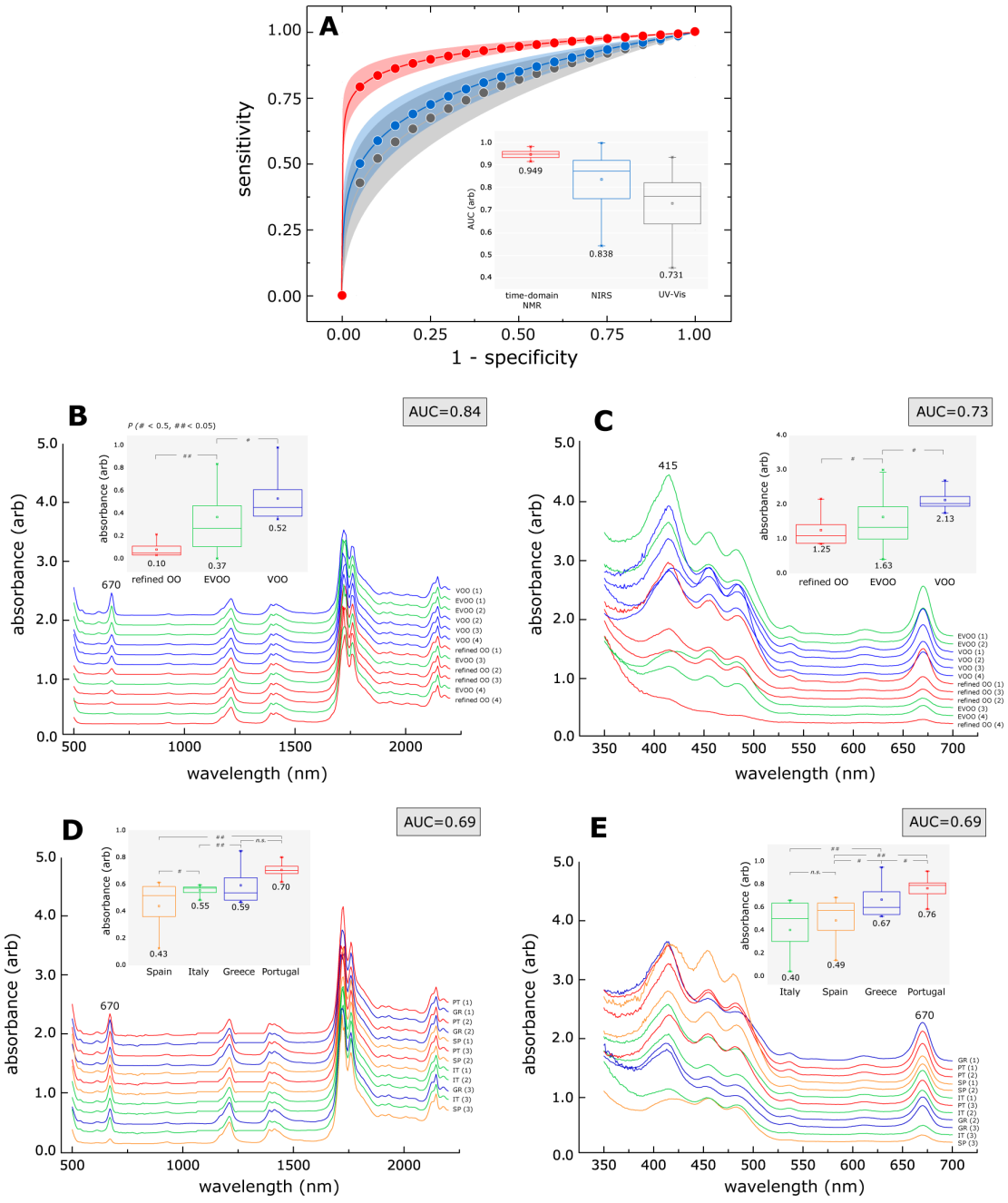

Figure 15: **Classification of olive oil grading and origin using gold-standard techniques (e.g., UV-Vis, NIR spectroscopy).** Classification of OO by its grade (e.g., EVOO, VOO, refined OO) and identification of region of origin (e.g., Portugal, Spain, Italy, Greece) against the gold-standard techniques. The color coding follows previous images. (A) The ROC curves for NMR-based traits (red), NIRS (blue) and UV-Vis (gray) calculated from a number of supervised models (Appendix, Figure A3). The fitting 99% confidence bands are displayed (see 'Methods'). (B) NIRS, and (C) UV-Vis spectra for classification by grade (e.g., refined OO (red), EVOO (green), VOO (blue)). (D) NIRS, and (E) UV-Vis spectra for identification of region of origin (e.g., Portugal (red), Spain (orange), Italy (green), Greece (blue)). Each experimental curve (i.e., NIRS or UV-Vis) represents the average of 3 measurements for each sample. The box plots represent the standard error of the median quantile of the entire measures. Sensitivity and specificity of each analysis (grey boxes) were calculated using the AUC of the ROC curve (Appendix, Table A9-A10). Peaks for ROC analysis were chosen on the most overlapped region.

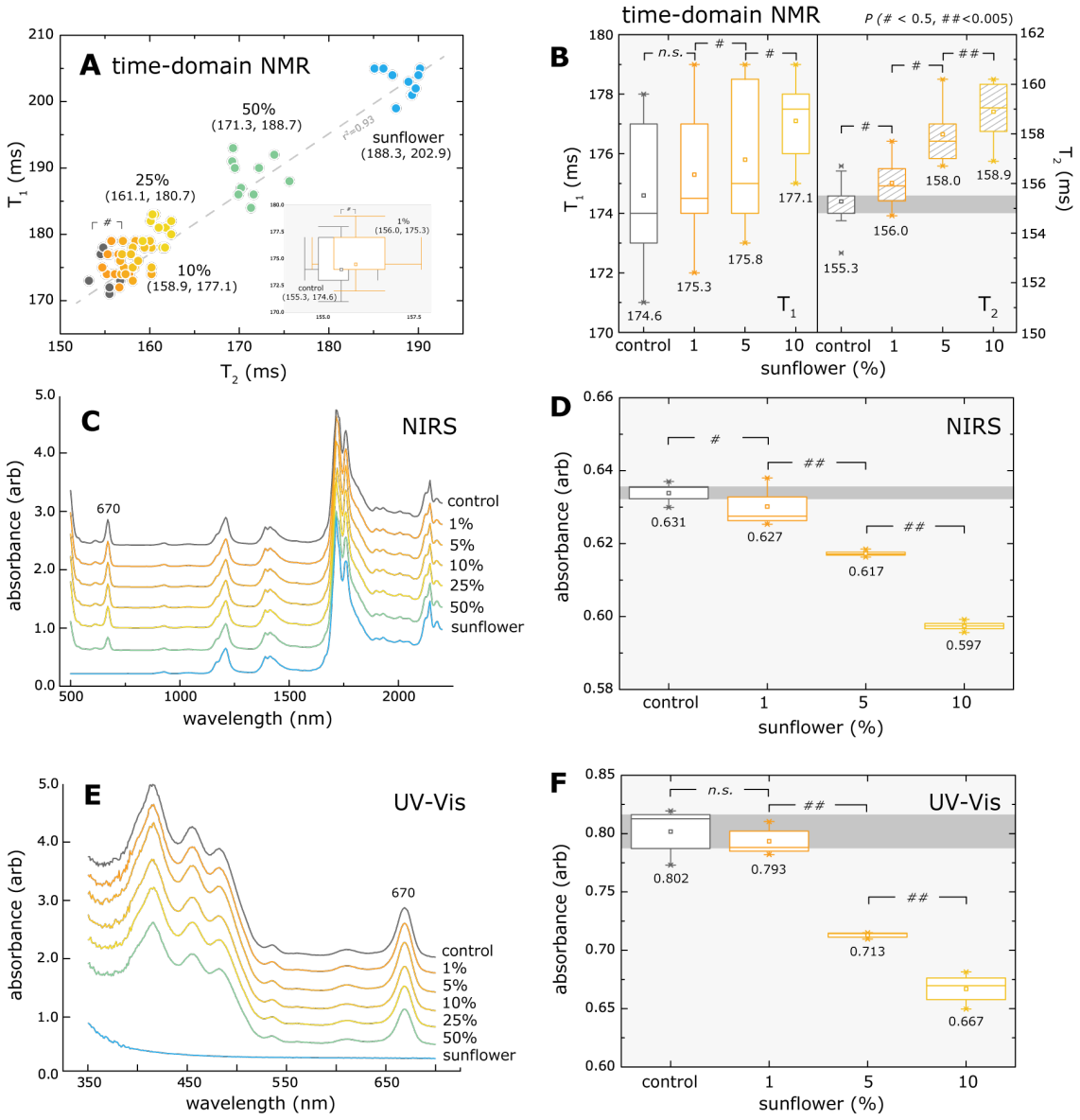

Figure 16: **Limit-of-detection of the time-domain NMR versus gold-standard (e.g., UV-Vis, NIR spectroscopy).** The EVOOs (as control) were mixed with sunflower oil in concentration of 1%, 5%, 10%, 25%, and 50% to mimic the cases of adulteration. (A) Two-dimensional relaxation pairs (T_2, T_1) of the EVOOs (grey) as a function of sunflower oils concentration (coloured). The mean relaxation pairs (10 per mixture) were denoted for each dilution. Data points were linearly fitted (R^2=0.93) with function y=0.85x+42.13. The zoom-in plot indicates the OOs (as control) and OOs with 1% adulteration, the box plots indicating 25% and 80% percentiles of the entire measurements. (B) The averaged T_1 (without strips, left) and T_2 (with strips, right) relaxations of the most overlapped region (e.g., control, 1%, 5% and 10%). (C) NIRS spectra taken from 500 nm to 2250 nm. (D) Multiple samplings were taken for each dilution. The maximum peak deviations were found to be at the 670 nm. (E) UV-Vis spectra taken from 250 nm to 700 nm. (F) Multiple samplings were taken for each dilution. UV-Vis- The most significant peak was at 670 nm. Each experimental curve (i.e., NIRS or UV-Vis) represents the average of 3 measurements for each sample. The box plots represent standard error of median quantile of the entire measurements. Two tailed Student's *t*-test was used to calculate the *P*-value. The LODs were 1%, 1%, and 5%, for time-domain NMR, NIRS and UV-Vis, respectively (see 'Methods').

Part III

DISCUSSION

7

LIPIDOMIC PROFILER

Is reported the employment of the methodology Lipidomic Profiler for fast, label-free and distinctive lipid profiling. This is essential for testing and reducing attempts of adulteration, assuring vegetable oil safety and quality. The NMR-phenotypic traits represent the intrinsic molecular relaxation dynamics due to the composite effect of the FA profile (e.g., unsaturation level) and/or the presence of FFA (e.g., acid value), which created the observed molecular environment differences.

Fatty acids in oil-phase consists of small domains of attractive forces (e.g., Van der Walls) predominant in well-packed hydrocarbon chains [31]. The degree of presence of these forces is proportional to viscosity (i.e., liquid friction) with a direct impact in NMR-based traits. Disrupting the packing 'efficiency' (i.e., weakening of Van der Walls forces) leads to a increase in degrees of freedom for molecular mobility (i.e., higher T_1 and T_2) [31]. Based in our results, this effect is clear in differentiating PUFA-rich species (e.g., sunflower, linseed, grapeseed) from MUFA-, SAFA-rich species. Similar mechanisms for packing disruption occurs in OO classification based on their grading (i.e., due to free FA content).

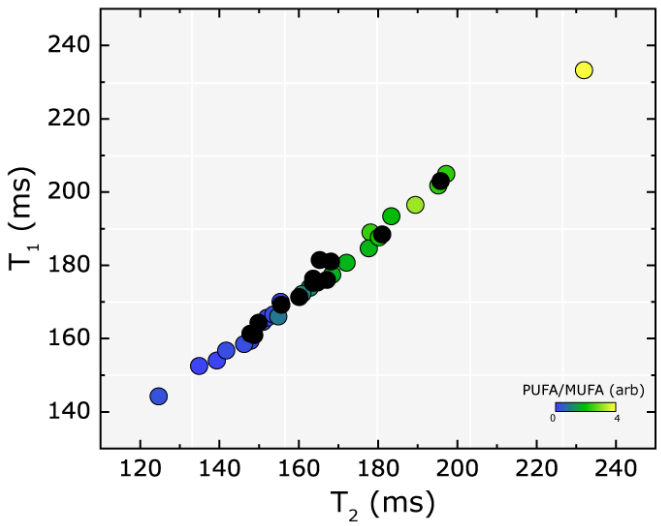

Figure 17: **Single-phase system as averaged for each vegetable oils.** The time-domain NMR measures as averaged per relaxation pairs (T_2, T_1) in 36 vegetable oils. Two-dimensional plot of relaxation pairs as averaged per sample. Color gradient resumes the PUFA/MUFA ratio value of the labelled (Appendix, Table A2) lipid profiles as disclosed by manufacturers (Appendix, Table A1), otherwise (i.e., not labelled samples) in black color. The Pearson correlation coefficient (r) between PUFA/MUFA ratio and relaxation pairs is (r>0.85).

In the characterization of the lipid profile, relaxation pairs indicate a major phenotypic deviation due to unsaturation level (i.e., amount of double bonds), more precisely the PUFA/MUFA ratio (Figure 17). For example, averaged relaxation pair (T_2, T_1) for palm oil (SAFA-rich) was (124.7, 144.3) ms, while linseed (PUFA-rich) obtained (232.0, 232.2) ms. Even though identification of vegetable oil type was 'successful' in validating NMR-based traits (i.e., prediction level of NMR-based traits increases with *n*-dimensions), clear-cut classification between vegetable oils should be more viable through bioactive compounds and not lipid profile changes (i.e., different oil may mimic lipid profile of another when looking at a macroscopic, SAFA, MUFA and PUFA, point of view). NMR-based traits (e.g., single-phase and biphasic systems) averaged an detection level with a R^2 of (0.86, 0.89) for the pair MUFA, PUFA with good accuracy (mean square error < 0.013) for all supervised models. SAFA detection was left out due to poor ML models evaluation metrics (R^2 of 0.65), in part by the natural bias for vegetable oils in MUFA-, PUFA-rich species (i.e., SAFA inter-variability is low). Thus, our approach lacks in the prediction ability in some regions of the lipid profile landscape.

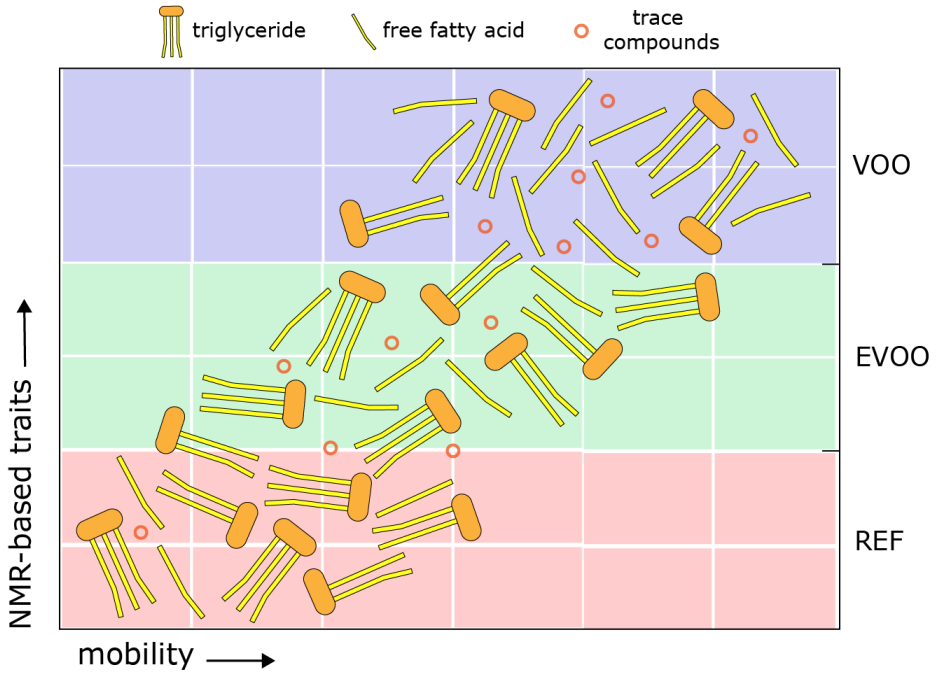

Figure 18: **Olive oil variation in NMR-based traits.** Molecular dynamics in vegetable oils are mapped by a function of the lipid profile (e.g., unsaturation level) and interactions with 'trace compounds' (e.g., tocopherols) and FFAs (e.g., acid value). Since refined OO display FFAs contents severely reduced when compared to its counterparts, VOOs and EVOOs, we hypothesize that an arise in FFAs is one of the mechanisms for packing disruption. Note: Content amount was chosen for illustration purposes and does not indicate reality.

In the classification of OO by grade, the process of making OOs has a direct impact on the FA profile, for instance, in the process of filtration, making of EVOOs (AV=0.52) and VOOs (AV=0.71), the AV of the oils is greater than for refined OO (AV=0.40) [102]. Since refined OO display FFAs contents severely reduced when compared to its counterparts with, VOOs and EVOOs, we observed and draw conclusion that a increment in FFAs is one of the mechanisms contribute to the packing disruption (Figure 18). Further support is made upon clustering pattern between AV, T_2 and T_1 relaxations(P<0.005). This is important since AV is one of the most important parameters related to the oil quality is, being key in grading OO [102–104]. In overall, performance of NMR-based traits were excellent, with a AUC of (0.96) in distinguish OO grade, when compared to NIRS (0.84) and UV-Vis (0.73) techniques. Yet, in classification of the region of origin NMR-based traits averaged an AUC of (0.71) and gold-standards (0.69), proving the versatility of the generalized information about the molecular environment (i.e., relaxation pairs). Further geographical clustering (i.e., neighbouring countries closer) was obtained by employing a AUC pair-wise comparison matrix has a distance matrix, however, due to undisclosed types of olive used in each vegetable oil, further conclusions can't be accomplished.

When compared with the current gold-standard techniques (spectroscopy [8–12]), similar or better properties were achieved (SWOT-like Table 3). The proposed NMR-based detection is cheaper per assay and user-friendly. Augmented by machine learning models, the concept of Lipidomic Profiler displayed high levels of accuracy and precision. The sensible and generalizable information presented within relaxation times proved to predict both qualitative (e.g., grade) and quantitative traits (e.g., lipid profile). Due to the tools used, validation of the vegetable oil (i.e., scientific cross-validation) lipid profile is the main goal for future market-implantation.

Table 3: **Qualitative performance of the Lipidomic Profiler against gold-standards (e.g., UV-Vis, NIRS).** SWOT-like analysis between the state-of-the-art technologies (e.g., Near-Infrared spectroscopy, UV-Visible) versus the Lipidomic Profiler proposed in this work (machine learning assisted time-domain NMR).

Features	**Lipidomic Profiler**	**NIRS, UV-Vis**
Sensitivity	very high	high/medium
Specificity	very high	high/medium
LOD	(1%)	(1%, 5%)
Extensive experience	not required	not required
Time to results	minutes	minutes
Sample processing	nil (no solvents needed)	nil, need specific solvents
Price per assay	ultra-cheap	expensive (cuvettes, solvents)
Equipment size	point-of-care testing	bench-top

8

CONCLUSION

In this thesis we have shown that the developed Lipidomic Profiler can detect the unsaturation level of vegetable oils, accurately grade OO, and can be useful to determine the region of origin of EVOOs (more work/samples is needed in this area). Thus proposed Lipidomic Profiler was extremely sensible to phenotypic deviations of lipid profiles, more specifically in detecting changes on the unsaturation level (i.e., MUFA/PUFA ratio) and free fatty content (i.e., acid value). Although information is not scientifically cross-validated, the precision of the results positions this concept as a powerful scientific tool. For market-implantation previous validations of manufacturer label would be crucial. Morevover, in future studies, the access of the lipid profile in more detail should be held with individual (e.g., pure FA samples) and complex mixtures (e.g., vegetable oils) with machine learning. Since prediction of the lipid profile landscape is improved with increasing NMR-based traits [16], its possible that inverse *Laplace* transform algorithm (i.e., relaxometry spectra) can provide other details based on FA variability. However, this transformation lacks the repeatability of exponential NMR-based traits (e.g., single-phase and biphasic systems) and the time per assay to obtain better spectra is largely increased.

Conventionally, chromatographic-based techniques, or high-field NMR, are time-consuming and require complicated sample preparation with expensive laboratory equipment. Some techniques is require complex data interpretation (i.e., chemometric studies), in comparison to the proposed NMR-based detection and other gold-standard technologies. With the introduction of EU Protected Designation of Origin registration and equivalents in other geographical locations, rapid classification of EVOOs, and vegetable oils in general, will be invaluable to industry and regulatory agencies alike. On the other hand and especially in grading the olive oils, the proposed Lipidomic Profiler (i.e., time-domain NMR-phenotypic traits augmented with machine learning models) provides rapid, precise, low-cost, label-free and accurate analysis. In addition, the introduction of machine learning, is now inexpensive to process large datasets running in almost real-time settings, opening the door to make predictions of unlabelled samples with much high sensitivity and specificity.

BIBLIOGRAPHY

[1] Abdelkhalek Oussama, Fatiha Elabadi, Stefan Platikanov, Fouzia Kzaiber, and Roma Tauler. Detection of olive oil adulteration using ft-ir spectroscopy and pls with variable importance of projection (vip) scores. *Journal of the American Oil Chemists' Society*, 89(10):1807–1812, 2012.

[2] Emilio Gelpí, Manuel Posada de la Paz, Benedetto Terracini, Ignacio Abaitua, Agustín Gómez de la Cámara, Edwin M Kilbourne, Carlos Lahoz, Bénoit Nemery, Rossanne M Philen, Luis Soldevilla, et al. The spanish toxic oil syndrome 20 years after its onset: a multidisciplinary review of scientific knowledge. *Environmental Health Perspectives*, 110(5):457–464, 2002.

[3] Enrico Casadei, Enrico Valli, Filippo Panni, James Donarski, Jordina Farrús Gubern, Paolo Lucci, Lanfranco Conte, Florence Lacoste, Alain Maquet, Paul Brereton, et al. Emerging trends in olive oil fraud and possible countermeasures. *Food Control*, 124:107902, 2021.

[4] Franz Ulberth and Manuela Buchgraber. Authenticity of fats and oils. *European Journal of Lipid Science and Technology*, 102(11):687–694, 2000.

[5] S Azadmard-Damirchi and M Torbati. Adulterations in some edible oils and fats and their detection methods. *Journal of food quality and hazards control*, 2(2):38–44, 2015.

[6] Kevin Lim, Kun Pan, Zhe Yu, and Rong Hui Xiao. Pattern recognition based on machine learning identifies oil adulteration and edible oil mixtures. *Nature communications*, 11(1):1–10, 2020.

[7] Hilary S Green and Selina C Wang. First report on quality and purity evaluations of avocado oil sold in the us. *Food Control*, 116:107328, 2020.

[8] S Ok. Detection of olive oil adulteration by low-field nmr relaxometry and uv-vis spectroscopy upon mixing olive oil with various edible oils. *Grasas Y Aceites*, 68(1):e173–e173, 2017.

[9] Georgia Fragaki, Apostolos Spyros, George Siragakis, Emmanuel Salivaras, and Photis Dais. Detection of extra virgin olive oil adulteration with lampante olive oil and refined olive oil using nuclear magnetic resonance spectroscopy and multivariate statistical analysis. *Journal of agricultural and food chemistry*, 53(8):2810–2816, 2005.

[10] Yuanpeng Li, Tao Fang, Siqi Zhu, Furong Huang, Zhenqiang Chen, and Yong Wang. Detection of olive oil adulteration with waste cooking oil via raman spectroscopy combined with ipls and sipls. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 189:37–43, 2018.

[11] Vincent Baeten, Marc Meurens, Maria T Morales, and Ramon Aparicio. Detection of virgin olive oil adulteration by fourier transform raman spectroscopy. *Journal of Agricultural and Food Chemistry*, 44(8):2225–2230, 1996.

[12] Francesca Guimet, Joan Ferré, and Ricard Boqué. Rapid detection of olive–pomace oil adulteration in extra virgin olive oils from the protected denomination of origin "siurana" using excitation–emission fluorescence spectroscopy and three-way methods of analysis. *Analytica Chimica Acta*, 544(1-2):143–152, 2005.

[13] Maninder Meenu, Qianxi Cai, and Baojun Xu. A critical review on analytical techniques to detect adulteration of extra virgin olive oil. *Trends in Food Science & Technology*, 91:391–408, 2019.

[14] Hazem Jabeur, Malika Drira, Ahmed Rebai, and Mohamed Bouaziz. Putative markers of adulteration of higher-grade olive oil with less expensive pomace olive oil identified by gas chromatography combined with chemometrics. *Journal of agricultural and food chemistry*, 65(26):5375–5383, 2017.

[15] Weng Kung Peng, Tian-Tsong Ng, and Tze Ping Loh. Machine learning assistive rapid, label-free molecular phenotyping of blood with two-dimensional nmr correlational spectroscopy. *Communications biology*, 3(1):1–10, 2020.

[16] Weng Kung Peng. Clustering nuclear magnetic resonance: Machine learning assistive rapid two-dimensional relaxometry mapping. *Engineering Reports*, page e12383, 2021.

[17] Xuewen Hou, Guangli Wang, Xin Wang, Xinmin Ge, Yiren Fan, Rui Jiang, and Shengdong Nie. Rapid screening for hazelnut oil and high-oleic sunflower oil in extra virgin olive oil using low-field nuclear magnetic resonance relaxometry and machine learning. *Journal of the Science of Food and Agriculture*, 101(6):2389–2397, 2021.

[18] Zhi-Ming Huang, Jia-Xiang Xin, Shan-Shan Sun, Yi Li, Da-Xiu Wei, Jing Zhu, Xue-Lu Wang, Jiachen Wang, and Ye-Feng Yao. Rapid identification of adulteration in edible vegetable oils based on low-field nuclear magnetic resonance relaxation fingerprints. *Foods*, 10(12):3068, 2021.

[19] Antoine Dupré, Ka-Meng Lei, Pui-In Mak, Rui P Martins, and Weng Kung Peng. Micro-and nanofabrication nmr technologies for point-of-care medical applications–a review. *Microelectronic Engineering*, 209:66–74, 2019.

[20] Expert consultation. Fats and fatty acids in human nutrition. *Food and Agriculture Organization of the United Nations*, 10:166, 2010.

[21] Narinder Kaur, Vishal Chugh, and Anil K Gupta. Essential fatty acids as functional components of foods-a review. *Journal of food science and technology*, 51(10):2289–2303, 2014.

[22] Cathriona R Monnard and Abdul G Dulloo. Polyunsaturated fatty acids as modulators of fat mass and lean mass in human body composition regulation and cardiometabolic health. *Obesity Reviews*, 22:e13197, 2021.

[23] Tewodros Shibabaw. Omega-3 polyunsaturated fatty acids: anti-inflammatory and anti-hypertriglyceridemia mechanisms in cardiovascular disease. *Molecular and Cellular Biochemistry*, 476(2):993–1003, 2021.

[24] I Reinders, RA Murphy, X Song, M Visser, MF Cotch, TF Lang, ME Garcia, LJ Launer, K Siggeirsdottir, G Eiriksdottir, et al. Polyunsaturated fatty acids in relation to incident mobility disability and decline in gait speed; the age, gene/environment susceptibility-reykjavik study. *European journal of clinical nutrition*, 69(4):489–493, 2015.

[25] A Gliszczynska-Swiglo, Ewa Sikorska, Igor Khmelinskii, and Marek Sikorski. Tocopherol content in edible plant oils. *Polish Journal of Food and Nutrition Sciences*, 57(4 [A]):157–161, 2007.

[26] Anna-Sophie von Hanstein, Sigurd Lenzen, and Thomas Plötz. Toxicity of fatty acid profiles of popular edible oils in human endoc-βh1 beta-cells. *Nutrition & diabetes*, 10(1):1–5, 2020.

[27] T Plötz, AS von Hanstein, B Krümmel, A Laporte, I Mehmeti, and S Lenzen. Structure-toxicity relationships of saturated and unsaturated free fatty acids for elucidating the lipotoxic effects in human endoc-βh1 beta-cells. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1865(11):165525, 2019.

[28] Bonnie Quinn, Fernanda Peyronel, Tyler Gordon, Alejandro Marangoni, Charles B Hanna, and David A Pink. Aggregation in complex triacylglycerol oils: coarse-grained models, nanophase separation, and predicted x-ray intensities. *Journal of Physics: Condensed Matter*, 26(46):464108, 2014.

[29] David P Cistola, James A Hamilton, David Jackson, and Donald M Small. Ionization and phase behavior of fatty acids in water: application of the gibbs phase rule. *Biochemistry*, 27(6):1881–1888, 1988.

[30] ING Wardana, Agung Widodo, and Widya Wijayanti. Improving vegetable oil properties by transforming fatty acid chain length in jatropha oil and coconut oil blends. *Energies*, 11(2):394, 2018.

[31] Michelle D Robinson and David P Cistola. Nanofluidity of fatty acid hydrocarbon chains as monitored by benchtop time-domain nuclear magnetic resonance. *Biochemistry*, 53(48):7515–7522, 2014.

[32] Jenny Lund and Arild C Rustan. Fatty acids: Structures and properties. 2020.

[33] Daniella Valeri and Antonio JA Meirelles. Viscosities of fatty acids, triglycerides, and their binary mixtures. *Journal of the American Oil Chemists' Society*, 74(10):1221–1226, 1997.

[34] PV Rao, Stephen Clarke, Richard Brown, and Kai-Shen Wu. Influence of iodine value on combustion and nox emission characteristics of a di diesel engine. *Proceedings of the Chemeca 2010*, 2010.

[35] Gerhard Knothe and Kevin R Steidley. Kinematic viscosity of biodiesel fuel components and related compounds. influence of compound structure and comparison to petrodiesel fuel components. *Fuel*, 84(9):1059–1065, 2005.

[36] Arild C Rustan and Christian A Drevon. Fatty acids: structures and properties. *e LS*, 2001.

[37] Ulrich Laufs, Klaus G Parhofer, Henry N Ginsberg, and Robert A Hegele. Clinical review on triglycerides. *European heart journal*, 41(1):99–109c, 2020.

[38] Kathleen L Wyne. Free fatty acids and type 2 diabetes mellitus. *The American journal of medicine*, 115(8):29–36, 2003.

[39] Guenther Boden and Gl Shulman. Free fatty acids in obesity and type 2 diabetes: defining their role in the development of insulin resistance and β-cell dysfunction. *European journal of clinical investigation*, 32:14–23, 2002.

[40] Steen Stender and Jørn Dyerberg. Influence of trans fatty acids on health. *Annals of nutrition and metabolism*, 48(2):61–66, 2004.

[41] Alberto Ascherio and Walter C Willett. Health effects of trans fatty acids. *The American journal of clinical nutrition*, 66(4):1006S–1010S, 1997.

[42] Vandana Dhaka, Neelam Gulia, Kulveer Singh Ahlawat, and Bhupender Singh Khatkar. Trans fats—sources, health risks and alternative approach-a review. *Journal of food science and technology*, 48(5):534–541, 2011.

[43] John W Gofman, Frank Lindgren, Harold Elliott, William Mantz, John Hewitt, Beverly Strisower, Virgil Herring, and Thomas P Lyon. The role of lipids and lipoproteins in atherosclerosis. *Science*, 111(2877):166–186, 1950.

[44] DE Lorgeril and E Michel. Essential polyunsaturated fatty acids, inflammation, atherosclerosis and cardiovascular diseases. *Inflammation in the Pathogenesis of Chronic Diseases*, pages 283–297, 2007.

[45] JZ Lu, SP Muench, M Allary, S Campbell, CW Roberts, E Mui, RL McLeod, DW Rice, and ST Prigge. Type i and type ii fatty acid biosynthesis in eimeria tenella: enoyl reductase activity and structure. *Parasitology*, 134(14):1949–1962, 2007.

[46] Stuart Smith. The animal fatty acid synthase: one gene, one polypeptide, seven enzymes. *The FASEB journal*, 8(15):1248–1259, 1994.

[47] BJ Nikolau, DJ Oliver, PS Schnable, and ES Wurtele. *Molecular biology of acetyl-CoA metabolism*. Portland Press Ltd., 2000.

[48] Joaquín J Salas, John L Harwood, and Enrique Martínez-Force. *Lipid metabolism in olive: Biosynthesis of triacylglycerols and aroma components*. Springer, 2013.

[49] Turgay Unver, Zhangyan Wu, Lieven Sterck, Mine Turktas, Rolf Lohaus, Zhen Li, Ming Yang, Lijuan He, Tianquan Deng, Francisco Javier Escalante, et al. Genome of wild olive and the evolution of oil biosynthesis. *Proceedings of the National Academy of Sciences*, 114(44):E9413–E9422, 2017.

[50] Seung Kon Hong, Kook Han Kim, Joon Kyu Park, Ki-Woong Jeong, Yangmee Kim, and Eunice EunKyeong Kim. New design platform for malonyl-coa-acyl carrier protein transacylase. *FEBS letters*, 584(6):1240–1244, 2010.

[51] John L Harwood. Fatty acid metabolism. *Annual Review of Plant Physiology and Plant Molecular Biology*, 39(1):101–138, 1988.

[52] RA Sidorov, AV Zhukov, VP Pchelkin, and VD Tsydendambaev. Palmitic acid in higher plant lipids. *Palmitic Acid: Occurrence, Biochemistry and Health Effects*, pages 125–143, 2014.

[53] Z Zhen, TF Xi, and YF Zheng. Surface modification by natural biopolymer coatings on magnesium alloys for biomedical applications. pages 301–333, 2015.

[54] Aleksandra Czumaj and Tomasz Śledziński. Biological role of unsaturated fatty acid desaturases in health and disease. *Nutrients*, 12(2):356, 2020.

[55] John Shanklin and Edgar B Cahoon. Desaturation and related modifications of fatty acids. *Annual review of plant biology*, 49(1):611–641, 1998.

[56] James G Wallis and John Browse. Mutants of arabidopsis reveal many roles for membrane lipids. *Progress in lipid research*, 41(3):254–278, 2002.

[57] Claude Alban, Dominique Job, and Roland Douce. Biotin metabolism in plants. *Annual review of plant biology*, 51(1):17–47, 2000.

[58] Lin Li, Hui Li, JiYing Li, ShuTu Xu, XiaoHong Yang, JianSheng Li, and JianBing Yan. A genome-wide survey of maize lipid-related genes: candidate genes mining, digital gene expression profiling and co-location with qtl for maize kernel oil. *Science China Life Sciences*, 53(6):690–700, 2010.

[59] Dongmei Yin, Yun Wang, Xingguo Zhang, Hemin Li, Xiang Lu, Jinsong Zhang, Wanke Zhang, and Shouyi Chen. De novo assembly of the peanut (arachis hypogaea l.) seed transcriptome revealed candidate unigenes for oil accumulation pathways. *Plos one*, 8(9):e73767, 2013.

[60] Brian K Benson, Glen Meades Jr, Anne Grove, and Grover L Waldrop. Dna inhibits catalysis by the carboxyltransferase subunit of acetyl-coa carboxylase: Implications for active site communication. *Protein Science*, 17(1):34–42, 2008.

[61] IS Bhatia, KL Ahuja, and PS Sukhija. Changes in the activity of acetyl coa carboxylase in germinating and ripening sunflower seeds. *Physiologia Plantarum*, 44(3):141–144, 1978.

[62] Patrick Bilder, Sandra Lightle, Graeme Bainbridge, Jeffrey Ohren, Barry Finzel, Fang Sun, Susan Holley, Loola Al-Kassim, Cindy Spessard, Michael Melnick, et al. The structure of the carboxyltransferase component of acetyl-coa carboxylase reveals a zinc-binding motif unique to the bacterial enzyme. *Biochemistry*, 45(6):1712–1722, 2006.

[63] Katayoon Dehesh, Heeyoung Tai, Patricia Edwards, James Byrne, and Jan G Jaworski. Overexpression of 3-ketoacyl-acyl-carrier protein synthase iiis in plants reduces the rate of lipid synthesis. *Plant physiology*, 125(2):1103–1114, 2001.

[64] Jun Li, Mei-Ru Li, Ping-Zhi Wu, Chang-En Tian, Hua-Wu Jiang, and Guo-Jiang Wu. Molecular cloning and expression analysis of a gene encoding a putative β-ketoacyl-acyl carrier protein (acp) synthase iii (kas iii) from jatropha curcas. *Tree physiology*, 28(6):921–927, 2008.

[65] Guo-Zhang Wu and Hong-Wei Xue. Arabidopsis β-ketoacyl-[acyl carrier protein] synthase i is crucial for fatty acid synthesis and plays a role in chloroplast division and embryo development. *The Plant Cell*, 22(11):3726–3744, 2010.

[66] John L Harwood. Recent advances in the biosynthesis of plant fatty acids. *Biochimica et Biophysica Acta (BBA)-Lipids and Lipid Metabolism*, 1301(1-2):7–56, 1996.

[67] Aejaz A Dar, Abhikshit R Choudhury, Pavan K Kancharla, and Neelakantan Arumugam. The fad2 gene in plants: occurrence, regulation, and role. *Frontiers in plant science*, 8:1789, 2017.

[68] Magdalena Maszewska, Anna Florowska, Elżbieta Dłużewska, Małgorzata Wroniak, Katarzyna Marciniak-Lukasiak, and Anna Żbikowska. Oxidative stability of selected edible oils. *Molecules*, 23(7):1746, 2018.

[69] Nurhan Turgut Dunford et al. Edible oil quality. 2016.

[70] Eunok Choe and David B Min. Mechanisms and factors for edible oil oxidation. *Comprehensive reviews in food science and food safety*, 5(4):169–186, 2006.

[71] Victoria Jackson and Meera Penumetcha. Dietary oxidised lipids, health consequences and novel food technologies that thwart food lipid oxidation: an update. *International Journal of Food Science & Technology*, 54(6):1981–1988, 2019.

[72] DAVID B Min, JEFFREY M Boff, et al. Lipid oxidation of edible oil. *Food lipids: chemistry, nutrition, and biotechnology*, (Ed. 2):335–363, 2002.

[73] Malcolm H Levitt. *Spin dynamics: basics of nuclear magnetic resonance*. John Wiley & Sons, 2013.

[74] Robert L Kleinberg. *9. Nuclear Magnetic Resonance*, volume 35. Elsevier, 1999.

[75] Herman Y Carr and Edward M Purcell. Effects of diffusion on free precession in nuclear magnetic resonance experiments. *Physical review*, 94(3):630, 1954.

[76] Saul Meiboom and David Gill. Modified spin-echo method for measuring nuclear relaxation times. *Review of scientific instruments*, 29(8):688–691, 1958.

[77] Manuel Arsenio Lores Guevara, Juan Carlos García Naranjo, and Carlos Alberto Cabal Mirabal. Mr relaxation studies of hemoglobin aggregation process in sickle cell disease: Application for diagnostics and therapeutics. *Applied Magnetic Resonance*, 50(4):541–551, 2019.

[78] Lidia Latanowicz. Spin-lattice nmr relaxation and second moment of nmr line in solids containing ch3 groups. *Concepts in Magnetic Resonance Part A*, 44(4):214–225, 2015.

[79] AL MacKay. A proton nmr moment study of the gel and liquid-crystalline phases of dipalmitoyl phosphatidylcholine. *Biophysical journal*, 35(2):301–313, 1981.

[80] August V Bailey and Robert A Pittman. Wide-line nmr spectra of some saturated and unsaturated long chain fatty acids. *Journal of the American Oil Chemists Society*, 48(12):775–777, 1971.

[81] Kenneth D Lawson and Thomas J Flautt. Nuclear magnetic resonance absorption in anhydrous sodium soaps. *The Journal of Physical Chemistry*, 69(12):4256–4268, 1965.

[82] RF Grant and BA Dunell. Proton magnetic resonance absorption in the c-form of stearic acid. *Canadian Journal of Chemistry*, 38(3):359–364, 1960.

[83] Hossein Noureddini, BC Teoh, and L Davis Clements. Densities of vegetable oils and fatty acids. *Journal of the American Oil Chemists Society*, 69(12):1184–1188, 1992.

[84] Iqbal H Sarker. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3):1–21, 2021.

[85] Andrew McCallum. Information extraction: Distilling structured data from unstructured text. *Queue*, 3(9):48–57, 2005.

[86] Mohssen Mohammed, Muhammad Badruddin Khan, and Eihab Bashier Mohammed Bashier. *Machine learning: algorithms and applications*. Crc Press, 2016.

[87] Ayon Dey. Machine learning algorithms: a review. *International Journal of Computer Science and Information Technologies*, 7(3):1174–1179, 2016.

[88] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[89] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[90] George H John and Pat Langley. Estimating continuous distributions in bayesian classifiers. *arXiv preprint arXiv:1302.4964*, 2013.

[91] Batta Mahesh. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9:381–386, 2020.

[92] Vinod Kumar Chauhan, Kalpana Dahiya, and Anuj Sharma. Problem formulations and solvers in linear svm: a review. *Artificial Intelligence Review*, 52(2):803–855, 2019.

[93] J Han, M Kamber, and J Pei. Data transformation and data discretization. data mining: Concepts and techniques, 2011.

[94] Happiness Ugochi Dike, Yimin Zhou, Kranthi Kumar Deveerasetty, and Qingtian Wu. Unsupervised learning based on artificial neural network: A review. pages 322–327, 2018.

[95] James MacQueen et al. Some methods for classification and analysis of multivariate observations. 1(14):281–297, 1967.

[96] Keinosuke Fukunaga and Larry Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on information theory*, 21(1):32–40, 1975.

[97] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. 96(34):226–231, 1996.

[98] Darren A Whitaker and Kevin Hayes. A simple algorithm for despiking raman spectra. *Chemometrics and Intelligent Laboratory Systems*, 179:82–84, 2018.

[99] ISO. Iso 660: animal and vegetable fats and oils: determination of acid value and acidity, 2009.

[100] Ricardo Fernández-Escobar. Olive nutritional status and tolerance to biotic and abiotic stresses. *Frontiers in plant science*, 10:1151, 2019.

[101] Manuel Tejada and Concepción Benítez. Effects of different organic wastes on soil biochemical properties and yield in an olive grove. *Applied Soil Ecology*, 146:103371, 2020.

[102] Cecilia Jimenez-Lopez, Maria Carpena, Catarina Lourenço-Lopes, Maria Gallardo-Gomez, Jose M Lorenzo, Francisco J Barba, Miguel A Prieto, and Jesus Simal-Gandara. Bioactive compounds and quality of extra virgin olive oil. *Foods*, 9(8):1014, 2020.

[103] Paola Conte, Costantino Fadda, Alessandra Del Caro, Pietro Paolo Urgeghe, and Antonio Piga. Table olives: an overview on effects of processing on nutritional and sensory quality. *Foods*, 9(4):514, 2020.

[104] MDG da Silva, Ana M Costa Freitas, MJ Cabrita, and Raquel Garcia. Olive oil composition: Volatile compounds. *Olive oil-constituents, quality, health properties and bioconversions*, 2012.

Part IV

APPENDIX

Table A1: **Quantitative traits (e.g., lipid profile) of all vegetable oils samples** used (e.g., palm, olive, avocado, peanut, sesame, corn, grapeseed, sunflower, linseed). Grey color represent the predominant specie of FA on this organism. Values with arbitrary units refer to the proportion of the FA specie on the lipid profile.

Oil type	Manufacturer brand	Energy (kJ / kcal)	SAFA (g)	MUFA (g)	PUFA (g)	Unsaturated (g)	Fatty acids (g)	SAFA (arb)	MUFA (arb)	PUFA (arb)	Unsaturated (arb)
Avocado	Graduva™	3397	11.0	67.0	14.0	81.0	92.0	0.120	0.728	0.152	0.880
Avocado	La Masia™	3700	15.0	67.0	18.0	85.0	100.0	0.150	0.670	0.180	0.850
Avocado	Woxiaoya	3696	-	-	-	-	99.9	-	-	-	-
Avocado	Three Squirrels-Deer Blue™	3696	17.2	70.8	11.9	82.7	99.9	0.172	0.709	0.119	0.828
Avocado	Meinongji™	3696	14.2	73.6	12.1	85.7	99.9	0.142	0.737	0.121	0.858
Corn	Fula™	3397	13.0	28.0	50.0	78.0	92.0	0.141	0.304	0.543	0.848
Corn	Cofco-Chucui™	3696	-	-	-	-	100.0	-	-	-	-
Corn	Xiwang	3696	15.0	30.8	54.1	84.9	99.9	0.150	0.308	0.542	0.850
Corn	Haitian-Yousling™	3694	17.7	26.0	56.0	82.0	99.7	0.178	0.261	0.562	0.822
Corn	Longevity Flower™	3700	15.0	30.6	54.5	85.1	100.0	0.150	0.306	0.545	0.851
Corn	Yihai Kerry-Arawana Brand™	3700	-	-	-	-	100.0	-	-	-	-
Grapeseed	Fula™	3404	11.0	19.0	61.0	80.0	91.0	0.121	0.209	0.670	0.879
Linseed	Nature Foods™	3760	11.0	18.0	71.0	89.0	100.0	0.110	0.180	0.710	0.890
Olive	Herdade do Esporão™	3375	13.2	71.2	6.3	77.5	90.7	0.146	0.785	0.069	0.854
Olive	Olivoila™	3700	14.0	79.0	7.0	86.0	100.0	0.140	0.790	0.070	0.860
Olive	Andasalusia™	3700	15.0	79.0	6.0	85.0	100.0	0.150	0.790	0.060	0.850
Palm	Guineas™	3700	48.0	39.0	13.0	52.0	100.0	0.480	0.390	0.130	0.520
Peanut	Fula™	3374	16.0	61.0	15.0	76.0	92.0	0.174	0.663	0.163	0.826
Peanut	Vitaquell™	3700	15.0	80.0	5.0	85.0	100.0	0.150	0.800	0.050	0.850
Peanut	Luhua™	3696	-	-	-	-	99.9	-	-	-	-
Peanut	Hujihua™	3700	-	-	-	-	100.0	-	-	-	-
Peanut	Longda™	3700	20.6	42.2	37.1	79.3	99.9	0.206	0.422	0.371	0.794
Peanut	Kingshare-Xianyoufang™	3696	-	-	-	-	99.9	-	-	-	-
Peanut	Moyanghua™	3700	-	-	-	-	100.0	-	-	-	-
Sesame	Emile Noel™	3700	16.0	42.0	42.0	84.0	100.0	0.160	0.420	0.420	0.840
Sesame	La Masia™	3700	14.0	40.0	46.0	86.0	100.0	0.140	0.400	0.460	0.860
Sesame	Yihai Kerry-Xiangmanyuan™	3700	-	-	-	-	100.0	-	-	-	-
Sesame	Fuyun™	3700	-	-	-	-	100.0	-	-	-	-
Sesame	Yihai Kerry-Arawana Brand™	3700	-	-	-	-	100.0	-	-	-	-
Sesame	Totole™	3696	-	-	-	-	99.9	-	-	-	-
Sesame	Luhua™	3696	-	-	-	-	99.9	-	-	-	-
Sunflower	Fula™	3397	10.0	28.0	53.0	81.0	91.0	0.110	0.308	0.582	0.890
Sunflower	Cofco-Chucui™	3696	-	-	-	-	99.9	-	-	-	-
Sunflower	Cofco-Fulinmen™	3696	13.0	26.4	60.5	86.9	99.9	0.130	0.264	0.606	0.870
Sunflower	Sinopharm-Lizzil™	3696	12.0	25.8	62.1	87.9	99.9	0.120	0.258	0.622	0.880
Sunflower	Mighty™	3700	12.0	26.0	62.0	88	100.0	0.120	0.260	0.620	0.880

Table A2: **Average vegetable oil measures obtained using NMR-based traits.** Mean longitudinal (T_1) and transversal (T_2) relaxation times, A-ratio, and unfolded biphasic traits (T_{2a}, T_{2b}) experimentally obtained for palm, olive, avocado, peanut, sesame, corn, grapeseed, sunflower, linseed oils. Values with arbitrary units refer to the proportion of the FA specie on the lipid profile.

Oil type	Manufacturer Brand	Saturated (arb)	MUFA (arb)	PUFA (arb)	PUFA/MUFA (arb)	T_1 (ms)	T_2 (ms)	A-ratio	T_{2a} (ms)	T_{2b} (ms)
Avocado	Graduva™	0.120	0.728	0.152	0.209	159.4	147.8	1.08	78.9	257.3
Avocado	La Masia™	0.150	0.670	0.180	0.269	158.5	146.3	1.08	77.2	251.3
Avocado	Woxiaoya™					161.4	147.8	1.09	78.9	247.0
Avocado	Three Squirrels-Deer Blue™	0.172	0.709	0.119	0.168	166.6	153.6	1.08	80.7	267.3
Avocado	Meinongji™	0.142	0.737	0.121	0.164	170.1	155.4	1.1	83.8	270.2
Corn	Fula™	0.141	0.304	0.543	1.786	180.7	172.1	1.1	84.9	293.4
Corn	Cofco-Chucui™					188.5	181.1	1.04	89.2	309.7
Corn	Xiwang™	0.150	0.308	0.542	1.756	177.5	168.4	1.05	83.8	288.7
Corn	Haitian-Yousling™	0.178	0.261	0.562	2.154	187.6	180.2	1.04	89.0	308.7
Corn	Longevity Flower™	0.150	0.306	0.545	1.781	184.6	177.7	1.04	88.5	304.3
Corn	Yihai Kerry-Arawana Brand™					176.1	167.1	1.05	84.3	288.9
Grapeseed	Fula™	0.121	0.209	0.670	3.211	196.5	189.4	1.04	89.6	330.2
Linseed	Nature Foods™	0.110	0.180	0.710	3.944	233.2	232.0	1.01	108.6	437.3
Olive	Herdade do Esporão™	0.146	0.785	0.069	0.088	154.1	139.3	1.11	74.1	235.0
Olive	Olivoila™	0.140	0.790	0.070	0.089	164.6	151.0	1.09	81.2	252.1
Olive	Andasalusia™	0.150	0.790	0.060	0.076	165.7	152.1	1.09	80.9	250.4
Palm	Guineas™	0.480	0.390	0.130	0.333	144.3	124.7	1.16	69.0	220.0
Peanut	Fula™	0.174	0.663	0.163	0.246	156.8	141.7	1.11	76.8	242.1
Peanut	Vitaquell™	0.150	0.800	0.050	0.063	152.6	134.9	1.13	73.8	223.5
Peanut	Luhua™					161.0	148.8	1.08	79.6	253.6
Peanut	Hujihua™					164.4	149.9	1.10	80.4	259.5
Peanut	Longda™	0.206	0.422	0.371	0.879	166.1	154.9	1.07	81.4	269.0
Peanut	Kingshare-Xianyoufang™					181.1	168.1	1.1	86.3	297.1
Peanut	Moyanghua™					176.4	163.7	1.08	85.3	288.0
Sesame	Emile Noël™	0.160	0.420	0.420	1.000	172.2	160.9	1.07	81.9	277.8
Sesame	La Masia™	0.140	0.400	0.460	1.150	173.9	162.8	1.07	82.9	280.9
Sesame	Yihai Kerry-Xiangmanyuan™					175.3	164.7	1.06	83.5	283.3
Sesame	Fuyun™					181.5	165.3	1.10	81.9	285.4
Sesame	Yihai Kerry-Arawana Brand™					169.3	155.6	1.09	82.0	269.8
Sesame	Totole™					171.4	160.1	1.07	81.6	274.9
Sesame	Luhua™					175.2	163.6	1.07	84.0	278.8
Sunflower	Fula™	0.110	0.308	0.582	1.893	193.4	183.4	1.06	88.9	318.8
Sunflower	Cofco-Chucui™					203.0	195.8	1.04	94.2	342.4
Sunflower	Cofco-Fulinmen™	0.130	0.264	0.606	2.292	201.8	195.2	1.03	92.5	339.8
Sunflower	Sinopharm-Lizzil™	0.120	0.258	0.622	2.407	189.0	178.1	1.06	87.9	315.5
Sunflower	Mighty™	0.120	0.260	0.620	2.385	204.9	197.2	1.04	94.9	344.5

Table A3: **Quantitative traits (lipid profile) of all olive oil samples.** The nutritional information (e.g., lipid profile, acidity) of each olive oil sample as disclosed by manufacturers (e.g., olive oil type, region of origin). Undisclosed countries are denotes as not defined (n.d.).

Olive oil type	Region	Manufacturer	Nutritional Information (per 100g)			
			SAFA (g)	MUFA (g)	PUFA (g)	Max acidity (%)
EVOO	Greece	Agric™	14.0	77.0	9.0	0.8
EVOO	Greece	Omega live™	13.6	70.7	7.3	0.8
EVOO	Greece	Molon™	12.8	70.5	8.3	0.4
EVOO	Italy	Antika™	-	-	-	-
EVOO	Italy	Costa'Oro™	15.0	-	-	-
EVOO	Italy	Ewen™	-	-	-	0.8
EVOO	Italy	Berio™	13.9	70.4	7.0	0.6
EVOO	Portugal	GALLO - Colheira Madura™	15.0	68.0	7.9	0.3
EVOO	Portugal	Oliveira da Serra - Gourmet™	15.0	69.0	6.9	0.3
EVOO	Portugal	Herdade do Esporão - Azeite DOP™	13.1	71.8	6.3	0.3
EVOO	Portugal	GALLO - Clássico™	15.0	68.0	7.9	0.7
EVOO	Portugal	GALLO - Reserva™	15.0	68.0	7.9	0.5
EVOO	Portugal	Vidigueira™	13.1	-	-	0.8
EVOO	Portugal	Chaparro - Origens™	13.3	-	-	0.7
EVOO	Portugal	Oliveira da Serra - Clássico™	13.0	72.0	6.0	0.5
EVOO	Portugal	Flor do Alentejo™	13.0	-	-	0.8
EVOO	Spain	Olivolia™	14.0	79.0	7.0	0.5
EVOO	Spain	Froiz™	13.0	-	-	-
EVOO	Spain	Mercadona™	13.2	71.2	6.3	0.4
EVOO	Spain	Rego - Arbequina™	-	-	-	-
EVOO	Spain	Rego™	-	-	-	-
VOO	*n.d.*	Continente™	13.1	-	-	2.0
VOO	Portugal	Oliveira da Serra - Versátil™	14.5	68.6	8.1	0.9
VOO	Portugal	Oliveira da Serra - Virgem™	15.0	69.0	8.1	0.9
VOO	Portugal	GALLO - Delicado™	15.0	68.0	7.9	1.0
VOO	Portugal	5 Soldos - Casto™	13.1	71.8	6.3	0.5
VOO	Portugal	Chaparro - Virgem™	14.5	-	-	0.9
VOO	Portugal	Vila Branca™	14.0	77.0	9.0	0.9
VOO	Portugal	Guia™	12.0	-	-	0.7
VOO+REF	Portugal	5 Soldos - Azeite™	13.1	71.8	6.3	1.0
VOO+REF	Portugal	Oliveira da Serra - Azeite™	13.0	72.0	6.0	1.0
VOO+REF	Portugal	Serrata™	14.0	-	-	1.0
VOO+REF	Portugal	Rustica™	14.0	-	-	1.0
VOO+REF	Spain	Froiz™	14.0	-	-	1.0
VOO+REF	Spain	Olearia del Olivar™	11.9	-	-	1.0
VOO+REF	Spain	La Española™	14.0	-	-	-

Table A4: **Average olive oil measures obtained using NMR-based traits.** Mean longitudinal (T_1) and transversal (T_2) relaxation times, A-ratio, and experimentally obtained acid value for refined OO, EVOOs and VOOs with respect to their type and origin. Undisclosed regions were denoted as not defined (n.d.)

Region	Olive oil type	T_1 (ms)	T_2 (ms)	A-ratio (arb)	Acid Value (mg KOH g^{-1})
Greece	EVOO	164.3	148.3	1.11	0.456
Greece	EVOO	167.1	146.2	1.14	0.578
Greece	EVOO	167.5	148.7	1.13	0.510
Italy	EVOO	167.8	148.7	1.13	0.671
Italy	EVOO	170.6	150.8	1.09	0.718
Italy	EVOO	171.5	153.1	1.12	0.414
Italy	EVOO	165.6	148.2	1.12	0.478
Portugal	EVOO	168.9	150.1	1.13	0.487
Portugal	EVOO	166.8	151.9	1.10	0.418
Portugal	EVOO	166.6	150.3	1.11	0.575
Portugal	EVOO	168.7	151.2	1.12	0.616
Portugal	EVOO	170.2	151.5	1.12	0.624
Portugal	EVOO	172.7	149.8	1.15	0.435
Portugal	EVOO	170.8	154.2	1.11	0.561
Portugal	EVOO	167.2	151.6	1.10	0.673
Portugal	EVOO	168.4	148.1	1.14	0.596
Spain	EVOO	165.2	152.1	1.09	0.409
Spain	EVOO	166.6	149.9	1.11	0.469
Spain	EVOO	166.8	153.1	1.09	0.391
Spain	EVOO	167.3	144.8	1.16	0.449
Spain	EVOO	167.6	150.6	1.14	0.368
n.d.	VOO	174.4	153.3	1.14	0.740
Portugal	VOO	172.9	149.9	1.15	0.628
Portugal	VOO	173.3	152.7	1.14	0.740
Portugal	VOO	174.4	152.1	1.15	0.648
Portugal	VOO	173.5	152.2	1.14	0.718
Portugal	VOO	174.6	154.8	1.13	0.592
Portugal	VOO	174.6	155.3	1.13	0.787
Portugal	VOO	177.1	155.1	1.14	0.860
Portugal	VOO+REF	164.1	146.6	1.12	0.435
Portugal	VOO+REF	163.9	146.1	1.12	0.350
Portugal	VOO+REF	163.5	144.3	1.13	0.364
Portugal	VOO+REF	162.3	150.4	1.08	0.504
Spain	VOO+REF	160.3	144.2	1.11	0.359
Spain	VOO+REF	163.5	146.2	1.12	0.397
Spain	VOO+REF	161.9	146.0	1.11	0.387

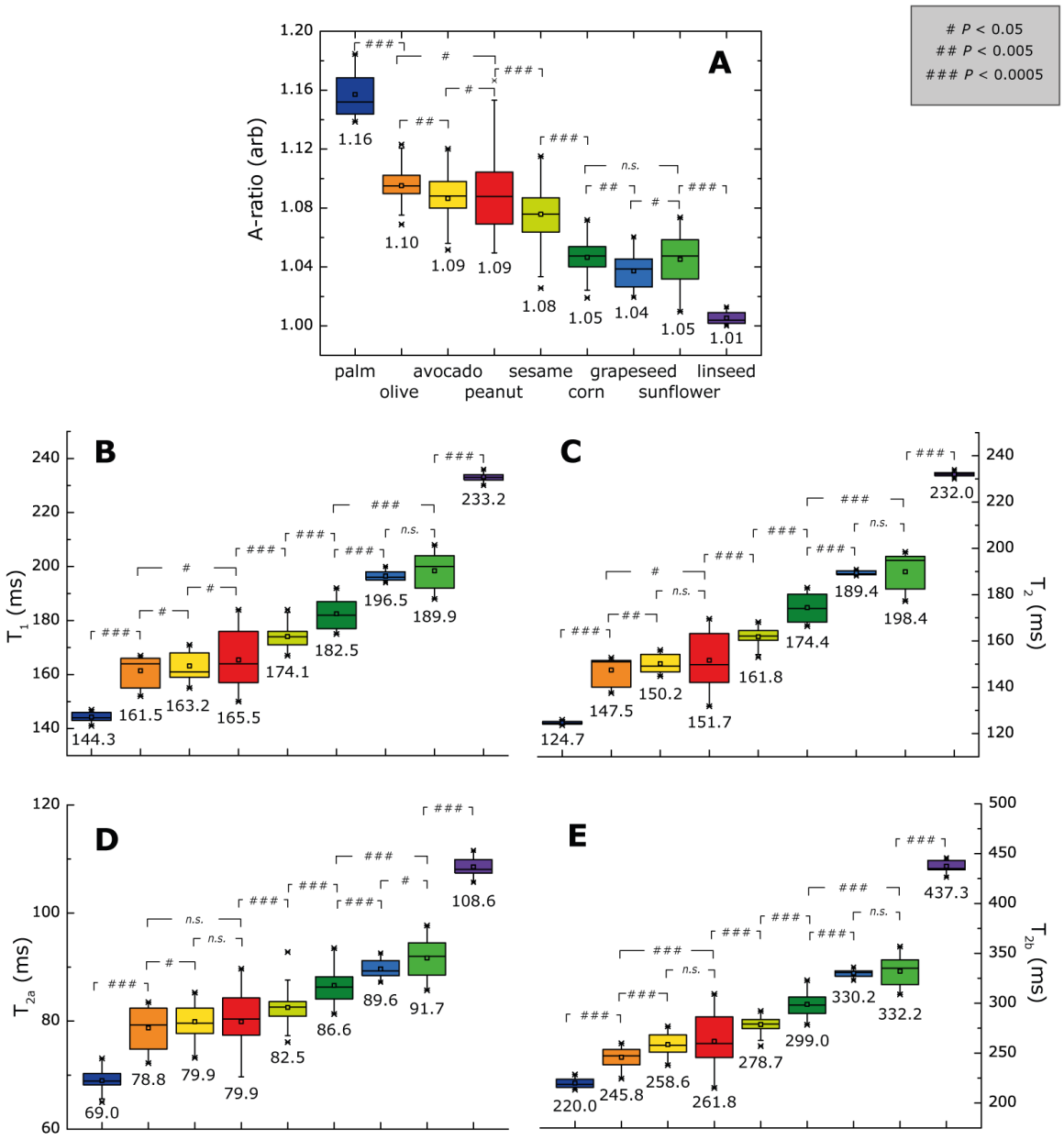

Figure A1: **One-dimensional plot of all the NMR-based traits.** (A) A-ratio, (B) T_1, (C) T_2, (D) T_{2a}, (E) T_{2b}. The legend for each vegetable oil is the same for each plot (top). The box plots represent 25% and 75% quantiles of the measurements. One tailed Student's *t*-test was used to calculate the *P*-value

Table A5: **Identification of vegetable oils using ROC with single-phase system.** AUC (range between 0 to 1) of the various supervised models evaluated to identify vegetable oil. Models were validated using Leave-one-out method using the single-phase system (A-ratio, T_1, T_2).

Single-phase	Model	AUC	CA	F1	Precision	Recall
(T_1)	kNN	0.857	0.633	0.620	0.642	0.633
	Logistic Regression	0.884	0.524	0.476	0.455	0.524
	Naïve Bayes	0.829	0.456	0.368	0.311	0.456
	Neural Network	0.922	0.599	0.585	0.592	0.599
	Random Forest	0.942	0.667	0.665	0.669	0.667
	average	**0.887**	**0.576**	**0.543**	**0.534**	**0.576**
(T_2)	kNN	0.982	0.849	0.846	0.851	0.849
	Logistic Regression	0.896	0.567	0.513	0.478	0.567
	Naïve Bayes	0.828	0.433	0.352	0.299	0.433
	Neural Network	0.937	0.732	0.717	0.745	0.732
	Random Forest	0.992	0.889	0.887	0.891	0.889
	average	**0.927**	**0.694**	**0.663**	**0.653**	**0.694**
(A-ratio)	kNN	0.902	0.569	0.558	0.562	0.569
	Logistic Regression	0.753	0.238	0.135	0.094	0.238
	Naïve Bayes	0.764	0.335	0.234	0.185	0.335
	Neural Network	0.825	0.438	0.407	0.387	0.438
	Random Forest	0.971	0.762	0.757	0.765	0.762
	average	**0.843**	**0.469**	**0.418**	**0.399**	**0.469**
(T_1, T_2)	kNN	0.991	0.897	0.896	0.898	0.897
	Logistic Regression	0.893	0.611	0.575	0.557	0.611
	Naïve Bayes	0.852	0.468	0.397	0.360	0.468
	Neural Network	0.964	0.768	0.757	0.782	0.768
	Random Forest	0.997	0.944	0.944	0.944	0.944
	average	**0.940**	**0.738**	**0.714**	**0.708**	**0.738**
(T_1, T_2, A-ratio)	kNN	0.991	0.897	0.896	0.898	0.897
	Logistic Regression	0.891	0.619	0.585	0.565	0.619
	Naïve Bayes	0.897	0.512	0.463	0.487	0.512
	Neural Network	0.967	0.756	0.746	0.750	0.756
	Random Forest	0.998	0.954	0.954	0.954	0.954
	average	**0.949**	**0.748**	**0.729**	**0.731**	**0.748**

Table A6: **Identification of vegetable oils using ROC with biphasic system.** AUC (range between 0 to 1) of the various supervised models evaluated to identify vegetable oils. Models were validated using Leave-one-out method using the biphasic system (T_{2a},T_{2b}) and in combination with single-phase traits (All).

Biphasic	Model	AUC	CA	F1	Precision	Recall
(T_{2a})	kNN	0.896	0.558	0.554	0.562	0.558
	Logistic Regression	0.813	0.407	0.356	0.323	0.407
	Naïve Bayes	0.789	0.421	0.340	0.288	0.421
	Neural Network	0.839	0.468	0.424	0.414	0.468
	Random Forest	0.934	0.651	0.647	0.649	0.651
	average	**0.854**	**0.501**	**0.464**	**0.447**	**0.501**
(T_{2b})	kNN	0.945	0.677	0.669	0.689	0.677
	Logistic Regression	0.882	0.548	0.500	0.467	0.548
	Naïve Bayes	0.830	0.444	0.358	0.302	0.444
	Neural Network	0.910	0.587	0.558	0.577	0.587
	Random Forest	0.982	0.823	0.824	0.825	0.823
	average	**0.910**	**0.616**	**0.582**	**0.572**	**0.616**
(T_{2a},T_{2b})	kNN	0.965	0.766	0.758	0.762	0.766
	Logistic Regression	0.886	0.587	0.550	0.532	0.587
	Naïve Bayes	0.835	0.460	0.367	0.427	0.460
	Neural Network	0.946	0.716	0.703	0.706	0.716
	Random Forest	0.994	0.909	0.908	0.908	0.909
	average	**0.925**	**0.688**	**0.657**	**0.667**	**0.688**
All	**Model**	**AUC**	**CA**	**F1**	**Precision**	**Recall**
(NMR-based traits)	kNN	0.989	0.891	0.891	0.892	0.891
	Logistic Regression	0.911	0.637	0.619	0.617	0.637
	Naïve Bayes	0.907	0.556	0.487	0.523	0.556
	Neural Network	0.984	0.861	0.861	0.862	0.861
	Random Forest	1.000	0.974	0.974	0.975	0.974
	average	**0.958**	**0.784**	**0.766**	**0.774**	**0.784**

Table A7: **Characterization of vegetables oils with NMR-based traits.** Determination coefficient (R^2, range between 0 to 1) of the various supervised models evaluated to characterize vegetable oils. ML models were trained with lipid profiles and NMR-based traits (from both single-phase and biphasic system) measures (Appendix, Table A2) in order to predict lipid profile (e.g., SAFA, MUFA, PUFA content). Validation was done using the Leave-one-out method.

Prediction of	**Model**	**MSE**	**RMSE**	**MAE**	**R^2**
(SAFA)	AdaBoost	0.000	0.012	0.005	0.972
	kNN	0.000	0.012	0.007	0.973
	Linear Regression	0.004	0.060	0.035	0.318
	Neural Network	0.002	0.043	0.026	0.644
	SGD	0.003	0.058	0.039	0.345
	average	**0.002**	**0.037**	**0.023**	**0.650**
(MUFA)	AdaBoost	0.002	0.049	0.013	0.952
	kNN	0.003	0.055	0.024	0.942
	Linear Regression	0.013	0.115	0.100	0.740
	Neural Network	0.005	0.071	0.054	0.901
	SGD	0.012	0.108	0.093	0.773
	average	**0.007**	**0.080**	**0.057**	**0.862**
(PUFA)	AdaBoost	0.002	0.040	0.011	0.970
	kNN	0.002	0.047	0.021	0.960
	Linear Regression	0.012	0.109	0.093	0.783
	Neural Network	0.004	0.060	0.046	0.933
	SGD	0.011	0.103	0.088	0.804
	average	**0.006**	**0.072**	**0.052**	**0.890**

Table A8: **Characterization landscape of vegetables oils in single-phase system.** Determination coefficient (R^2, range between 0 to 1) of the various supervised models evaluated to characterize the phenotypic landscape of vegetable oils. ML models were trained with full database (i.e., 504 points) lipid profiles (Appendix, Table A2) to predict individually T_1 and T_2 (i.e., single-phase system) phenotypic landscape. Validation was done using the random sampling (50%) method repeated 5 times.

Prediction of	Model	MSE	RMSE	MAE	R^2
(T_1)	AdaBoost	6.593	2.568	1.963	0.978
	kNN	6.611	2.571	2.026	0.978
	Neural Network	9.668	3.109	2.456	0.968
	average	**7.624**	**2.749**	**2.148**	**0.975**
(T_2)	AdaBoost	3.583	1.893	1.358	0.991
	kNN	5.880	2.425	1.716	0.986
	Neural Network	6.423	3.551	2.712	0.970
	average	**5.295**	**2.623**	**1.929**	**0.982**

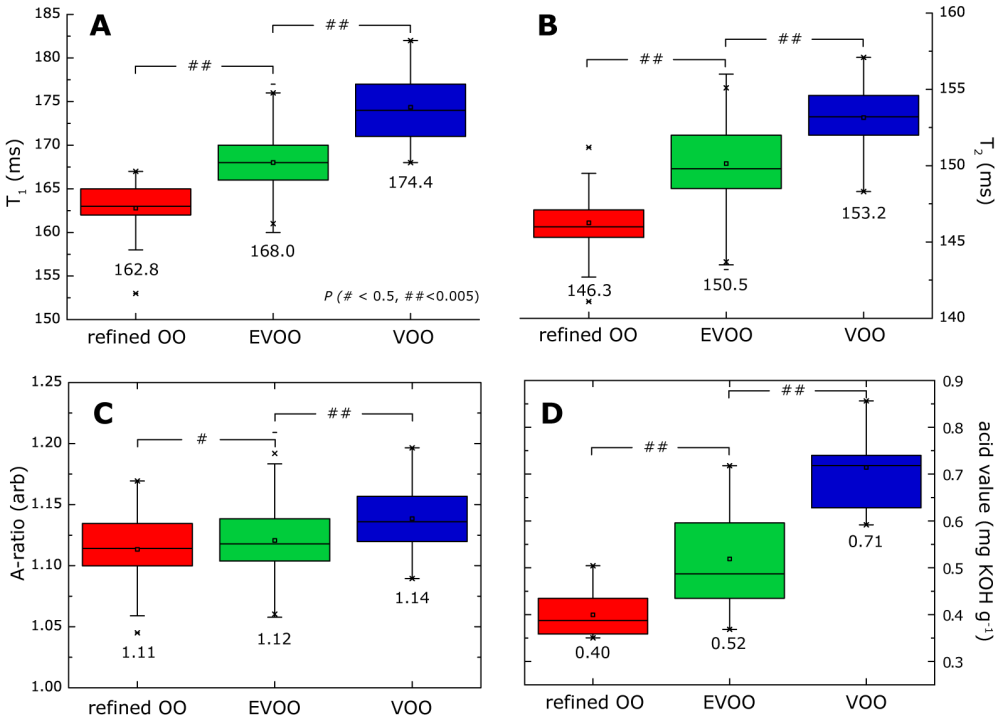

Figure A2: **Classification of olive oil using single-phase system.** (A) T_1 and (B) T_2 relaxations, (C) A-ratio, and (D) acid value obtained for the different commercial brands (7 refined OOs, 21 EVOOs and 8 VOOs). The box plots represent 25% and 75% quantile of the entire measurements. Two tailed Student's *t*-test was used to calculate the P-value.

Table A9: **Classification of olive oils using ROC analysis.** Area Under the Curve (range between 0 to 1) of the various supervised models evaluated to predict OO grading. Models were validated using Leave-one-out method with averaged single-phase system (e.g., T_1, T_2 and A-ratio). The wavelength (λ) used for UV-Vis spectroscopy and NIRS was chosen upon the region of bigger differentiation, being it the 415nm and 670nm peaks, respectively.

Olive oil type	Model	AUC	CA	F1	Precision	Recall
UV-Vis	kNN	0.898	0.833	0.833	0.835	0.833
(λ=415 nm)	Logistic Regression	0.444	0.417	0.370	0.333	0.417
	Naïve Bayes	0.615	0.472	0.479	0.495	0.472
	Neural Network	0.762	0.667	0.663	0.672	0.667
	Random Forest	0.937	0.778	0.781	0.790	0.778
	average	**0.731**	**0.633**	**0.625**	**0.625**	**0.633**
NIRS	kNN	1.000	1.000	1.000	1.000	1.000
(λ=670nm)	Logistic Regression	0.542	0.417	0.362	0.333	0.417
	Naïve Bayes	0.771	0.750	0.743	0.778	0.750
	Neural Network	0.875	0.833	0.822	0.889	0.833
	Random Forest	1.000	1.000	1.000	1.000	1.000
	average	**0.838**	**0.800**	**0.785**	**0.800**	**0.800**
NMR-based traits	kNN	0.974	0.889	0.889	0.889	0.889
(T_1, T_2, A-ratio)	Logistic Regression	0.984	0.889	0.889	0.889	0.889
	Naïve Bayes	0.950	0.861	0.864	0.878	0.861
	Neural Network	0.918	0.889	0.889	0.889	0.889
	Random Forest	0.919	0.833	0.831	0.834	0.833
	average	**0.949**	**0.872**	**0.872**	**0.876**	**0.872**

Table A10: **Classification by ROC analysis for regions of origin.** Area Under the Curve (range between 0 to 1) of the various supervised models evaluated to predict region of origin. Models were validated using Leave-one-out method with averaged single-phase system traits(e.g., T_1, T_2 and A-ratio). The wavelength (λ) used for UV-Vis spectroscopy and NIRS was chosen upon the region of bigger differentiation (670nm peaks), full data is not shown.

Olive oil region	Model	AUC	CA	F1	Precision	Recall
UV-VIS	kNN	0.856	0.667	0.646	0.679	0.667
(λ=670 nm)	Logistic Regression	0.403	0.179	0.203	0.274	0.179
	Naïve Bayes	0.674	0.385	0.387	0.422	0.385
	Neural Network	0.751	0.641	0.62	0.684	0.641
	Random Forest	0.786	0.694	0.685	0.704	0.694
	average	**0.694**	**0.513**	**0.508**	**0.553**	**0.513**
NIRS	kNN	0.856	0.667	0.646	0.679	0.667
(λ=670nm)	Logistic Regression	0.43	0.179	0.131	0.127	0.179
	Naïve Bayes	0.687	0.385	0.383	0.382	0.385
	Neural Network	0.753	0.436	0.433	0.456	0.436
	Random Forest	0.752	0.641	0.625	0.635	0.641
	average	**0.696**	**0.462**	**0.444**	**0.456**	**0.462**
NMR-based traits	kNN	0.718	0.538	0.539	0.541	0.538
(T_1, T_2, A-ratio)	Logistic Regression	0.658	0.433	0.365	0.331	0.433
	Naïve Bayes	0.667	0.433	0.383	0.348	0.433
	Neural Network	0.788	0.576	0.561	0.571	0.576
	Random Forest	0.699	0.5	0.497	0.497	0.5
	average	**0.706**	**0.496**	**0.469**	**0.458**	**0.496**

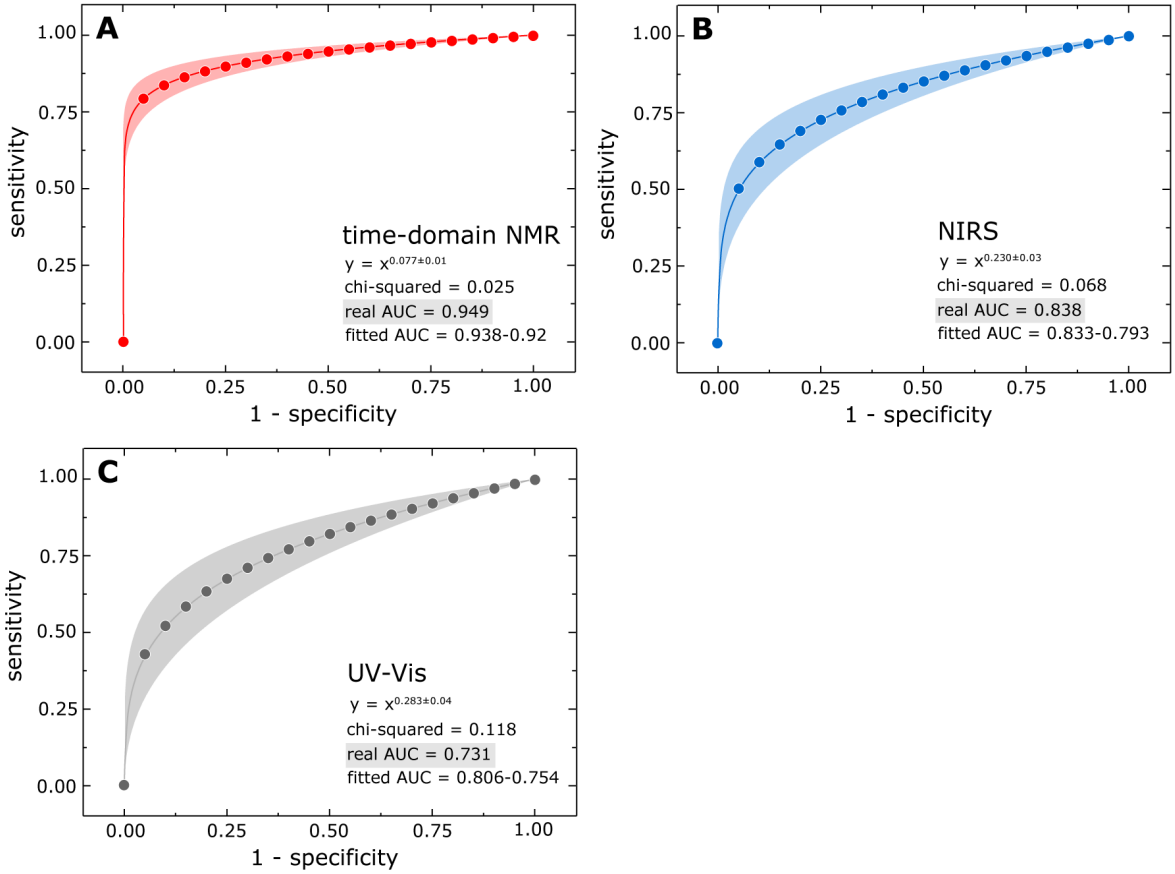

Figure A3: **Performance of olive oil classification by ROC analysis.** Classification of of live oil types (e.g., VOO, EVOO, refined OO) with (A) time-domain NMR (red), (B) NIRS (blue), and (C) UV-Vis spectroscopy (grey) techniques assisted by supervised models (e.g., kNN, Logistic Regression, Naïve Bayes, Neural Network and Random Forest) in Appendix, Table A9. The models were trained using the NMR-based traits (e.g., T_1, T_2 and A-ratio), NIRS (670nm peak) and UV-Vis (415nm peak) values of each sample. Power function fitting curves with confidence levels of 99% were used.