UNIVERSITY OF CONCEPCIÓN
FACULTY OF PHARMACY

NUCLEAR MAGNETIC RESONANCE (NMR) SPECTROSCOPY WITH MULTIVARIATE ANALYSIS AS A TRACEABILITY TOOL FOR RAW AND COMMERCIAL MATERIAL FOR AUTHENTICATION AND CLASSIFICATION PURPOSES IN THE SUGAR INDUSTRY

By: CRISTIAN ANDRES FUENTES ZEPEDA

Thesis submitted in accordance with the requirements of the Faculty of Pharmacy at the University of Concepción for the degree of doctor in Analytical Science and Technology

August 2025
Concepción, Chile

Professor Advisor: Dra. Rosario del Pilar Castillo Felices
Professor Co-Advisor: Dr. Mecit Halil Öztop

ACKNOWLEDGMENTS

I would like to exprees my deepest gratitude to my supervisors, Rosario del P. Castillo and Mecit Öztop, whose guidance, critical insights, and constant support have been crucial to the completion of this thesis.

With the support of the SuChAQuality project [#101008228], FONDECYT [1221287], FOVI [220172], and Beca Doctorado Nacional [21240532], this work was made possible through access to key resources and collaborative opportunities.

Also, the progress made in this research was also the result of the effort and insightful contributions of my colleagues and the lab team, Macarena, Pamela, David, Martin, and Mario.

Numerous discussions and suggestions were fundamental in shaping the direction of this thesis and I am sincerely grateful for them.

Not least, I owe a great to my family, especially my mother Cristina, for their continuous emotional support, patience, and understanding during the most challenging moments of this journey.

All the moments of doubt were alleviated by their encouragement and belief in me.

During this process, I also appreciated the small, but significant acts of kindness and support from those around me.

In particular, researches and professors who assisted me were essential to finish this work.

Every achievement reflected in this thesis would not have been possible without the effort and support of all involved.

Abstract

In the context of an increasingly globalized food system, the authenticity and traceability of food products have become critical factors for ensuring quality and origin. This is particularly relevant in the sugar industry, a key sector of the agro-industrial economy, which is highly susceptible to adulteration and production variability. Traditional quality parameters used in the evaluation of raw materials, commercial brown sugars, and derivative products such as pectin show a limited capacity to discriminate samples from different cultivation fields, botanical origin, or production method, and do not provide a detailed chemical profile.

The main objective of this Ph.D. thesis was to evaluate the potential of proton nuclear magnetic resonance (^{1}H NMR) spectroscopy combined with multivariate analysis as an alternative methodology for the quality assessment and authentication of sugar beet roots and commercial brown sugars. Complementarily, the potential of time-domain NMR supported by multivariate analysis was explored to differentiate pectin samples produced by distinct production methods. To reduce spectral complexity, extract meaningful spectral components, enhance predictive model performance, preserve essential spectroscopic information (e.g., signal multiplicity and coupling constants), and identify relevant variables, a processing strategy based on multivariate curve resolution (MCR) was implemented. The latter approach was complemented by exploratory and supervised models, which allowed effective discrimination according to the cultivation field of sugar beet roots, the botanical origin of commercial brown sugars and the method of production of pectin samples, even in the presence of simulated adulteration and highly similar spectral profiles. Furthermore, a novel graphical user interface, termed interval resonance analysis (InRA), was developed to flexibly automate critical processing steps in ^{1}H NMR spectroscopy (e.g., automated signal detection and signal decomposition via MCR), thus reducing manual intervention and analysis time, while promoting accessibility and for non-expert operators.

The combined use of ^{1}H NMR spectroscopy and TD-NMR with multivariate analysis proved to be a robust, selective and reproducible analytical tool for chemical characterization and authentication in the sugar industry, offering an effective alternative to conventional quality control and traceability methods within the agro-industrial sector.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Global overview of sugar production

Sugar plays a fundamental role in human nutrition and food product formulation. Among the naturally occurring simple sugars, the most relevant are glucose, fructose, and sucrose. Sucrose, commonly known as *table sugar* or *white sugar*, is the most widely consumed. Chemically, it is a low molecular weight disaccharide with the empirical formula $C_{12}H_{22}O_{11}$, composed of equimolar units of α -D-glucopyranosyl and β -D-fructofuranose linked by a glycosidic bond [1]. Sucrose is naturally present in a wide range of fruits and vegetables, where it serves as a major carbohydrate reserve, and is extensively extracted for use in the food industry.

Commercial sugar production is derived primarily from two botanical sources: sugar beet (*Beta vulgaris* L.) and sugarcane (*Saccharum officinarum* L.). In total, approximately 115 countries are involved in sugar production (Fig. 1.1.1), of which 67 produce sugar exclusively from sugarcane, 39 from sugar beet and 9 from both sources [2]. According to the International Sugar Organization (ISO), on average about 80% of global sugar production originates from sugarcane, while the remaining 20% are obtained from sugar beet [3, 4]. The five leading sugar-producing countries (India, Brazil, Thailand, Australia, and China) represent almost 40% of the total global production, Brazil being the largest producer and exporter of sugar worldwide [5].

Approximately, 95% of the sugar produced worldwide is intended for food-related applications, primarily for domestic consumption or as an ingredient in processed foods. In the food industry, sugar serves multiple functions, not only as a sweetener but also as a flavor enhancer, texture modifier, and fermentable substrate [6]. The remaining 5% of global sugar production is directed at non-food industrial uses, where it functions as a raw material in the manufacture of bio-ethanol, acetone or acetaldehyde [7]. Furthermore, due to its substantial caloric contribution ($\backsim$ 4 kilocalories per gram), human sugar consumption has increased steadily over the years. As a result, sugarcane and sugar beet have become two of the most important biomass sources in terms of production volume, collectively generating approximately 64 million tons annually [8].

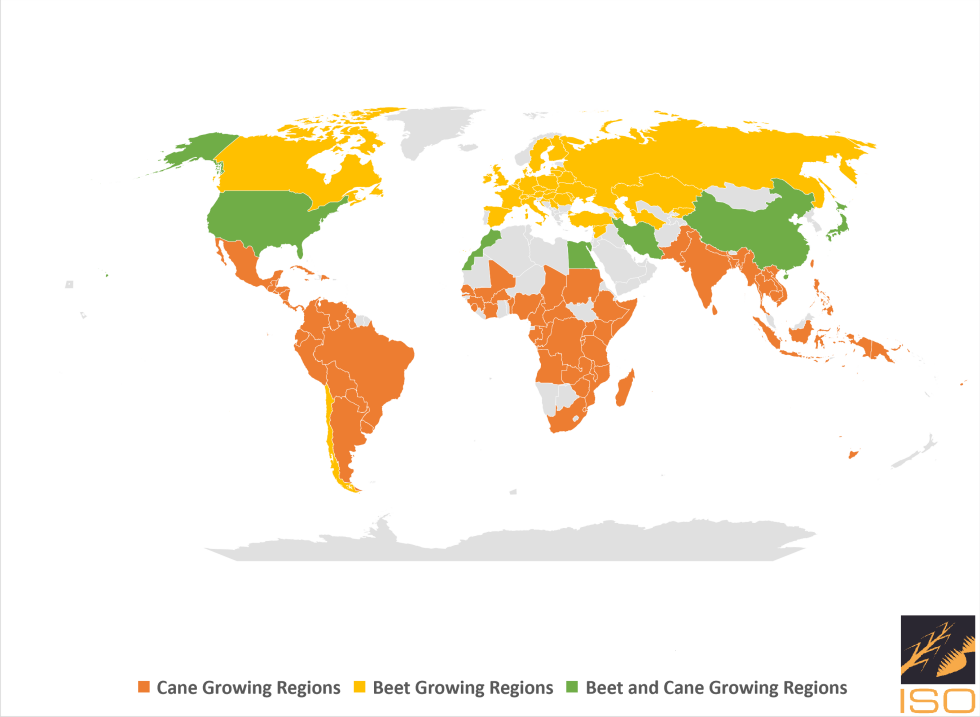

Figure 1.1.1: World sugarcane and sugar beet growing countries according to the *International Sugar Organization* (https://www.isosugar.org/sugarsector/sugar).

1.2 Sugar beet as a raw material

Despite the predominance of sugarcane in global sucrose production, sugar beet represents approximately 20% of total worldwide production. Currently, the European Union (EU) is the largest sugar beet producer, with 46% of the global sugar supply derived from beet [9]. In South America, Chile stands out as the only country that cultivates and processes sugar beet for industrial sugar production. In the national sense, sugar beet cultivation covers approximately 1,107 hectares per year, representing 15% of the total cultivated area of the country and is concentrated primarily in the south-central regions of Ñuble, Maule, Biobío and Araucanía [10].

As a cold-tolerant crop adapted to temperate climates, sugar beet offers several agronomic advantages over sugarcane, including a shorter growing cycle, a lower irrigation demand and a slightly higher sucrose content [11]. From a compositional point of view, sugar beet typically consists of 75% water, 18-21% sucrose, 2.5% non-sugar compounds and 5% fiber. In contrast, sugarcane contains approximately 72% water, 13-18% sucrose, 5% non-sugar compounds, and 10% fiber [7].

Sugar beet processing involves a standardized industrial workflow that includes mechanical harvesting, washing, cutting into "cossettes", and countercurrent extraction to obtain raw juice and beet pulp [12]. The juice is purified using milk of lime and CO_2 through liming, carbonation, and sulphitation, which partially removes non-sugar compounds. Subsequently, the juice is concentrated by multi-effect evaporation and crystallized to isolate sugar crystals, while molasses is recovered as a by-product [13]. However, in the context of sustainable agriculture and economic optimization of crop sources, a growing emphasis has been placed on exploiting the entire plant beyond its traditional role in sugar production. Therefore, sugar beet also serves as a valuable source of co-products, as illustrated in Fig. 1.2.1.

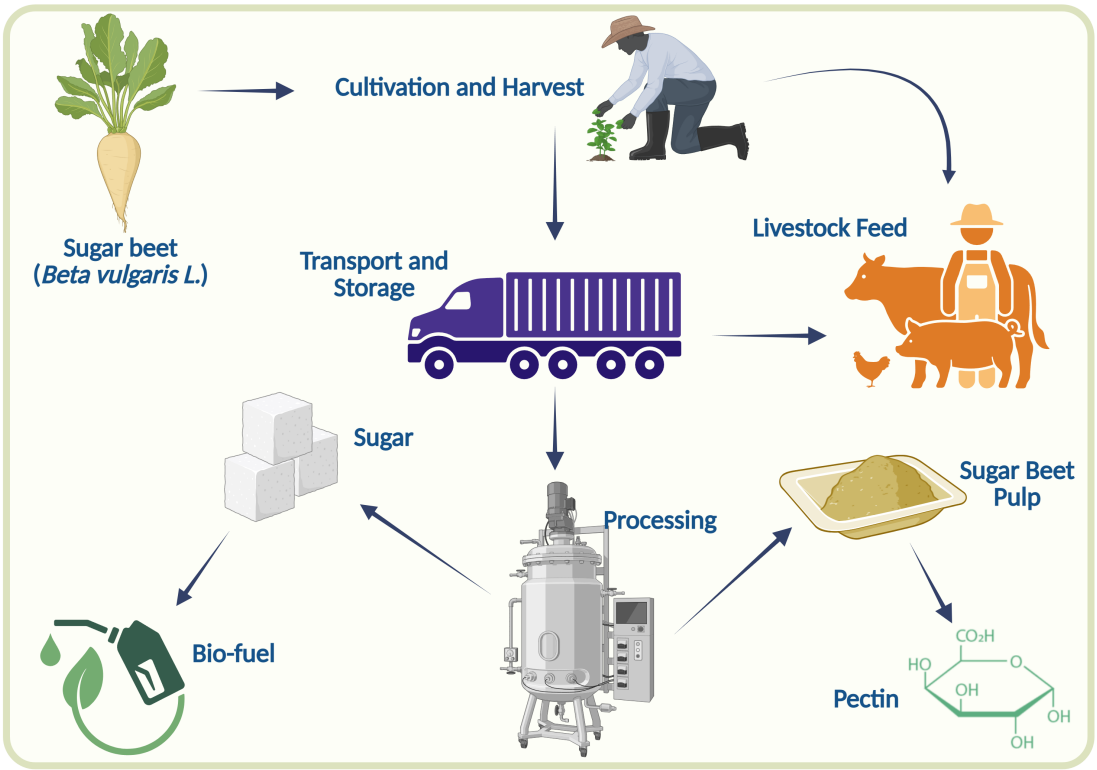

Figure 1.2.1: Schematic representation of the general uses of sugar beet.

One of the most widespread alternative uses of sugar beet is a livestock feed [14]. The residual sugar beet pulp (SBP), a by-product of sugar extraction, is especially valued due to its nutritional content, which includes approximately 9% proteins [15]. In addition, SBP has been explored as a source of bioactive compounds (e.g., phenolic compounds), which can be extracted and used for their antioxidant properties in food applications [16].

A crucial polysaccharide component of SBP is pectin, which possesses notable gelling and emulsifying properties. Pectin can be extracted and used in stabilizing systems, for example, in the suspension of anthocyanin pigments derived from berries [17, 18]. Furthermore, sugar beet has shown potential in renewable energy production, particularly through direct fermentation of sucrose into ethanol. This bio-fuel pathway often involves integrated processes that combine SBP storage, hydrolysis, and fermentation [19]. Moreover, cellulose isolated from SBP has been used in the paper industry, where the incorporation of micro-fibrillated cellulose into a gel matrix based on pectin improves the internal bond strength of paper products [20, 21].

1.3 Non-refined sugars

Refined sugar or white sugar (composed of 99.9% sucrose) is characterized by an exceptionally high degree of purity. Although conducive to its widespread application, it confers limited nutritional value [22]. Sucrose remains an indispensable component in the human diet, not only due to its organoleptic properties, but also to the preservation of food [23]. Although refined sugar continues to dominate consumption patterns in Europe and North America, non-refined sugars (commonly referred to as brown sugars) are still extensively produced and consumed in regions such as South America, Asia, and Africa [24]. The presence of these non-refined sugars in global markets has notably increased, a trend attributed to the growing demand for natural food products.

Brown sugar is a traditional sugar-derivative widely used in the food industry and cooking. Beyond its sweetness, it offers distinctive organoleptic properties, with "sweet" and "toasted" aromas playing a key role in consumer acceptability and palatability [25]. From a botanical perspective, brown sugar is derived mainly from sugarcane through non-centrifugal extraction and crystallization processes, giving rise to non-centrifugal sugar (NCS), in which sucrose crystals remain with molasses and, therefore, preserve natural minerals, vitamins and phytochemicals [26]. The main varieties of brown sugar include brown granulated sugar (BGS) and muscovado sugar (MS), as shown in Fig. 1.3.1. In contrast to NCS, BGS is a sugar obtained by a centrifugal process in which the crystallized sucrose is separated from the molasses. However, it retains a light yellow color due to the presence of small residual amounts of molasses that adhere to the crystals [27]. On the other hand, MS is considered a non-centrifugal sugar, characterized by a prolonged heat treatment that darkens the cane juice through caramelization and the Maillard reaction, giving it a significantly darker color compared to NCS [28].

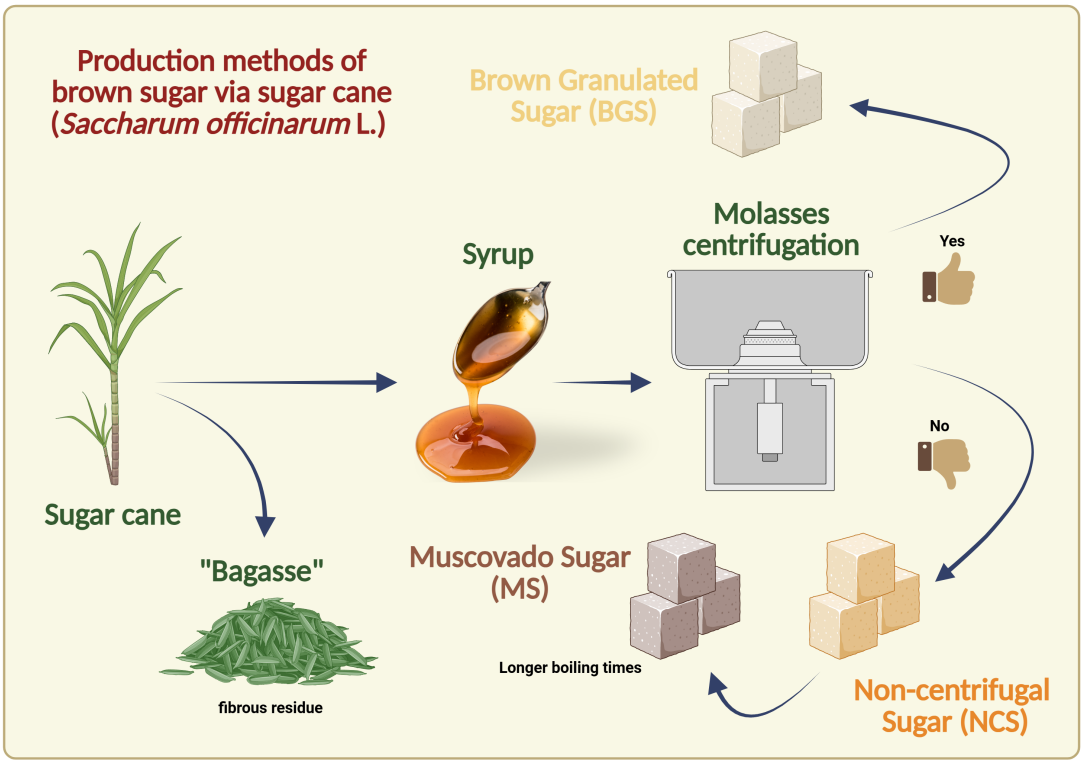

Figure 1.3.1: Typical methods of brown sugar production derived from sugarcane.

Alternatively, brown sugar can also be obtained by mixing refined beet sugar with cane molasses, giving organoleptic properties and a nutritional profile comparable to NCS. Furthermore, coconut blossom sugar (*Cocos nucifera* L.) has been adopted by consumers as an alternative to brown sugar, driven by the growing interest in healthy eating [29]. Coconut blossom is traditionally produced by evaporating the sap until it reaches its saturated liquid and forms a crystalline structure [30]. Its similarity in appearance and aroma profile, along with a glycemic index of approximately 35 versus brown sugar of 82, suggests a healthier option as it is a natural unrefined product [31].

1.4 Importance of traceability and authenticity in the sugar industry

1.4.1 Raw material

The agronomic and industrial performance of sugar beet is fundamentally influenced by two principal factors: root weight and sucrose concentration. The interplay of these variables determines the sugar yield per hectare, a crucial parameter that directly influences the economic returns of producers and the operational efficiency of the sugar processing industries [32]. From a technological standpoint, elevated sucrose concentrations facilitate more efficient extraction, reducing energy requirements and maximizing the amount of recoverable sugar [33].

However, sustaining high productivity in sugar beet cultivation remains a critical challenge under increasingly unpredictable environmental conditions. Variations in soil fertility, irrigation strategies, temperature and precipitation regimes can significantly affect the chemical composition of the crop, compromising the yield and processing quality [34]. Furthermore, heterogeneity in cultivation practices between regions impedes product traceability, particularly in the absence of standardized agronomic protocols.

A central issue in processing efficiency lies in the accumulation of non-sugar constituents in the root. Compounds such as amino acids and betaine, typically intensified by excessive nitrogen fertilization or abiotic stressors (e.g., drought), contribute significantly to the formation of molasses during juice purification.

These nitrogenous compounds are poorly removed during the clarification and crystallization steps, thereby reducing the recovery of sucrose and the overall process yield [35].

In addition, the presence of inorganic cations, e.g., sodium (Na^+) and potassium (K^+) further exacerbates the processing losses. These ions are classified as "melassigenic substances " due to their propensity to increase the volume of molasses and inhibit the crystallization of sucrose [35]. Their concentration in beet roots is closely linked to soil mineral composition and nutrient management strategies, underscoring the importance of balanced fertilization regimes and the appropriate selection of cultivars to optimize sugar recovery [36]. Post-harvest handling also plays a decisive role in the preservation of the technological quality of sugar beets [37]. Prolonged storage durations or exposure to elevated temperatures can catalyze the degradation of sucrose into invert sugars (i.e., glucose and fructose) and increase the concentrations of fermentation-promoting compounds, reducing the extraction efficiency [38].

Traditional quality parameters used in the evaluation of sugar beet, for example, the content of sucrose, soluble solids, moisture, dry matter, and mechanical properties, are limited in their ability to distinguish between samples originating from different geographic regions or cultivation fields [39]. Moreover, these methods do not provide a detailed chemical profile that would allow the identification of minority compounds associated with such differences. The latter shortcoming represents a critical barrier to effective product traceability and quality assurance. In an increasingly competitive global market, there is a pressing need to incorporate advanced analytical methodologies capable of assessing the chemical complexity of agri-products.

1.4.2 Brown sugar

In the context of an increasingly globalized food system, the authenticity of food products has become critical factors in ensuring transparency throughout the supply chain [40]. Brown sugar, in its various forms and botanical origins, exemplifies a product that could be highly susceptible to fraudulent practices, e.g., adulteration and mislabeling, often driven by economic incentives [41]. A documented case of brown sugar adulteration involves the deliberate addition of small quantities of cane sugar during the production of coconut sugar, to increase

the yield and facilitate the crystallization process [42]. In addition, coconut sugar has a high commercial value due to its distinctive nutritional profile and traditional processing methods. Its market price ranges from €15 to €46 per kilogram, in stark contrast to conventional refined sugar, which was priced at approximately €0.88 per kilogram in 2021 [43]. Although no consolidated official records are available on the prevalence of sugar adulteration in regions such as Latin America and Europe, internationally, sugar has been identified as the third most adulterated food product, with a reported incidence rate of 37% [44]. These practices not only undermine consumer trust, but also pose significant risks to food safety and compromise fair trade within the global marketplace.

Consumer demand for authentic products has increased significantly, particularly in markets where geographic origin directly influences perceived quality, commercial value, and cultural identity of food. Brown sugars derived from sugarcane (non-centrifugal sugar or muscovado sugar) and coconut blossom sugar retain phytochemicals, minerals and flavonoid compounds with antioxidant properties that confer distinctive organoleptic and nutritional characteristics compared to refined sugar (white) [22, 45]. This unique chemical profile constitutes a distinguishing attribute that may serve as a reliable marker of authenticity.

Moreover, brown sugars originating from organic sources or specific regions have gained commercial relevance through certification schemes such as geographical indications (GI), which recognize and protect the uniqueness of a product with specific characteristic, quality, or reputation attributed esentially to its geographical origin [46]. Furthermore, GI not only improves market value, but also plays a crucial role in mitigating food fraud [47]. Therefore, the availability of analytical methodologies capable of discriminating sugars based on their botanical origin is essential to substantiate authenticity claims on product labeling and to inform strategic decisions in producing countries, particularly those that cultivate sugar beet, sugarcane, or coconut sugar and face economic pressure to import more cost-effective raw materials.

1.4.3 Pectin

In the current context of the sugar industry, functional ingredients such as pectin have gained significant importance from a commercial perspective. Pectin is a high-value plant-derived biopolymer, recognized as safe by the Food and Drug Administration (FDA) and widely accepted under the concept of "clean-label" ingredients [48]. Sugar beet pectin, derived from SBP is characterized by its macromolecular complexity. It comprises distinct structural regions, including homogalacturonan (HG), rhamnogalacturonan I (RG-I) and rhamnogalacturonan II (RG-II), as well as a backbone of galacturonic acid substituted with side chains containing galactose, arabinose, rhamnose, glucose and xylose [49]. In contrast to other types of pectin (e.g., apple-derived), sugar beet pectin exhibits higher levels of methyl and acetyl groups, conferring enhanced emulsification capacity [50].

However, the functional properties of pectin are strongly influenced by the extraction conditions applied during its production. The most commonly used methods involve acid hydrolysis using inorganic acids (e.g., HCl, HNO_3), usually carried out at elevated temperatures ($70 - 100°C$) under controlled pH conditions ($1.5 - 2.0$) [51, 52]. Following extraction, pectin is recovered through alcohol precipitation, followed by successive washing and drying steps. Each of these stages directly affects key characteristics of the final product, including degree of esterification, solubility, molecular size, and structural conformation [53]. For instance, the pH and type of acid used govern the extent of side-chain hydrolysis, while extraction time and temperature influence the retention of functional groups such as methoxyl and acetyl moieties [54]. Additionally, the incorporation of processing aids, such as maltodextrin, into the drying method may alter the molecular mobility and solubility behavior of pectin in solution [55].

Structural and compositional variations resulting from different production methods can introduce substantial batch-to-batch variability, which poses challenges to industrial standardization and consistent product quality. Therefore, it is essential to adopt analytical strategies that not only characterize pectin at the molecular level, but also enable discrimination based on the specific extraction and purification methods employed. In this context, traceability becomes a key tool for verifying the integrity of the production process and for evaluating the impact of varying processing conditions on the final pectin properties.

1.5 Scientific need and analytical justification

Based on the background information, the molecular composition of food products is inherently complex and influenced by agronomic and environmental factors, which poses significant challenges in the development of representative chemical profiles. This complexity is particularly critical for the identification of minor compounds that are essential for assessing product quality, authenticity, and traceability. High-performance analytical techniques represent a suitable approach for developing sensitive, selective, and robust methodologies capable of addressing complex matrices.

The high concentration of sucrose in sugar beet roots and brown sugar represents a major analytical limitation, as it masks signals from compounds of low abundance. Although sucrose can be chemically removed, doing so may introduce bias, lead to loss of information, and increase the complexity of sample preparation. Therefore, it is essential to adopt a strategy that incorporates an analytical technique capable of capturing and preserving the full composition of the matrix while simultaneously enhancing the selectivity and sensitivity toward minor chemical compounds without the need to remove sucrose as the primary interference.

The following sections provide a general overview of nuclear magnetic resonance (NMR) spectroscopy in food analysis and highlight its relevance as an analytical technique for the development of this research.

1.5.1 NMR spectroscopy as a reliable analytical technique in food analysis

The most commonly used techniques for food quality evaluation are mass spectrometry (MS), typically coupled with liquid chromatography (LC) or gas chromatography (GC), near-infrared (NIR) spectroscopy, mid-infrared (MIR) spectroscopy, and NMR spectroscopy [56, 57, 58, 59, 60].

Nowadays, spectroscopic techniques have become well-established as rapid, non-destructive, and cost-effective tools to assess food quality. NIR and MIR spectroscopy has been widely applied in various analytical contexts because of their operational simplicity and the ability to analyze multiple samples with minimal or no preparation [61]. However, both techniques present limitations

in terms of sensitivity, especially for low-abundance compounds [62]. From a spectral perspective, MIR provides more specific functional information due to its reliance on fundamental vibrational transitions, while NIR, based on overtones and combination bands, produces broader and less specific signals [63]. The latter limits the direct assignment of spectral bands to individual molecules (particularly in NIR), reducing its suitability for the identification of low-abundance compounds with high structural specificity.

In contrast, NMR spectroscopy is a non-destructive analytical technique with adequate sensitivity (typically in the range of 1-10 μM) that enables the simultaneous detection and quantification of a wide range of compounds within a single assay [64]. Although NMR spectroscopy exhibits a sensitivity 10 to 100 times lower than LC-MS or GC-MS, it offers certain advantages over MS [65]. NMR is characterized by its excellent reproducibility on different spectrometers (regardless of the field strength), it requires little to no chromatographic separation, it does not require chemical derivatization, and it operates with minimal consumption of reagents, thus reducing the generation of hazardous waste [66]. Moreover, it is particularly suitable for the characterization of compounds that are analytically challenging for LC-MS, that is, sugars, organic acids, alcohols, polyols and other highly polar compounds [67].

The principle of NMR spectroscopy is based on the magnetic properties of certain atomic nuclei (e.g. ^{31}P), which, when exposed to a strong magnetic field, can absorb and emit radio-frequency radiation [68]. The resulting signals are processed via the Fourier transform to generate spectra that reflect the molecular composition of the sample. Therefore, NMR spectroscopy is particularly well-suited for the analysis of complex mixtures, i.e., food or agri-products, as it provides detailed structural information about chemical species and their molecular environment [69]. In addition, low-field NMR, specifically in the time domain (TD-NMR), has emerged as a complementary alternative. TD-NMR operates under lower magnetic fields, enabling rapid and accessible analysis of the physicochemical composition and molecular dynamics of complex systems [70]. Due to its lower cost, simple instrumentation, operational simplicity, and short acquisition times, TD-NMR has become a valuable tool for food quality control and process monitoring applications [71]. More detailed information on this technique is provided in Chapter 6.

NMR spectroscopy can be applied to solid or liquid samples; however, in food science, its application has predominantly focused on liquid-state analysis using high-resolution spectroscopy with superconducting magnets [72]. This approach produces high-representative spectra that allow the simultaneous and accurate detection of various chemical compounds. To date, proton NMR (^{1}H NMR) is the most widely used methodology to investigate chemical variability in food matrices due to its inherent ability to provide detailed structural profiles [73]. Consequently, ^{1}H NMR has played a key role in the analysis of several different food systems, e.g, fruits or dairy products, serving as a comprehensive fingerprinting tool to identify sources of variation and potential biomarkers for authenticity, traceability, and nutritional quality purposes [74, 75, 76].

To achieve accurate and reproducible interpretations in NMR studies, high-quality spectra are essential. The quality relies not only on the experimental design and instrumentation, but also on the implementation of appropriate spectral processing steps. Fig. 1.5.1 illustrates the standard workflow along with the computational tools commonly used in ^{1}H NMR analysis. The workflow includes initial operations applied directly to the spectrum, such as apodization, phase correction, baseline correction, and reference to an internal standard [77]. Subsequently, more general processing procedures are applied, including spectral alignment, normalization, binning (spectral segmentation) and/or scaling [78]. These steps aim to minimize artifacts, enhance spectral resolution, reduce spectral dimensionality, and ensure the comparability of resonance signals across samples.

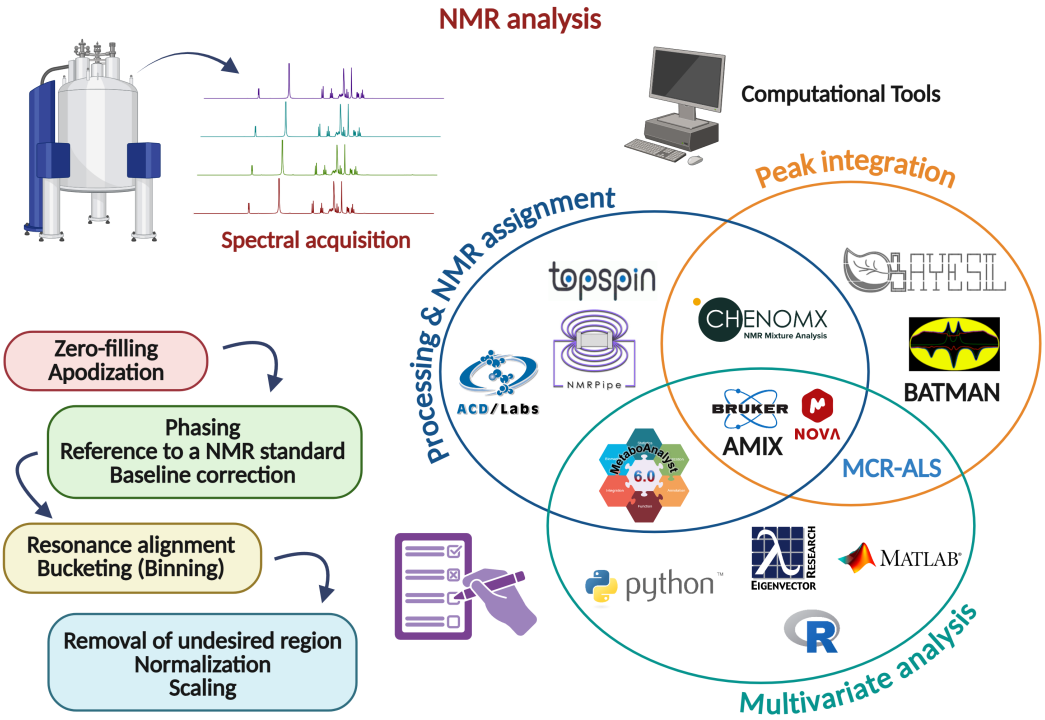

Figure 1.5.1: Common workflow of an ^{1}H NMR experiment applied to complex samples.

Moreover, the inherent complexity of ^{1}H NMR spectra, due to their high dimensionality, signal overlap, spectral noise, and frequency shifts poses challenges to direct interpretation. In response, several computational tools have been developed to streamline and automate the workload. As shown in Fig. 1.5.1, these tools can be broadly classified into three main categories: **Processing & NMR Assignment**, **Peak integration**, and **Multivariate Analysis**. Platforms such as Python and MATLAB have led to the widespread adoption of multivariate analysis as a crucial strategy to extract meaningful chemical information, which improves the sensitivity, selectivity and general applicability of the technique in the analysis of complex matrices [79]. Meanwhile, MestreNova and AMIX remain the most comprehensive tools currently available, both capable of spectral processing and signal integration, although their multivariate analysis capabilities remain limited to basic methods [80]. However, most of these tools address only a specific segment of the analytical workflow, necessitating the use of multiple platforms. This limitation has prompted the development of new open-access solutions that aim to enable comprehensive analyses within a unified environment, thereby enhancing the efficiency and reproducibility of the general process.

1.5.2 Multivariate analysis applied to ^{1}H NMR spectroscopy

Chemometrics (also known as multivariate analysis) has significantly transformed the application of NMR spectroscopy in the field of food science. The methodological integration mentioned above has elevated NMR from a purely compositional technique to a powerful analytical tool capable of providing holistic and unbiased evaluations, facilitating the exploration of complex relationships between the chemical profile of food matrices and their functional, nutritional, or sensory properties [69].

Multivariate analysis encompasses a wide range of mathematical and statistical methods applicable in various domains of chemistry, ranging from optimization of experimental parameters and experimental design to signal processing and the extraction of meaningful information through pattern recognition techniques [81]. In this context, multivariate analysis serves as a critical bridge between experimentally acquired spectral information and the derivation of interpretable and chemically relevant knowledge.

Fig. 1.5.2 shows the general organization of the spectral framework of multivariate techniques based on bilinear models, classified according to their analytical objectives. In a ^{1}H NMR study, it is common to apply a method called *binning* (or bucketing) to reduce spectral dimensionality, whereby each spectrum is divided into a defined number of "bins" of equal width (typically 0.04 ppm) [82].

Once binning has been applied, the spectral information is organized in a matrix $\mathbf{X}$ of dimensions $n \times k$, where n denotes the number of samples analyzed and k corresponds to the number of variables measured, such as the chemical shift points in each spectrum [79]. Each row represents a single NMR spectrum corresponding to a sample, while each column reflects a specific spectral variable. Alternatively, the matrix can be structured according to relative concentrations, with rows representing the samples and columns corresponding to the compound detected [83].

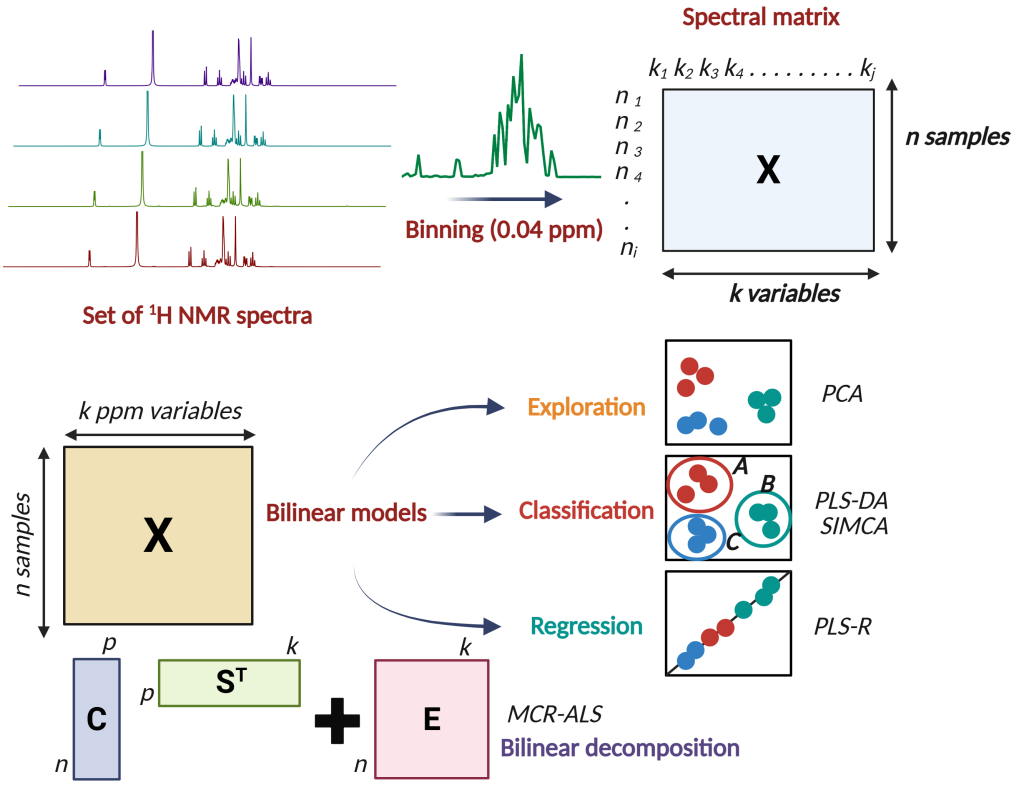

Figure 1.5.2: Framework of multivariate analysis techniques based on bilinear models categorized according to their analytical objectives.

Several methods are applied to the **X** matrix for the exploration, modeling, and classification of spectral information (Fig. 1.5.2). Principal component analysis (PCA) is a widely used unsupervised method applied during the exploratory phase. It enables dimensionality reduction, facilitates sample variability visualization, uncovers latent patterns, and identifies potential outliers, without requiring prior knowledge of class membership [84]. On the other hand, when response variables are available, supervised techniques are applied. Partial least-squares regression (PLS-R) is used to model linear relationships between spectral information and quantitative variables, making it suitable for constructing robust predictive models. For classification purposes, partial least squares discriminant analysis (PLS-DA) or soft independent modeling of class analogy (SIMCA) allows the discrimination of samples into predefined groups and the identification of spectral variables that contribute the most significantly to class separation [85, 86].

Finally, resonance signal decomposition through multivariate curve resolution with alternating least squares (MCR-ALS) has recently been adopted as a robust alternative for the resolution of spectral mixtures, without requiring prior knowledge of the pure component profiles [87]. As shown in Fig. 1.5.2, spectral information can be mathematically described by a bilinear model, in which **X** (or **D**) is decomposed into the product of the concentration matrix of the constituents **C**, a transposed matrix of pure component spectra $\mathbf{S}^T$, and a residual matrix **E** that captures information not explained by the model. **E** encompasses all disturbances not present in ideal ^{1}H NMR spectra, such as noise, instrumental artifacts, and solvent impurities.

1.6 Specialized food and sugar-related studies

According to a report by Lolli et al. (2024), the most studied food categories include honey and edible oils and fats (16%), followed by dairy products (14%), as well as fruits, vegetables, and beverages such as wine and fruit juices (12%) [88]. In the dairy sector, various studies have employed ^{1}H NMR fingerprinting in combination with multivariate analysis to assess critical variables such as composition, producing species, and feeding regime [89, 90]. In addition, this approach has been proven to be effective in authenticating the geographical origin of high-value food products such as coffee, wine, or honey [91, 92, 93].

Table 1.6.1 summarizes relevant studies focused on food and sugar-related analysis using multivariate approaches, highlighting the application of exploratory techniques such as PCA to assess the intrinsic variance of spectral datasets, as well as supervised classification models including PLS-DA, orthogonal PLS-DA (OPLS-DA), and k-nearest neighbors (k-NN) for purposes of authentication, geographical origin differentiation, adulteration detection, and evaluation of storage or stress conditions. Moreover, the most commonly used preprocessing method is binning, due to its simplicity and dimensionality reduction capabilities, whereas the use of resonance signal decomposition through MCR-ALS remains comparatively limited in these applications.

Table 1.6.1: Summary of relevant studies focused on foods and sugar-related through multivariate analysis.

Sample	Objective	Technique	Method	Model	Reference
Foods					
Walnuts	Geographical origin	^{1}H NMR	Binning	PCA	[94]
Onion	Storage comparison	^{1}H NMR	Binning	PCA, PLS-DA	[95]
Cocoa beans	Fermentation level	^{1}H NMR	Binning	PCA, PLS-DA	[96]
Lentils	Geographical origin	^{1}H NMR	–	PCA, k-NN	[97]
Vostizza currants	Traceability	^{1}H NMR	Binning	PCA, PLS-DA	[98]
Honey	Adulteration	^{1}H NMR	Binning	PCA, PLS-DA; k-NN	[99]
Persimmon	Geographical origin	^{1}H NMR	Binning	PCA, OPLS-DA; PLS-R	[100]
Yam	Genotype differentiation	^{1}H NMR	Binning	PCA, OPLS-DA	[101]
Beer	Product characterization	^{1}H NMR	MCR-ALS	PCA	[102]
Cheese	Geographical origin	^{1}H NMR	MCR-ALS	PCA	[103]
Trichosanthes	Product characterization	^{1}H NMR	MCR-ALS	PCA	[104]
Sugar-related					
Sugarcane	Geographical origin	^{1}H NMR	–	PCA	[105]
Sugar beet	Temporary stress	^{1}H NMR	Binning	PCA	[106]
Sugar beet	Resistance to *Cercospora*	^{1}H NMR	Binning	PCA, OPLS-DA	[107]
Red beetroot	Harvest and seasonality conditions	^{1}H NMR	Binning	PCA	[108]
Brown sugars	Adulterations	^{1}H NMR; UHPLC-Q-TOF-MS	Binning	PCA; SVM	[109]
Brown sugars	Type differentiation	GC-O-MS	–	PCA	[27]
Brown sugars	Geographical origin	3-D Fluorescence	–	PCA; LDA	[110]

Studies employing NMR spectroscopy for the characterization of sugar beet and commercially available brown sugars remain limited, mainly due to the analytical challenges posed by their high sucrose content. The high concentration of a dominant molecule often masks signals from minor but potentially informative compounds. However, selected investigations have demonstrated the applicability of NMR-based approaches in specific experimental contexts. For instance, Wedeking et al. (2018) used NMR spectroscopy to monitor metabolic changes in the leaves and roots of sugar beet plants subjected to a drought-re-watering cycle (13 days of water deprivation followed by 12 days of re-hydration). PCA revealed

that drought-induced alterations in primary metabolism, along with deterioration of the plant's water status, were largely reversible after re-watering. However, the recovered plants showed root metabolic profiles that remained distinguishable from those of non-stressed controls [106]. In a related study, Sekiyama et al. (2017) applied an NMR-based metabolomic strategy to assess differential resistance to *Cercospora* leaf spot (CLS) between 12 sugar beet genotypes in four stages of development. Through the PCA and OPLS-DA models, the authors identified significant variations in key metabolites, such as sucrose, betaine, and choline, establishing a correlation between the profiles of foliar metabolites and resistance levels to CLS [107]. Furthermore, Giampaoli et al. (2021) employed NMR spectroscopy to investigate how environmental conditions, harvest timing, and seasonality influence the phytochemical composition of red beet roots (*Beta vulgaris*). According to PCA, they observed that the choline and betaine content increased depending on the year of production, while the amino acid concentrations decreased. These findings suggest that such factors could serve as modulators of the phytochemical profile [108].

In the field of brown sugars, certain studies have investigated their metabolomic and sensory characterization using multivariate analysis in combination with various analytical techniques. Chen et al. (2021) compared volatile compounds in NCS, MS and BGS using gas chromatography-olfactometry coupled with mass spectrometry (GC-O-MS). Using PCA and OPLS-DA models, they successfully distinguished between the three sample groups, attributing the observed differences primarily to variations in Maillard reaction products, i.e., pyrazines, furanones and organic acids (formic and acetic) [27]. In a subsequent study, Chen et al. (2022) used three-dimensional fluorescence spectroscopy to examine brown sugars produced from sugarcane grown in three different provinces of China. A total of 80 active compounds were identified and OPLS-DA models were used to determine key compounds that contribute to the geographical differentiation of the samples [110]. In addition, Bachmann et al. (2022) analyzed the metabolic profiles of beet-derived brown sugar, unrefined cane sugar, and coconut blossom sugar using ^{1}H NMR spectroscopy in conjunction with UHPLC-Q-TOF-MS. This combined approach allowed the identification of sugar-specific metabolites: betaine for beet brown sugar, *trans*-aconitic acid for unrefined cane sugar, and pyroglutamic acid for coconut blossom sugar [109].

Reported studies have demonstrated that, when combined with multivariate analysis techniques, ^{1}H NMR spectroscopy holds considerable potential for the chemical characterization and authentication of high-value food products. However, its application to specific matrices, such as sugar beet roots and brown sugars, remains limited in terms of authentication and traceability.

Although the binning method has been widely used due to its efficiency in rapidly reducing spectral dimensionality, it also presents significant limitations. This approach entails a substantial loss of spectral resolution, which directly affects critical parameters such as resonance multiplicity. In addition, it may introduce undesired artifacts, such as the inclusion of signals from different resonances within the same bucket, thereby compromising the fidelity of the chemical information. It is important to note that each bucket cannot be assumed to represent an accurate resonance integral, as spectral noise and nearby signals are also captured within the interval.

Moreover, while resonance signal decomposition through MCR-ALS enables independent noise modeling and improves the contribution of meaningful signals independently, its application for classification purposes beyond PCA remains underexplored. The latter is due in part to the fact that the decomposition process can be labor-intensive and may not be readily accessible to less experienced users of multivariate resolution techniques.

1.7 References

[1] Yves Queneau, Slawomir Jarosz, Bartosz Lewandowski, and Juliette Fitremann. Sucrose Chemistry and Applications of Sucrochemicals. In *Advances in Carbohydrate Chemistry and Biochemistry*, volume 61, pages 217–292. 2007.

[2] Manish Kumar Chauhan, Varun, Sachin Chaudhary, Suneel Kumar, and Samar. Life cycle assessment of sugar industry: A review. *Renewable and Sustainable Energy Reviews*, 15(7):3445–3453, 2011.

[3] Balancing selection contributed to domestication of autopolyploid sugarcane (Saccharum officinarum L.). *Euphytica*, 209(2):477–493, 2016.

[4] Isa C. Ribeiro, Carla Pinheiro, Carla M. Ribeiro, Maria M. Veloso, Maria C. Simões-Costa, Isabel Evaristo, Octávio S. Paulo, and Cândido P. Ricardo. Genetic diversity and physiological performance of portuguese wild beet (beta vulgaris spp. maritima) from three contrasting habitats. *Frontiers in Plant Science*, 7(AUG2016):1–14, 2016.

[5] Hafiz Imran Fakhar, Adam Kasparek, Karol Kolodziejski, Leonid Grunin, Mecit Halil Öztop, Muhammad Qasim Hayat, Hussnain A. Janjua, and Danuta Kruk. Universal 1H Spin–Lattice NMR Relaxation Features of Sugar—A Step towards Quality Markers. *Molecules*, 29(11):1–20, 2024.

[6] H.-D. Belitz; W. Grosch; P. Schieberle. *Food Chemistry*. 2014.

[7] Ramesh Duraisam, Ketemaw Salelgn, and Abiyu Kerebo Berekete. Production of Beet Sugar and Bio-ethanol from Sugar beet and it Bagasse: A Review. *International Journal of Engineering Trends and Technology*, 43(4):222–233, 2017.

[8] Origins and evolution of the Western diet: Health implications for the 21st century. *American Journal of Clinical Nutrition*, 81(2):341–354, 2005.

[9] B M Muir. Development and Diversification of Sugar Beet in Europe. *Sugar Tech*, 24(4):992–1009, 2022.

[10] González Wong C. H, Fuentes-Barría, Muñoz Peña D, Aguilera Eguía R. Influencia de los compuestos bioactivos de betarraga (Beta vulgaris L)

sobre el efecto cardio-protector: Una revisión narrativa. *Revista chilena de nutrición*, 45(2):178–182, 2018.

[11] A Review on the Complete Utilization of the Sugarbeet. *Sugar Tech*, 16(4):339–346, 2014.

[12] O Rezbova, H; Belova, Anna; Skubna. Sugar beet production in the European Union and their future trends. *Agris on-line Papers in Economics and Informatics*, 5(4):165–178, 2015.

[13] M Rafik, H Qabli, S Belhamidi, F Elhannouni, A Elkhedmaoui, and A Elmidaoui. Review Article Membrane separation in the sugar industry. *Journal of Chemical and Pharmaceutical Research*, 7(9):653–658, 2015.

[14] A. Teimouri Yansari. Physically effectiveness of beet pulp-based diets in dairy cows as assessed by responses of feed intake, digestibility, chewing activity and milk production. *Journal of Animal Physiology and Animal Nutrition*, 98(1):158–168, 2014.

[15] Nutrient digestibility in sheep fed diets containing Roundup Ready or conventional fodder beet, sugar beet, and beet pulp. *Journal of Animal Science*, 83(2):400–407, 2005.

[16] Adel A.A. Mohdaly, Mohamed A. Sarhan, Iryna Smetanska, and Awad Mahmoud. Antioxidant properties of various solvent extracts of potato peel, sugar beet pulp and sesame cake. *Journal of the Science of Food and Agriculture*, 90(2):218–226, 2010.

[17] Marshall L. Fishman, Hoa K. Chau, Peter H. Cooke, Madhav P. Yadav, and Arland T. Hotchkiss. Physico-chemical characterization of alkaline soluble polysaccharides from sugar beet pulp. *Food Hydrocolloids*, 23(6):1554–1562, 2009.

[18] M. Buchweitz, A. Nagel, R. Carle, and D. R. Kammerer. Characterisation of sugar beet pectin fractions providing enhanced stability of anthocyanin-based natural blue food colourants. *Food Chemistry*, 132(4):1971–1979, 2012.

[19] Yi Zheng, Chaowei Yu, Yu Shen Cheng, Christopher Lee, Christopher W. Simmons, Todd M. Dooley, Ruihong Zhang, Bryan M. Jenkins, and Jean S.

VanderGheynst. Integrating sugar beet pulp storage, hydrolysis and fermentation for fuel ethanol production. *Applied Energy*, 93:168–175, 2012.

[20] Mária Fišerová, Juraj Gigac, and Rudolf Butaš. Influence of sugar beet pulp on bond strength and structure of paper. *Wood Research*, 52(3):59–74, 2007.

[21] Juraj Gigac, Mária Fišerová, and Michal Rosenberg. Improvement of paper strength via surface application of sugar beet pectin. *Chemical Papers*, 62(5):509–515, 2008.

[22] Jong Suk, Srinivasan Ramalingam, Il Guk, Ye Som, and Ashutosh Bahuguna. Comparative study of the physicochemical , nutritional , and antioxidant properties of some commercial re fi ned and non-centrifugal sugars. *Food Research International*, 109(December 2017):614–625, 2018.

[23] M A Harish Nayaka, U V Sathisha, M P Manohar, K B Chandrashekar, and Shylaja M Dharmesh. Cytoprotective and antioxidant activity studies of jaggery sugar. *Food Chemistry*, 115(1):113–118, 2009.

[24] Laura Calabuig-jim, Noelia Betoret, and Pedro Fito. Physicochemical and antioxidant properties of non-refined sugarcane alternatives to white sugar. *International Journal of Food Science and Technology*, pages 2579–2588, 2015.

[25] Jie Liu, Peng Wan, Caifeng Xie, and De Wei Chen. Brown sugar aroma: Key aroma-active compounds, formation mechanisms and influencing factors during processing. *Journal of Food Composition and Analysis*, 128(February):106076, 2024.

[26] Monthana Weerawatanakorn, Yonathan Asikin, Makoto Takahashi, Hajime Tamaki, Koji Wada, Chi Tang Ho, and Raweewan Chuekittisak. Physicochemical properties, wax composition, aroma profiles, and antioxidant activity of granulated non-centrifugal sugars from sugarcane cultivars of Thailand. *Journal of Food Science and Technology*, 53(11):4084–4092, 2016.

[27] Erbao Chen, Huanlu Song, Shuna Zhao, Chen Liu, Long Tang, and Yu Zhang. Comparison of odor compounds of brown sugar, muscovado sugar, and brown granulated sugar using GC-O-MS. *Lwt*, 142(January):111002, 2021.

[28] Yanjing Ge, Kai Li, Caifeng Xie, Yongshi Xu, Changrong Shi, Fangxue Hang, and William O.S. Doherty. Formation of volatile and aroma compounds during the dehydration of membrane-clarified sugarcane juice to non-centrifugal sugar. *Foods*, 10(7), 2021.

[29] Jasmin Wrage, Stephanie Burmester, Jürgen Kuballa, and Sascha Rohn. Coconut sugar (Cocos nucifera L.): Production process, chemical characterization, and sensory properties. *Lwt*, 112(February):108227, 2019.

[30] Bambang Nurhadi, Nandi Sukri, Wahyu Kristian Sugandi, Annisa Puteri Widanti, Resi Restiani, Ziske Noflianrini, Bayu Rezaharsamto, and Marleen Herudiyanto. Comparison of crystallized coconut sugar produced by traditional method and amorphous coconut sugar formed by two drying methods: Vacuum drying and spray drying. *International Journal of Food Properties*, 21(1):2339–2354, 2018.

[31] Muhammad Tuseef Asghar, Yus Aniza Yusof, Mohd Noriznan Mokhtar, Mohammad Effendy Ya'acob, Hasanah Mohd. Ghazali, Lee Sin Chang, and Yanty Noorzianna Manaf. Coconut (Cocos nucifera L.) sap as a potential source of sugar: Antioxidant and nutritional properties. *Food Science and Nutrition*, 8(4):1777–1787, 2020.

[32] Radosław Michał Gruska, Andrzej Baryga, Alina Kunicka-Styczyńska, Stanisław Brzeziński, Justyna Rosicka-Kaczmarek, Karolina Miśkiewicz, and Teresa Sumińska. Fresh and Stored Sugar Beet Roots as a Source of Various Types of Mono- and Oligosaccharides. *Molecules*, 27(16):1–10, 2022.

[33] Mahdi Hassani, Bahram Heidari, Ali Dadkhodaie, and Piergiorgio Stevanato. Genotype by environment interaction components underlying variations in root , sugar and white sugar yield in sugar beet (Beta vulgaris L .). *Euphytica*, 214(4):1–21, 2018.

[34] Olimpia Masetti, Angela Sorbo, and Luigi Nisini. Nmr tracing of food geographical origin: The impact of seasonality, cultivar and production year on data analysis. *Separations*, 8(12), 2021.

[35] Christa M Hoffmann and Bernward Märländer. Composition of harmful nitrogen in sugar beet (Beta vulgaris L .)— amino acids , betaine , nitrate

— as affected by genotype and environment. *European Journal of Agronomy*, 22:255–265, 2005.

[36] Christine Kenter and Christa M. Hoffmann. Changes in the processing quality of sugar beet (Beta vulgaris L.) during long-term storage under controlled conditions. *International Journal of Food Science and Technology*, 44(5):910–917, 2009.

[37] Radosław Nowicki and Edward Wilczewski. The Timing of Sugar Beet Harvesting Significantly Influences Roots Yield and Quality Characteristics. *Agronomy*, 15(3):1–20, 2025.

[38] K. Vukov and K. Hangypal. Sugar beet storage. *Sugar Technology Review*, 12:143–265, 1985.

[39] Dry matter and sugar content as parameters to assess the quality of sugar beet varieties for anaerobic digestion. *Sugar Industry*, 139(4):232–240, 2014.

[40] Vural Gökmen. Importance of Food Authentication and Origin Testing. *Food Chemistry: X*, 18(May):100708, 2023.

[41] Shuocong Li, Xiwen Yu, Zhenpeng Zhen, Minxing Huang, and Jianhua Lu. Geographical origin traceability and identification of refined sugar using UPLC-QTof-MS analysis. *Food Chemistry*, 348(March 2020):128701, 2021.

[42] Susanto B Sulistyo, Arief Sudarmaji, Pepita Haryanti, and Purwoko H Kuncoro. Information Processing in Agriculture A novel approach for detection of granulated coconut sugar adulteration using LED-based spectrometer and machine learning. *Information Processing in Agriculture*, (June), 2024.

[43] Ariana Saraiva, Conrado Carrascosa, Fernando Ramos, Dele Raheem, and Maria Lopes. Coconut Sugar : Chemical Analysis and Nutritional Profile ; Health Impacts ; Safety and Quality Control ; Food Industry Applications. *International Journal of Environmental Research and Public Health*, 20(3671), 2023.

[44] Kingsley O Iwuozor, Valentine Ugochukwu Anyanwu, Bashir Oladapo Olaniyi, Pamela S Mbamalu, and Adewale George Adeniyi. Adulteration of Sugar : A Growing Global Menace. *Sugar Tech*, 24(3):914–919, 2022.

[45] Simin Feng, Zisheng Luo, Yanbing Zhang, Zhou Zhong, and Baiyi Lu. Phytochemical contents and antioxidant capacities of different parts of two sugarcane (Saccharum officinarum L .) cultivars. *FOOD CHEMISTRY*, 151:452–458, 2014.

[46] Katerina Katerinopoulou, Achilleas Kontogeorgos, Constantinos E Salmas, Angelos Patakas, and Athanasios Ladavos. Geographical Origin Authentication of Agri-Food Products : A Review. *Foods*, pages 1–16, 2020.

[47] Xianqiao Hu, Lin Lu, Shuimei Li, Weixing Zhang, Yuntao He, and Mingxue Chen. Journal of Food Composition and Analysis Comparison of appearance quality , cooking quality , and nutritional quality of geographical indication rice and their application in geographical indication discrimination. *Journal of Food Composition and Analysis*, 135(July):106668, 2024.

[48] Varucha Misra, Santeshwari Srivastava, and Ashutosh Kumar Mall. *Sugar Beet Cultivation, Management and Processing*, volume 1. 2022.

[49] Yedidya Zagury, Shlomit David, Ravit Edelman, Roni Hazan Brill, and Yoav D. Livney. Sugar beet pectin as a natural carrier for curcumin, a water-insoluble bioactive for food and beverage enrichment: Formation and characterization. *Innovative Food Science and Emerging Technologies*, 74(October):102858, 2021.

[50] Pectin at the oil-water interface: Relationship of molecular composition and structure to functionality. *Food Hydrocolloids*, 68:211–218, 2017.

[51] Combined effects of independent variables on yield and protein content of pectin extracted from sugar beet pulp by citric acid. *Carbohydrate Polymers*, 129:108–114, 2015.

[52] Extraction, characterization and spontaneous emulsifying properties of pectin from sugar beet pulp. *Carbohydrate Polymers*, 98(1):750–753, 2013.

[53] Controlling the Degree of Esterification of Citrus Pectin for Demanding Applications by Selection of the Source. *ACS Omega*, 2(11):7991–7995, 2017.

[54] Chao Ai, Hecheng Meng, Jiawei Lin, Xiangyi Tang, and Xiaoming Guo. Emulsification properties of alkaline soluble polysaccharide from sugar

beet pulp: Effect of acetylation and methoxylation. *Food Hydrocolloids*, 124(PB):107361, 2022.

[55] Nikola Maravić, Zita Šereš, Ivana Nikolić, Petar Dokić, Szabolcs Kertész, and Ljubica Dokić. Emulsion stabilizing capacity of sugar beet fibers compared to sugar beet pectin and octenyl succinate modified maltodextrin in the production of O/W emulsions: individual and combined impact. *Lwt*, 108(February):392–399, 2019.

[56] Special Issue Reprint and Gianfranco Picone. *Development of Analytical Methods in the Field of Food Analysis*. 2024.

[57] Overview of the applications of liquid chromatography-mass spectrometry interfacing systems in food analysis: Naturally occurring substances in food. *Journal of Chromatography A*, 794(1-2):263–297, 1998.

[58] Huanlu Song and Jianbin Liu. GC-O-MS technique and its applications in food flavor analysis. *Food Research International*, 114(April):187–198, 2018.

[59] Gloria Domínguez-Rodríguez, Lidia Montero, Miguel Herrero, Alejandro Cifuentes, and María Castro-Puyana. Capillary electromigration methods for food analysis and Foodomics: Advances and applications in the period March 2021 to March 2023. *Electrophoresis*, 45(1-2):8–34, 2024.

[60] Marietta Fodor, Anna Matkovits, Eszter Luca Benes, and Zsuzsa Jókai. The Role of Near-Infrared Spectroscopy in Food Quality Assurance: A Review of the Past Two Decades. *Foods*, 13(21), 2024.

[61] Di Wu, Xiaojing Chen, Fang Cao, Da Wen Sun, Yong He, and Yanhui Jiang. Comparison of Infrared Spectroscopy and Nuclear Magnetic Resonance Techniques in Tandem with Multivariable Selection for Rapid Determination of ω-3 Polyunsaturated Fatty Acids in Fish Oil. *Food and Bioprocess Technology*, 7(6):1555–1569, 2014.

[62] Rizki Rachmad Saputra, Mokhamat Ariefin, Meiyanti Ratna Kumalasari, Unita Dongoran, Mulani Jeni Lestari Tampubolon, Putri Sulistiawati, Sri Yulandari Simangunsong, Risya Ariska, Pandu Gizta Rapi P, Amelia Siska, Jeddah Yanti, and Luluil Maknun. Advancements in NMR and IR Spectroscopy: Enhancing Metabolomics and Disease Diagnostics in the

Health Sector: A Comprehensive Review. *IJCA (Indonesian Journal of Chemical Analysis)*, 7(2):77–88, sep 2024.

[63] Lisa C. Heather, Xinzhu Wang, James A. West, and Julian L. Griffin. A practical guide to metabolomic profiling as a discovery tool for human heart disease. *Journal of Molecular and Cellular Cardiology*, 55(1):2–11, 2013.

[64] Study on infrared spectroscopy technique for fast measurement of protein content in milk powder based on LS-SVM. *Journal of Food Engineering*, 84(1):124–131, 2008.

[65] Ulrike Holzgrabe and Myriam Malet-Martino. NMR spectroscopy in pharmaceutical and biomedical analysis. *Journal of Pharmaceutical and Biomedical Analysis*, 93:1–2, 2014.

[66] Nicholas Dias Robert Cronin Yung Peng, Rose Khavari and 2 Dominique C. Hinshaw1, Lalita A. Shevde1. The integration of LC-MS and NMR for the analysis of low molecular weight trace analytes in complex matrices. *Mass Spectrometry Reviews*, 176(3):139–148, 2017.

[67] Katherine Huang, Natalie Thomas, Paul R. Gooley, and Christopher W. Armstrong. Systematic Review of NMR-Based Metabolomics Practices in Human Disease Research. *Metabolites*, 12(10), 2022.

[68] W. S. Veeman. Nuclear magnetic resonance, a simple introduction to the principles and applications. *Geoderma*, 80(3-4):225–242, 1997.

[69] Fenfen Tang, Morgan Vasas, Emmanuel Hatzakis, and Apostolos Spyros. *Magnetic resonance applications in food analysis*, volume 98. Elsevier Ltd., 1 edition, 2019.

[70] Donatella Capitani, Anatoly P. Sobolev, Valeria Di Tullio, Luisa Mannina, and Noemi Proietti. Portable NMR in food analysis. *Chemical and Biological Technologies in Agriculture*, 4(1):1–14, 2017.

[71] J. van Duynhoven, A. Voda, M. Witek, and H. Van As. *Time-domain NMR applied to food products*, volume 69. Elsevier Ltd., 1 edition, 2010.

[72] Ruge Cao, Xinru Liu, Yuqian Liu, Xuqing Zhai, Tianya Cao, Aili Wang, and Ju Qiu. Applications of nuclear magnetic resonance spectroscopy to the

evaluation of complex food constituents. *Food Chemistry*, 342(29):128258, 2021.

[73] Cristian A. Fuentes, David Montoya, Mecit Öztop, Macarena Rojas-Rioseco, Martin Bravo, Francisca González, and Rosario del P. Castillo. Interval resonance analysis (InRA): A versatile tool for automated untargeted 1H NMR fingerprinting – A case study in sugar beet field authentication. *Analytica Chimica Acta*, 1363(January), 2025.

[74] Soeren Wenck, René Bachmann, Sarah Marie Barmbold, Anna Lena Horns, Nele Paasch, and Stephan Seifert. Authentication of apples (Malus × domestica BORKH.) according to geographical origin, variety and production method using 1H NMR spectroscopy and random forest. *Food Control*, 167(August 2024), 2025.

[75] An untargeted metabolomics investigation of milk from dairy cows with clinical mastitis by 1H-NMR. *Foods*, 10(8):1–10, 2021.

[76] A. Caligiani, D. Acquotti, G. Palla, and V. Bocchi. Identification and quantification of the main organic components of vinegars by high resolution 1H NMR spectroscopy. *Analytica Chimica Acta*, 585(1):110–119, 2007.

[77] NMR-based metabolomics: From sample preparation to applications in nutrition research. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 83:42–49, 2014.

[78] Katherine Huang, Natalie Thomas, Paul R. Gooley, and Christopher W. Armstrong. Systematic Review of NMR-Based Metabolomics Practices in Human Disease Research. *Metabolites*, 12(10), 2022.

[79] Chemometric Analysis of NMR Spectroscopy Data: A Review. *Annual Reports on NMR Spectroscopy*, 54:41–80, 2004.

[80] NMR-based metabolomics: Where are we now and where are we going? *Progress in Nuclear Magnetic Resonance Spectroscopy*, 150-151(August 2024):101564, 2025.

[81] Barry Lavine and Jerry Workman. Chemometrics. *Analytical and Bioanalytical Chemistry*, 78(12):4137–4145, 2006.

[82] Tim De Meyer, Davy Sinnaeve, Bjorn Van Gasse, Elena Tsiporkova,

Ernst R Rietzschel, Marc L De Buyzere, Thierry C Gillebert, Sofie Bekaert, José C Martins, and Wim Van Criekinge. NMR-Based Characterization of Metabolic Alterations in Hypertension Using an Adaptive , Intelligent Binning Algorithm. *Analytical Chemistry*, 80(10):3783–3790, 2008.

[83] Francesc Puig-Castellví, Ignacio Alfonso, and Romà Tauler. Untargeted assignment and automatic integration of 1H NMR metabolomic datasets using a multivariate curve resolution approach. *Analytica Chimica Acta*, 964:55–66, 2017.

[84] Guy Beavis, Martin Rusilowicz, James Donarski, Adrian Charlton, and Julie Wilson. *Chemometrics Applied to Nuclear Magnetic Resonance Analysis*. 2019.

[85] Raúl González-Domínguez, Ana Sayago, and Ángeles Fernández-Recamales. An Overview on the Application of Chemometrics Tools in Food Authenticity and Traceability. *Foods*, 11(23):1–13, 2022.

[86] Vi Khanh Truong, Madeleine Dupont, Aaron Elbourne, Sheeana Gangadoo, Piumie Rajapaksha Pathirannahalage, Samuel Cheeseman, James Chapman, and Daniel Cozzolino. From academia to reality check: A theoretical framework on the use of chemometric in food sciences. *Foods*, 8(5), 2019.

[87] Francesc Puig-Castellví, Ignacio Alfonso, and Romà Tauler. Untargeted assignment and automatic integration of 1H NMR metabolomic datasets using a multivariate curve resolution approach. *Analytica Chimica Acta*, 964:55–66, 2017.

[88] Veronica Lolli and Augusta Caligiani. How nuclear magnetic resonance contributes to food authentication: current trends and perspectives. *Current Opinion in Food Science*, 58:101200, 2024.

[89] Application of metabolomics to assess milk quality and traceability. *Current Opinion in Food Science*, 40:168–178, 2021.

[90] Ilaria Lanza, Veronica Lolli, Severino Segato, Augusta Caligiani, Barbara Contiero, Alessandro Lotto, Gianni Galaverna, Luisa Magrin, and Giulio Cozzi. Use of GC–MS and 1H NMR low-level data fusion as an advanced and comprehensive metabolomic approach to discriminate milk from dairy chains

based on different types of forage. *International Dairy Journal*, 123:105174, 2021.

[91] Sabur Badmos, Seung Hun Lee, and Nikolai Kuhnert. Comparison and quantification of chlorogenic acids for differentiation of green Robusta and Arabica coffee beans. *Food Research International*, 126(April), 2019.

[92] Louis Gougeon, Gregory da Costa, Tristan Richard, and François Guyon. Wine authenticity by quantitative1H NMR versus multitechnique analysis: A case study. *Food Analytical Methods*, 12(4):956–965, 2019.

[93] Chantelle Spiteri, Frederick Lia, and Claude Farrugia. Determination of the geographical origin of maltese honey using 1H NMR fingerprinting. *Foods*, 9(10):1–18, 2020.

[94] Caroline Schmitt, Tobias Schneider, Laura Rumask, Markus Fischer, and Thomas Hackl. Food Profiling: Determination of the Geographical Origin of Walnuts by 1H NMR Spectroscopy Using the Polar Extract. *Journal of Agricultural and Food Chemistry*, 68(52):15526–15534, 2020.

[95] Gabriella Saviano, Debora Paris, Dominique Melck, Francesca Fantasma, Andrea Motta, and Maria Iorizzi. Metabolite variation in three edible Italian Allium cepa L. by NMR-based metabolomics: a comparative study in fresh and stored bulbs. *Metabolomics*, 15(8), 2019.

[96] Augusta Caligiani, Luigi Palla, Domenico Acquotti, Angela Marseglia, and Gerardo Palla. Application of 1H NMR for the characterisation of cocoa beans of different geographical origins and fermentation levels. *Food Chemistry*, 157:94–99, 2014.

[97] Francesco Longobardi, Valentina Innamorato, Annalisa Di Gioia, Andrea Ventrella, Vincenzo Lippolis, Antonio F. Logrieco, Lucia Catucci, and Angela Agostiano. Geographical origin discrimination of lentils (Lens culinaris Medik.) using1H NMR fingerprinting and multivariate statistical analyses. *Food Chemistry*, 237:443–448, 2017.

[98] Maria Eleni Dimitrakopoulou, Konstantina Matzarapi, Styliani Chasapi, Apostolos Vantarakis, and Georgios A. Spyroulias. Nontargeted 1H NMR fingerprinting and multivariate statistical analysis for traceability of Greek PDO Vostizza currants. *Journal of Food Science*, 86(10):4417–4429, 2021.

[99] Elisangela F. Boffo, Leila A. Tavares, Antonio C.T. Tobias, Márcia M.C. Ferreira, and Antonio G. Ferreira. Identification of components of Brazilian honey by 1H NMR and classification of its botanical origin by chemometric methods. *Lwt*, 49(1):55–63, 2012.

[100] Shoraku Ryu, Tomonari Muramatsu, Kazuo Furihata, Feifei Wei, Masanori Koda, Takuya Miyakawa, and Masaru Tanokura. NMR-based metabolic profiling and comparison of Japanese persimmon cultivars. *Scientific Reports*, 9(1):1–10, 2019.

[101] Qiang Wang, Xiaohui Wang, Xujin Wu, Yanli Wang, Yuli Zhang, Yaling Jiang, Chaonan Zhang, Xiaowan Huang, Li An, Huan Ma, and Kaihong Xu. 1H NMR-based metabolic profiling approach to identify the geo-authentic Chinese yam (Dioscorea polystachya Turczaninow cv. Tiegun). *Journal of Food Composition and Analysis*, 98(January), 2021.

[102] Nicola Cavallini, Francesco Savorani, Rasmus Bro, and Marina Cocchi. A Metabolomic Approach to Beer Characterization. *Molecules*, 26(5):1–15, 2021.

[103] N Cavallini, L Strani, P P Becchi, V Pizzamiglio, S Michelini, and F Savorani. Tracing the identity of Parmigiano Reggiano " Prodotto di Montagna - Progetto Territorio " cheese using NMR spectroscopy and multivariate data analysis. *Analytica Chimica Acta*, 1278(April):341761, 2023.

[104] Jiayu Yang, Wenzhu Li, Jinguo Ding, Ying Dong, Fang Zhao, Jianyang Pan, and Haibin Qu. A multivariate curve resolution-alternating least squares (MCR-ALS) technology assisted 1 H-NMR methodology for multi-component quantitation of Trichosanthis Pericarpium injection. *Pythochemical Analysis*, 34(August 2022):40–47, 2023.

[105] Elenilson G. Alves Filho, Lorena M.A. Silva, Rafael Choze, Luciano M. Lião, Neli K. Honda, and Glaucia B. Alcantara. Discrimination of sugarcane according to cultivar by 1H NMR and chemometric analyses. *Journal of the Brazilian Chemical Society*, 23(2):273–279, 2012.

[106] 1H-NMR metabolomic profiling reveals a distinct metabolic recovery response in shoots and roots of temporarily drought-stressed sugar beets. *PLoS ONE*, 13(5):1–21, 2018.

[107] Yasuyo Sekiyama, Kazuyuki Okazaki, Jun Kikuchi, and Seishi Ikeda. NMR-based metabolic profiling of field-grown leaves from sugar beet plants harbouring different levels of resistance to Cercospora leaf spot disease. *Metabolites*, 7(1), 2017.

[108] Ottavia Giampaoli, Fabio Sciubba, Giorgia Conta, Giorgio Capuani, Alberta Tomassini, Giorgio Giorgi, Elisa Brasili, Walter Aureli, and Alfredo Miccheli. Red beetroot's nmr-based metabolomics: Phytochemical profile related to development time and production year. *Foods*, 10(8):1–12, 2021.

[109] René Bachmann, Anna Lena Horns, Nele Paasch, Robbin Schrieck, Markus Weidner, Iris Fransson, and Jan Philipp Schrör. Minor metabolites as chemical marker for the differentiation of cane, beet and coconut blossom sugar. From profiling towards identification of adulterations. *Food Control*, 135(December 2021), 2022.

[110] Jui Yi Chen, Xin Wei Chen, Yu Yu Lin, Gow Chin Yen, and Jer An Lin. Authentication of dark brown sugars from different processing using three-dimensional fluorescence spectroscopy. *Lwt*, 150(June):111959, 2021.

Chapter 2

Hypothesis and Objectives

2.1 Hypothesis

Spectral decomposition through multivariate analysis could enhance the selectivity and discriminating capacity of ^{1}H NMR spectroscopy towards highly similar spectral profiles for traceability purposes in the sugar industry.

2.2 Objectives

2.2.1 Main objective

Evaluate the potential of ^{1}H NMR spectroscopy combined with multivariate analysis as an alternative methodology for quality assessment and authentication related to the discrimination of sugar beet (*Beta vulgaris L.*) roots and commercial brown sugars.

2.2.2 Specific objectives

1. To propose and optimize an MCR-based processing methodology aimed at improving the selectivity and analytical applicability of ^{1}H NMR spectroscopy in the development of robust unsupervised and supervised models.
2. To establish exploratory and classification models capable of revealing intrinsic clustering patterns and effectively discriminating between samples from sugar beet cultivation fields and commercial brown sugars.
3. To define the representative spectral profile of hydrophilic extracts by characterizing the predominant polar chemical compounds present in sugar beet roots and commercial brown sugars using ^{1}H NMR spectroscopy.
4. To identify and assign the compounds statistically responsible for sample discrimination and to assess their potential relevance as markers of quality or origin.

Chapter 3

Analytical Strategy

3.1 Scope of the study

To address the objectives outlined in this Ph.D. thesis, a comprehensive analytical strategy was developed to enhance the applicability of ^{1}H NMR spectroscopy in terms of selectivity and sensitivity. This approach is proposed as a versatile alternative for addressing analytical challenges related to authentication and traceability in the sugar industry, as well as in other potentially related applications.

The analytical strategy focused primarily on two study matrices: **sugar beet roots** harvested from three cultivation fields with comparable pedoclimatic conditions and **commercial brown sugars** derived from three distinct botanical origins (sugar beet, sugarcane, and coconut) through ^{1}H NMR spectroscopy. In addition, **pectin** samples produced using various industrial-scale methods were included as a complementary study using TD-NMR. Details on sample acquisition criteria are provided in Sections 4.2.2, 5.3.2, and 6.2.1, respectively.

For sugar beet roots and commercial brown sugars, spectral sets of ^{1}H NMR were generated and combined with multivariate analysis. To maximize the extraction of relevant chemical differences between samples, a solid-liquid extraction focused on stable polar compounds (hydrophilic), i.e., carbohydrates, organic acids, and amino acids, was chosen, given their higher signal dispersion and stability under neutral pH conditions compared to those of low-polarity compounds. Furthermore, deuterium oxide (D_2O) was used as a solvent to ensure spectral reproducibility and minimize unwanted variations related to chemical shifts. Furthermore, sets of relaxation curves acquired by two complementary pulse sequences (Carr-Purcell-Meiboom-Gill [CPMG] and Solid Echo–Magic Sandwich Echo [SE-MSE]) were analyzed through multivariate analysis to improve sensitivity to structural heterogeneity associated with the different pectin samples.

Figure 3.1.1 illustrates a schematic summary of the analytical strategy adopted to address each objective and its link to the corresponding scientific publications derived from the study.

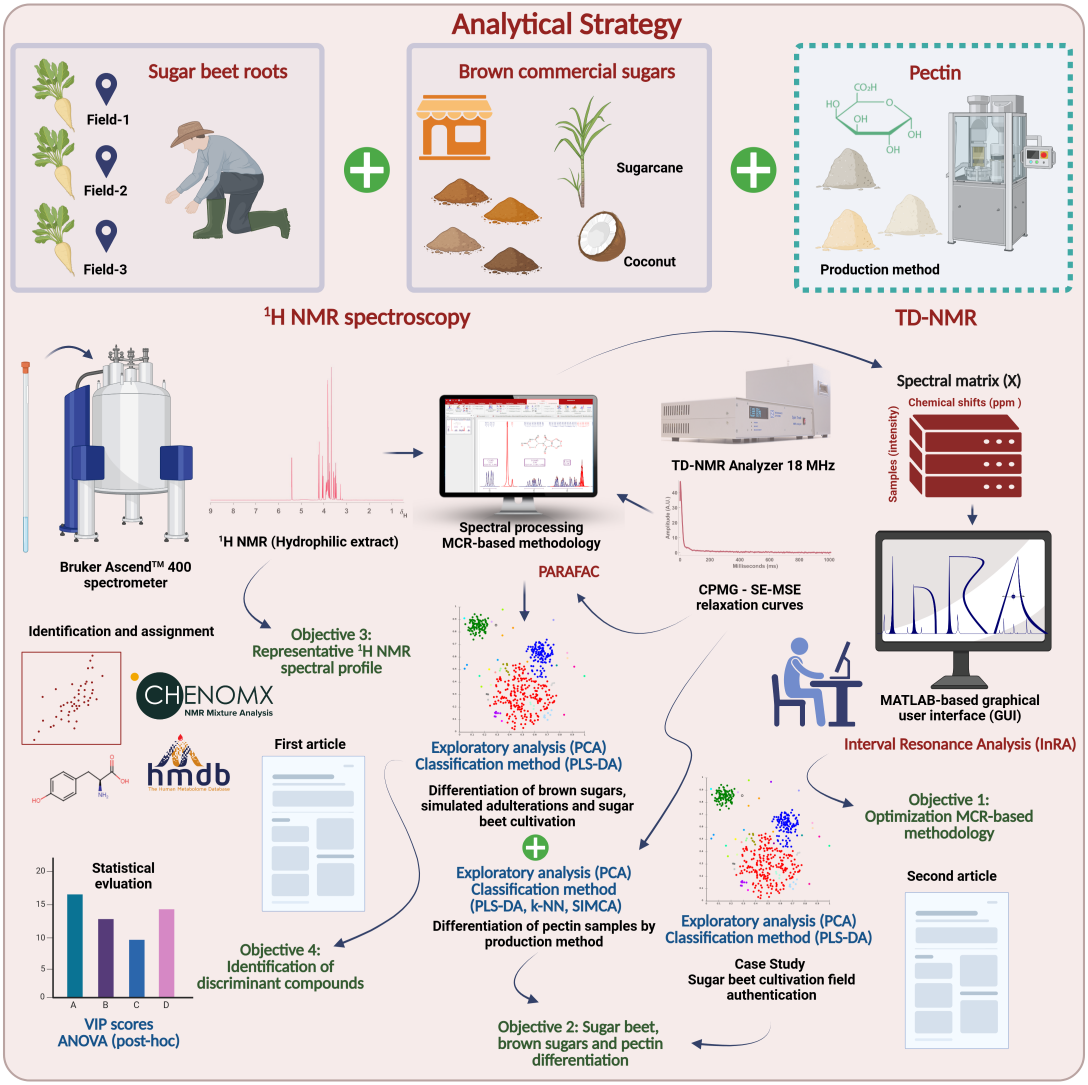

Figure 3.1.1: Schematic overview of the analytical strategy implemented to address the specific objectives of the Ph.D. thesis.

3.2 MCR-based processing methodology

In order to improve the selectivity and analytical applicability of ^{1}H NMR spectroscopy in the development of robust unsupervised and supervised models, an alternative approach based on the bilinear decomposition of individual resonance signals was adopted using the MCR-ALS algorithm. Pure spectral and concentration profiles associated with the individual resonance signals were resolved, preserving the structural relationships between correlated signals and thus improving the chemical interpretation of the system.

Due to the high dependence on manual intervention and time-consuming procedures in the workflow, a graphical user interface (GUI) called Interval Resonance Analysis (InRA) was developed to optimize and automate crucial stages of spectral processing. Specialized algorithms for alignment, automatic detection of resonance signals, simultaneous decomposition of MCR models, and exploratory analysis were incorporated into InRA. Its performance was validated using benchmark tools and its applicability was demonstrated through a case study focused on the discrimination of sugar beet root from different cultivation fields. Section 5.2 details the operational framework of InRA throughout its different stages.

3.3 Differentiation of sugar beet, brown sugar and pectin samples

To investigate patterns of variation, promote the between-group variability, and assess the discriminative power of extracted chemical profiles, a structured approach was implemented in two complementary phases: an unsupervised exploratory phase and a supervised classification phase. In the first phase, the concentration profiles obtained through MCR-ALS (and PARAFAC applied exclusively to TD-NMR) were analyzed using PCA to explore the clustering patterns among the different sample groups and identify trends associated with the botanical origin of brown sugars, the cultivation field of sugar beet roots, and the production method of the pectin samples. In the second phase, classification models were built through PLS-DA using as input variables the same concentration profiles generated by MCR (and PARAFAC) to promote better separation and assess discriminatory capacities between predefined classes. For TD-NMR, additional supervised models, including PLS-DA, k-NN, and SIMCA, were developed.

The obtained results were contrasted with conventional methodologies based on spectral binning for ^{1}H NMR spectroscopy and direct analysis of relaxation curves for TD-NMR, to qualitatively and quantitatively evaluate the performance of the models in terms of clustering and classification rate and to highlight the improvements obtained through the proposed methodology based on MCR-ALS (or PARAFAC).

3.4 Representative ^{1}H NMR spectral profile

Characterization of the predominant compounds present in hydrophilic extracts of sugar beet roots and commercial brown sugars with the aim of establishing a representative spectroscopic profile and identifying possible relevant chemical compounds for authentication and quality assessment purposes.

Spectral signal assignment was conducted through an integrative strategy that included a comparison with experimental ^{1}H NMR spectra previously reported in relevant studies, the use of two-dimensional (2D-NMR) spectra to confirm structural correlations, and the application of statistical analysis by spectral correlation (STOCSY) to link resonance signals with specific compounds. The Chenomx NMR Suite 8.3 reference database (https://www.chenomx.com) was also used to validate assignments by direct matching with standard spectra. In addition, the Human Metabolome Database (HMDB, https://www.hmdb.ca) was consulted to corroborate the presence of compounds commonly reported in the plant metabolome (e.g., amino acids).

3.5 Identification of discriminant compounds

To identify the most relevant variables for discrimination between samples, an analysis was performed based on the variable importance in the projection (VIP) obtained from the PLS-DA models to evaluate the contribution of each variable to the predictive model, considering those that exceeded a predefined threshold as significant. Furthermore, to support the statistical validity of the identified variables, a one-way analysis of variance (ANOVA) with *post-hoc* tests was applied in order to select only those variables that were statistically representative and, at the same time, control the possible errors associated with multiple comparisons.

The interpretation and assignment of statistically significant variables were related to specific chemical compounds present in the analyzed extracts, evaluating their potential as quality or traceability indicators based on their origin and possible influences derived from the cultivation or production processes.

Chapter 4

Application of segmented analysis via multivariate curve resolution with alternating least squares to ^{1}H-nuclear magnetic resonance spectroscopy to identify different sugar sources

This chapter discusses the results that served as the basis for the first article, titled "Application of segmented analysis via multivariate curve resolution with alternating least squares to ^{1}H-nuclear magnetic resonance spectroscopy to identify different sugar sources", published in July 2023 in the journal *Food Chemistry* (https://doi.org/10.1016/j.foodchem.2023.136817).

The article addressed specific objectives 3 and 4 by performing a chemical characterization of sugar beet roots and commercial brown sugars. The crucial compounds involved in the differentiation of each sample group were also identified and their potential as quality-related markers was evaluated.

In addition, objectives 1 and 2 were partially included by demonstrating the feasibility of using MCR-ALS as a spectral data processing tool to enhance the selectivity of supervised models aimed at sample discrimination (rather than classification), which produced promising results. However, the latter approach also demonstrated a highly manual and labor-intensive process, which may compromise the reproducibility and efficiency of the analysis as a viable alternative methodology.

Application of segmented analysis via multivariate curve resolution with alternating least squares to ^{1}H-nuclear magnetic resonance spectroscopy to identify different sugar sources

Cristian A. Fuentes[a,b], Mecit Halil Öztop[c], Macarena Rojas-Rioseco[a,b], Martín Bravo[a,b], Aylin Özgür Göksu[d], Marena Manley[e,*], Rosario del P. Castillo[a,b]

[a]Departamento de Análisis Instrumental, Facultad de Farmacia, Universidad de Concepción, Concepción 4070386, Chile
[b]Laboratorio de Bioespectroscopia y Quimiometría (BioSpeQ), Centro de Biotecnología, Universidad de Concepción, Concepción 4070386, Chile
[c]Department of Food Engineering, Middle East Technical University, Ankara 06800, Türkiye
[d]Kayseri Sugar R&D Center, Kayseri Sugar Factory, Kayseri 38070, Türkiye
[e]Deparment of Food Science, Stellenbosch University, Private Bag X1, Matieland (Stellenbosch) 7602, South Africa

***Corresponding author:**
Marena Manley
Tel: +27 21 808 3511
Email: mman@sun.ac.za

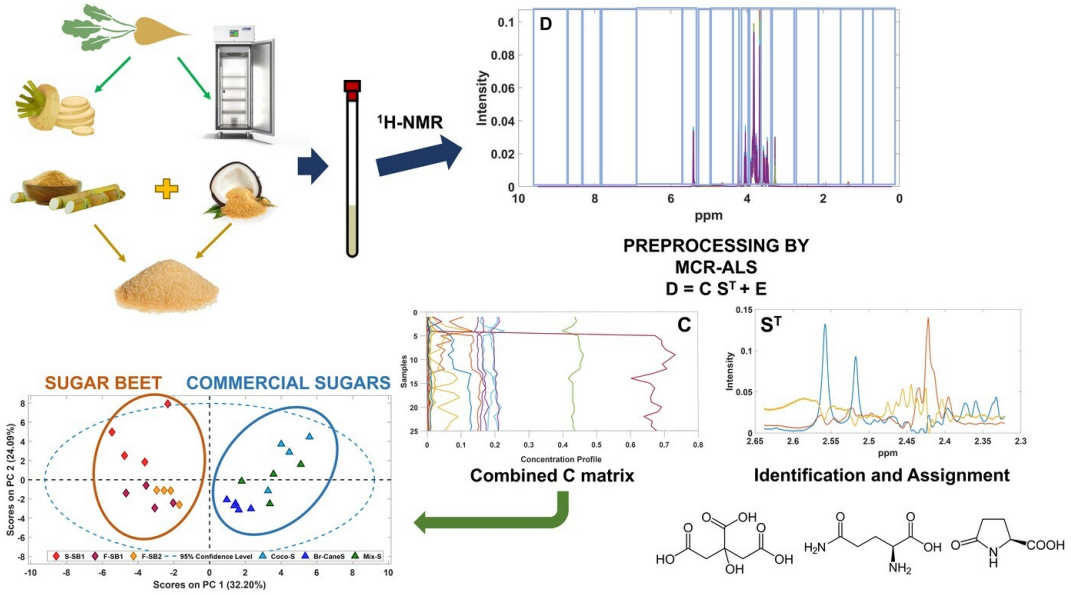

Figure 4.0.1: Graphical abstract

Abstract

The different types of sugar employed in the food industry exhibit chemical similarity and are mostly dominated by sucrose. Owing to the sugar origin and differences in production, the presence of certain minor organic compounds differs. To differentiate between sugars based on their botanical source, geographical origin, or storage conditions, commercial brown sugars and sugar beet extracts were analyzed by ^{1}H NMR spectroscopy applying a segmented analysis by means of multivariate curve resolution–alternating least squares (MCR–ALS). Principal component analysis and partial least squares–discriminant analysis yielded excellent differentiation between sugars from different sources after the application of this preprocessing strategy; without loss of chemical information and with direct interpretation of the results. By applying a segmented analysis via MCR–ALS to ^{1}H NMR sugar data, similar spectroscopic profiles could be differentiated. This improved the selectivity of ^{1}H NMR spectroscopy for sugar source differentiation which can be useful for industrial sugar authentication purposes.

Keywords – Sugar beet, brown sugar, multivariate analysis, ^{1}H NMR spectroscopy, preprocessing

4.1 Introduction

Sugar constitutes one of the main ingredients in food formulations. Sucrose is the most commonly used from of sugar. It is a non-reducing sugar comprising glucose and fructose monomers [1]. The anomeric carbons of both monomers are involved in glycosidic linkages, which prevents their mutarotation [2]. Having only one form in the aqueous solution results in easier crystallization of sucrose compared to other sugars. Sucrose is primarily produced from sugar beet (*Beta vulgaris*) or sugar cane (*Saccharum officinarum*). Worldwide, 80% of the annual sugar production is from sugar cane while 20% is from sugar beet [3, 4].

In sugar beet, sodium, potassium, betaine, amino acids, and nitrate affect the quality of sugar as they cannot be eliminated from the juice of sugar beet during the purification process [5]. Moreover, factors such as storage conditions of the raw material, climatic temperature during the harvest, and soil composition of the sugar beet growing region can affect the efficiency of sugar extraction [6, 7]. Furthermore, the quality of sugar depends on the composition of sugar beet and, therefore, on the growing area and environmental conditions. Hence, the control and characterization of sugar beet, based on geographical region or cultivation area, are of great importance to the food industry to achieve high product quality. Several approaches using proton nuclear magnetic resonance (^{1}H NMR) spectroscopy-based metabolic profiling, have been described. These include evaluation of differences in leaves and roots of sugar beets subjected to different durations of intermittent drought [8], different degrees of resistance to Cercospora leaf spot in sugar beets [9], and differences in the phytochemical profile of red beetroot from three different harvests [10].

Sucrose is mainly produced as white crystals; however, brown crystalline sugar (referred to as brown sugar hereafter) is an important commercial sugar that is widely produced and consumed globally [11].

Brown sugar is either prepared from sugar cane juice by means of thermal processing or by mixing refined beet sugar with cane molasses [11]. Different types of edible brown sugar, including non-centrifugal cane sugar (NCS), muscovado sugar (MS), or brown granulated sugar (BGS) can be obtained depending on the manufacturing process [12]. In addition, coconut sugar is similar in appearance and aromatic profile to that of brown sugar produced from either beet or cane

sugar [13]. Coconut sugar is produced by heating the coconut sap till the juices are saturated and sugar crystals are obtained [14]. Owing to the similarity in color and flavor, it is difficult to distinguish the origin of these brown sugars by means of only sensory evaluation. Furthermore, because brown sugar is usually sold at a higher price than white sugar, the former is susceptible to adulteration. For example, with sugar can be coated with caramelized sugar or synthetic dyes and sold as brown sugar.

Recently, gas chromatography-olfactometry-mass spectrometry has been used to differentiate between different types of brown cane sugars (NCS, MS, and BGS) as well as brown sugars produced in three provinces of China by comparing the constituent odor compounds [12, 15]. Also, three-dimensional fluorescence spectroscopy was used to identify and differentiate between natural and commercial brown cane sugars [11]. Similarly, the metabolic profile of brown beet, unrefined cane, and coconut blossom sugar was analyzed with ^{1}H NMR [13].

In food analysis, NMR spectroscopy is considered similar to a fingerprinting technique [16] because it provides chemical information regarding the composition of the samples. However, elucidating the information in an NMR spectrum can be challenging given the strong overlap of the signals and complexity of its spectral interpretation [6]. To overcome these problems, multivariate analysis techniques can be applied to ^{1}H NMR data to reduce its dimensionality and enable the extraction of relevant information to identify similarities or differences between groups of samples. This has been illustrated by classifying samples from a number of different commodities based on geographical origin [17, 18, 19, 20], harvest time [10], temporary stress [21, 22, 8] and adulterations [13, 23].

Generally, the chemometric models applied to ^{1}H NMR spectroscopy data use *"data binning"* to eliminate small variations between different samples due to chemical shifts produced by fluctuations in the pH, temperature, or concentration [24]. However, binning the data across a defined frequency width decreases the spectral resolution. Also, grouping of overlapping signals in the same bin makes the direct interpretation of the results challenging. To improve spectral interpretation, Pérez et al., [25] developed a methodology based on multivariate curve resolution–alternating least squares (MCR–ALS) [26]. This method was derived from the Decision Tree Correlation methodology reported by Puig-Castellví et al., [27]. The latter authors applied it as an independent preprocessing method

to resolve the concentration **(C)** and spectral **($\mathbf{S}^T$)** profiles in an ^{1}H NMR dataset of zebrafish samples. Furthermore, this method improves sample clustering when using PCA. In addition, Cavallini et al., [28] proposed a strategy based on evaluating intervals via MCR–ALS as a resolution technique to characterize beer. The authors highlighted that the resolution of $\mathbf{S}^T$ profiles simplified the information, making it easily interpretable.

To the best of our knowledge, the application of segmented analysis via MCR–ALS to the ^{1}H NMR spectra for the differentiation based on geographical origin and storage conditions of sugar beet, and the botanical source of commercial sugars has not been explored before. Herein, segmented analysis via MCR–ALS was applied to ^{1}H NMR spectroscopy data to differentiate sugars based on their sources. MCR–ALS allow analysis of the ^{1}H NMR spectra of sugars, avoiding the issues associated with loss of chemical information that are observed when data binning preprocessing is used. It also removes spectral noise and identifies the main chemical compounds responsible for the differences and the subsequent classification. Thus, segmented analysis via MCR–ALS was evaluated as a novel alternative method to data binning for classification or authentication purposes in the sugar industry.

4.2 Materials and methods

4.2.1 Reagent and chemicals

Deuterium oxide (D_2O, 99.9%), 3-(Trimethylsilyl)-propionic-2,2,3,3-d_4 acid sodium salt (TSP-d_4, 98%), potassium phosphate monobasic anhydrous (KH_2PO_4, $>$ 99%), potassium phosphate dibasic anhydrous (K_2HPO_4, $>$ 98%), and methanol (high-performance liquid chromatography grade) were purchased from Sigma-Aldrich (Darmstadt, Germany).

4.2.2 Sample collection and preparation

In 2022, twelve sugar beet samples, that were harvested from two geographical regions of the center-south of Chile (San Carlos and Los Ángeles) and three cultivation fields (Santa Isabel, Luciana, and Santa Laura) were kindly provided by the National Federation of Beet Growers (FENARE, Chile). Detailed information

regarding the origin of the samples is provided in Table 4.2.1. From these samples, four fresh sugar beets from San Carlos (F-SB1) and Los Ángeles (F-SB2) were selected, washed, cut into slices, freeze-dried, pulverized in a mortar, and stored at −80°C. In parallel, to establish the robustness of the segmented analysis and assess the discrimination capacity of the chemometric models by adding a new variable, four samples obtained from San Carlos (S-SB1) were stored in a growth chamber under controlled conditions at 20°C (without humidity control) for 1 month. After this period, the samples were collected, washed, cut into slices, freeze-dried, pulverized in a mortar, and stored at −80°C. In addition, nice commercial sugars, four coconut sugars (Coco-S), and five brown cane sugars (Br-CaneS) were purchased from local markets in different countries. These commercial sugars were pulverized in a mortar, and, subsequently, intentionally mixed in different proportions for simulated adulterations (Mix-S) (Table 4.2.1).

4.2.3 Extraction of carbohydrates from the sugar beet samples

The extraction of carbohydrates from F-SB1, F-SB2, and S-SB1 samples was conducted using the methanol/water protocol reported by Yang et al., [29] with some modifications. Briefly, 150 mg freeze-dried sugar beet was suspended in 1000 μL cold (4°C) MeOH:H_2O (1:1 v/v) solution, vortexed for 1 min, and sonicated for 15 min. Following centrifugation at 17,000 × g for 10 min, the supernatant was collected and transferred to a 2-mL Eppendorf tube. The extraction process for each sample was performed twice, and subsequently, both supernatants were combined, freeze-dried, and stored at −80°C until analyzed.

4.2.4 NMR sample preparation

The sugar beet extracts were dissolved in 800 μL of a D_2O solution containing 100 mM K_2HPO_4/KH_2PO_4 pH 7.4 as a buffer and TSP-d_4 0.1% w/w as an internal standard. For Coco-S, Br-CaneS, and Mix-S, 300 mg of each sample was dissolved using the same solution. The samples were centrifuged at 17,000 × g for 5 min, and 600 μL supernatant was transferred to a 5-mm NMR tube.

Table 4.2.1: Information of sugar beet and commercial sugar samples, including geographical region, country, cultivation field, brand, and code.

Source of sugar	Geographical region or country	Cultivation field or commercial brand	N° of samples	Code
Fresh sugar beet	San Carlos Chile	Santa Laura ($-36.460742°$, $-71,878242°$)	4	F-SB1
	Los Ángeles Chile	Santa Isabel ($-37.4337°$, $-72,29983°$)	2	F-SB2
	Los Ángeles Chile	Luciana ($-37.43191°$, $-72,31375°$)	2	F-SB2
Stored sugar beet	San Carlos Chile	Santa Laura ($-36.460742°$, $-71,878242°$)	4	F-SB1
Coconut sugar	Chile	NN	1	Coco-S
	Chile	NO	1	Coco-S
	Colombia	SL	1	Coco-S
	United States	MT	1	Coco-S
Brown cane sugar	Türkiye	KY	1	Br-CaneS
	Chile	IA	1	Br-CaneS
	Colombia	DL	1	Br-CaneS
	Colombia	CO	1	Br-CaneS
	Perú	DU	1	Br-CaneS
Proportion mixtures between coconut and brown cane sugar (% *w/w*)		NK:KY 15:15	1	Mix-S
		NO:IA 20:10	1	Mix-S
		SL:DL 10:20	1	Mix-S
		MT:CO 25:5	1	Mix-S
		Total number of samples	**25**	

4.2.5 NMR acquisition

All ^{1}H NMR spectra were acquired using a Bruker 400 MHz spectrometer model Ascend^{TM} (Bruker Biospin, Germany) operating at 400.13 MHz and equipped with a PABBI 1H/D-BB-Z-GRD liquid probe. For the ^{1}H NMR spectra, 96 number of scans were recorded for each sample with 65,536 data points over a spectral width of 6393.862 Hz, with an acquisition time of 1.26 s, receiver gain of 57.0, and relaxation delay of 1.00 s using the *zg30* pulse sequence at 293K. A representative ^{1}H NMR spectrum was obtained for each sample by averaging these 96 scans.

4.2.6 NMR data preprocessing using chemometric tools

4.2.6.1 Processing ^{1}H NMR dataset

All ^{1}H NMR spectra were manually phased, baseline corrected, and referenced to the TSP-d_4 resonance signal ($\delta_H = 0.00$) using MestreNova v.12.0 (MestreLab Research, Spain) software and imported into MATLAB R2021a (TheMathWorks Inc., Natick, MA, USA) as a spectral data matrix. The spectral regions of $\delta_H = 3.32 - 3.38$ and $\delta_H = 4.68 - 5.17$ caused by the presence of residual resonance signals from methanol and water, respectively, were excluded from the data. The spectral regions with chemical shifts $< \delta_H = 0.20$ and $> \delta_H = 9.50$ were also removed. The final ^{1}H NMR dataset comprised a data matrix with 25 spectra in rows with $\delta_H = 35,886$ values in columns. To minimize any variability due to chemical shifts that may affect the chemometrics models, the resonance misalignments were corrected using the *icoshift* algorithm [30], and rows of the dataset were normalized to the total area. The corrected and normalized ^{1}H NMR dataset (processed ^{1}H NMR dataset) was subjected to two independent preprocessing methods before the unsupervised and supervised analysis, i.e., ^{1}H NMR data binning and segmented analysis via MCR–ALS, respectively.

4.2.6.2 ^{1}H NMR data binning

The processed ^{1}H NMR dataset was used to make a binning dataset with a bucket width of 0.04 ppm using MestreNova v.12.0. The result consisted of a matrix with 25 spectra in rows and 219 buckets in columns.

4.2.6.3 Segmented analysis via MCR–ALS

Alternative to data binning, a segmented analysis via MCR–ALS was applied to the processed ^{1}H NMR dataset using the methodology developed by Pérez et al., [25]. The ^{1}H NMR dataset was manually divided into 29 subarrays (segments) to maintain the multiplicity of the resonance signals. Each subarray encompassed one or more resonance signals with distinct intensities in relation to the spectral noise. The dimensions of the 29 subarrays are reported in detail in supplementary Table 4.6.1. The 29 subarrays were independently analyzed using MCR–ALS GUI 2.0 Toolbox [31] via MATLAB R2021a. Every subarray was resolved via MCR–ALS using the single value decomposition (SVD) to find the optimal number

of components. Non-negativity constrains were applied to the $\mathbf{C}$ and $\mathbf{S}^T$ profiles in the ALS optimization. The quality of all MCR–ALS models was determined by evaluating the values of lack of fit (LOF) and percentage of explained variance. The combination of the 29 concentration profiles matrices with different number of components produced a combined $\mathbf{C}$ matrix with 25 samples in rows and 37 components in columns.

4.2.7 Unsupervised and supervised analyses

Exploratory analysis by PCA using data binning and combined $\mathbf{C}$ matrix as an independent $\mathbf{X}$ matrix, composed by the 25 samples (described in Table 4.2.1), was applied to evaluate the maximization sample clustering according to the intrinsic variance in the dataset. In addition, mean centering in combination with variance or *Pareto* scaling was used as a pretreatment for the different PCA models to evaluate the influence of the pretreatment on the sample clustering.

To improve the separation between the sample groups and evaluate the variables that produced considerable differences in the discrimination, two PLS–DA models were applied to the combined $\mathbf{C}$ matrix. The first PLS–DA model was used to discriminate between stored sugar beets and fresh sugar beets and between the different geographical regions of cultivation, considering 12 samples. The second PLS–DA model was applied to discriminate between coconut sugar, brown cane sugar, and their simulated adulterations, with a total of 13 samples. For each PLS–DA model, mean centering with variance scaling was used as a pretreatment.

All the PCA and PLS–DA models were constructed using PLS_Toolbox 9.0 (Eigenvector Research Inc.) via MATLAB R2021a. The quality of the PCA and PLS–DA models was evaluated using the Venetian blinds cross-validation method with 5 and 6 blinds, respectively. The optimal number of components (and latent variables [LVs]) was determined based on the percentage of explained variance. To identify the most discriminant variables from the PLS–DA models, the values from the variable importance in the projection (VIP) were analyzed. The variables with a VIP of ≥ 1 were considered relevant for discrimination. To assess the statistical significance of each relevant variable, a one-way analysis of variance (ANOVA) test with Bonferroni correction was applied to adjust the significance level, and for pairwise differences, t-test were performed using OriginPro 2016 (OriginLab Corporation, Northampton, MA, USA). VIP values with $p < 0.05$

statistical significance were recognized as discriminant.

4.3 Results and discussion

The processed ^{1}H NMR dataset of the 25 samples is shown in Fig. 4.3.1A. Similarities were observed mainly in the spectral region of carbohydrates with $\delta_H = 3.50 - 6.00$. In addition, lower resonance signals in the spectral regions of aliphatic compounds with $\delta_H = 0.50 - 3.50$ and those of aromatic/aldehyde compounds with $\delta_H = 6.00 - 9.50$ were detected in all the ^{1}H NMR spectra (Fig. 4.3.1B and 4.3.1C), suggesting similar spectroscopic profiles in all the samples. The identification of these resonance signals in the processed ^{1}H NMR dataset will be discussed on detail in Section 4.3.4.

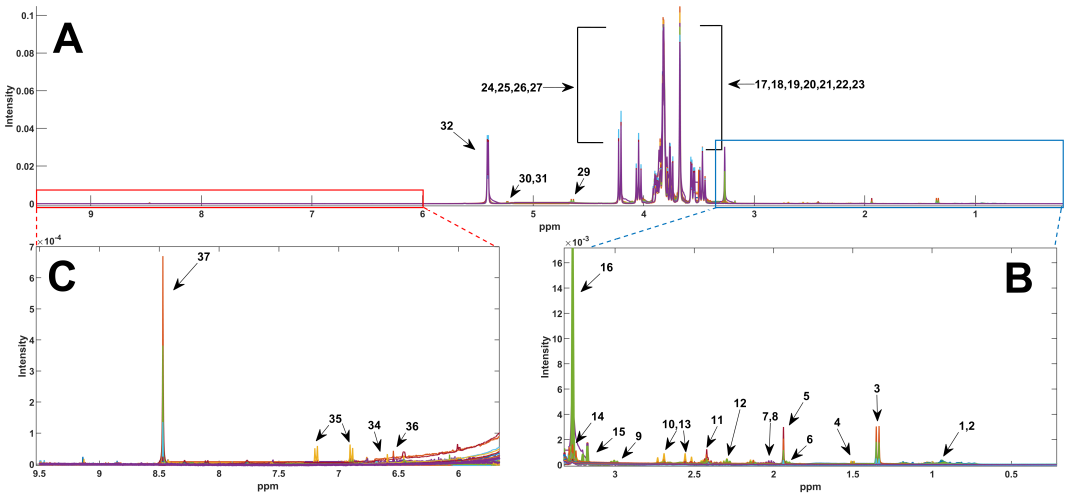

Figure 4.3.1: Processed ^{1}H NMR dataset (400.13 MHz) from the 25 samples (carbohydrates extracted from eight fresh sugar beets, four storage sugar beets, four coconut sugars, five brown cane sugars, and four simulated adulterated samples) used for the segmented analysis via MCR–ALS. **A)** Complete range of resonance signals $\delta_H = 0.20 - 9.50$. **B)** Expanded aliphatic spectral region $\delta_H = 0.20 - 3.50$. **C)** Expanded aromatic/aldehyde spectral region $\delta_H = 6.00 - 9.50$. Numbers 1-37 represent resonance signals (Table 4.3.2) associated to: (1): Leucine; (2): Valine + Isoleucine; (3): Lactate; (4): Alanine; (5): Acetate; (6-9-14): GABA; (7-8): Glutamate; (10-13): Acetate + Malate; (11): Pyroglutamate; (12): Glutamine; (15): Choline; (16): Betaine; (17-18-19-20-21-22-23-25-27-32): Sucrose; (24-26): Fructose; (29): β-glucose; (30): α-glucose; (31): α-xylose; (34): *trans*-aconitate; (35): Tyrosine; (36): Fumarate; (37): Formate.

4.3.1 Resolution of resonance signals via MCR–ALS

The 29 manually selected subarrays across the full range of resonances ($\delta_H = 0.20-9.50$) were independently resolved via MCR–ALS. For example, two different subarrays in the processed ^{1}H NMR dataset, containing resonance signals in the ranges of $\delta_H = 4.16 - 4.25$ and $\delta_H = 1.80 - 1.96$ are presented in supplementary Fig. 4.6.1A and 4.6.1B, respectively. The optimal number of components required to resolve $\mathbf{C}$ and $\mathbf{S}^T$ profiles for each resonance signal was determined using the SVD method using MCR–ALS. The resonance signal of the first subarray was successfully resolved with only one component (supplementary Fig. 4.6.1A), whereas for the second subarray, two component were found to be optimal for accurately resolve resonance signals with good symmetry with a certain degree of overlap (supplementary Fig. 4.6.1B). For this preprocessing, 1-3 components were required to describe the $\mathbf{C}$ and $\mathbf{S}^T$ profiles for each subarray, obtaining a total of 45 components with a convergence between 4 and 500 iterations. Of these 45 components, 8 were discarded because they contained only noise or were incorrectly resolved. The remaining 37 components had an LOF of $\backsim$ 4.46% and a total explained variance of 98.4%. Owing to their chemical shifts, components 1-17, 18-33, and 34-37 described resonance signals associated with the aliphatic, carbohydrate, and aromatic/aldehyde spectral regions, respectively. More precise representation of all the chemical shifts according to each $\mathbf{C}$ and $\mathbf{S}^T$ profiles is shown in supplementary Table 4.6.2.

4.3.2 Effect of the pretreatment on the exploratory analysis using PCA

4.3.2.1 Binning data

The PCA score plot, obtained using mean centering with variance scaling as pretreatment, is shown in Fig. 4.3.2A. The first two principal components (PCs) explained 45.29% of the total variance. No clear clustering trend was observed with respect to the sugar source (Fig. 4.3.2A); however, separations between S-SB1 and F-SB2 and Br-CaneS in PC1 (25.79% explained variance) were observed. PC2 (19.50%) showed lesser differences between Br-CaneS and Coco-S compared to PC1. In this case, the data binning did not allow a good differentiation between the sample groups. Using mean centering and *Pareto* scaling as a pretreatment,

the total explained variance could be increase, based on the contribution of lower-intensity resonance signals; providing equal weighting to each variable [32]. Thus, the differentiation was not biased toward variables with greater magnitude, as seen in the case of resonance signals present in the carbohydrate spectral region (Fig. 4.3.1A), which maximized the differentiation as in the case of the S-SB1 samples (Fig. 4.3.2B). In this case, the first two PCs explained 79.94% of the variance. The first source of variation was attributed to storage, in which S-SB1 differed from F-SB2, Coco-S, Br-CaneS, and Mix-S in PC1 (56.13% of the explained variance). PC2 (23.81%) demonstrated greater differentiation between F-SB1 and F-SB2 with respect to S-SB1, Coco-S, and Mix-S than PC1. Despite these observations, satisfactory results in differentiating sugar beet extracts and commercial sugars were not achieved.

4.3.2.2 Combined C matrix

PCA applied to the combined **C** matrix using mean centering and variance scaling as a pretreatment, is shown in Fig. 4.3.2C. The first two PCs explained 56.29% of the total variance. Variance scaling uses the standard deviation of each variable as scaling factor [33], thereby rendering each variable equally important (same variance). However, erroneous estimations can be observed for those variables that do not contain chemical information or a certain degree of overlap, as in the case of data binning. While binning data minimizes variations caused by chemical shifts, it also results in loss of spectral resolution and superposition of resonance signals in the same bin, making sample clustering challenging (Fig. 4.3.2A and 4.3.2B). However, this does not occur with the combined **C** matrix, because each **C** profile is described by an $\mathbf{S}^T$ profile regardless of the degree of overlap. The spectral noise is modelled in an independent matrix of residuals, and even small variations within the variables can be detected, thereby maximizing clustering in PCA (Fig. 4.3.2C). Thus, with mean centering and *Pareto* scaling used as a pretreatment (Fig. 4.3.2D), MCR–ALS preprocessing achieved clustering of similar samples. These results agreed with those reported by Pérez et al., [25] and Khakimov et al., [34] in zebrafish and human urine samples, respectively. MCR–ALS maximized the between-group variability and decreased the within-group variability. Therefore, different pretreatments can be applied depending on the objectives of the study.

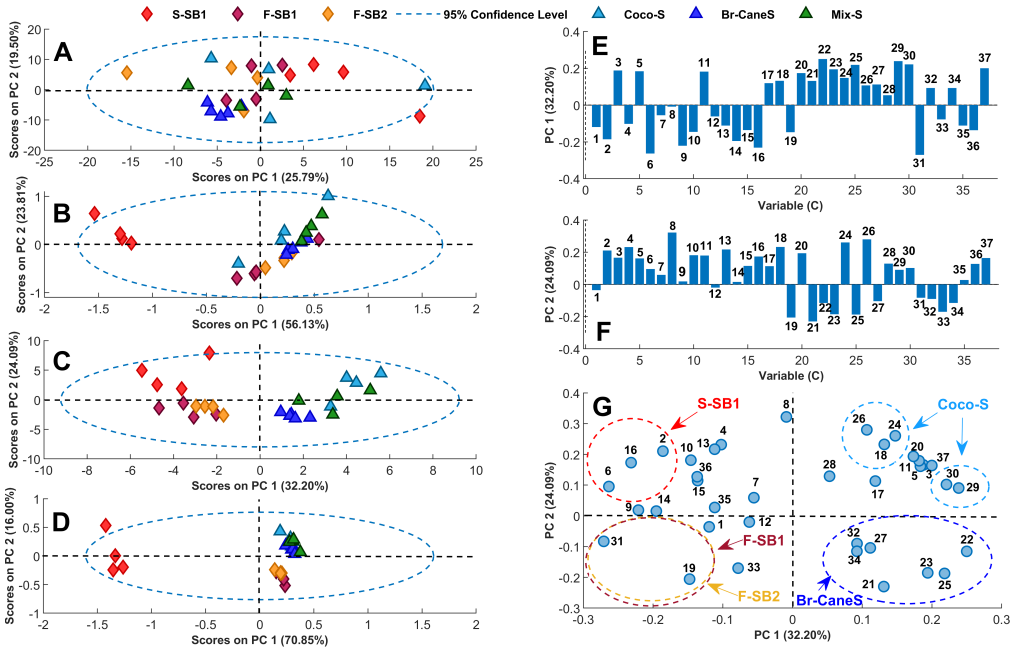

Figure 4.3.2: PCA plots derived from the binning data (A,B) and the combined **C** matrix obtained via MCR–ALS (C,D) for the differentiation of sugar beet extracts and commercial sugars. **A)** PCA score plot (PC1 vs PC2) using mean centering and variance scaling on binning data. **B)** PCA score plot (PC1 vs PC2) using man centering and *Pareto* scaling on binning data. **C)** PCA score plot (PC1 vs PC2) using mean centering and variance scaling on the combined **C** matrix. **D)** PCA score plot (PC1 vs PC2) using mean centering and *Pareto* scaling on the combined **C** matrix. **E)** and **F)** Loading bar plots of PC1 and PC2 from PCA described in **C)**, respectively. **G)** Scatter loading plots of PC1 vs PC2 of PCA described in **C)**. Variables were represented as a bar plot and blue dots in which each number represented a concentration profile (**C**) that is associated to its respective spectral profile ($\mathbf{S}^1$). S-SB1: storage sugar beet from San Carlos (red); F-SB1: fresh sugar beet from San Carlos (brown); F-SB2: fresh sugar beet from Los Ángeles (yellow); Coco-S: coconut sugar (light blue); Br-CaneS: brown cane sugar (blue); Mix-S (green).

4.3.2.3 Differentiation of sugar beet extracts and commercial sugars using the combined C matrix

The score plot in Fig. 4.3.2C shows that the first source of variation corresponded with sugar beet extracts or commercial sugar samples, in which separation between S-SB1, F-SB1, and F-SB2 and Coco-S, Br-CaneS, and Mix-S was achieved with 32.20% of the variance explained. The second source of variation allowed the differentiation between the stored and fresh sugar beet samples S-SB1, F-SB1, and F-SB2, as well as differentiation of Coco-S and Br-CaneS with 24.09% of variance

explained. Notably, the clustering of samples based on the sugar source was maximized when using the combined **C** matrix. Also, the simulated adulteration samples (Mix-S) were separated from the Coco-S and Br-CaneS clusters, with one sample of Coco-S clustering with the group of Mix-S samples. Most samples remained inside the 95% confidence level, except for one sample from S-SB1. The samples that showed Q residuals over the 95% confidence level and did not exhibit significant improvement in the percentage of explained variance following their exclusion were not considered outliers.

The variables that contributed to the clustering according to the sugar source can be observed in the loading plots of both PCs in Fig. 4.3.2E and 4.3.2F. These loading plots are shown as a bar plot, in which each component represents **C** that was described by its respective **S** profile. The variables 3, 5, 11, 17-32 (except 19 and 31), 34, and 37 were most relevant for the clustering of the commercial sugars Coco-S, Br-CaneS, and Mix-S (Fig. 4.3.2E). The variables 1-16 (except 3, 5, and 11), 19, and 31 contributed to the clustering of the stored and fresh sugar beet samples S-SB1, F-SB1, and F-SB2, followed by less relevant variables 33, 35, and 36. Regarding the spectral regions of each variable (Table S4.6.2), the carbohydrates region made the greatest contribution to the differentiation of commercial sugars from sugar beet extracts, whereas the aliphatic region exhibited a high relevance for differentiating sugar beet extracts under storage conditions, except for certain resonance signals. The spectral region of aromatic/aldehyde compounds also exhibited specific resonance signals that contributed to the differentiation of sugar beet extracts and commercial sugars. The loading plot for PC2 (Fig. 4.3.2F) illustrated that the variables from the carbohydrate spectral region 18, 20, 24, and 26 were more relevant for clustering S-SB1, Coco-S, and some samples of Mix-S, whereas the variables 19, 21, 22, 23, 25, 27, 31, 32, 33, and 34 were relevant for F-SB1, F-SB2, Br-CaneS, Mix-S, and one sample of Coco-S, implying the possible influence of resonance signals associated with different sugars contributing to the differentiation. Similarly, the loading plot for PC2 shows that the variables 2-16 (except 1 and 12) of the aliphatic spectral region made a clear contribution to the clustering of S-SB1, Coco-S and some Mix-S samples. According to the scatter plot of the loadings (Fig. 4.3.2G) variables 18, 24, 26, 29, and 30 characterized the clustering of Coco-S samples, while stored samples were characterized by variables 2, 6, and 16. Variables 19 and 31 characterized the samples F-SB1 and F-SB2,

while variables 21, 22, 23, 25, 27, 32, and 34 had a high contribution for Br-CaneS.

4.3.3 Discrimination of a group of samples using PLS–DA

4.3.3.1 Sugar beet under storage conditions and geographical origin

Samples coming from sugar beet were analyzed by PLS–DA to evaluate differences by storage conditions and geographical origin. The first two LVs explained 55.62% of the total variance. The discrimination of S-SB1 from F-SB1 and F-SB2 was achieved with the first LV (LV1), with 40.27% of the variance explained. This discrimination is depicted in the score plot of the PLS–DA model in Fig. 4.3.3A. The PCA score plot revealed that the same differences were obtained with PC1, with 41.47% of the variance explained (supplementary Fig. 4.6.2A). Discrimination based on geographical origin (F-SB1 from F-SB2) was possible with LV2 (15.35% of explained variance). In the PCA score plot, a similar differentiation was observed but to a lesser extent in the fourth PC, which explained 7.64% of the variance (Supplementary Fig.4.6.2).

The VIP plot (Fig. 4.3.3B) revealed that the highest contributions associated with the discrimination between S-SB1 and F-SB1 and F-SB2 were from the variables 2, 3, 8, 10, 16, 18, 21, 22, 23, 25, and 36. After the analysis of the statistical significance of each relevant variable (VIP $\geq$ 1), only variables 2, 8, 10, 16, and 36 were observed to contribute to differentiate S-SB1 from the fresh samples. This was expected due to the other variables being related to the resonance signals of the sucrose molecule. Table 4.3.1 summarized the VIP values obtained by the PLS–DA model for these variables. Based on the identification of resonance signals (Table 4.3.2), these variables were associated with valine, isoleucine, glutamate, citrate, malate, betaine, and fumarate. The score plot for LV1 (Fig. 4.3.3A) revealed that storage time affected sugar beet, characterized by an increase in the intensity of concentration profiles associated with amino acids and nitrogenous compounds (Supplementary Fig. 4.6.5). This could be explained by a possible tendency toward the hydrolysis of proteins to amino acids and their catabolism via the tricarboxylic acid (TCA) cycle (citrate, malate, fumarate), thereby increasing the concentration of drought stress-induced N-amino compounds [8]. Furthermore, an increase in the intensity of betaine concentration profiles was observed (Supplementary Fig. 4.6.5E). This, could be related to a

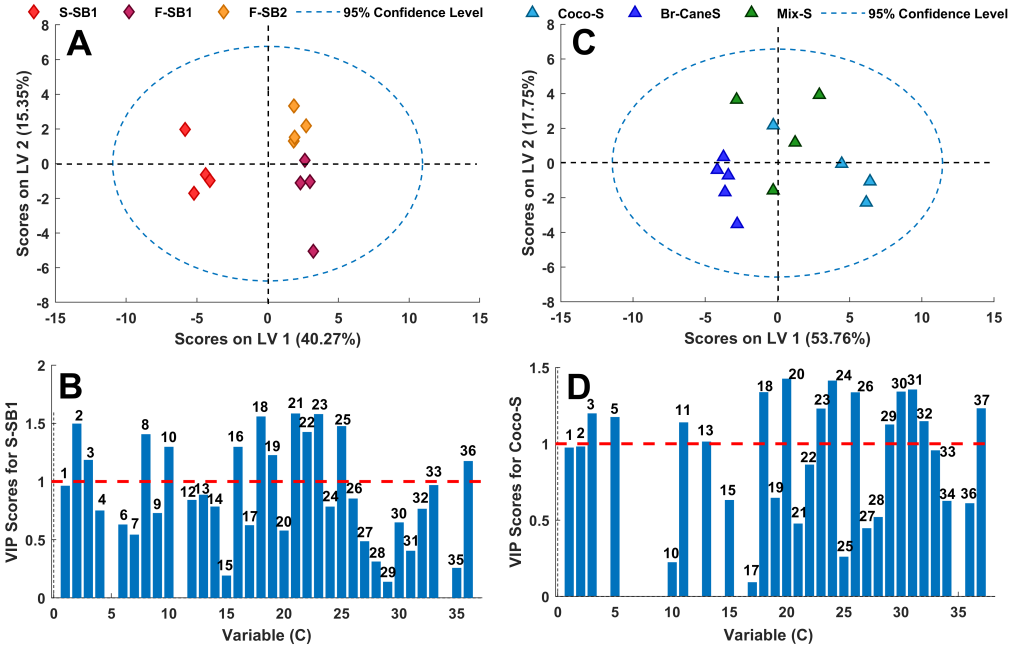

Figure 4.3.3: PLS–DA plots derived from the combined **C** matrix obtained via MCR–ALS for the discrimination of sugar beets and commercial sugars. **A)** PLS–DA score plot (LV1 vs LV2). **B)** VIP plot for S-SB1. **C)** PLS–DA score plot (LV1 vs LV2). **D)** VIP plot for Coco-S. Variables with VIP ≥ 1 were considered relevant for discrimination. S-SB1: storage sugar beet from San Carlos (red); F-SB1: fresh sugar beet from San Carlos (brown); F-SB2: fresh sugar beet from Los Ángeles (yellow).

decrease in the intensity of resonance signals associated with sucrose, because its biosynthesis requires energy from sucrose. Regarding discrimination according to geographical region, the VIP plots (supplementary Fig. 4.6.3) indicate the variables with more relevance are associated with the concentration profiles of amino acids. Based on this, only leucine showed significant discrimination for F-SB1 (Supplementary Fig. 4.6.5A), suggesting possible differences in soil type per cultivation field, generating variations in the concentration of free amino acids.

4.3.3.2 Coconut sugar, brown cane sugar, and simulated adulterations

A separated analysis was performed for the commercial brown sugars to identify compounds responsible for discrimination based on botanical source using PLS–DA. The PLS–DA model enabled the discrimination between Coco-S and Br-CaneS, with the first two LVs explaining 71.51% of the total variance. The PLS–DA score plot (Fig. 4.3.3C) revealed that the discrimination between Coco-S and Br-CaneS was achieved with LV1, explaining 53.76% of the variance. The same

Table 4.3.1: Variables that contribute to the discrimination of sugar beet extracts and commercial sugars associated with the VIP values ($p < 0.05$) obtained by PLS–DA.

Source of sugar	Variable (C) and (S^T)	Compound	VIP value
S-SB1	2	Valine-Isoleucine	1.50
	8	Glutamate	1.41
	10	Citrate-Malate	1.30
	16	Betaine	1.30
	36	Fumarate	1.18
F-SB1	1	Lecuine	1.82
Coco-S	3	Lactate	1.05
	5	Acetate	1.03
	11	Pyroglutamate	1.03
	24	Fructose	1.31
	26	Fructose	1.26
	29	β-glucose	1.42
	30	α-glucose	1.24
	37	Formate	1.07
Br-CaneS	31	α-xylose	1.22
	32	Sucrose	1.03

differentiation was possible with PCA, with 54.34% of the explained variance in PC1 (Supplementary Fig. 4.6.2C). LV2 showed that discrimination of Mix-S from Coco-S and Br-CaneS was possible to a certain extent, with 17.75% of the variance explained (Fig. 4.3.3C), whereas Mix-S was mostly differentiated from both commercial sugars with PC1 (Supplementary Fig. 4.6.2C). Comparing these results with the PCA score plot of Fig. 4.3.2C, it was observed that in all models, the Mix-S samples were clustered with Coco-S and Br-CaneS, with one sample of Coco-S considered as Mix-S, suggesting a slight difference from the other Coco-S samples. The identification of resonance signals (Table 4.3.2) revealed that only the variables 15 (choline), 5 (acetate), 3 (lactate), 37 (formate), 17-32 (sucrose), 24-26 (fructose), 29 (β-glucose), and 30 (α-glucose) were detected in all the samples of Coco-S, whereas some resonance signals associated with amino/organic acids were only detected in three samples of Coco-S. This indicated that the use of the combined **C** matrix obtained from segmented analysis via MCR–ALS represents the pure resonance signals resolved in the processed ^{1}H NMR dataset. This allows, supervised or unsupervised models to present a less biased contribution to each variable and, therefore, be able to differentiate the one sample of Coco-S from the other three Coco-S samples.

The VIP plot (Fig. 4.3.3D and Table 4.3.1) showed that the contribution associated with the discrimination of Coco-S and Br-CaneS were from the variables 3, 5, 11, 24, 25, 29, 30 ,31, 32, and 37. Regarding the identification of resonance signals (Table 4.3.2), these variables corresponded to lactate, acetate, pyroglutamate, fructose, β-glucose, α-glucose, α-xylose, sucrose and formate. The high intensities of variables 3 (lactate), 5 (acetate), and 37 (formate) in Coco-S (Supplementary Fig. 4.6.6) may be partly due to fermentation after harvesting and the Maillard reaction [13, 12], contributing to is discrimination. The formation of pyroglutamate through the intramolecular cyclization of glutamate or glutamine [35] due to high temperatures and pressure during the production of coconut sugar contributed to its discrimination. In addition, the **C** of the variables 24-26 (fructose), 29 (β-glucose), and 30 (α-glucose) exhibited higher intensities in Coco-S than in Br-CaneS (Supplementary Fig. 4.6.6). Conversely, the high intensities of **C** associated with resonance signals of sucrose and possibly related to α-xylose contributed to the differentiation of Br-CaneS (Supplementary Fig. 4.6.7).

4.3.4 Tentative identification of resonance signals and assignment for sugar beet extracts and commercial sugars through the spectral profiles ($\mathbf{S}^T$) obtained via MCR–ALS

The 37 $\mathbf{S}^T$ profiles obtained via MCR–ALS according to each subarray were evaluated to determine the spectroscopic parameters as chemical shifts (δ_H), multiplicity, and coupling constants (J) using MestreNova v.12.0. The spectral regions of aliphatic, carbohydrate, and aromatic/aldehyde compounds were determined by comparing the spectroscopic parameters with values from previously reported studies [13, 36, 17, 10, 22, 8, 29], using reference library Chenomx NMR suite 8.3, and the online available Human Metabolome Database (HMDB) [37]. The latter was used for some primary compounds present also in plant metabolome. Overall, 22 polar compounds were identified in the sugar beet extracts and commercial sugar samples, i.e., 9 amino acids, 2 amino compounds, 7 organic acids, and 4 sugars. The identification, assignment, chemical shifts, coupling constants, and multiplicities of all the compounds are presented in detail in Table 4.3.2. The identification of the resonance signals through $\mathbf{S}^T$ profiles was consistent with the results of the scores and loading plots obtained with PCA (Fig. 4.3.2C,

4.3.2E, and 4.3.2F). The specific resonance signals for sugar beet extracts and commercial sugars, i.e., variables 4 (alanine), 6, 9, 14 (GABA), 8, 7, (glutamate), 11 (pyroglutamate), 12 (glutamine), 16 (betaine), 35 (tyrosine), 5 (acetate), 37 (formate), and 34 (*trans*-aconitate), demonstrated accurate contribution to the clustering and differentiation according to the sugar source.

4.3.4.1 Carbohydrates

As anticipated, most of the intense resonance signals present in the ^{1}H NMR dataset (Fig. 4.3.1A) corresponded to the resonance signals of sucrose in the carbohydrate spectral region. Glucose anomers (α and β), α-xylose, and sucrose were represented by the doublets of protons bonded to the anomeric carbon at $\delta_H = 5.23$ (d, $J = 3.76$ Hz), $\delta_H = 4.65$ (d, $J = 7.93$ Hz), $\delta_H = 5.20$ (d, $J = 3.82$ Hz), and $\delta_H = 5.41$ (d, $J = 3.87$ Hz), respectively. Resonance signals from fructose were detected as multiplets at $\delta_H = 4.01$ and 4.11.

4.3.4.2 Amino acids and amino compounds

The compounds in the aliphatic region associated with the branched-chain amino acids signals of valine $\delta_H = 1.00, 1.04$ (d, $J = 6.64$, 6.86 Hz), leucine $\delta_H = 0.93$ (t, $J = 6.80 - 6.30$ Hz), and isoleucine $\delta_H = 1.07$ (d, $J = 7.06$ Hz) were identified in S-SB1, F-SB1, F-SB2, and only in three Coco-S samples. Alanine $\delta_H = 1.50$ (d, $J = 7.22$ Hz) and GABA $\delta_H = 1.94$ (m), 2.29 (t, $J = 7.42$ Hz), and 3.01 (t, $J = 7.95 - 7.84$ Hz) were detected only in sugar beet extracts. The resonance signal associated with betaine $\delta_H = 3.27$ (s) was found only in the sugar beet extracts, as betaine is a specific molecule of sugar beet [13, 38]. The multiplets of glutamate at $\delta_H = 2.04$, 2.14 were detected only in S-SB1, and the multiplet of glutamine $\delta_H = 2.44$ was found in F-SB1 and F-SB2. Pyroglutamate $\delta_H = 2.42$, 2.52 (m) was present in only three samples of Coco-S, possibly due to the heat-degradation of glutamate or glutamine during sugar production [8]. The resonance signal belonging to choline $\delta_H = 3.17$ (s) was detected in S-SB1 and Coco-S.

4.3.4.3 Organic acids and derivates

Malate $\delta_H = 2.37, 2.68$ (dd, $J = 15.1 - 9.53, 15.4 - 3.53$ Hz), acetate $\delta_H = 1.94$ (s), citrate $\delta_H = 2.54, 2.71$ (d, $J = 16.2 - 16.3$ Hz), and lactate $\delta_H = 1.34$ (d, $J = 6.93$ Hz) were identified as the most prominent organic acids in Coco-S (except for lactate), suggesting the occurrence of mixed acid fermentation between the harvesting of the sugar juice and heating [13]. In sugar beet extracts, only malate and citrate were detected in S-SB1, whereas lactate was identified in F-SB1, F-SB2, and S-SB1. In the aromatic/aldehyde spectral region, tyrosine at $\delta_H = 6.89 - 7.19$ (d, $J = 8.73, 8.88$ Hz) was detected only in three samples of the sugar beet extracts, i.e., one fresh and two stored samples. Furthermore, the resonance signal of fumarate $\delta_H = 6.54$ (s), an intermediate of the TCA cycle with citrate and malate [8], was detected in S-SB1, F-SB1, F-SB2, and three samples of Coco-S. Formate $\delta_H = 8.47$ (s) was identified only in Coco-S and Br-CaneS samples. The identification of formate and acetate only in Coco-S and Br-CaneS could be related to the formation of these compounds due to the degradation of sugars at high temperatures and the Maillard reaction [12]. Finally, the presence of trans-aconitate $\delta_H = 6.59$ (s) was detected only in Br-CaneS, consistent with the results of Bachmann et al., [13] and Palmonari et al., [38]. Aconitate is a specific acid produced by sugar cane, and its predominant form is *trans*-aconitate [39], formed by the isomerization of *cis*-aconitate.

Table 4.3.2: Identification and assignments of each spectral profile ($\mathbf{S}^T$) obtained by MCR–ALS for ^{1}H NMR dataset from sugar beets and commercial sugars. S-SB1: Storage sugar beet from San Carlos; F-SB1: Fresh sugar beet from San Carlos; F-SB2: Fresh sugar beet from Los Ángeles; Coco-S: Coconut sugar; Br-CaneS: Brown cane sugar; s: singlet; d: doublet, t: triplet; dd: doublet of doublets; dq: doublet of quartets; m: multiplet.

Variable $\mathbf{S}^T$	**Compound**	**$\delta_\mathbf{H}$ in ppm (mult, J in Hz)**	**Group**	**Detected in sugar beet extracts**	**Detected in sugar beet extracts**
Amino acids and amino compounds					
2	Valine	1.00 (d, J = 6.64)	γCH_3	+	Only in 3 samples
		1.04 (d, J = 6.86)	γ'CH_3	+	of Coco-S
1	Leucine	0.93 (d, J = 6.86-6.30)	δCH_3	+	Only in 3 samples
				+	of Coco-S
2	Isoleucine	1.07 (d, J = 7.06)	γCH_3	+	Only in 3 samples
				+	of Coco-S
4	Alanine	1.50 (d, J = 7.22)	βCH_3	+	-
6-9-14	GABA	1.91 (m)	γCH_2	+	-
		2.29 (t, J = 7.42)	αCH_2	+	-
		3.01 (t, J = 7.95-7.84)	γCH_2	+	-
8-7	Glutamate	2.04 (m)	βCH_2	Only in S-SB1	
		2.14 (m)	β'CH_3	Only in S-SB1	-
11	Pyroglutamate	2.42 (m)	γCH_2	-	Only in 3 samples
		2.52 (m)	βCH_2	-	of Coco-S
12	Glutamine	2.44 (m)	γCH_2	Only in S-SB1 and F-SB2	-
15	Choline	3.17 (s)	$N(CH_3)_3$	Only in S-SB1	Only in Coco-S
16	Betaine	3.27 (s)	$N(CH_3)_3$	+	-
35	Tyrosine	6.89 (d, J = 8.73)	3.5-CH	Only in 3	-
		7.19 (d, J = 8.88)	2.6-CH	samples	-
Organic acids					
10-13	Malate	2.37 (dd, J = 15.19.53)	βCH	Only in S-SB1	Only in 3 samples
		2.68 (dd, J = 15.4-3.53)	β'CH_3	Only in S-SB1	of Coco-S
5	Acetate	1.94 (s)	CH_3	-	+
10-13	Citrate	2.54 (d, J = 16.2)	CH_2	Only in S-SB1	Only in 3 samples
		2.71 (d, J = 16.3)	CH_2	Only in S-SB1	of Coco-S
3	Lactate	1.34 (d, J = 6.93)	βCH_3	+	+
36	Fumarate	6.54 (s)	$(CH=)_2$	+	Only in 3 samples of Coco-S
37	Formate	8.47 (s)	CH	-	+
34	*trans*-aconitate	6.59 (s)	CH=	-	Only in Br-CaneS
Carbohydrates					
17	Sucrose	3.47 (d, J = 9.38)	G_4H	+	+
18		3.56 (dd, J = 9.97-3.85)	G_2H	+	+
19		3.67 (s)	F_1H	+	+
20		3.76 (t, J = 9.56)	G_3H	+	+
21		3.82 (m)	F_6H	+	+
22		3.85 (m)	G_5H	+	+
23		3.39 (dq, J = 8.30-4.01)	F_5H	+	+
25		4.05 (t, J = 8.76)	F_4H	+	+
27		4.22 (d, J = 8.54)	F_3H	+	+
32		5.41 (d, J = 3.87)	G_1H	+	+
24-26	Fructose	4.01 (m)	C_6H_2	+	+
		4.11 (m)	C_3H+C_4H	+	+
31	α-xylose	5.20 (d, J = 3.82)	CH	+	Only in Br-CaneS
30	α-glucose	5.23 (d, J = 3.76)	CH	+	+
29	β-glucose	4.65 (d, J = 7.93)	CH	+	+
Not identified					
28	Unknown	4.39 (d, J = 8.77)	-	+	+
33	Unknown	5.62 (d, J = 3.82)	-	+	+

4.4 Conclusion

The application of a segmented analysis via MCR–ALS to ^{1}H NMR spectral data was effective in discriminating sugar beet extracts and commercial sugars with respect to botanical source, geographical region, and storage. This had not been possible under other preprocessing methods such as data binning. In addition, the discrimination was successful without the loss of chemical information required for identifying compounds. The spectral assignment, based on the $\mathbf{S}^T$ profiles using MCR–ALS, allowed the identification of 22 compounds present in the sugar beet extracts and commercial sugars. Valine, isoleucine, glutamate, betaine, acetate, malate, and fumarate were responsible for discriminating stored from fresh sugar beet, whereas lactate, acetate, pyroglutamate, fructose, α-glucose, β-glucose, and formate were responsible for discriminating coconut sugar from brown cane sugar. Finally, this study demonstrated that sugar from different sources, with similar spectroscopic profiles and which are dominated by sucrose, can be differentiated independent of their degree of spectral overlap. The use of MCR–ALS enables greater selectivity in chemometric models. It is important to note that the number of samples used for the analysis was small, and more samples are required in both groups (sugar beet and commercial sugars) to improve robustness of the models and ensure that the methodology is more generalizable in future.

Credit authorship contribution statement

Cristian A. Fuentes: Conceptualization, Methodology, Formal analysis, Investigation, Writing – original draft. **Mecit Halil Öztop**: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing. **Macarena Rojas-Rioseco**: Formal analysis, Methodology, Writing – review & editing. **Martín Bravo**: Conceptualization, Methodology, Writing – review & editing. **Aylin Özgur Göksu**: Conceptualization, Methodology, Writing – review & editing. **Marena Manley**: Methodology, Supervision, Writing – review & editing. **Rosario del P. Castillo**: Conceptualization, Methodology, Supervision, Writing – review & editing

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank FONDECYT 1221387 project (ANID, Chile), National Federation of Beet Growers (FENARE, Chile) for supplying the Chilean sugar beet samples. Cristian Fuentes thanks the Postgraduate Direction of the University of Concepción Ph.D. scholarship.

Funding

This study received funding from the European Union's Horizon 2020 Research and Innovation program–MSCA RISE under grant agreement #101008228 and FONDECYT 1221387 (ANID, Chile).

4.5 References

[1] Yves Queneau, Slawomir Jarosz, Bartosz Lewandowski, and Juliette Fitremann. Sucrose Chemistry and Applications of Sucrochemicals. In *Advances in Carbohydrate Chemistry and Biochemistry*, volume 61, pages 217–292. 2007.

[2] J. P. Kamerling and J. F. G. Vliegenthart. Chapter 4. Carbohydrates. *Mass Spectrometry*, pages 175–264, 2021.

[3] Jie Arro, Jong Won Park, Ching Man Wai, Robert VanBuren, Yong Bao Pan, Chifumi Nagai, Jorge da Silva, and Ray Ming. Balancing selection contributed to domestication of autopolyploid sugarcane (Saccharum officinarum L.). *Euphytica*, 209(2):477–493, 2016.

[4] Isa C. Ribeiro, Carla Pinheiro, Carla M. Ribeiro, Maria M. Veloso, Maria C. Simões-Costa, Isabel Evaristo, Octávio S. Paulo, and Cândido P. Ricardo. Genetic diversity and physiological performance of portuguese wild beet (beta vulgaris spp. maritima) from three contrasting habitats. *Frontiers in Plant Science*, 7(AUG2016):1–14, 2016.

[5] Christine Kenter and Christa M. Hoffmann. Changes in the processing quality of sugar beet (Beta vulgaris L.) during long-term storage under controlled conditions. *International Journal of Food Science and Technology*, 44(5):910–917, 2009.

[6] Olimpia Masetti, Angela Sorbo, and Luigi Nisini. Nmr tracing of food geographical origin: The impact of seasonality, cultivar and production year on data analysis. *Separations*, 8(12), 2021.

[7] K. Vukov and K. Hangypal. Sugar beet storage. *Sugar Technology Review*, 12:143–265, 1985.

[8] Rita Wedeking, Mickaël Maucourt, Catherine Deborde, Annick Moing, Yves Gibon, Heiner E. Goldbach, and Monika A. Wimmer. 1H-NMR metabolomic profiling reveals a distinct metabolic recovery response in shoots and roots of temporarily drought-stressed sugar beets. *PLoS ONE*, 13(5):1–21, 2018.

[9] Yasuyo Sekiyama, Kazuyuki Okazaki, Jun Kikuchi, and Seishi Ikeda. NMR-based metabolic profiling of field-grown leaves from sugar beet plants

harbouring different levels of resistance to Cercospora leaf spot disease. *Metabolites*, 7(1), 2017.

[10] Ottavia Giampaoli, Fabio Sciubba, Giorgia Conta, Giorgio Capuani, Alberta Tomassini, Giorgio Giorgi, Elisa Brasili, Walter Aureli, and Alfredo Miccheli. Red beetroot's nmr-based metabolomics: Phytochemical profile related to development time and production year. *Foods*, 10(8):1–12, 2021.

[11] Jui Yi Chen, Xin Wei Chen, Yu Yu Lin, Gow Chin Yen, and Jer An Lin. Authentication of dark brown sugars from different processing using three-dimensional fluorescence spectroscopy. *Lwt*, 150(June):111959, 2021.

[12] Erbao Chen, Huanlu Song, Shuna Zhao, Chen Liu, Long Tang, and Yu Zhang. Comparison of odor compounds of brown sugar, muscovado sugar, and brown granulated sugar using GC-O-MS. *Lwt*, 142(January):111002, 2021.

[13] René Bachmann, Anna Lena Horns, Nele Paasch, Robbin Schrieck, Markus Weidner, Iris Fransson, and Jan Philipp Schrör. Minor metabolites as chemical marker for the differentiation of cane, beet and coconut blossom sugar. From profiling towards identification of adulterations. *Food Control*, 135(December 2021), 2022.

[14] Bambang Nurhadi, Nandi Sukri, Wahyu Kristian Sugandi, Annisa Puteri Widanti, Resi Restiani, Ziske Noflianrini, Bayu Rezaharsamto, and Marleen Herudiyanto. Comparison of crystallized coconut sugar produced by traditional method and amorphous coconut sugar formed by two drying methods: Vacuum drying and spray drying. *International Journal of Food Properties*, 21(1):2339–2354, 2018.

[15] Erbao Chen, Shuna Zhao, Huanlu Song, Yu Zhang, and Wanyao Lu. Analysis and Comparison of Aroma Compounds of Brown Sugar in Guangdong, Guangxi and Yunnan Using GC-O-MS. *Molecules*, 27(18), 2022.

[16] A. Caligiani, D. Acquotti, G. Palla, and V. Bocchi. Identification and quantification of the main organic components of vinegars by high resolution 1H NMR spectroscopy. *Analytica Chimica Acta*, 585(1):110–119, 2007.

[17] Maria Eleni Dimitrakopoulou, Konstantina Matzarapi, Styliani Chasapi, Apostolos Vantarakis, and Georgios A. Spyroulias. Nontargeted 1H NMR

fingerprinting and multivariate statistical analysis for traceability of Greek PDO Vostizza currants. *Journal of Food Science*, 86(10):4417–4429, 2021.

[18] Francesco Longobardi, Valentina Innamorato, Annalisa Di Gioia, Andrea Ventrella, Vincenzo Lippolis, Antonio F. Logrieco, Lucia Catucci, and Angela Agostiano. Geographical origin discrimination of lentils (Lens culinaris Medik.) using1H NMR fingerprinting and multivariate statistical analyses. *Food Chemistry*, 237:443–448, 2017.

[19] Caroline Schmitt, Tobias Schneider, Laura Rumask, Markus Fischer, and Thomas Hackl. Food Profiling: Determination of the Geographical Origin of Walnuts by 1H NMR Spectroscopy Using the Polar Extract. *Journal of Agricultural and Food Chemistry*, 68(52):15526–15534, 2020.

[20] Qiang Wang, Xiaohui Wang, Xujin Wu, Yanli Wang, Yuli Zhang, Yaling Jiang, Chaonan Zhang, Xiaowan Huang, Li An, Huan Ma, and Kaihong Xu. 1H NMR-based metabolic profiling approach to identify the geo-authentic Chinese yam (Dioscorea polystachya Turczaninow cv. Tiegun). *Journal of Food Composition and Analysis*, 98(January), 2021.

[21] Rabelani Munyai, Maropeng Vellry Raletsena, and David Mxolisi Modise. LC-MS Based Metabolomics Analysis of Potato (Solanum tuberosum L.) Cultivars Irrigated with Quicklime Treated Acid Mine Drainage Water. *Metabolites*, 12(3), 2022.

[22] Gabriella Saviano, Debora Paris, Dominique Melck, Francesca Fantasma, Andrea Motta, and Maria Iorizzi. Metabolite variation in three edible Italian Allium cepa L. by NMR-based metabolomics: a comparative study in fresh and stored bulbs. *Metabolomics*, 15(8), 2019.

[23] Luciana Marçal Ravaglia, Ana Beatriz Coriguazi Pizzotti, and Glaucia Braz Alcantara. NMR-based and chemometric approaches applicable to adulteration studies for assessment of the botanical origin of edible oils. *Journal of Food Science and Technology*, 56(1):507–511, 2019.

[24] Todd M. Alam and M. Kathleen Alam. Chemometric Analysis of NMR Spectroscopy Data: A Review. *Annual Reports on NMR Spectroscopy*, 54:41–80, 2004.

[25] Yolanda Pérez, Marta Casado, Demetrio Raldúa, Eva Prats, Benjamín Piña,

Romà Tauler, Ignacio Alfonso, and Francesc Puig-Castellví. MCR-ALS analysis of 1H NMR spectra by segments to study the zebrafish exposure to acrylamide. *Analytical and Bioanalytical Chemistry*, 412(23):5695–5706, 2020.

[26] Anna De Juan, Joaquim Jaumot, and Romà Tauler. Multivariate Curve Resolution (MCR). Solving the mixture analysis problem. *Analytical Methods*, 6(14):4964–4976, 2014.

[27] Francesc Puig-Castellví, Ignacio Alfonso, and Romà Tauler. Untargeted assignment and automatic integration of 1H NMR metabolomic datasets using a multivariate curve resolution approach. *Analytica Chimica Acta*, 964:55–66, 2017.

[28] Nicola Cavallini, Francesco Savorani, Rasmus Bro, and Marina Cocchi. A Metabolomic Approach to Beer Characterization. *Molecules*, 26(5):1–15, 2021.

[29] Seung Ok Yang, Yoo Soo Shin, Sun Hee Hyun, Sayeon Cho, Kyong Hwan Bang, Dongho Lee, Seung Phill Choi, and Hyung Kyoon Choi. NMR-based metabolic profiling and differentiation of ginseng roots according to cultivation ages. *Journal of Pharmaceutical and Biomedical Analysis*, 58(1):19–26, 2012.

[30] F. Savorani, G. Tomasi, and S. B. Engelsen. icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal of Magnetic Resonance*, 202(2):190–202, 2010.

[31] Joaquim Jaumot, Anna de Juan, and Romà Tauler. MCR-ALS GUI 2.0: New features and applications. *Chemometrics and Intelligent Laboratory Systems*, 140:1–12, 2015.

[32] Robert A. van den Berg, Huub C.J. Hoefsloot, Johan A. Westerhuis, Age K. Smilde, and Mariët J. van der Werf. Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics*, 7:1–15, 2006.

[33] Parvaneh Ebrahimi, Nanna Viereck, Rasmus Bro, and Søren B Engelsen. Modern Magnetic Resonance. *Modern Magnetic Resonance*, pages 1–20, 2017.

[34] Bekzod Khakimov, Nabiollah Mobaraki, Alessia Trimigno, Violetta Aru, and

Søren Balling Engelsen. Signature Mapping (SigMa): An efficient approach for processing complex human urine 1H NMR metabolomics data. *Analytica Chimica Acta*, 1108:142–151, 2020.

[35] Behzad Gazme, Ruth T. Boachie, Apollinaire Tsopmo, and Chibuike C. Udenigwe. Occurrence, properties and biological significance of pyroglutamyl peptides derived from different food sources. *Food Science and Human Wellness*, 8(3):268–274, 2019.

[36] Elisangela F. Boffo, Leila A. Tavares, Antonio C.T. Tobias, Márcia M.C. Ferreira, and Antonio G. Ferreira. Identification of components of Brazilian honey by 1H NMR and classification of its botanical origin by chemometric methods. *Lwt*, 49(1):55–63, 2012.

[37] David S. Wishart, Craig Knox, An Chi Guo, Roman Eisner, Nelson Young, Bijaya Gautam, David D. Hau, Nick Psychogios, Edison Dong, Souhaila Bouatra, Rupasri Mandal, Igor Sinelnikov, Jianguo Xia, Leslie Jia, Joseph A. Cruz, Emilia Lim, Constance A. Sobsey, Savita Shrivastava, Paul Huang, Philip Liu, Lydia Fang, Jun Peng, Ryan Fradette, Dean Cheng, Dan Tzur, Melisa Clements, Avalyn Lewis, Andrea De souza, Azaret Zuniga, Margot Dawe, Yeping Xiong, Derrick Clive, Russ Greiner, Alsu Nazyrova, Rustem Shaykhutdinov, Liang Li, Hans J. Vogel, and Ian Forsythei. HMDB: A knowledgebase for the human metabolome. *Nucleic Acids Research*, 37(SUPPL. 1):603–610, 2009.

[38] A. Palmonari, D. Cavallini, C. J. Sniffen, L. Fernandes, P. Holder, L. Fagioli, I. Fusaro, G. Biagi, A. Formigoni, and L. Mammi. Short communication: Characterization of molasses chemical composition. *Journal of Dairy Science*, 103(7):6244–6249, 2020.

[39] Guillermo Montoya, July Londono, Paola Cortes, and Olga Izquierdo. Quantitation of trans -Aconitic Acid in Di ff erent Stages of the Sugar-Manufacturing Process. pages 6–10, 2014.

4.6 Supplementary Material

Table 4.6.1: Processed ^{1}H NMR dataset manually divided according to their resonance signals for the application of segmented analysis via MCR–ALS.

Subarray	Matrix size	Subarray	Matrix size
1	25 × 4400	16	25 × 180
2	25 × 400	17	25 × 85
3	25 × 1700	18	25 × 265
4	25 × 750	19	25 × 365
5	25 × 950	20	25 × 285
6	25 × 500	21	25 × 280
7	25 × 1300	22	25 × 440
8	25 × 600	23	25 × 800
9	25 × 1150	24	25 × 940
10	25 × 650	25	25 × 360
11	25 × 400	26	25 × 900
12	25 × 500	27	25 × 4250
13	25 × 350	28	25 × 3350
14	25 × 470	29	25 × 8886
15	25 × 380		

Table 4.6.2: Spectral regions for each concentration ($\mathbf{C}$) and spectral ($\mathbf{S}^T$) profiles resolved by MCR–ALS due to their chemical shifts.

C-S^T	**Chemical shift ($\delta_\mathbf{H}$)**	**Spectral region**	**C-S^T**	**Chemical shift ($\delta_\mathbf{H}$)**	**Spectral region**
		Aliphatic			**Carbohydrates**
1	0.20 – 1.20		20	3.71 – 3.79	
2	0.20 – 1.20		21	3.79 – 3.84	
3	1.26 – 1.38		22	3.84 – 3.86	
4	1.40 – 1.75		23	3.86 – 3.92	
5	1.80 – 1.96		24	3.92 – 4.02	
6	1.80 – 1.96		25	4.02 – 4.08	
7	1.95 – 2.20		26	4.09 – 4.15	
8	1.95 – 2.20		27	4.14 – 4.26	
9	2.20 – 2.32		28	4.26 – 4.46	
10	2.30 – 2.65		29	4.50 – 5.10	
11	2.30 – 2.65		30	5.18 – 5.26	
12	2.30 – 2.65		31	5.18 – 5.26	
13	2.64 – 2.78		32	5.25 – 5.50	
14	2.80 – 3.05		33	5.50 – 6.50	
15	3.08 – 3.22				**Aromatic/Aldehyde**
16	3.24 – 3.38		34	6.50 – 7.40	
17	3.40 – 3.50		35	6.50 – 7.40	
		Carbohydrates	36	6.50 – 7.40	
18	3.50 – 3.59		37	7.00 – 9.50	
19	3.60 – 3.70				

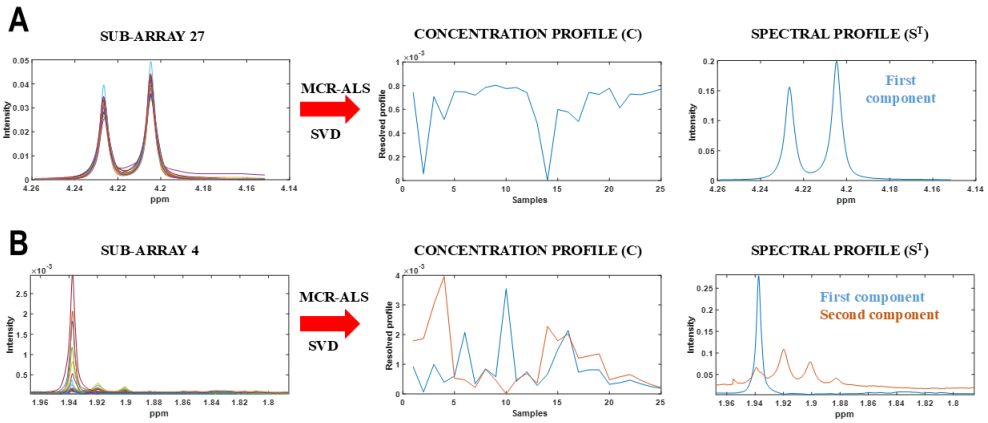

Figure 4.6.1: Example of subarrays manually selected for segmented analysis via MCR–ALS. The segmentation of resonance signals in the processed ^{1}H NMR dataset was resolved using the SVD method to determinate the optimal number of components that resolved the $\mathbf{C}$ and $\mathbf{S}^T$ profiles according to their eigenvalues. **A)** Subarray containing a resonance signal in an interval of $\delta_H = 4.16 - 4.25$. The $\mathbf{C}$ and $\mathbf{S}^T$ profiles were resolved by the first component according to the eigenvalues obtained by SVD. **B)** Subarray containing resonance signals in an interval of $\delta_H = 1.80 - 1.96$. Two $\mathbf{C}$ and $\mathbf{S}^T$ were clearly resolved by the first and second component according to the eigen values obtained by SVD.

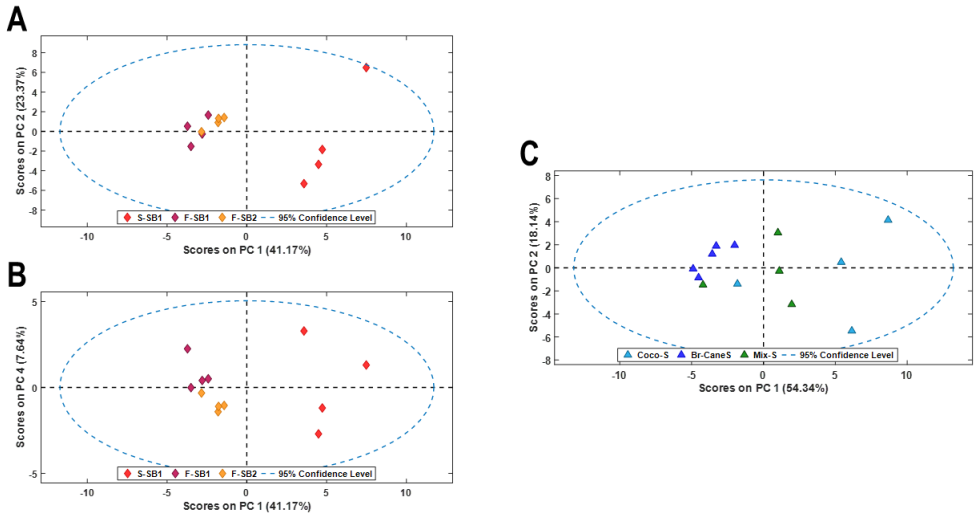

Figure 4.6.2: PCA score plots derived from the combined C matrix obtained via MCR–ALS using mean centering and variance scaling as a pretreatment. **A)** and **B)** score plots of PC1 vs PC2 and PC1 vs PC4, for the differentiation of sugar beets. **C)** PCA score plot (PC1 vs PC2) for the differentiation of commercial sugars. S-SB1: storage sugar beet from Santa Carlos (red); F-SB1: fresh sugar beet from San Carlos (brown); F-SB2: fresh sugar beet from Los Ángeles (yellow); Coco-S: coconut sugar (light blue); Br-CaneS: brown cane sugar (blue); Mix-S (green).

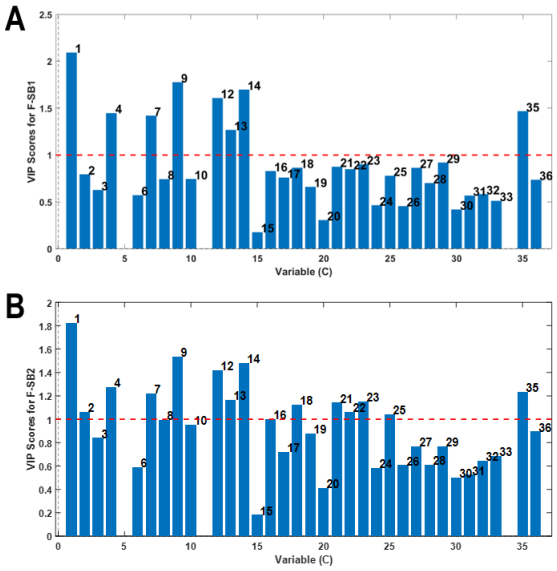

Figure 4.6.3: VIP plots obtained by PLS–DA derived from the combined **C** matrix for fresh sugar beet samples from San Carlos (F-SB1) and Los Ángeles (F-SB2). **A)** VIP plot for F-SB1. **B)** VIP plot for F-SB2. After the analysis of the statistical significance of each relevant variable (VIP $\geq$ 1), only variable 1 (leucine) was considered discriminant to differentiate F-SB1 from F-SB2.

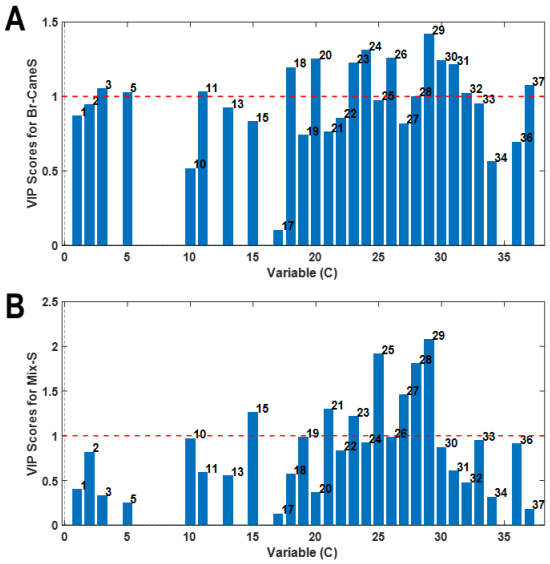

Figure 4.6.4: VIP plots obtained by PLS–DA derived from the combined **C** matrix brown coconut sugar (Br-CaneS) and simulated adulterations (Mix-S). **A)** VIP plot for Br-CaneS. **B)** VIP plot for Mix-S. After the analysis of the statistical significance of each relevant variable (VIP$\geq$ 1), variables 31 (α-xylose) and 32 (sucrose) were considered discriminant to differentiate Br-CaneS from coconut sugar (Coco-S).

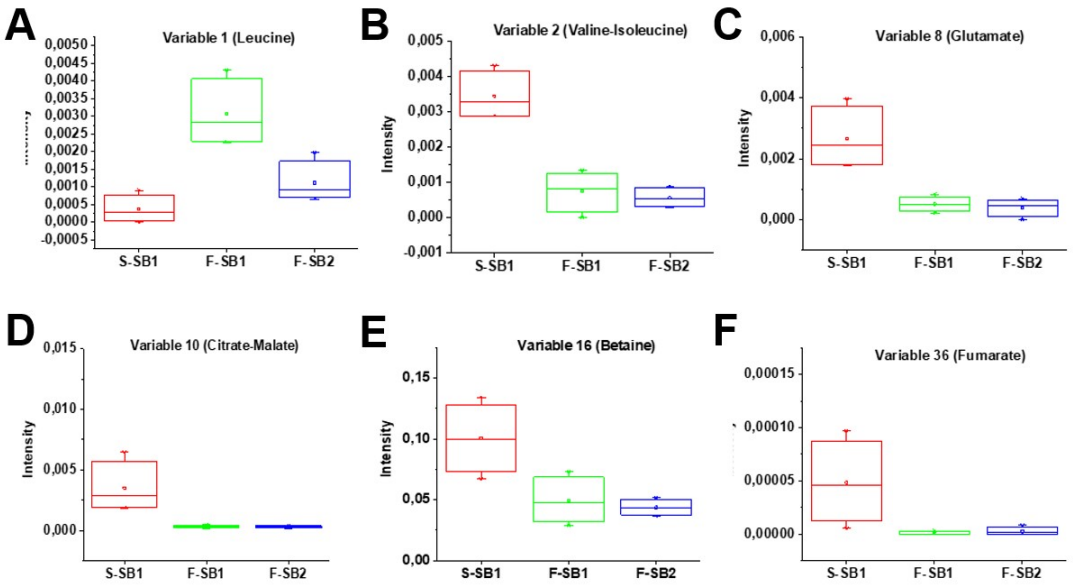

Figure 4.6.5: Box whisker plots of variables that are relevant to discriminate stored sugar beet from San Carlos (S-SB1) with fresh sugar beet from San Carlos (F-SB1) and Los Ángeles (F-SB2) under statistical significance by a one-way ANOVA test with Bonferroni correction. Variable 1 (leucine) was relevant for the differentiation of F-SB1 samples. Variables 2 (valine-isoleucine), 8 (glutamate), 10 (citrate-malate), 16 (betaine), and 36 (fumarate), showed higher intensities of concentration profiles in S-SB1 samples, and therefore relevant for the differentiation of S-SB1 samples with F-SB1 and F-SB2.

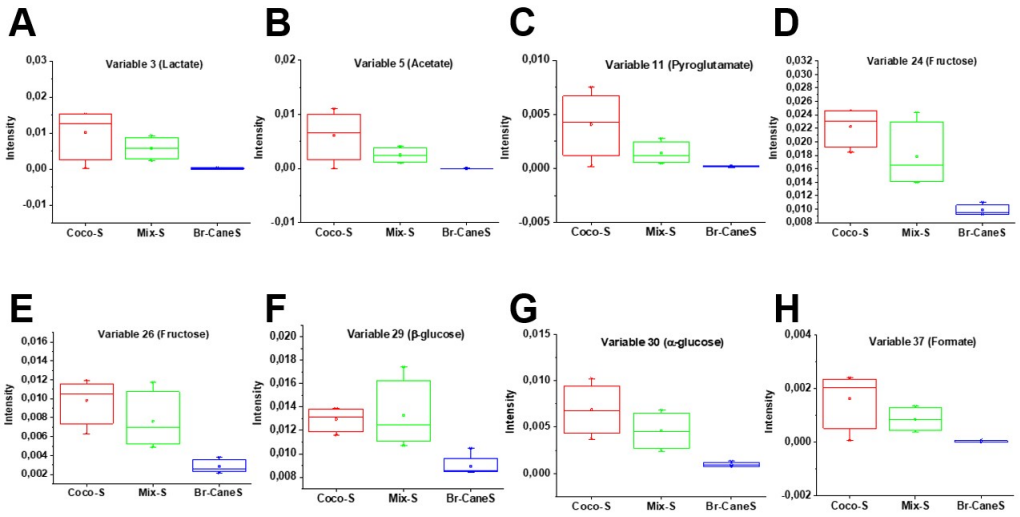

Figure 4.6.6: Box whisker plots of variables that are relevant to discriminate coconut sugar (Coco-S) with brown cane sugar (Br-CaneS) and simulated adulterations (Mix-S) under statistical significance by a one-way ANOVA test with Bonferroni correction. All the variables showed higher intensities of concentration profiles in Coco-s compared to the Br-CaneS samples, and therefore relevant for the differentiation of Coco-S.

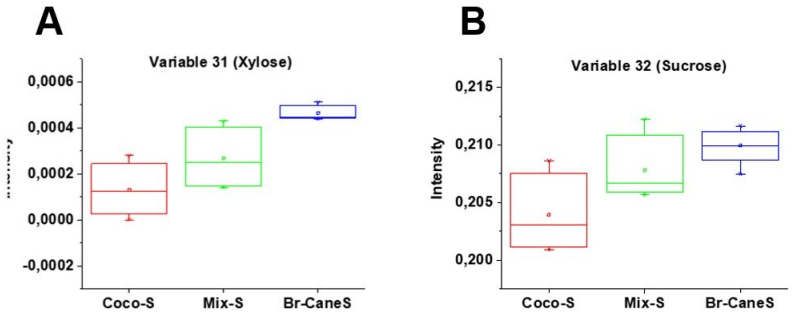

Figure 4.6.7: Box whisker plots of variables that are relevant to discriminate brown cane sugar (Br-CaneS) with coconut sugar (Coco-S) and simulated adulterations (Mix-S) under statistical significance by a one-way ANOVA test with Bonferroni correction. Variables 31 (α-xylose) and 32 (sucrose) showed higher intensities of concentration profiles Br-CaneS, and therefore relevant for the differentiation of Coco-S.

Chapter 5

Interval Resonance Analysis (InRA): A versatile tool for automated untargeted ^{1}H NMR fingerprinting – A case study in sugar beet field authentication

This chapter presents and discusses the results that served as the basis for the second article, titled "Interval Resonance Analysis (InRA): A versatile tool for automated untargeted ^{1}H NMR fingerprinting – A case study in sugar beet field authentication" published in May 2025 in the journal *Analytica Chimica Acta* (https://doi.org/10.1016/j.aca.2025.344175).

This article addressed specific objectives 1 and 2. An MCR-based automated analytical workflow was developed to reduce reliance on manual spectral processing. The proposed methodology was implemented within a newly designed graphical user interface (GUI), which consolidated the analytical pipeline into a unified platform, thereby eliminating the need for multiple software tools.

The performance of this new tool was validated by discriminating sugar beet root samples originating from different cultivation fields. Several classification models were constructed, which demonstrate strong performance and highlight the effectiveness of the optimized approach relative to conventional analytical methodologies.

Additionally, objectives 3 and 4 were partially addressed through the chemical characterization of the predominant constituents in the root samples and the statistical identification of compounds that contributed significantly to sample differentiation.

Interval Resonance Analysis (InRA): A versatile tool for automated untargeted ^{1}H NMR fingerprinting – A case study in sugar beet field authentication

Cristian A. Fuentes[a,b], David Montoya[b,c], Mecit Halil Öztop[d], Macarena Rojas-Rioseco[a,b], Martín Bravo[a,b], Francisca González[b], Rosario del P. Castillo[a,b*]

[a]Departamento de Análisis Instrumental, Facultad de Farmacia, Universidad de Concepción, Concepción 4070386, Chile
[b]Laboratorio de Bioespectroscopia y Quimiometría (BioSpeQ), Centro de Biotecnología, Universidad de Concepción, Concepción 4070386, Chile
[c]Departamento de Física, Facultad de Ciencias Físicas y Matemáticas, Universidad de Concepción, Concepción 4070386, Chile
[c]Department of Food Engineering, Middle East Technical University, Ankara 06800, Türkiye

***Corresponding author:**
Rosario del P. Castillo
Tel: +56 41 220 7298
Email: rosariocastillo@udec.cl

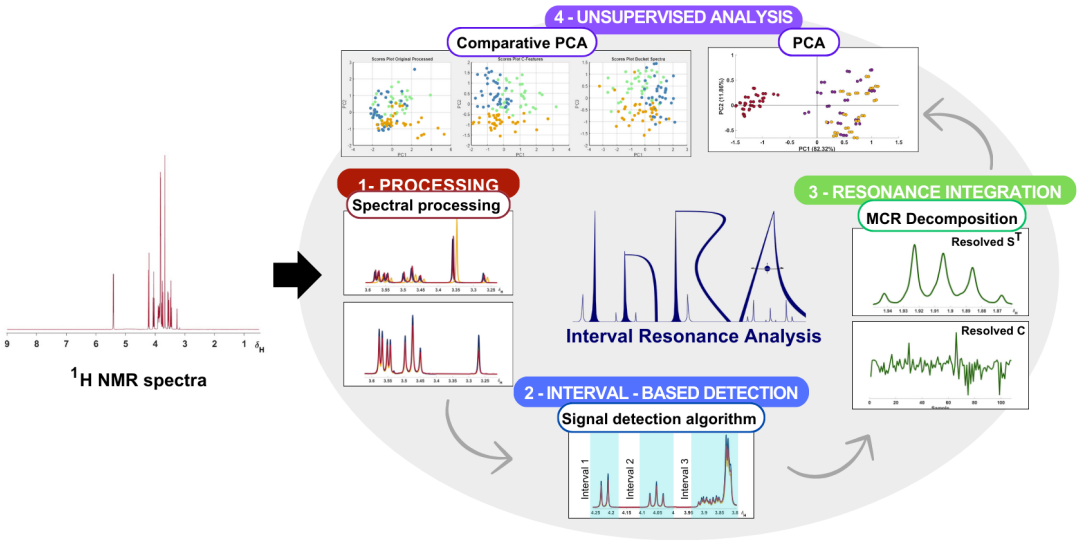

Figure 5.0.1: Graphical abstract

Abstract

Background: The extraction of relevant information from proton nuclear magnetic resonance (^{1}H NMR) spectra through preprocessing and multivariate analysis requires integrating multiple software tools and extensive manual intervention, compromising efficiency and reproducibility when the technique is used. Consequently, the development of automated, versatile, and reliable methodologies has become imperative to streamline workflows, improve analytical performance, and broaden the applicability of multivariate methods for the analysis of diverse sample types and experimental conditions.

Results: This work presents the development and application of Interval Resonance Analysis (InRA), an alternative software tool focused on ^{1}H NMR multivariate analysis. InRA includes a novel algorithm for resonance signal detection (intervals), specifically designed to operate with flexibility across diverse ^{1}H NMR spectra. All intervals are integrated using multivariate curve resolution with alternating least squares (MCR–ALS) and analyzed by exploratory analysis. The performance of InRA was tested by evaluating the ^{1}H NMR spectra of hydrophilic sugar beet root extracts cultivated in three different fields and their discrimination by partial least squares – discriminant analysis (PLS–DA). The workflow provided by InRA yielded consistent results regarding the distribution of samples according to their field, enabling identification of subtle sources of variation and achieving classification accuracies $\geq 88.9\%$.

Significance: The proposed methodology represents an advancement in the multivariate analysis of ^{1}H NMR spectra for untargeted studies and enhances analytical efficiency by reducing manual intervention and reliance on analyst experience. InRA is versatile and can be applied to various sample types and analytical objectives, as it is not restricted by specific experimental conditions.

Keywords – ^{1}H NMR spectroscopy, food profiling, multivariate curve resolution, preprocessing, sugar beet

5.1 Introduction

The use of nuclear magnetic resonance (NMR) spectroscopy on different matrices (e.g., foods, tissues, and biofluids) for untargeted analyses has gained prominence due to its reproducibility, quantitative capability, minimal sample preparation, and ability to perform multi-detection in a single experiment [1, 2, 3]. In particular, proton NMR (^{1}H NMR) spectroscopy is the predominant method to asses chemical variability in complex matrices due to its inherent capacity to provide detailed information on the molecular structure of numerous compounds. Therefore, ^{1}H NMR spectroscopy is invaluable for the evaluation of spectroscopic profiles by untargeted analysis, as it serves as a comprehensive fingerprint technique to facilitate the identification of potential sources of variation and biomarkers [4]. However, ^{1}H NMR spectra are often too complex for direct interpretation; thus, rigorous preprocessing combined with multivariate analysis (chemometrics) is necessary to address and reduce variability arising from the inherent complexity of spectra produced by frequency shifts, spectral noise, and signal overlap [5].

Despite the well-established workflow to process and analyze ^{1}H NMR spectra for untargeted studies (i.e., alignment, normalization, and bucketing), in recent years, an alternative methodology based on the decomposition of small intervals containing one or more resonance signals via multivariate curve resolution (MCR) has been successfully applied to reduce spectral complexity, improve the quality of chemometric models, identify significant variables preserving essential spectroscopic information (multiplicity and coupling constants), separate overlapped signals, and provide a robust option for quantification purposes [6, 7, 8, 9, 10, 11]. Nevertheless, the applicability of this particular methodology often involves a predominantly manual labor-intensive process. The initial detection and selection of the intervals is especially challenging for resonance signals with low signal-to-noise ratios (SNR). Furthermore, each interval must be independently decomposed via MCR, requiring the generation and evaluation of separate MCR models. The latter yields a substantial amount of spectral information, demanding significant input and expertise from the analyst to ensure accurate results. Moreover, to perform a comprehensive analysis, the integration of multiple software tools is necessary, i.e., processing the ^{1}H NMR spectra, performing MCR decomposition, and carrying out exploratory/discriminant

analyses, which increases the complexity of the process and poses challenges in reproducibility and efficiency. These limitations highlight the need to develop versatile solutions capable of streamlining and optimizing the workflow for broader applicability and improved analytical performance [12].

Important software tools have been developed to solve particular automation requirements for ^{1}H NMR spectra, e.g., processing and quantification in complex biofluid samples [13, 14], unsupervised and supervised analyses for metabolomics [15], or peak extraction [16]. However, their development has also underscored the growing need for versatile approaches to handle various type of samples and experimental conditions, with a specific emphasis on automated resonance detection, interval decomposition via MCR, and exploratory analysis within a single integrated platform.

This work presents Interval Resonance Analysis (InRA), a MATLAB-based graphical user-friendly interface (GUI) designed for untargeted ^{1}H NMR fingerprinting analysis. InRA features specialized algorithms to perform a comprehensive analytical pipeline, including spectral alignment, normalization, interval-based resonance detection, feature integration via MCR, parameter optimization, principal component analysis (PCA), and proper figures of merit for accessible interpretation of results. The spectroscopic ^{1}H NMR profile of sugar beet roots (*Beta vulgaris L.*) was used as a case study to validate InRA. NMR studies on sugar beet remain limited, partly due to the analytical challenges posed by its high sucrose content, which typically ranges from $18-20\%$ [17, 18]. Existing research has primarily addressed metabolic variations under intermittent drought conditions and resistance to *Cercospora* leaf spots [19, 20]. Moreover, the potential of ^{1}H NMR spectroscopy to differentiate sugars based on storage conditions and botanical sources was previously reported using manual analysis [21]. In this study, the authentication of three distinct sugar beet cultivation fields was used to evaluate the performance of InRA as a versatile tool designed for multivariate analysis of ^{1}H NMR spectra, streamlining the analytical workflow, from spectral processing to result interpretation, by employing specialized algorithms for identifying sources of variation while preserving relevant spectral information and providing an effective alternative for characterization, authentication, or traceability purposes.

5.2 Interval Resonance Analysis (InRA) workflow

InRA GUI (v1.0) consists of m. files developed under MATLAB R2023b (TheMathWorks Inc., Natick, MA, USA) environment and is compatible with MATLAB R2020b and later. InRA is freely available at https://github.com/InRA-Software/InRA and it does not require any additional toolboxes for proper operation. The GUI analytical workflow is organized into four main tabs: **I) Processing, II) Interval-based detection, III) Resonance Integration** and **IV) Unsupervised Analysis**. Fig. 5.2.1 provides a schematic overview of the key functionalities integrated into InRA, and a detailed description of each tab is outlined below.

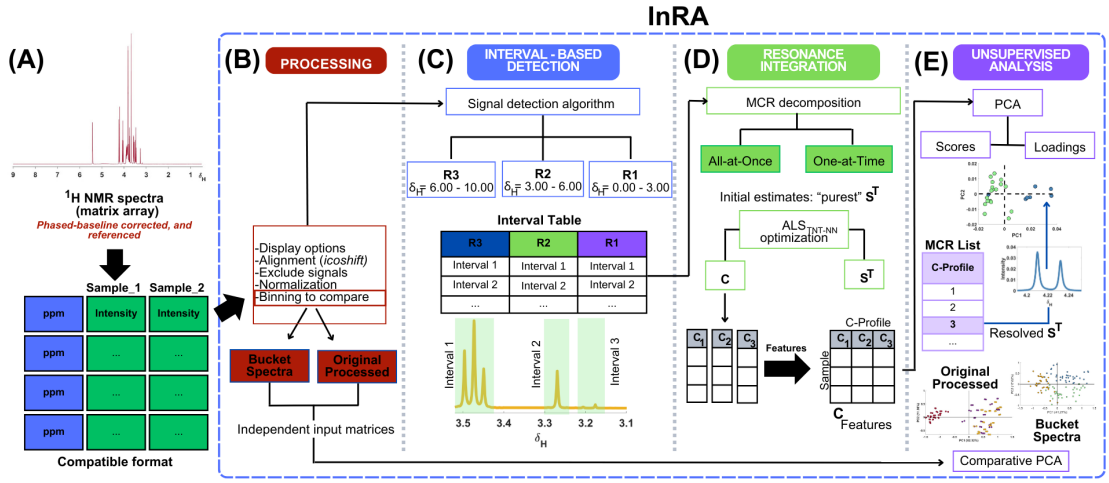

Figure 5.2.1: Illustrative workflow of the untargeted ^{1}H NMR fingerprint analysis using InRA.

5.2.1 Processing

The matrix array required to import a set of ^{1}H NMR spectra into the current version of InRA (v1.0) must be previously phased-baseline corrected, referenced, and could be in .csv, .xlsx, or .xls format. The first column must contain the ppm values, and the subsequent columns should include the intensity values for each sample as is shown in Fig. 5.2.1A. The import process offers different visualization options (e.g., each spectrum individually), displays the size of the spectral matrix and the possibility to delete samples before preprocessing as needed.

Regarding the processing options (Fig. 5.2.1B), the selection of a defined spectral range and undesired resonance signals (e.g., water after presaturation) can be removed by entering chemical shift values adjustable in increments of ± 0.01 ppm. In both cases, a lower and higher ppm value must be specified, e.g., $\delta_H = 0.20$ and $\delta_H = 0.85$, respectively. Given that frequency shifts in resonance signals can occur among different samples, a thorough spectral alignment is necessary to ensure the reliability of the results remains unaffected by variations unrelated to the intrinsic variance of the samples under study. Moreover, well-aligned spectra are essential to avoid spectral artifacts that can significantly reduce the accuracy of resonance interval detection and their subsequent integration via MCR [22]. Therefore, the resonance misalignments are corrected using the *icoshift* algorithm developed by Savorani et al., which has been adapted and implemented into InRA [23]. Two selectable configurations have been incorporated to perform alignment. The first defines the target vector for the alignment (i.e., reference): *average*, *median*, *maximum* (spectrum with the most intense features), or *average2* (the average of the average multiplied by a specified value) spectrum. The second determines the alignment mode: *whole* to apply the correction across the entire spectral range or *interval*, which requires selecting specific segments containing a group of signals to optimize results. If interval-based alignment is chosen, intervals can be easily added, removed, or edited to suit the characteristics of the ^{1}H NMR spectra. Furthermore, two standard normalization methods have been integrated: *Norm-1* (total sum norm) and *Norm-2* (Euclidean norm).

The option to apply the "binning" technique according to analytical needs has also been incorporated. The equidistant binning (equal size) method has been implemented, which requires the prior specification of the bucket size (± 0.01 ppm). The latter functionality can be employed through two distinct approaches into InRA. The first approach is to apply binning to reduce the number of variables in the processed ^{1}H NMR spectra and continue with the subsequent steps of the workflow. However, the use of excessively large bin sizes (e.g., 0.04 ppm) is not recommended, as the core methodology implemented in InRA was designed to preserve essential spectroscopic information (i.e., avoid loss of spectral resolution). The second approach corresponds to a functionality called *Binning to compare*. When the latter option is enabled, an optimal bin size must be specified, and the resulting matrix is internally stored under the name *Bucket Spectra* (hereafter,

$\mathbf{B}_{Spectra}$). In parallel, once the *Binning to compare* option is activated, the processed matrix without binning (i.e., the ^{1}H NMR spectra previously range specified, aligned and normalized) is also automatically stored with the name of *Original Processed* (hereafter, $\mathbf{P}_{Spectra}$). $\mathbf{B}_{Spectra}$ and $\mathbf{P}_{Spectra}$ serve as independent input matrices that can be used for their comparison via PCA once the analytical workflow has been completed. Further details on this functionality are provided in Section 5.2.3.5. Finally, the processed ^{1}H NMR spectra (or bucket ^{1}H NMR spectra) can be exported to the MATLAB workspace as a spectral matrix or saved as a .csv file for other specific analyses or applications.

5.2.2 Interval-based detection

improve feature extraction in ^{1}H NMR spectra, a novel automated algorithm for interval-based resonance signal detection was developed and integrated into the InRA framework. Following appropriate processing of ^{1}H NMR spectra (i.e., signal removal, alignment, and normalization), the next step involves an automated signal detection performed within three predefined spectral regions: $\delta_H = 0.00 - 3.00$ (low frequencies), $\delta_H = 3.00 - 6.00$ (medium frequencies), and $\delta_H = 6.00 - 10.00$ (high frequences) as shown in Fig. 5.2.1C. Predefined limits, namely $\delta_H = 3.00$ and $\delta_H = 6.00$, may be modified according to the specific requirements and characteristics of the ^{1}H NMR spectra. The detection works independently among these three regions and only requires an input threshold (by default 0.1000) for each spectral region. The threshold can be adjusted between 0.0001 (0.01%) and 0.7500 (75%) and three settings parameters have been implemented to optimize the interval-based detection: *detection window (D)*, *distance of consecutive maxima (L)*, and *separation distance (H)*, with default values of 0.01, 0.03, and 0.005 ppm, respectively. After the detection settings and best threshold values are selected, the detected intervals are presented by three distinct colors (purple, green, and blue) and organized into three independent tables, labeled R1, R2 and R3. These tables provide information about the chemical shift range and the identification number of each interval based on their respective spectral region. Upon completing the interval-based resonance signal detection, options to delete, edit or add intervals are available as optimization parameters. If necessary, the intervals can be exported to the MATLAB workspace in cell format as *IntervalsRegion1*, *IntervalsRegion2*, and *IntervalsRegion3*, respectively. Each cell contains two columns: one for the

ppm values and the other for the associated intensities.

The operational framework of the developed algorithm is described in Section 5.2.3, and a detailed explanation of the required conditions is provided in the flowcharts of supplementary Fig. 5.7.1, 5.7.2,5.7.3, respectively.

5.2.3 Framework of the interval-based detection algorithm

The proposed algorithm for automated signal detection is composed of three essential sub-algorithms called: **maxima and minima**, **interval delimitation**, and **interval correction**. Each sub-algorithm was designed under predefined conditions to minimize errors from baseline drift, variations in the SNR, and non-representative fluctuations to ensure precise and reliable delimitation of the intervals. These conditions were specifically established to operate solely based on spectral information, i.e., considering the shape and multiplicity of the resonance signals, without dependence on chemical shifts due to experimental conditions (e.g., pH, temperature, pulse sequence, or extraction method) used for sample preparation and spectral acquisition. As a result, the detection is not restricted to any particular sample type and can be applied to a wide range of ^{1}H NMR spectra, outlining the robustness and versatility of the algorithm for feature extraction. A schematic representation of the operational workflow of the proposed algorithm is shown in Fig. 5.2.2.

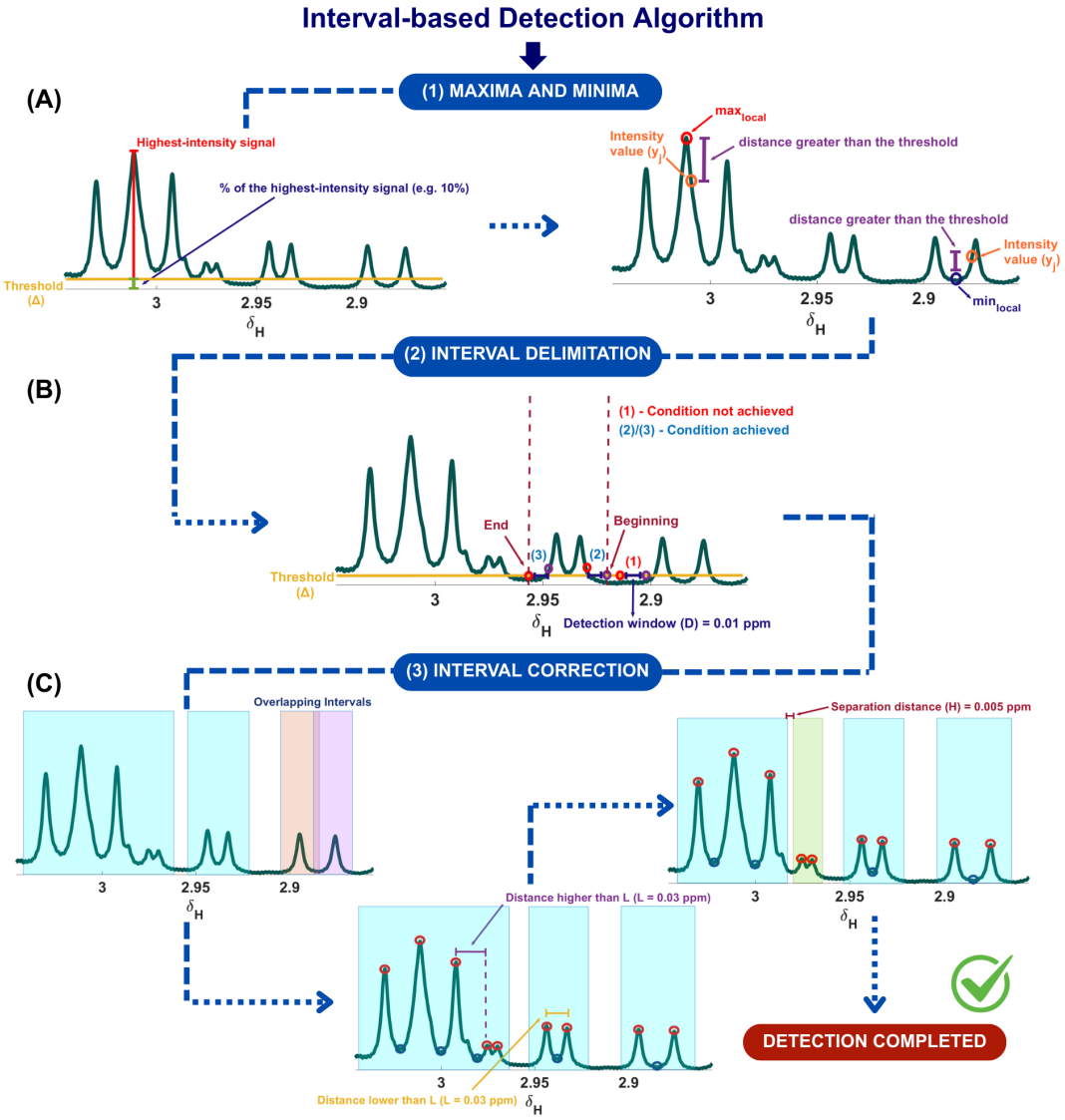

Figure 5.2.2: Schematic representation of the operational framework for the interval-based detection algorithm. **A)** Identification of $\max_{local}$ and $\min_{local}$ according to the comparative distance with the threshold value. **B)** Example of interval delimitation with a detection window (D) of 0.01 ppm. **C)** Correction of overlapping intervals.

5.2.3.1 Maxima and minima algorithm

The fundamental basis of the main algorithm is rooted in the detection of maxima ($\max_{local}$) and minima ($\min_{local}$) of the signal across a given spectral region. For this purpose, two elements are considered: a reference vector with the average intensity of the resonance signals $\bar{\mathbf{y}}$ and the threshold value (Δ). Δ is defined as a percentage of the highest-intensity resonance signal within the analyzed spectral

region (e.g., 10%). A visual representation of both elements, i.e., $\bar{\mathbf{y}}$ and Δ, are shown in Fig. 5.2.2A. To identify a $\max_{local}$ or $\min_{local}$, an iterative process is carried out through the vector $\bar{\mathbf{y}}$, where intensity values ($\bar{y}_j$) are dynamically assigned as temporary maximum ($\max_{temp}$) and temporary minimum ($\min_{temp}$). If the distance between a given $\bar{y}_j$ and the current $\max_{temp}$ or $\min_{temp}$ exceeds a predefined Δ, the value of $\bar{y}_j$ is updated and assigned as a $\max_{local}$ or $\min_{local}$. The latter procedure is illustrated in Fig. 5.2.2A and evaluated according to the following expressions:

$$\max_{temp} - \bar{y}_j > \Delta \quad \longrightarrow \quad \max_{local} = \max_{temp} \tag{5.2.1}$$

$$\bar{y}_j - \min_{temp} > \Delta \quad \longrightarrow \quad \min_{local} = \min_{temp} \tag{5.2.2}$$

One a $\max_{local}$ and $\min_{local}$ are detected based on Eq. 5.2.1 and 5.2.2, the identified values are saved, and the cycle is repeated iteratively until the last element of the vector $\bar{\mathbf{y}}$ is reached.

5.2.3.2 Interval delimitation algorithm

The resonance signals detection procedure is based on the correct delimitation of intervals where the signal returns to the baseline. To achieve the latter, the interval delimitation algorithm incorporates the following elements: the maxima and minima algorithm, Δ, a *detection window (D)*, a modified vector ($\bar{\mathbf{u}}'$), and a $\mathbf{p}$ vector. $\bar{\mathbf{u}}'$ is defined as the average intensity vector where values below Δ are systematically substituted by the Δ value. $\mathbf{p}$ contains the ppm values. To initiate signal detection and subsequent interval delimitation, a difference of D ppm (default $D = 0.01$ ppm) is set as a conditional parameter to perform a sequential scan across the components of the vector $\bar{\mathbf{u}}'$. If the values compared within the 0.01 ppm window are equal and do not exceed the defined threshold, the condition is not achieved, and the presence of a signal is not established. Conversely, if at least one of the values within the window differs from the threshold, the condition is addressed at a specific index d, as illustrated in Fig. 5.2.2B and using the following mathematical expression:

$$D = p_{1+d} - p_1 \tag{5.2.3}$$

Based on Eq. 5.2.3, the initial delimitation of an interval is established to begin at index b and end at index e where the values of the components of vector $\bar{\mathbf{u}}'$ are equal to the threshold within a detection window of 0.01 ppm (Fig. 5.2.2B), when the following conditions presented in Eq. 5.2.4 and 5.2.5 are satisfied:

$$\bar{u}'_b > \Delta \quad ; \quad \bar{u}'_{b-d} = \Delta \tag{5.2.4}$$

$$\bar{u}'_e = \Delta \quad ; \quad \bar{u}'_{e-d} > \Delta \tag{5.2.5}$$

5.2.3.3 Interval correction algorithm

While initial delimitation can be successfully achieved, overlaps between intervals may occur. This phenomenon arises when the final boundary of one interval coincides with the initial boundary of an adjacent interval, as shown in Fig. 5.2.2C. The presence of overlaps can compromise the accuracy of the resonance signal detection analysis, making it essential to apply a corrective strategy. Therefore, overlapping intervals are merged into a single interval with the new boundaries defined by the initial index of the first interval and the final index of the second. The $\max_{local}$ and $\min_{local}$ values are then re-identified within the resulting unified interval Fig. 5.2.2C). However, it is necessary to redefine the intervals to ensure individual representation of signals with different multiplicity. Therefore, the vector $\mathbf{p}$ is used to measure the distance between consecutive maxima, as a verifying analysis of the resonance J-couplings according to the following condition:

$$p_{\mathrm{idx}^{\max}_{i,\,k+1}} - p_{\mathrm{idx}^{\max}_{i,\,k}} > L \tag{5.2.6}$$

where $\mathrm{idx}^{\max}_{i,\,k}$ represents the index of the component in vector that corresponds to the ppm value of the *k-th* $\max_{local}$ within the *i-th* detected interval. If the condition stated in Eq. 5.2.6 is satisfied, i.e., the distance between consecutive maxima exceeds a specific value of L ppm (with $L = 0.03$ ppm by default), a division is made at the position of the minlocal located between two consecutive maxima. The interval split is performed with a separation distance (H) of 0.005 ppm (by default) between the end of the new interval and the beginning of the adjacent interval, as can be seen in Fig. 5.2.2C.

5.2.3.4 Resonance integration

All the intervals previously detected can be represented as an independent matrix **D**, where the rows (I) denote the intensities of the resonance signals for each sample, and the columns (J) corresponds to the chemical shift values (δ_H). **D** can be decomposed as a bilinear model via MCR as described by Puig-Castellví et al. [6]. MCR is a mathematical decomposition method that separates an overall mixed instrumental response into individual components within a system according to the following equation [24]:

$$\mathbf{D} = \mathbf{C}\mathbf{S}^T + \mathbf{E} \tag{5.2.7}$$

where **C** ($I \times N$) is the matrix of the relative concentrations (integrated features), $\mathbf{S}^T$ ($N \times J$) is the matrix of the pure resolved spectra of the resonance signals, N is the number of components estimated for the decomposition, and **E** ($I \times J$) is the residual matrix with the variance of the unmodeled data. Eq. 5.2.7 is iteratively resolved using an alternating least squares (ALS) algorithm, which computes the **C** and $\mathbf{S}^T$ matrices to achieve the best fist for the experimental matrix **D** [25].

Therefore, in the next stage, MCR is applied to decompose each interval. To streamline the process quickly and easily, InRA incorporates two iterative initialization strategies: *All-at-Once* and *Once-at-Time* (Fig. 5.2.1D). In the *All-at-Once* approach, a fixed N (e.g., 3) must be defined (which is automatically assigned to all **D** matrices) and a required number of iterations, providing a simplified and straightforward process. In contrast, for the *One-at-Time* option, it is necessary to individually determine the optimal N value to explain the variance of each matrix **D**. The latter is supported by a module called *purest* $\boldsymbol{S}^T$, where N is selected based on the evaluation of the magnitude of the eigenvalues obtained via singular value decomposition (SVD). The SVD displays automatically up to a default maximum of 10 components for each interval. Once N is defined, the initial spectral profiles estimates are obtained using the most dissimilar rows ("purest" $\mathbf{S}^T$) through an analogous method to the SIMPLISMA algorithm [26]. After the initial estimates are evaluated, and the appropriate number of iterations has been entered, an optimized multi-modelling approach is applied to integrate the resonance signals within each interval (features).

Due to rotational ambiguities in the solution of $\mathbf{C}$ and $\mathbf{S}^T$, the MCR model implemented in InRA employs the TNT-NN algorithm (an improved version of FNNLS) proposed by Myre et al., [27, 28] to enforce non-negativity constraints in both directions ($\mathbf{C}$ and $\mathbf{S}^T$), restrict the range of feasible solutions, and reduces the analysis time of multiple MCR models being computed simultaneously during the ALS optimization. To assess model quality, the percent of lack of fit (%LOF), which evaluates the difference between the original matrix $\mathbf{D}$ and the reproduced $\hat{\mathbf{D}}$ from the $\mathbf{CS}^T$, the percent of explained variance (%R^2), and the standard deviation of residuals relative to the experimentally calculated data (σ) are displayed. The specific equations can be found in Supplementary Table 5.7.1. The convergence criteria were defined when the variability of σ is below 0.1 or when there is no variability of σ between 25 iterations in a row. To ensure that only chemically relevant signals are included, pure resonance signals that are poorly resolved, contain artifacts, or are identified as noise, can be removed from the $\mathbf{C}$ and $\mathbf{S}^T$ matrices for each MCR model. Once only the relevant components remain, the set of integrated features ($\mathbf{C}$) is automatically merged column-wise to generate a new matrix called $\mathbf{C}_{Features}$ ($I \times K$) for further unsupervised analysis. K represents the number of solved pure resonance signals in a specific interval. Additionally, the option to export the MCR models to the MATLAB workspace as a structure (*ModelsRegion1*, *ModelsRegion2*, and *ModelsRegion3*) is available. Each structure contains seven independent cells with all relevant information, i.e., ppm, $\mathbf{C}$, $\mathbf{S}^T$, N, %LOF, %R^2, and σ for every individual interval.

5.2.3.5 Unsupervised analysis

When the $\mathbf{C}_{Features}$ matrix has been defined, the integrated features are ready for exploratory analysis via PCA. The PCA algorithm included in InRA employs SVD to calculate scores, loadings, and explained variance. Several pre-treatments have been incorporated, including Log_{10} transformation, *Pareto* and variance scaling, and mean centering [29]. Furthermore, up to five distinct samples types can be color-coded to enhance interpretability. The number of principal components (PCs) is preset to 5, though the selection of an appropriate number of PCs can be adjusted as needed. A block-based cross-validation method is used to estimate the model errors by the root mean square error of calibration (RMSEC) and the root mean square error of cross-validation (RMSECV) [30]. For equations see Supplementary Table 5.7.1. The explained variance, cumulative variance, RMSEC,

and RMSECV are provided in a *Statistical Results* table. To assist in identifying possible outliers, the option to display the 95% confidence ellipse on the scores plot is offered. For a comprehensive interpretation of the crucial spectral region responsible for sample clustering, the contribution of the loadings is represented as bar plots with distinct color patterns to their spectral region (i.e., purple, green, and blue), in which each bar represents a resolved **C** (hereafter, **C**-Profiles). One of the main advantages of employing InRA for PCA is the capacity to visualize every resolved $\mathbf{S}^T$ associated with the corresponding **C**-profile directly, quickly, and easily through an MCR component list (Fig. 5.2.1E). The latter allows for a detailed assessment of individual resonance signals, thereby streamlining the interpretative complexity of the obtained results.

If the *Binning to compare* option was previously selected in the Processing tab, and additional functionality labeled *Comparative PCA* becomes available. To access this functionality, independent PCA models must first be conducted for the *Bucket Spectra* ($\mathbf{B}_{Spectra}$) and *Original Processed* ($\mathbf{P}_{Spectra}$) matrices. The latter is achieved via the *Comparative Pretreatment* module, to select the desired input matrix, apply an appropriate pre-treatment, and construct the PCA model. After the PCA models for each matrix have been carried out, the *Comparative PCA* option will be enabled, displaying the score plots, loading plots, and statistical results of the different input matrices (i.e., $\mathbf{C}_{Features}$, $\mathbf{P}_{Spectra}$, and $\mathbf{B}_{Spectra}$) in a unified new window. The *Comparative PCA* option allows for a direct and efficient assessment of different processing strategies, thereby facilitating the selection of the most appropriate analytical approach according to the specific characteristics of the ^{1}H NMR spectra.

5.3 Methodology for the case study

5.3.1 Reagents and chemicals

Deuterium oxide (D_2O, 99.9% D) and 3-(Trimethylsilyl)-propionic-2,2,3,3-d_4 acid sodium salt (TSP-d_4, 98.0% D) were purchased from Sigma-Aldrich (St. Louis, MO, USA). Monobasic potassium phosphate (KH_2PO_4, $\geq$ 99.0%) and dibasic potassium phosphate (K_2HPO_4, $\geq$ 98.0%) were obtained from Merck-Sigma Group (Merck KGaA, Darmstadt, Germany) and used in the preparation of a phosphate buffer solution (K_2HPO_4/KH_2PO_4, 200 mM at pH 7.0) in D_2O. Methanol (LiChrosolv®)

purchased from Supelco (Merck KGaA, Darmstadt, Germany) and ultrapure water from Simplicity® UV apparatus (Merck KGaA, Darmstadt, Germany) equipped with a 0.22 µm filter membrane were used throughout the hydrophilic extraction.

5.3.2 Experimental strategy

The applicability of the methodology implemented in InRA was validated through the evaluation of the spectroscopic ^{1}H NMR profile of sugar beet roots for cultivation field authentication. Three different fields affiliated with the "Federación Nacional de Remolacheros" (National Federation of Beet Growers, Chile) were selected for sample collection: "Santa Isabel", "Luciana", and "Santa Laura" (Supplementary Fig. 5.7.4 illustrated the sampled locations over each field). These fields were selected to demonstrate the practical relevance, efficiency, and selectivity of InRA in the discrimination of samples cultivated under comparable agronomic, environmental, and pedoclimatic conditions.

32 healthy samples were collected from each field between the last week of March and the first week of April 2023. The selection criteria was based on visual inspection, ensuring no visible symptoms of damage or infection in the roots and leaves. Sampling was carried out randomly to reflect representativeness. After sampling, the roots of each plant were separated from the leaves at the crown. The roots were then washed, sliced, crushed, and freeze-dried for 48 h before being stored in airtight bags at −80°C until further analysis. Before freeze-drying, the moisture content of each sugar beet root was measured after drying at 105°C until constant weight (± 0.5% of previous weight) using a Sartorius MA35 moisture analyzer (Germany). A list of the sugar beet samples analyzed is described in Supplementary Table 5.7.2.

A total of 108 ^{1}H NMR spectra were acquired. The analysis was based on stable polar compounds, i.e., amino acids, carbohydrates, and organic acids, due to their greater signal dispersion in ^{1}H NMR spectra compared to low-polarity compounds, enhancing the suitability of the hydrophilic fraction for resonance signal extraction (features) and multivariate analysis purposes [31].

5.3.3 Sample preparation

The sugar beet samples were subjected to a hydrophilic extraction using a methanol/water solution, as reported in a previous study with slight modifications [21]. Briefly, 150.0 ± 0.2 mg of freeze-dried root were individually weighed into two 2.0 mL Eppendorf tubes and suspended in 1000 μL of cold (4°C) MeOH:H_2O (1:1 *v/v*) solution. The mixtures were vortex-homogenized for 30 s, sonicated for 15 min, and centrifugated at 13,000 *g* for 10 min. Both supernatants were collected, transferred, and combined in a new 2.0 mL Eppendorf tube. Then, the hydrophilic extracts were freeze-dried for 48 h and stored at −4°C until analyzed. The extraction process of each sample was performed randomly to avoid bias in the analyses.

The hydrophilic extracts were dissolved in 900 μL of a phosphate buffer solution and 100 μL of TSP-d_4 58 mM (prepared in D_2O) as an internal standard. The solution was vortex-homogenized for 30 s, sonicated for 10 min, and centrifugated at 13,000 *g* for 15 min. Finally, 600 μL of supernatant was transferred to a 5-mm NMR tube for the ^{1}H NMR spectra acquisition. To check the robustness and reproducibility of the extraction process, one sample of each field was used to monitor analysis performance for every 8 analyzed samples, giving a total of 12 replicates (1 sample × 4 replicates × 3 fields).

5.3.4 ^{1}H NMR spectra acquisition

The ^{1}H NMR spectra were acquired on a Bruker Ascend^{TM} 400 spectrometer (Bruker BioSpin, Germany) operating at a proton frequency of 400.13 MHz and equipped with a 5-mm PABBI 1H/D-BB ZGRD inverse, multinuclear, and automatic tunning liquid probe. The *zg30* pulse sequence was used for spectra acquisition, and each Free Induction Decay (FID) was collected with a total of 80 scans, after 4 dummy scans, and into 64k data points across a spectral width of 15 ppm (6393.9 Hz). All experiments were performed without sample rotation at a temperature of 296.0 ± 0.1 K, featuring an acquisition time of 1.28 s, a receiver gain of 45.2, and a relaxation delay of 1.00 s. Fourier transform was applied to the FIDs with an exponential line-broadening of 0.3 Hz. Baseline and phase corrections were automatically applied to all ^{1}H NMR spectra, with the reference set to the TSP-d_4 singlet ($\delta_H = 0.00$) using MestreNova v.12.0 software (MestreLab

Research, Spain). When strictly necessary, additional manual adjustments were performed. Lastly, the set of spectra was exported in .csv format for further processing and analysis through InRA.

5.3.5 ^{1}H NMR spectral analysis through InRA

The 108 ^{1}H NMR spectra (.csv) were imported, processed, and analyzed into InRA GUI v1.0 as a spectral matrix. The analytical workflow was initiated in the processing tab with the definition of a specific chemical shift range. The lower ppm was set at $\delta_H = 0.50$ and the higher ppm at $\delta_H = 9.00$. Therefore, spectral regions outside the previously selected range were excluded due to the absence of significant signal intensities. The chemical shift misalignments were corrected by intervals to optimize the alignment. The reference was set to *average2*, and the mode was established to *interval*. A factor of 15 was defined for the multiplication of the averaged spectrum, and 30 small intervals were manually selected, each containing a specific group of signals. The resonance signals associated with the residual singlets of methanol and water were removed. For this purpose, the lower ppm was set at $\delta_H = 3.30$ and the higher ppm at $\delta_H = 3.40$ for methanol, while for water, the lower ppm was defined at $\delta_H = 4.70$ and the higher ppm at $\delta_H = 5.00$ (Supplementary Fig. 5.7.5). The normalization of the ^{1}H NMR spectra was configured at *Norm-2* and spectral bucketing was not applied ($\mathbf{P}_{Spectra}$). To evaluate the performance of the InRA methodology in enhancing sample clustering by cultivation field, the *Binning to compare* option was applied, and a bucket size of 0.02 ppm was set for the generation of the $\mathbf{B}_{Spectra}$ matrix.

Subsequently, the analysis proceeded to the interval-based detection tab, where three predefined spectral regions were delineated: $\delta_H = 0.50 - 3.00$, $\delta_H = 3.00 - 6.00$, and $\delta_H = 6.00 - 9.00$. The detection resonance signals was optimized based on the spectral characteristics of the hydrophilic extract of sugar beet root. Threshold values were set at 0.10 (10%) for the spectral regions of $\delta_H = 0.50 - 3.00$, 0.001 (0.1%) for $\delta_H = 3.00 - 6.00$, and 0.15 (15%) between $\delta_H = 6.00 - 9.00$ (Supplementary Fig. 5.7.6). The detection parameters were used with their default settings, i.e., $D = 0.01$ ppm, $L = 0.03$ ppm, and $H = 0.005$ ppm. According to the obtained results, the removal, modification, or manual addition of certain intervals was performed when absolutely required. The SNR for all detected intervals was determined using MestreNova v.12.0.

At this stage, the workflow was advanced to the decomposition of intervals via MCR in the resonance integration tab. The N for every individual interval was defined within the *purest* $\boldsymbol{S}^T$ module. Based on the magnitude of the eigenvalues, followed by an assessment of the shape and consistency of the initial spectral profiles estimates (initial spectrum), the optimal N was achieved and saved. A total of 150 iterations were set for the ALS optimization and each MCR model was built using the *One-at-Time* option (Supplementary Fig. 5.7.7). Only components with clear chemical significance were preserved, and $\mathbf{S}^T$ profiles that were strictly poorly resolved were removed.

The analytical workflow was concluded with an exploratory analysis carried out via PCA to obtain an overview of the intrinsic variation of the $\mathbf{C}_{Features}$, $\mathbf{P}_{Spectra}$, and $\mathbf{B}_{Spectra}$ matrices (Supplementary Fig. 5.7.8). Samples were color-coded according to their cultivation field and their sequence in the spectral matrix: samples 1-36 (Santa Isabel), 37-72 (Luciana), and 73-108 (Santa Laura). For the Comparative PCA analysis, the PCs were set to 10 and Log_{10} transformation with mean center was selected as pre-treatment to the $\mathbf{C}_{Features}$ and $\mathbf{B}_{Spectra}$, while *Pareto* scaling with mean center was used for the $\mathbf{P}_{Spectra}$.

5.3.6 Supervised Analysis

The effectiveness and robustness of sugar beet field discrimination were assessed via partial least squares discriminant analysis (PLS–DA). Each field was considered an independent class, resulting in a three-class model (model A). To enhance classification accuracy, a pairwise binary strategy was implemented, given three distinct two-class models: model B (Santa Isabel – Luciana), model C (Luciana – Santa Laura), and model D (Santa Isabel – Santa Laura). The results of this strategy were compared with the results of model A. The aforementioned pre-treatments for PCA were applied to the PLS–DA models. All models were carried out by PLS_Toolbox 9.2 (Eigenvector Research Inc.) within the MATLAB R2023b environment.

To prevent overfitting in the PLS–DA models due to the limited number of samples per class, a two-level cross-validation (CV) scheme was adopted. Specifically, an external k-fold CV strategy was applied to the $\mathbf{C}_{Features}$, $\mathbf{P}_{Spectra}$, and $\mathbf{B}_{Spectra}$ matrices (k = 4). In each iteration, three parts were used as the training set, while the fourth subset was reserved as a test set. In this way, it was ensured

that all samples participated in the test set at least once. Within each training set, the leave-one-out cross-validation (LOOCV) method was further applied as an internal CV during model calibration. The number of latent variables (LVs) was selected based on the RMSECV rather than on maximizing classification accuracy. Mean values from the received test sets were compiled into a final confusion matrix to evaluate model performance. The discrimination capacity was evaluated according to the following figures of merit: sensitivity (Sn), specificity (Sp), error rate (ER), classification accuracy (CA), and precision. In addition, **C**-profiles with variable importance in the projection (VIP) ≥ 1.2 were considered relevant for discrimination. A one-way ANOVA with Bonferroni-Holm correction was performed using OriginPro 2016 (OriginLab Corporation, Northampton, MA, USA) to assess the statistical significance of relevant **C**-profiles. VIP values with $p < 0.05$ were recognized as discriminant.

5.4 Results and discussion

5.4.1 Applicability of interval analysis for resonance signal detection in sugar beet roots

Supplementary Fig. 5.7.9 shows a representative ^{1}H NMR spectrum (400.13 MHz) of the hydrophilic extract from sugar beet, segmented into three regions: $\delta_H = 0.50 - 3.00$, $\delta_H = 3.00 - 6.00$, and $\delta_H = 6.00 - 9.00$. Due to sugar beet is an essential raw material owing to is ability to accumulate high concentrations of sucrose in the root, highly intense resonance signals predominantly associated with carbohydrates, primarily sucrose, were observed in the spectral region $\delta_H = 3.00 - 6.00$ (Supplementary Fig. 5.7.9B), whereas less intense signals particularly relevant for the detection of protons associated with free amino acids, organic acids, aromatic, and aldehydes were detected between $\delta_H = 0.50 - 3.00$ and $\delta_H = 6.00 - 9.00$ (Supplementary Fig. 5.7.9A and 5.7.9C).

The interval-based resonance signal detection performed by InRA identified successfully 35 intervals automatically, while 6 additional intervals were manually incorporated due to their slightly complex resonance patterns, resulting in 41 intervals: 15 in $\delta_H = 0.50 - 3.00$, 16 between $\delta_H = 3.00 - 6.00$, and 10 across $\delta_H = 6.00 - 9.00$. These results demonstrated the high efficiency of the automated

detection algorithm, covering 85.4% of the total intervals with minimal manual intervention. Out of the 35 intervals automatically detected, only 12 required minor adjustments to ensure better accuracy at their limits. Detailed information about the chemical shift ranges, SNR, and the detection mode (automatic or manual) for all intervals is provided in Table 5.4.1.

In particular, the intricate $\delta_H = 6.50 - 9.00$ range, defined by low-SNR resonance signals, was frequently masked by the dominance of stronger resonances, i.e., signals related to carbohydrates. Nevertheless, the implementation of a specific detection threshold tailored exclusively to the characteristics of the $\delta_H = 6.50-9.00$ range allowed for the complete automatic detection of all relevant resonance signals (Table 5.4.1), significantly improving the sensitivity and selectivity of the analysis.

Table 5.4.1: Summary of interval-based resonance signal detection obtain through InRA.

Spectral region ($\delta_{\mathbf{H}}$)	Interval	Resonance interval detection	SNR[a]	Chemical shift range ($\delta_{\mathbf{H}}$)
0.50 – 3.00				
	1	Automated[b]	96	0.699 – 0.852
	2	Automated[b]	106	0.859 – 1.071
	3	Automated	83	1.305 – 1.357
	4	Automated	150	1.431 – 1.508
	5	Automated[b]	145	1.846 – 1.960
	6	Automated[b]	162	2.033 – 2.184
	7	Automated[b]	155	2.268 – 2.329
	8	Manual	88	2.331 – 2.418
	9	Automated[b]	171	2.420 – 2.495
	10	Automated[b]	125	2.508 – 2.781
	11	Automated	108	2.787 – 2.816
	12	Automated	71	2.822 – 2.863
	13	Automated	74	2.868 – 2.912
	14	Automated	75	2.920 – 2.961
	15	Manual	170	2.962 – 3.060
3.00 – 6.00				
	16	Manual	133	3.061 – 3.105
	17	Automated[b]	>2000	3.162 – 3.196
	18	Manual	854	3.197 – 3.220
	19	Automated[b]	83	3.250 – 3.287
	20	Automated	>2000	3.412 – 3.521
	21	Automated	>2000	3.526 – 3.604
	22	Automated	>2000	3.609 – 3.722
	23	Automated[b]	>2000	3.723 – 3.796
	24	Manual	>2000	3.802 – 3.934
	25	Automated	>2000	4.022 – 4.124
	26	Automated	>2000	4.166 – 4.312
	27	Automated[b]	138	4.366 – 4.430
	28	Automated	95	5.182 – 5.219
	29	Manual	100	5.222 – 5.250
	30	Automated	>2000	5.368 – 5.464
	31	Automated	113	5.608 – 5.642
6.00 – 9.00				
	32	Automated	48	6.511 – 6.541
	33	Automated	45	6.876 – 6.928
	34	Automated	45	7.165 – 7.238
	35	Automated	30	7.306 – 7.354
	36	Automated	25	7.400 – 7.448
	37	Automated	40	7.520 – 7.570
	38	Automated	28	7.716 – 7.767
	39	Automated[b]	26	8.194 – 8.270
	40	Automated	43	8.446 – 8.476
	41	Automated	27	8.821 – 8.868

[a]Signal Noise Ratio. [b]Intervals with minor manual adjustments.

Furthermore, as shown in Fig.5.4.1, the algorithm effectively demonstrated its ability to distinguish between distinct spin systems based on the selected detection parameters (D, H, and L). Fig. 5.4.1A and Fig. 5.4.1C illustrate the average intensity of resonance signals with different J-coupling (between $\delta_H = 3.50 - 4.30$) prior to interval correction, using the same threshold (0.001) but different values for the D parameter. Specifically, Fig. 5.4.1A corresponds to a D value of 0.01 ppm (default), whereas Fig. 5.4.1C shows the outcome with D set to 0.02 ppm. In both cases, resonance signals corresponding to a triplet and a doublet were clearly delineated as individual intervals; however, regions containing a higher density of resonance signals exhibited overlap during their initial delimitation, leading to their consolidation into a single (Fig. 5.4.1A) or double (Fig. 5.4.1C) large interval encompassing two or more resonance signals. To enhance the correction and delimitation of overlap intervals, the effect of modifying the L and H parameters can be seen in Fig. 5.4.1B and 5.4.1D. Fig. 5.4.1B represents the default detection settings ($L = 0.03$ ppm and $H = 0.005$ ppm), while Fig. 5.4.1D shows a modified configuration with $L = 0.08$ ppm and $H = 0.01$ ppm. Under default conditions (Fig. 5.4.1B), the initially merged intervals were subdivided into four new corrected intervals, each containing different coupling patterns, triplet, singlet, or more complex multiplicities (i.e., doublets of doublets and multiplets). In contrast, Fig. 5.4.1D displays the generation of only two new broader intervals, suggesting a lower specific segmentation outcome. These findings underscore the algorithm's capacity to adapt the interval detection process through precise adjustment of the detection parameters. By modulating the values of D, L, and H, the algorithm enables the selective refinement of spectral segmentation, allowing a distinct spin systems delineation or the strategic consolidation of a group of resonance signals based on the complexity of the evaluated spectral region.

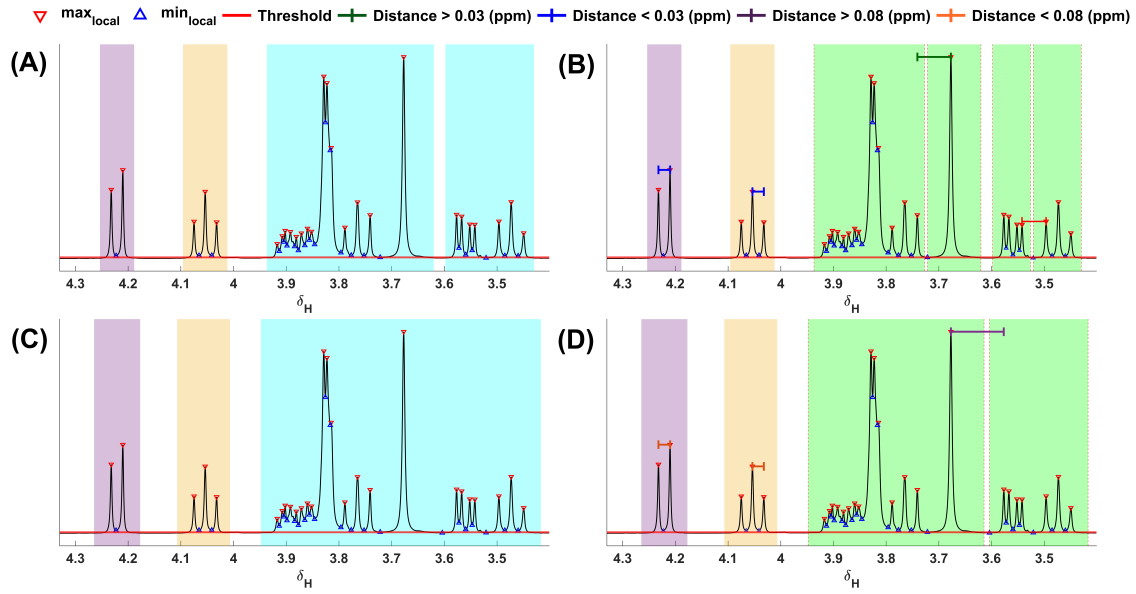

Figure 5.4.1: Practical performance of the interval-based algorithm with different detection settings. **A)** and **C)** illustrated the initial interval delimitation performed on the average intensity of resonance signals with $D = 0.01 - 0.02$ ppm, respectively. Intervals marked in light blue represent overlapping detection merged into a single interval. Over the signals, $\max_{local}$ are represented with inverted red triangles, while $\min_{local}$ are indicated by blue triangles. **B)** and **D)** display the final interval delimitation after correction performed with $L = 0.03 - 0.08$ ppm and $H = 0.005 - 0.01$ ppm, respectively.

5.4.2 MCR integration of resonance features

The 41 intervals were integrated (features) through an MCR multi-model approach conducted by InRA with an N ranging from 1 to 4. The simultaneous decomposition for all intervals was completed in $\backsim$ 66 s through the ALS optimization, explaining 96.3% of the total variance and achieving an average LOF of 16.1%. These findings emphasize the efficiency of InRA in handling multiple MCR models with a significant reduction in time analysis. MCR components associated with poorly resolved spectral features were removed, yielding a total of 44 well-defined pure resonance signals. A tentative identification was performed for the 44 $\mathbf{S}^T$ profiles, 26 were assigned to individual compounds, 14 were grouped as signals from the same compounds, and 4 were classified as "unknown" resonance signals. Besides carbohydrates, the predominant identifications corresponded to amino acids (e.g., alanine, glutamine, valine) and certain organic acids (e.g., acetate, malate, formate), compounds that have

crucial roles in metabolic and physiological processes essential for growth, stress resistance, and high agronomic performance [32]. The assignment, chemical shifts (δ_H), multiplicity (m), and coupling constants (J) of the main compounds tentatively identified in the hydrophilic sugar beet extract are provided in Table 5.4.2. Our spectral characterization was consistent with experimental results from ^{1}H NMR spectra reported in relevant studies and the Chenomx NMR Suite 8.3 reference library (https://www.chenomx.com, last access: September 12, 2024) [19, 20, 33, 34, 35, 36, 37, 38]. Additionally, the Human Metabolome Database (HMDB, https://www.hmdb.ca, last access: September 22, 2024) was consulted to identify essential compounds (e.g., amino acids) present in the plant metabolome [39].

Table 5.4.2: List of resolved resonance signals tentatively identified in hydrophilic sugar beet extracts (D_2O with K_2HPO_4/KH_2PO_4 180 mM at pH 7.0) by ^1H NMR (400.13 MHz) analysis.

Abbreviation	Compound	δ_H in ppm (mult[a], J[b] in Hz)	Assignment
FA	Fatty acids	0.78-0.93 (m)	-CH_3
Leu	Leucine	0.96 (d, J = 5.66)	δCH_3
		0.98 (d, J = 6.08)	δ'CH_3
		1.70 (m)	βCH_3
Ile	Isoleucine	1.01 (d, J = 7.18	γCH_3
Val	Valine	0.99 (d, J = 7.09)	γCH_3
		1.05 (d, J = 7.07)	γ'CH_3
Thr	Threonine	1.33 (d, J = 6.62)	γCH_3
Ala	Alanine	1.48 (d, J = 7.23)	βCH_3
GABA	γ-aminobutyrate	1.90 (m)	βCH_2
		2.30 (t, J = 7.37)	αCH_2
		3.01 (d, J = 7.31)	γCH_2
Ace	Acetate	1.92 (s)	CH_3
Glu	Glutamate	2.10 (m)	βCH_2
		2.36 (m)	γCH_2
Gln	Glutamine	2.14 (m)	βCH_2
		2.45 (m)	γCH_2
Suc	Succinate	2.41 (s)	CH_2
Cit	Citrate	2.56 (d, J = 15.5)	CH_2
		2.70 (d, J = 15.4)	CH_2
Mal	Malate	2.66 (dd, J = 15.3-3.52)	β'CH_3
Asp	Aspartate	2.82 (dd, J = 17.3-3.94)	β'CH
Asn	Asparagine	2.91 (dd, J = 21.4-5.90)	δ'CH
Cho	Choline	3.19 (s)	$N(CH_3)_3$
Bet	Betaine	3.27 (s)	$N(CH_3)_3$
Sur	Sucrose	3.47 (t, J = 9.04)	C_3H
		3.56 (dd, J = 9.99-3.87)	C_2H
		3.68 (s)	C'$_1$H
		3.76 (t, J = 9.57)	C_4H
		3.82 (m)	C'$_6$H
		3.88 (m)	C'$_5$H
		4.05 (t, J = 8.56)	C'$_4$H
		4.22 (d, J = 8.78)	C'$_3$H
		5.24 (d, J = 3.87)	C_1H
Xyl	α-xylose	5.20 (d, J = 3.76)	C_1H
Glu	α-glucose	5.23 (t, J = 3.76)	C_1H
Fum	Fumarate	6.52 (s)	$(CH=)_2$
Tyr	Tyrosine	6.90 (d, J = 8.54)	3,5-CH
		7.20 (d, J = 8.49)	2,6-CH
Phe	Phenylalanine	7.31 (m)	2,6-CH
		7.40 (m)	3,5-CH
Trp	Tryptophan	7.54 (d, J = 8.21)	C_4H
		7.74 (d, J = 7.96)	C_7H
For	Formate	8.46 (s)	CH
Tri	Trigonelline	8.84 (m)	4,6-CH
Unk	Unknown	3.09 (s) - 3.21 (s)	-
		4.40 (d, J = 8.83)	-
		5.62 (d, J = 3.79)	-
		8.21 (s) - 8.25 (s)	-

[a]Multiplcity: d(doublet); dd(doublets of doublets); m(multiplet); s(singlet); t(triplet).
[b]Coupling constant.

A key aspect of the MCR decomposition is illustrated in Fig. 5.4.2, where resolved $\mathbf{S}^T$ obtained from intervals with signal overlap are highlighted. As an example, Fig. 5.4.2A and 5.4.2C display resonance signals with complexity in the direct visual distinction of individual species. The MCR decomposition over both matrices successfully resolved four independent spectral profiles of meaningful chemical compounds. Among them, one of the multiplets of γ-aminobutyrate ($\delta_H = 1.90$), the overlapping singlet of acetate ($\delta_H = 1.92$), and the multiplets of glutamate ($\delta_H = 2.10$) with glutamine ($\delta_H = 2.14$) were identified, as shown in Fig. 5.4.2B and 5.4.2D, respectively. These results demonstrate the effectiveness of the interval-based multivariate strategy, as spectral complexity was reduced by limiting variability within each segment. Through the segmentation, the relative variability of meaningful spectral profiles was accurately resolved via MCR using a reduced number of components. Consequently, the analysis of overlapped regions was facilitated without prior of knowledge of resonance signals present in the interval.

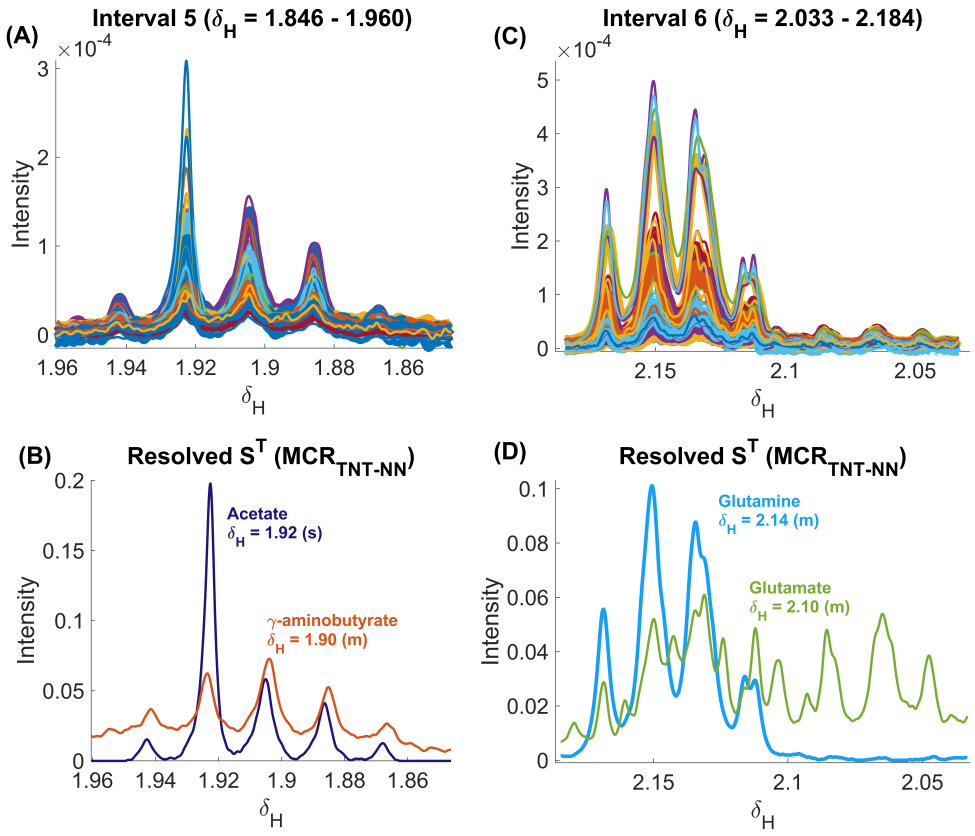

Figure 5.4.2: MCR models of intervals 5 and 6 performed through InRA. **A)** and **C)** selected intervals with distinct spectral overlapping. **B)** and **D)** resolved $\mathbf{S}_T$ obtained after ALS_{TNT-NN} optimization.

In order to ensure the reproducibility of the MCR results obtained through InRA, a systematic comparison was conducted against the well-established FNNLS method available in the MCR-ALS GUI 2.0 Toolbox [40]. This benchmarking aimed to validate the reliability of the ALS optimization performed by TNT-NN to confirm its capability to provide consistent and accurate outcomes. As example, Fig. 5.4.3 shows the MCR decomposition by both approaches (MCR_{TNT-NN} – MCR_{FNNLS}) of the triplet associated with the methylene groups (αCH_2) in γ-aminobutyrate (interval 7) and the doublet of the equivalent protons attached to the carbons at positions 3 and 5 in the aromatic ring (3,5-CH) of tyrosine (interval 33). A notably high level of spectral component resolution was achieved, as evidenced by the complete overlap of the resolved $\mathbf{S}^T$ with the representative resonance signal in identical consistency and shape, as shown in Fig. 5.4.3A. The correlation plots presented in Fig. 5.4.3B provide a quantitative evaluation of TNT-NN and FNNLS performed in terms of relative concentrations, showing a strong linear correlation ($r^2 = 0.999$) between the $\mathbf{C}$ values in both intervals for the 108 samples. Furthermore, the accuracy of the $\mathbf{C}$ values was evaluated by testing the homogeneity of variances through Levene's test, conducted in OriginPro 2016. The obtained p-values were 0.0993 for interval 7 and 0.0998 for interval 33 at a significance level of $\alpha = 0.05$. Therefore, no statistically significant difference in the population variance between the compared groups was found. Our results demonstrate the equivalent performance of MCR_{TNT-NN} with MCR_{FNNLS} regarding spectral resolution ($\mathbf{S}^T$) and signal quantification ($\mathbf{C}$). The strong agreement confirms the reliability of the TNT-NN-based approach in accurately resolving distinct signals in applications requiring an optimized decomposition of simple or complex ^{1}H NMR spectra.

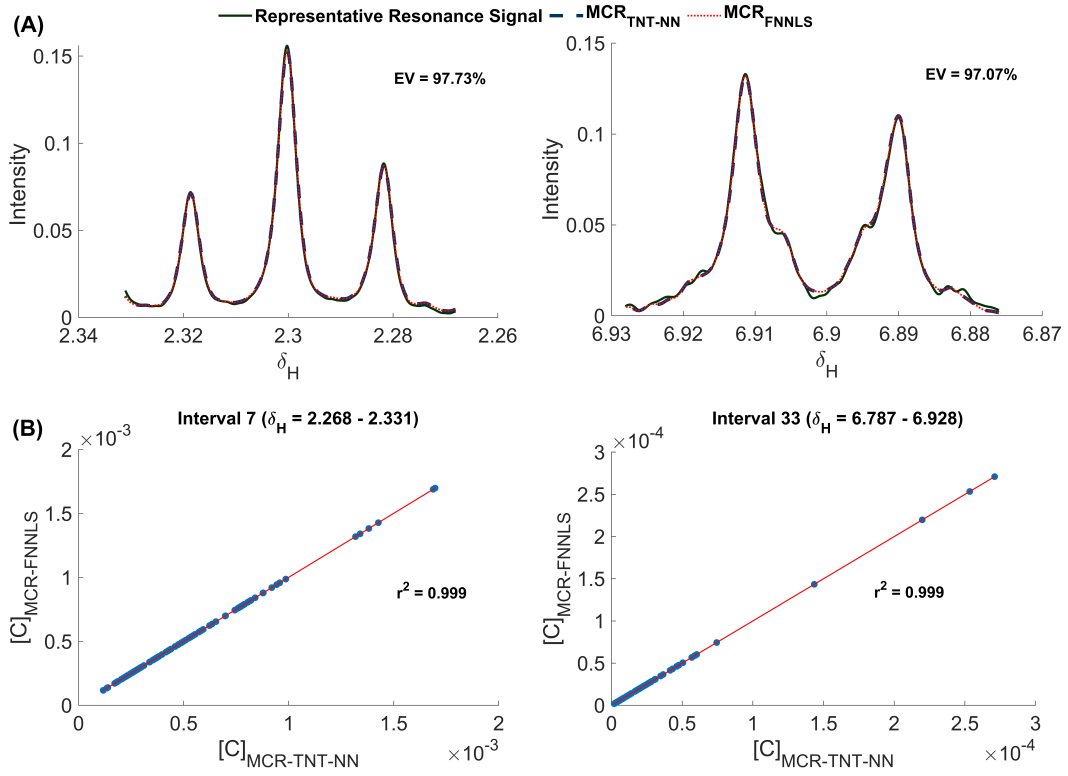

Figure 5.4.3: MCR decomposition of interval 7 and 33 through the MCR$_{TNT-NN}$ and MCR$_{FNNLS}$ approaches. **A)** Overlay resolved $\mathbf{S}^T$ with the reference resonance signal: a triplet from alanine and a doublet from tyrosine. The signals (reference and $\mathbf{S}^T$) were normalized to enhance visualization. **b)** Correlation plots of $\mathbf{C}$ values between both ALS optimization for each interval.

5.4.3 Sample clustering and sugar beet field discrimination

The score plot derived from the autoscale $\mathbf{C}_{Features}$ indicated that the first two PCs explained 50.86% of the total variance, as can be seen in Fig. 5.4.4A. Although no distinct clustering pattern was observed, PC1 (34.42%) suggested a slight separation of samples from Santa Isabel, while PC2 (16.44%) revealed a tendency for Luciana samples to cluster apart from Santa Isabel and Santa Laura. Fig. 5.4.4B shows the loading plots with primary sources of variation in PC1 distributed across the aliphatic (purple) and aromatic regions (blue), respectively. To address the dominance of high-intensity sucrose signals, a Log_{10} transformation was applied to mitigate the imbalance variance. The new PCA score plot is illustrated in Fig. 5.4.4C, explaining 59.02% of the total variance with the first two PCs, revealed distinct field-based clustering. Sugar beet samples from Santa Isabel tended to be separated from those of Santa Laura and Luciana along PC1 (36.88%), while PC2

(22.14%) accentuated the differentiation of Sana Laura relative to Luciana and Santa Isabel. These findings suggested that PC1 captured subtle variability, while PC2 accounted for more pronounced differences between fields. Regarding the loading plots (Fig. 5.4.4D), for PC1, the aliphatic region exhibited a prominent role in the clustering. On the other hand, in PC2, key contributions from resonance signals of citrate with malate, glutamate, multiplets, and the tyrosine doublets (**C**-profiles, 8, 10, 12, 36 and 37) arose for sugar beet samples from Santa Laura. Furthermore, the singlet of acetate, the resonance signals of γ-aminobutyrate, and the singlet of fumarate (**C**-profiles 5, 9, 18, and 35) were critical for the clustering of Santa Isabel and Luciana.

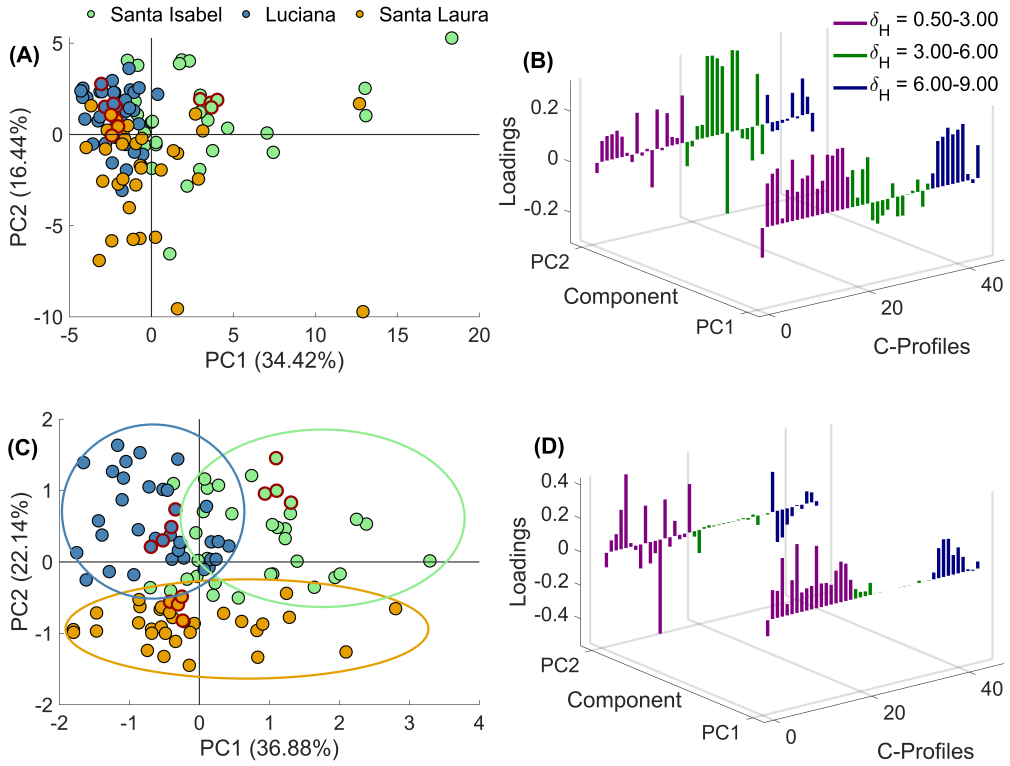

Figure 5.4.4: PCA results (extracted from InRA) obtained through $\mathbf{C}_{Features}$. **A)** and **C)** display score plots (PC1 vs PC2) obtained with autoscaling and Log_{10} transformation with mean center, respectively. Replicates are encircled in red for each field. **B)** and **D)** shows the respective loading plots (PC1 and PC2) for each pretreatment. The purple, green, and blue colors represent the contributions of the 44 **C**-profiles, categorized according to their respective spectral regions.

To evaluate the influence of different spectral processing strategies on sample clustering, the score plots derived from the comparative PCA of the $\mathbf{P}_{Spectra}$, $\mathbf{C}_{Features}$, and $\mathbf{B}_{Spectra}$ matrices are presented in Fig. 5.4.5. In the case of the $\mathbf{P}_{Spectra}$, 64.88% of the total variance with the first two PCs was achieved. However, considerable overlap among the sample groups, particularly between Luciana and Santa Laura was observed. The latter outcome was attributed to the high dimensionality and intrinsic spectral complexity of the $\mathbf{P}_{Spectra}$ matrix. When PCA was applied to the $\mathbf{B}_{Spectra}$, a more defined (although not fully separate) clustering was observed among Santa Isabel and Luciana along PC1 (28.33%), while PC3 (10.45%) was relevant for Santa Laura. These results demonstrated that the $\mathbf{C}_{Features}$ yielded enhanced clustering performance only with the two first PCs, due to the MCR capability to resolve components associated with relevant resonance signals and to reduce spectral complexity without loss of spectral resolution. Consequently, an effective separation of sample groups was achieved compared to standard processing strategies.

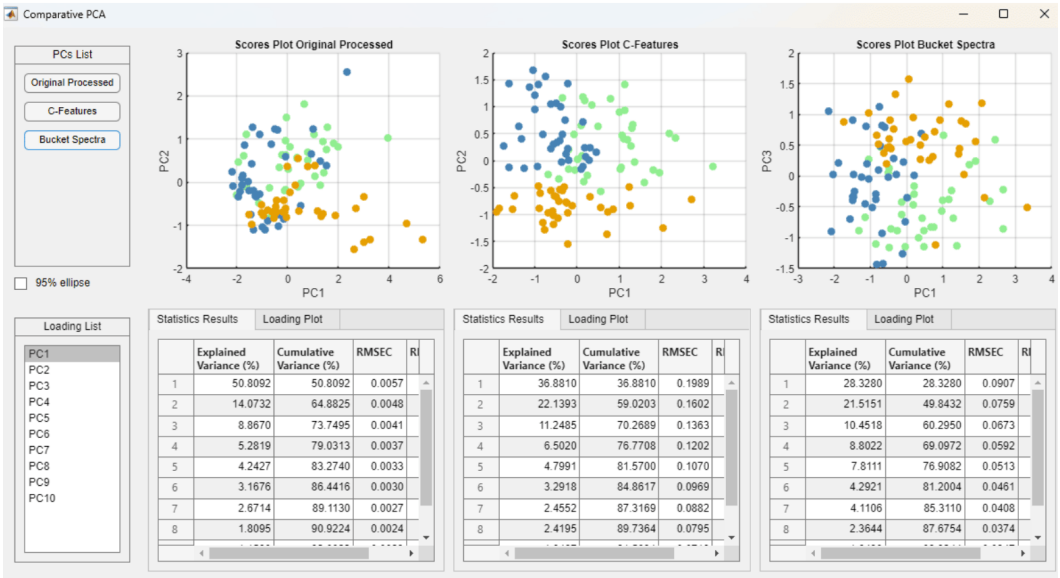

Figure 5.4.5: Comparative PCA results of the $\mathbf{C}_{Features}$, $\mathbf{P}_{Spectra}$ (Original Processed), and $\mathbf{B}_{Spectra}$ (Bucket Spectra) carried out through InRA. Each scores plot are color-coded based on the cultivation field, i.e., Santa Isabel (green), Luciana (blue), and Santa Laura (orange).

Based on the preliminary clustering of sugar beet samples observed via PCA, four PLS–DA models were performed to determine the discrimination potential based on sugar beet field authentication. The effectiveness of the PLS–DA results obtained via $\mathbf{C}_{Features}$ was evaluated by comparison with those achieved using the $\mathbf{B}_{Spectra}$ and $\mathbf{P}_{Spectra}$ matrices. The confusion matrices of three and two-class models with the precision percentage for all modelled classes are depicted in supplementary Fig. 5.7.10, 5.7.11, 5.7.12, respectively. Moreover, Table 5.4.3 shows the figures of merit and the optimized number of LVs for the four PLS–DA models using the different input matrices. The three-class model obtained via $\mathbf{C}_{Features}$ (Supplementary Fig. 5.7.10A) showed consistent discrimination, with Santa Laura achieving the highest sensitivity (100%) and an overall classification accuracy of 96.3%. Despite the observed misclassifications within the Santa Isabel group attributed to its higher variability and similarity to Luciana samples, distinctive spectral characteristics were maintained, resulting in a specificity of 95.8% in contrast with the 88.9% - 81.9% of the $\mathbf{B}_{Spectra}$ and $\mathbf{P}_{Spectra}$ matrices (Table 5.4.3). Regarding the two-class models, model C achieved a classification accuracy of 100% (supplementary Fig. 5.7.10B), indicating a clear discrimination between Luciana and Santa Laura. The PLS–DA score plot shown in Supplementary Fig. 5.7.13 and PCA in Fig. 5.4.4C, further support the results, reflecting the intra-group homogeneity and distinct spectral profiles differences. Model D obtained via $\mathbf{C}_{Features}$ exhibited a classification accuracy of 94.4%, revealing robust discrimination compared to $\mathbf{B}_{Spectra}$ (84.7%) and $\mathbf{P}_{Spectra}$ (77.8%), while model B displayed the lowest performance, with a sensitivity of 88.9% and a classification accuracy of 90.3% for Santa Isabel (Table 5.4.3). In general, the application of the $\mathbf{C}_{Features}$ matrix improved the performance of the PLS–DA models, surpassing the classification rates achieved with the $\mathbf{B}_{Spectra}$ and $\mathbf{P}_{Spectra}$, which ranged from 81.9% - 97.2% and 77.8% - 93.1%, respectively. These results highlight that $\mathbf{C}_{Features}$ enhance the discrimination capability, confirming its utility as an effective strategy for authentication and traceability purposes.

Table 5.4.3: PLS–DA figures of merit for the discrimination of sugar beet by two and three-class classification models corresponding to the fields of Santa Isabel, Luciana, and Santa Laura through $\mathbf{C}_{Features}$, $\mathbf{B}_{Spectra}$, and $\mathbf{P}_{Spectra}$ matrices.

Input matrix	Model	#LV	Class	CV[a]%	Sn[b]%	Sp[c]%	ER[d]%	CA[e]%
$\mathbf{C}_{Features}$								
	A	3	Santa Isabel	67.2	75.0	95.8	14.6	88.9
			Luciana		91.7	93.1	7.60	92.6
			Santa Laura		100	94.4	2.78	96.3
	B	3	Santa Isabel	58.2	88.9	91.7	9.70	90.3
			Luciana		91.7	88.9	9.70	90.3
	C	2	Luciana	54.4	100	100	0.00	100
			Santa Laura		100	100	0.00	100
	D	3	Santa Isabel	60.8	94.4	94.4	5.56	94.4
			Santa Laura		94.4	94.4	5.56	94.4
$\mathbf{B}_{Spectra}$								
	A	4	Santa Isabel	85.4	80.6	88.9	15.3	86.1
			Luciana		80.6	90.3	14.6	87.0
			Santa Laura		94.4	98.6	3.47	97.2
	B	3	Santa Isabel	64.1	77.8	86.1	18.1	81.9
			Luciana		86.1	77.8	18.1	81.9
	C	4	Luciana	70.6	94.4	97.2	4.20	95.8
			Santa Laura		97.2	94.4	4.20	95.8
	D	5	Santa Isabel	77.3	80.6	88.9	15.3	84.7
			Santa Laura		88.9	80.6	15.3	84.7
$\mathbf{P}_{Spectra}$								
	A	3	Santa Isabel	70.2	75.0	81.9	21.5	79.6
			Luciana		66.7	86.1	23.6	79.6
			Santa Laura		75.0	90.3	17.4	85.2
	B	3	Santa Isabel	65.4	75.0	97.2	13.9	86.1
			Luciana		97.2	75.0	13.9	86.1
	C	2	Luciana	72.4	91.7	94.4	6.90	93.1
			Santa Laura		94.4	91.7	6.90	93.1
	D	3	Santa Isabel	69.8	75.0	80.6	22.2	77.8
			Santa Laura		80.6	75.0	22.2	77.8

[a]Cumulative variance. [b]Sensitivity. [c]Specificity. [d]Error rate. [e]Classification accuracy.

Through the analysis of the VIP scores (Supplementary Fig. 5.7.14), relevant resonance signals that contributed significantly to the discrimination were recognized. These signals included the methyl (βCH_3) doublet of alanine, the methyl (CH_3) singlet of acetate, the methylene (αCH_2) triplet of γ-aminobutyrate, the methylene (γCH_2) multiplet of glutamine, the methylene ($CH_2/\beta CH_2$) doublets of citrate-malate, the methine (CH=) singlet of fumarate, and the doublets of the equivalent protons (3,5-CH/2,6-CH) attached to the aromatic ring in tyrosine. Notably, citrate, malate, and fumarate – intermediates in the tricarboxylic acid (TCA) cycle – displayed field-specific variations, suggesting their relative

concentrations are influenced by possible differences in soil composition and environmental conditions. Additionally, alanine and glutamine, were more prominent in Santa Isabel and Luciana samples, whereas acetate was predominantly detected in Santa Isabel, potentially reflecting unique stress responses.

5.5 Conclusions

This work introduced Interval Resonance Analysis (InRA), a versatile tool designed to provide an alternative analytical workflow for untargeted studies using ^{1}H NMR spectroscopy. InRA consolidates specialized algorithms into an effective, intuitive, friendly, and unified platform, specifically optimized for processing, resonance signal detection, integration via MCR, and exploratory analysis, addressing the challenges of manual analysis that are often labor-intensive and time-consuming. Furthermore, its intuitive GUI promotes the integration of users with different experience levels, thereby reducing the barrier to advanced multivariate ^{1}H NMR analysis

The applicability of InRA to hydrophilic extracts from sugar beet roots, with inherent spectral similarities, demonstrated substantial robustness and selectivity in differentiating cultivation fields, despite the suggested interference from sucrose signals. The results obtained with the developed interval-based detection algorithm were promising, enabling the automated identification of 85.4% of relevant resonance signals, even with low SNR, through spectral region segmentation. The comparison of ALS optimization via TNT-NN yielded consistent and reproducible results with FNNLS, highlighting the reliability of the algorithm for the simultaneous decomposition of multiple spectral intervals. The combination of the integrated features and multivariate models proved effective for distinguishing cultivation fields. Classification results by PLS–DA demonstrated the effectiveness in identifying crucial compounds responsible for accurate discrimination, primarily linked to amino acids, with classification rates ranging from 88.9% to 100%.

Nonetheless, the importance of incorporating the initial stages of spectral processing into the analytical workflow is necessary. Accordingly, the development roadmap includes the implementation of new modules for phase, baseline, and reference corrections. Future updates also aim to enhance analytical capabilities of the InRA through the integration of supervised algorithms, e.g., PLS–DA.

Author contributions

Cristian A. Fuentes: Conceptualization, Methodology, Formal analysis, Investigation, Software, Writing – original draft and Writing - review & editing. **David Montoya**: Conceptualization, Methodology, Formal analysis, Software, Visualization, Writing - review & editing. **Mecit Halil Öztop**: Conceptualization, Project Administration, Resources, Supervision, Writing - review & editing. **Macarena Rojas-Rioseco**: Conceptualization, Methodology, Formal analysis, Investigation, Visualization,, Writing – review & editing. **Martín Bravo**: Conceptualization, Formal analysis, Investigation, Visualization. **Francisca González**: Conceptualization, Methodology, Formal analysis. **Rosario del P. Castillo**: Conceptualization, Methodology, Formal analysis, Investigation, Visualization, Resources, Writing – review & editing

Conflict of interest

The authors declare that they have not conflict of interest.

Acknowledgements

This research was funded by the European Union's Horizon 2020 Research and Innovation program-MSCA RISE under grant agreement #101008228 and by ANID Chile: Beca Doctorado Nacional [21240532] and FONDECYT [1221287]. The authors would like to express their gratitude to FENARE (Federación Nacional de Remolacheros) for their continued support and supply of the necessary samples

5.6 References

[1] Maria Eleni Dimitrakopoulou, Konstantina Matzarapi, Styliani Chasapi, Apostolos Vantarakis, and Georgios A. Spyroulias. Nontargeted 1H NMR fingerprinting and multivariate statistical analysis for traceability of Greek PDO Vostizza currants. *Journal of Food Science*, 86(10):4417–4429, 2021.

[2] Christina Hoerterer, Jessica Petereit, Gisela Lannig, Christian Bock, and Bela H. Buck. 1H-NMR-Based Metabolic Profiling in Muscle and Liver Tissue of Juvenile Turbot (Scophthalmus maximus) Fed with Plant and Animal Protein Sources. *Metabolites*, 13(5), 2023.

[3] Inês Laíns, Daniela Duarte, António S. Barros, Ana Sofia Martins, Tatiana J. Carneiro, João Q. Gil, John B. Miller, Marco Marques, Tânia S. Mesquita, Patrícia Barreto, Ivana K. Kim, Maria Da Luz Cachulo, Demetrios G. Vavvas, Isabel M. Carreira, Joaquim Neto Murta, Rufino Silva, Joan W. Miller, Deeba Husain, and Ana M. Gil. Urine Nuclear Magnetic Resonance (NMR) Metabolomics in Age-Related Macular Degeneration. *Journal of Proteome Research*, 18(3):1278–1288, 2019.

[4] A. Caligiani, D. Acquotti, G. Palla, and V. Bocchi. Identification and quantification of the main organic components of vinegars by high resolution 1H NMR spectroscopy. *Analytica Chimica Acta*, 585(1):110–119, 2007.

[5] Todd M. Alam and M. Kathleen Alam. Chemometric Analysis of NMR Spectroscopy Data: A Review. *Annual Reports on NMR Spectroscopy*, 54:41–80, 2004.

[6] Francesc Puig-Castellví, Ignacio Alfonso, and Romà Tauler. Untargeted assignment and automatic integration of 1H NMR metabolomic datasets using a multivariate curve resolution approach. *Analytica Chimica Acta*, 964:55–66, 2017.

[7] Yolanda Pérez, Marta Casado, Demetrio Raldúa, Eva Prats, Benjamín Piña, Romà Tauler, Ignacio Alfonso, and Francesc Puig-Castellví. MCR-ALS analysis of 1H NMR spectra by segments to study the zebrafish exposure to acrylamide. *Analytical and Bioanalytical Chemistry*, 412(23):5695–5706, 2020.

[8] Nicola Cavallini, Francesco Savorani, Rasmus Bro, and Marina Cocchi. A Metabolomic Approach to Beer Characterization. *Molecules*, 26(5):1–15, 2021.

[9] R Martínez, Francesc Puig-castellví, Rodolfo M Rasia, Paula Burdisso, and Alejandro G García-reiriz. Multivariate curve resolution- based data fusion approaches applied in 1 H NMR metabolomic analysis of healthy cohorts. *Analytica Chimica Acta*, 1309(April), 2024.

[10] Jiayu Yang, Wenzhu Li, Jinguo Ding, Ying Dong, Fang Zhao, Jianyang Pan, and Haibin Qu. A multivariate curve resolution-alternating least squares (MCR-ALS) technology assisted 1 H-NMR methodology for multi-component quantitation of Trichosanthis Pericarpium injection. *Pythochemical Analysis*, 34(August 2022):40–47, 2023.

[11] N Cavallini, L Strani, P P Becchi, V Pizzamiglio, S Michelini, and F Savorani. Tracing the identity of Parmigiano Reggiano " Prodotto di Montagna - Progetto Territorio " cheese using NMR spectroscopy and multivariate data analysis. *Analytica Chimica Acta*, 1278(April):341761, 2023.

[12] David S Wishart, Leo L Cheng, Valérie Copié, Arthur S Edison, Hamid R Eghbalnia, Jeffrey C Hoch, Goncalo J Gouveia, Wimal Pathmasiri, Robert Powers, Tracey B Schock, Lloyd W Sumner, and Mario Uchimiya. NMR and Metabolomics—A Roadmap for the Future, 2022.

[13] Bekzod Khakimov, Nabiollah Mobaraki, Alessia Trimigno, Violetta Aru, and Søren Balling Engelsen. Signature Mapping (SigMa): An efficient approach for processing complex human urine 1H NMR metabolomics data. *Analytica Chimica Acta*, 1108:142–151, 2020.

[14] Manoj Rout, Matthias Lipfert, Brian L. Lee, Mark Berjanskii, Nazanin Assempour, Rosa Vazquez Fresno, Arnau Serra Cayuela, Ying Dong, Mathew Johnson, Honeya Shahin, Vasuk Gautam, Tanvir Sajed, Eponine Oler, Harrison Peters, Rupasri Mandal, and David S. Wishart. MagMet: A fully automated web server for targeted nuclear magnetic resonance metabolomics of plasma and serum. *Magnetic Resonance in Chemistry*, 61(12):681–704, 2023.

[15] Bradley Worley and Robert Powers. MVAPACK: A complete data handling

package for NMR metabolomics. *ACS Chemical Biology*, 9(5):1138–1144, 2014.

[16] Yue Liu, Ji Cheng, Huili Liu, Yinghua Deng, Jie Wang, and Fuqiang Xu. NMRSpec: An integrated software package for processing and analyzing one dimensional nuclear magnetic resonance spectra. *Chemometrics and Intelligent Laboratory Systems*, 162(January):142–148, 2017.

[17] Ramesh Duraisam, Ketemaw Salelgn, and Abiyu Kerebo Berekete. Production of Beet Sugar and Bio-ethanol from Sugar beet and it Bagasse: A Review. *International Journal of Engineering Trends and Technology*, 43(4):222–233, 2017.

[18] Dariush Taleghani, Abazar Rajabi, Saeed Sadeghzadeh Hemayati, and Ali Saremirad. Improvement and selection for drought-tolerant sugar beet (Beta vulgaris L.) pollinator lines. *Results in Engineering*, 13(February):100367, 2022.

[19] Rita Wedeking, Mickaël Maucourt, Catherine Deborde, Annick Moing, Yves Gibon, Heiner E. Goldbach, and Monika A. Wimmer. 1H-NMR metabolomic profiling reveals a distinct metabolic recovery response in shoots and roots of temporarily drought-stressed sugar beets. *PLoS ONE*, 13(5):1–21, 2018.

[20] Yasuyo Sekiyama, Kazuyuki Okazaki, Jun Kikuchi, and Seishi Ikeda. NMR-based metabolic profiling of field-grown leaves from sugar beet plants harbouring different levels of resistance to Cercospora leaf spot disease. *Metabolites*, 7(1), 2017.

[21] Cristian A. Fuentes, Mecit Halil Öztop, Macarena Rojas-Rioseco, Martín Bravo, Aylin Özgür Göksu, Marena Manley, and Rosario del P. Castillo. Application of segmented analysis via multivariate curve resolution with alternating least squares to 1H-nuclear magnetic resonance spectroscopy to identify different sugar sources. *Food Chemistry*, 428(June), 2023.

[22] Jakob Forsberg, Christian Tihic Rasmussen, Frans W.J. van den Berg, Søren Balling Engelsen, and Violetta Aru. Fermentation Analytical Technology (FAT): Monitoring industrial E. coli fermentations using absolute quantitative 1H NMR spectroscopy. *Analytica Chimica Acta*, 1311:342722, 2024.

[23] F. Savorani, G. Tomasi, and S. B. Engelsen. icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal of Magnetic Resonance*, 202(2):190–202, 2010.

[24] Anna De Juan, Joaquim Jaumot, and Romà Tauler. Multivariate Curve Resolution (MCR). Solving the mixture analysis problem. *Analytical Methods*, 6(14):4964–4976, 2014.

[25] Anna de Juan and Romà Tauler. Multivariate Curve Resolution (MCR) from 2000: Progress in concepts and applications. *Critical Reviews in Analytical Chemistry*, 36(3-4):163–176, 2006.

[26] W. Windig, C. E. Heckler, F. A. Agblevor, and R. J. Evans. Self-modeling mixture analysis of categorized pyrolysis mass spectral data with the SIMPLISMA approach. *Chemometrics and Intelligent Laboratory Systems*, 14(1-3):195–207, 1992.

[27] J. M. Myre, E. Frahm, D. J. Lilja, and M. O. Saar. TNT-NN: A Fast Active Set Method for Solving Large Non-Negative Least Squares Problems. *Procedia Computer Science*, 108:755–764, 2017.

[28] Joseph M. Myre, Erich Frahm, David J. Lilja, and Martin O. Saar. TNT: A solver for large dense least-squares problems that takes conjugate gradient from bad in theory, to good in practice. *Proceedings - 2018 IEEE 32nd International Parallel and Distributed Processing Symposium Workshops, IPDPSW 2018*, pages 987–995, 2018.

[29] Robert A. van den Berg, Huub C.J. Hoefsloot, Johan A. Westerhuis, Age K. Smilde, and Mariët J. van der Werf. Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics*, 7:1–15, 2006.

[30] R. Bro, K. Kjeldahl, A. K. Smilde, and H. A.L. Kiers. Cross-validation of component models: A critical look at current methods. *Analytical and Bioanalytical Chemistry*, 390(5):1241–1251, 2008.

[31] Thorsten Mix, Jasmin Janneschütz, Rami Ludwig, Julia Eichbaum, Markus Fischer, and Thomas Hackl. From Nontargeted to Targeted Analysis: Feature Selection in the Differentiation of Truffle Species (Tuber spp.) Using 1H NMR

Spectroscopy and Support Vector Machine. *Journal of Agricultural and Food Chemistry*, 71(46):18074–18084, 2023.

[32] Anna Lena Gippert, Silvia Madritsch, Patrick Woryna, Sandra Otte, Martina Mayrhofer, Herbert Eigner, Adriana Garibay-Hernández, John C. D'Auria, Eva M. Molin, and Hans Peter Mock. Unraveling metabolic patterns and molecular mechanisms underlying storability in sugar beet. *BMC Plant Biology*, 22(1):1–18, 2022.

[33] Ottavia Giampaoli, Fabio Sciubba, Giorgia Conta, Giorgio Capuani, Alberta Tomassini, Giorgio Giorgi, Elisa Brasili, Walter Aureli, and Alfredo Miccheli. Red beetroot's nmr-based metabolomics: Phytochemical profile related to development time and production year. *Foods*, 10(8):1–12, 2021.

[34] Antonia Garrido Frenich, Araceli Rivera-p, and Roberto Romero-gonz. Untargeted 1 H NMR-based metabolomics and multi-technique data fusion : A promising combined approach for geographical and processing authentication of thyme by multivariate statistical analysis. 420(April):0–2, 2023.

[35] Soyeon Jo, Yuyoung Song, Ji Ho Jeong, Junghyun Hwang, and Yongae Kim. Geographical discrimination of Allium species (garlic and onion) using 1H NMR spectroscopy with multivariate analysis. *International Journal of Food Properties*, 23(1):241–254, 2020.

[36] Caroline Schmitt, Tobias Schneider, Laura Rumask, Markus Fischer, and Thomas Hackl. Food Profiling: Determination of the Geographical Origin of Walnuts by 1H NMR Spectroscopy Using the Polar Extract. *Journal of Agricultural and Food Chemistry*, 68(52):15526–15534, 2020.

[37] Junfeng Lei, Lili Shen, Wei Zhang, Fangchao Ma, Jingchen Wang, Tingting Wei, and Chengping Xie. Comparative Chemical Characterization of Potato Powders Using 1 H NMR Spectroscopy and Chemometrics. *Plant Foods for Human Nutrition*, (78):590–596, 2023.

[38] Paraskevi Tsermoula, Niels Bastian Kristensen, Nabiollah Mobaraki, and Bekzod Khakimov. Efficient Quantification of Milk Metabolites from 1 H NMR Spectra Using the Signature Mapping (SigMa) Approach: Chemical Shift Library Development for Cows' Milk and Colostrum. 2024.

[39] David S. Wishart, Craig Knox, An Chi Guo, Roman Eisner, Nelson Young,

Bijaya Gautam, David D. Hau, Nick Psychogios, Edison Dong, Souhaila Bouatra, Rupasri Mandal, Igor Sinelnikov, Jianguo Xia, Leslie Jia, Joseph A. Cruz, Emilia Lim, Constance A. Sobsey, Savita Shrivastava, Paul Huang, Philip Liu, Lydia Fang, Jun Peng, Ryan Fradette, Dean Cheng, Dan Tzur, Melisa Clements, Avalyn Lewis, Andrea De souza, Azaret Zuniga, Margot Dawe, Yeping Xiong, Derrick Clive, Russ Greiner, Alsu Nazyrova, Rustem Shaykhutdinov, Liang Li, Hans J. Vogel, and Ian Forsythei. HMDB: A knowledgebase for the human metabolome. *Nucleic Acids Research*, 37(SUPPL. 1):603–610, 2009.

[40] Joaquim Jaumot, Anna de Juan, and Romà Tauler. MCR-ALS GUI 2.0: New features and applications. *Chemometrics and Intelligent Laboratory Systems*, 140:1–12, 2015.

5.7 Supplementary Material

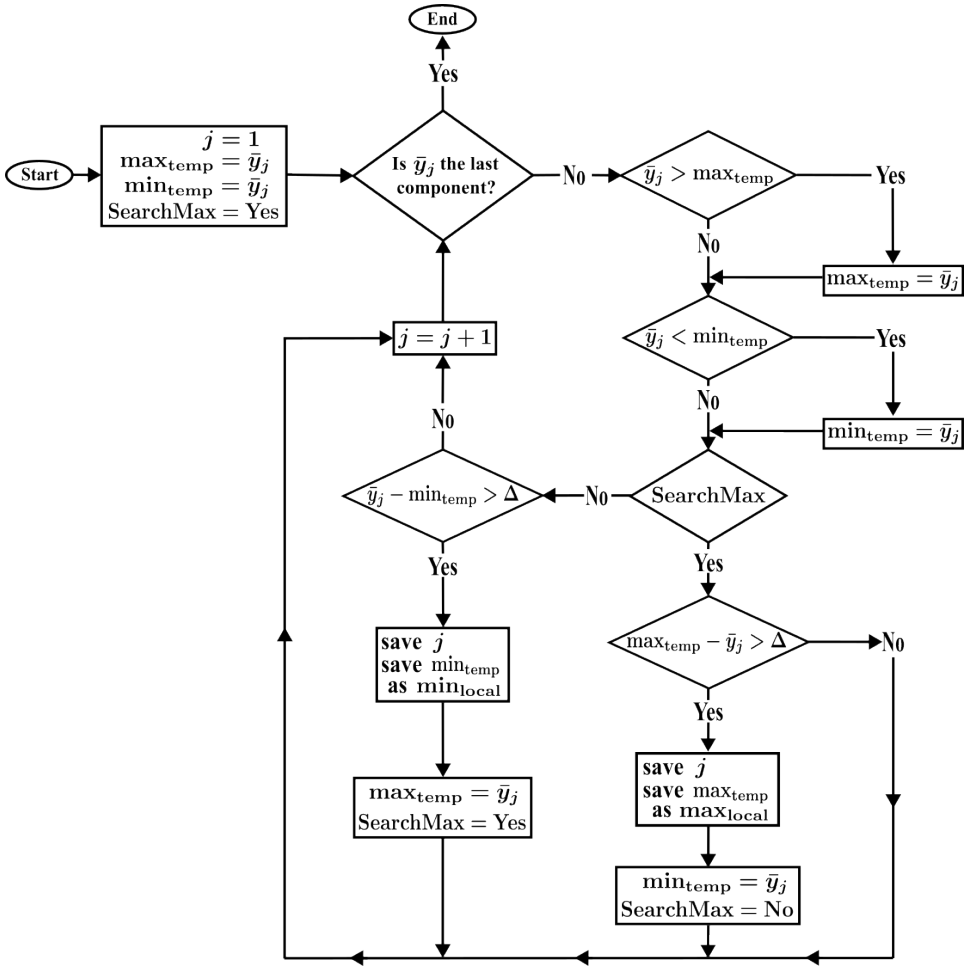

Figure 5.7.1: Flowchart of the maxima and minima algorithm. $\bar{\mathbf{y}}$ is the average intensity of the resonance signals (reference) and Δ corresponds to the selected threshold value.

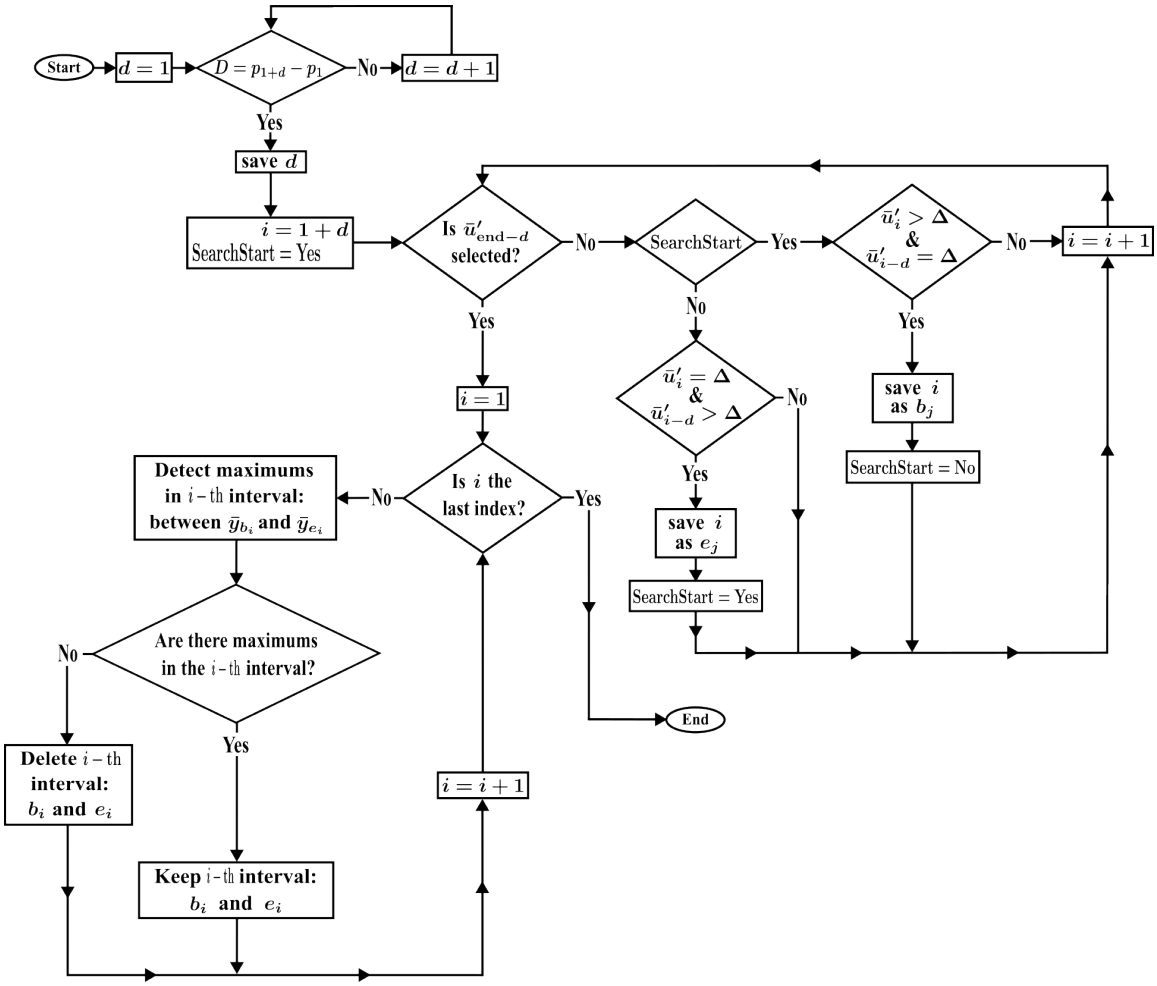

Figure 5.7.2: Flowchart of the detailed interval delimitation algorithm. $\bar{\mathbf{u}}'$ (modified vector) represents the average intensity with values below Δ being replaced by Δ. Index b and e correspond to the beginning and end of an interval, respectively.

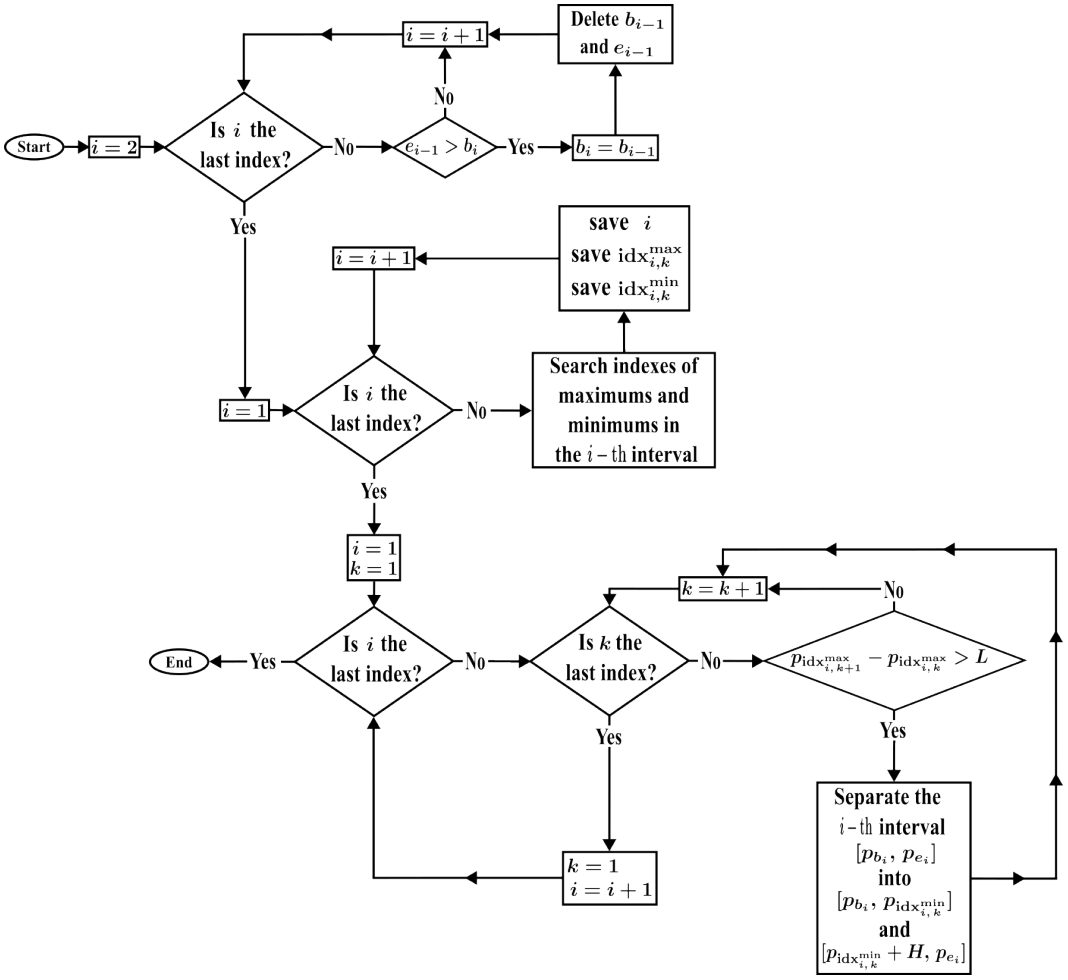

Figure 5.7.3: Flowchart of the detailed interval correction algorithm. To split (redefine) an overlapping interval, the distance between consecutive maxima must exceed a specified value of L ppm with a separation of H ppm.

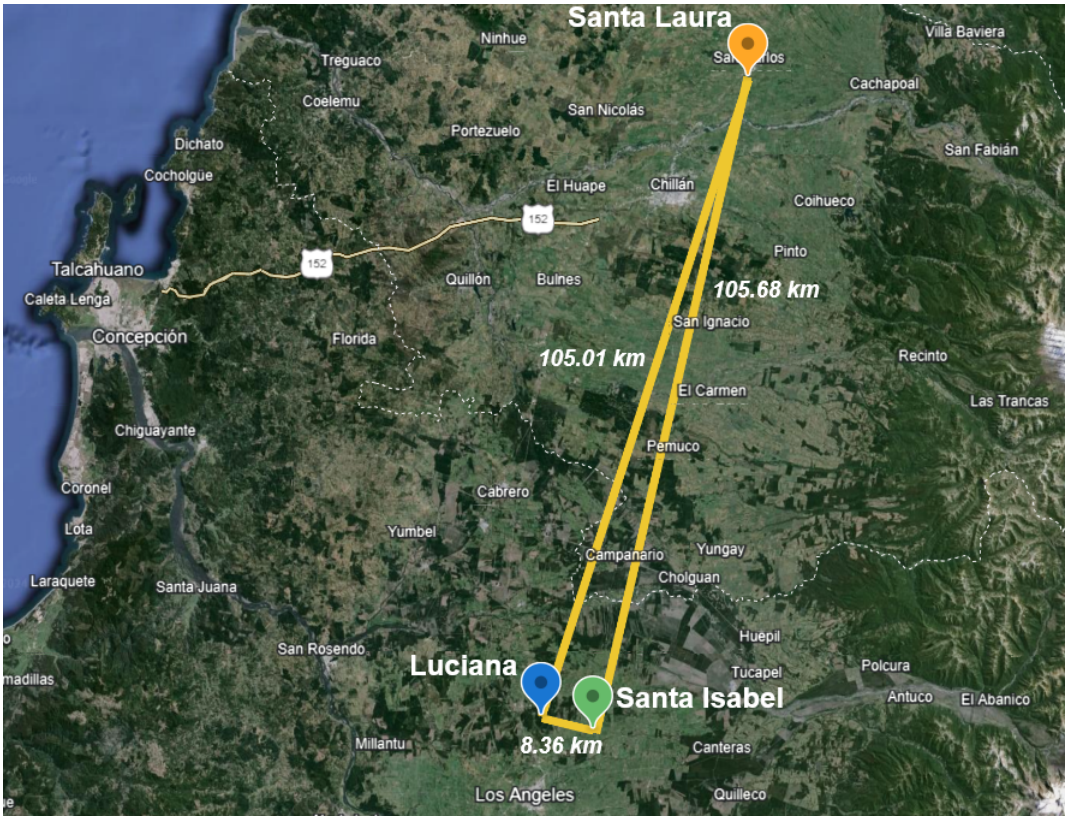

Figure 5.7.4: Map of sampling points associated with the three different sugar beet (*Beta vulgaris L.*) fields evaluated during the 2023 harvest.

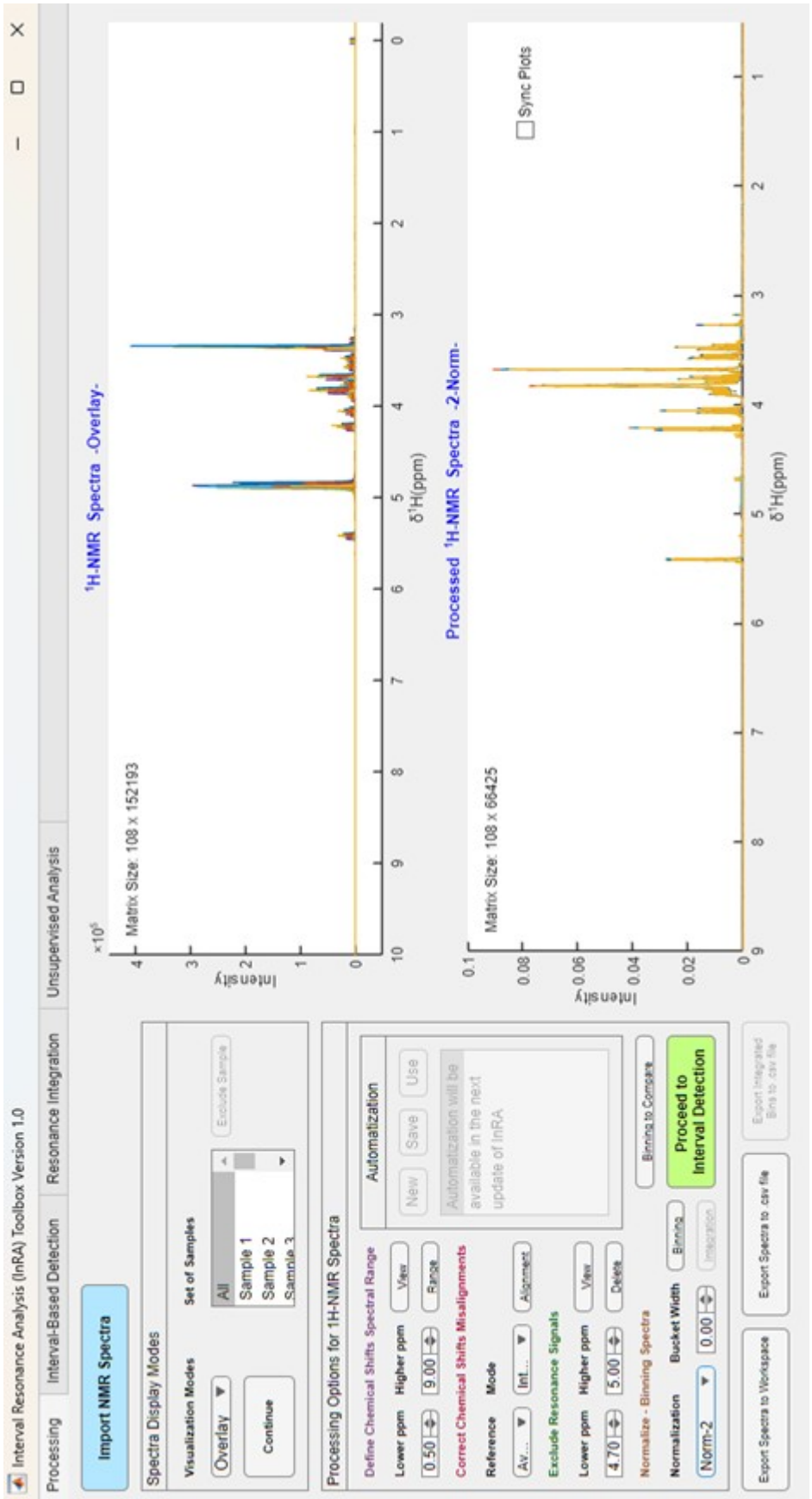

Figure 5.7.5: Processing tab.

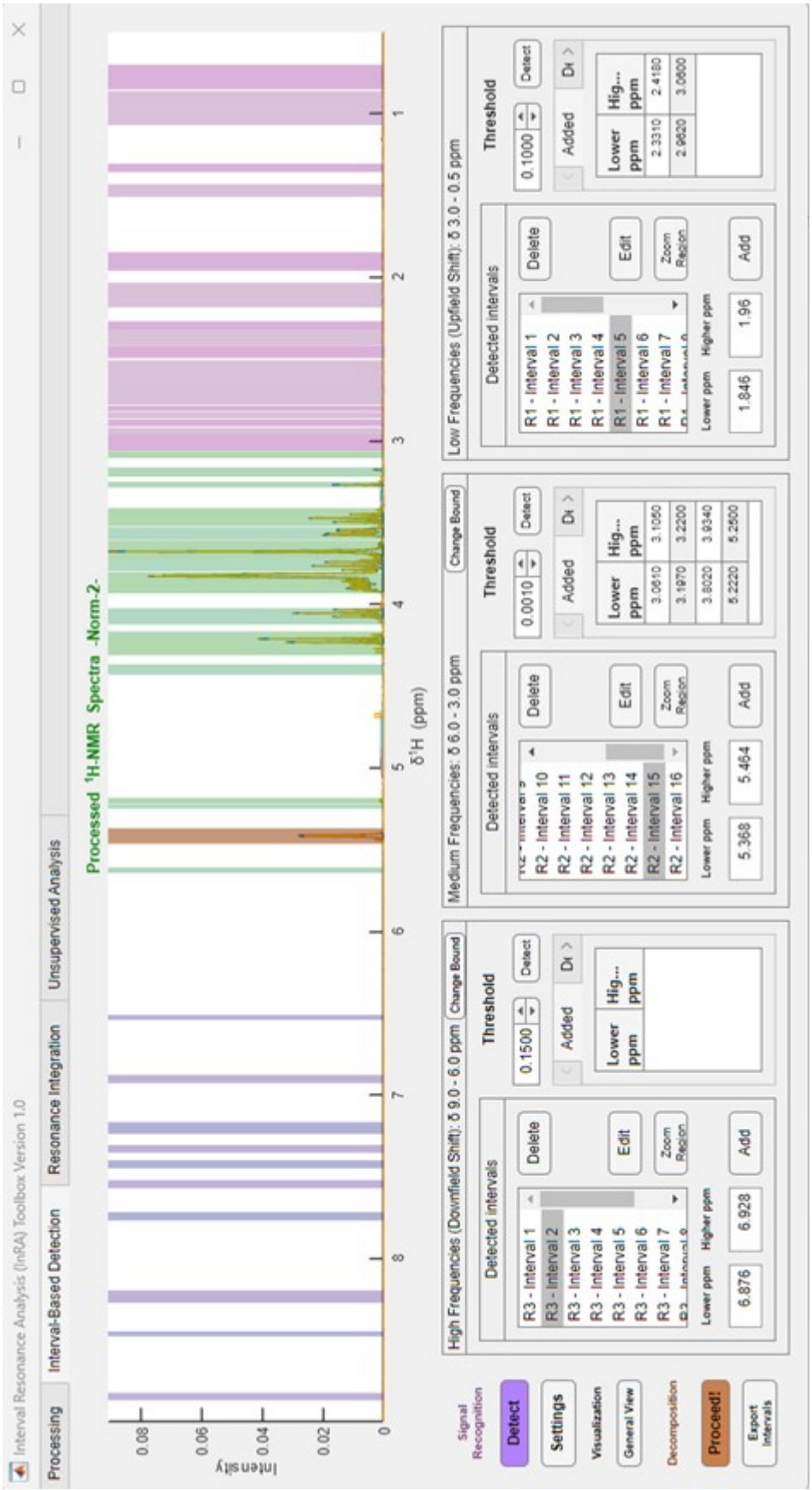

Figure 5.7.6: Interval-based detection tab.

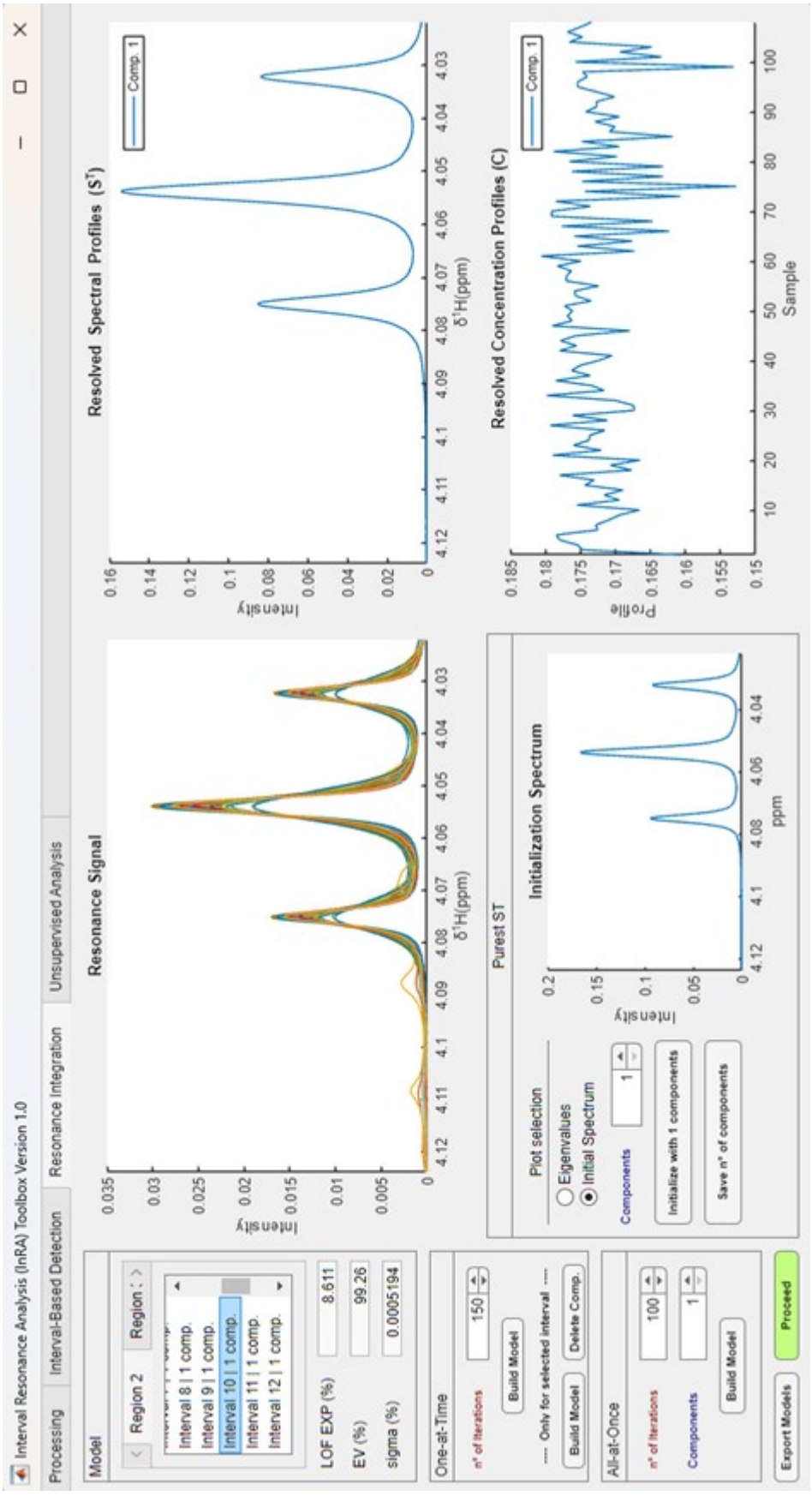

Figure 5.7.7: Resonance integration tab.

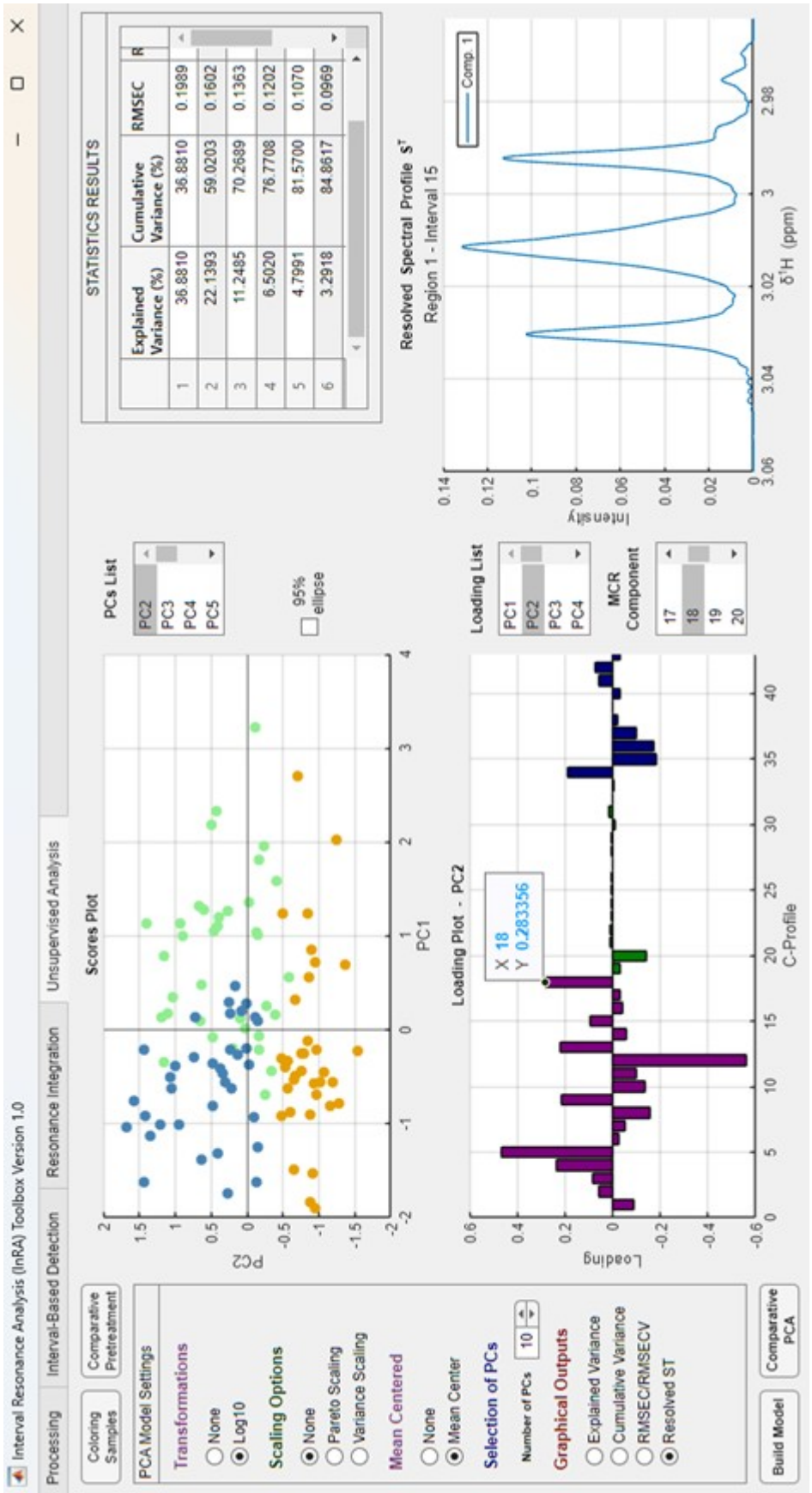

Figure 5.7.8: Unsupervised analysis tab.

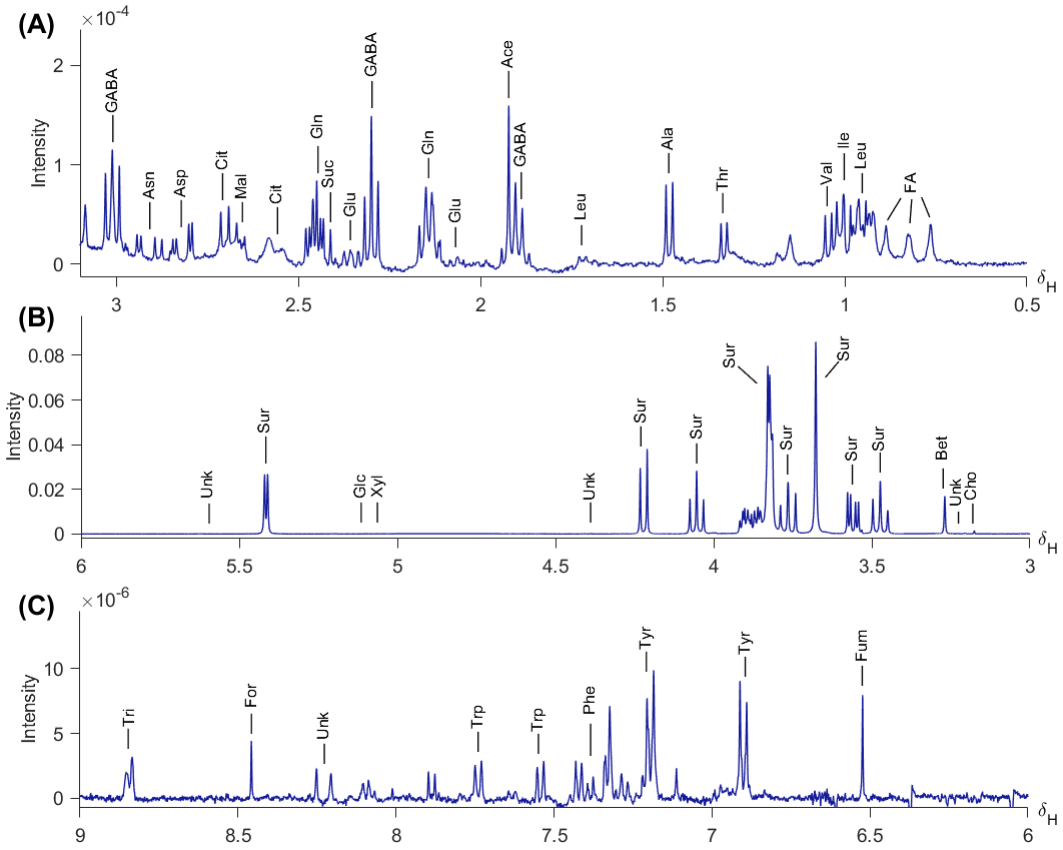

Figure 5.7.9: Representative ^{1}H NMR spectrum (400.13 MHz) of the hydrophilic extract from sugar beet root referenced to TSP-d_4 5.8 Mm. **A)** $\delta_H = 0.50 - 3.00$, **B)** $\delta_H = 3.00 - 6.00$, **C)** $\delta_H = 6.00 - 9.00$. FA: Fatty acids; Leu: Leucine, Ile: Isoleucine; Val: Valine; Thr: Threonine; Ala: Alanine; GABA: γ-aminobutyrate; Ace: Acetate; Glu: Glutamate; Gln: Glutamine; Suc: Succinate; Cit: Citrate; Mal: Malate; Asp: Aspartate; Asn: Asparagine; Cho: Choline; Unk: Uknown; Bet: Betaine; Sur: Sucrose; Xyl: α-xylose; Glc: α-glucose; Fum: Fumarate; Tyr: Tyrosine; Phe: Phenylalanine; Trp: Tryptophan; For: Formate; Tri: Trigolline.

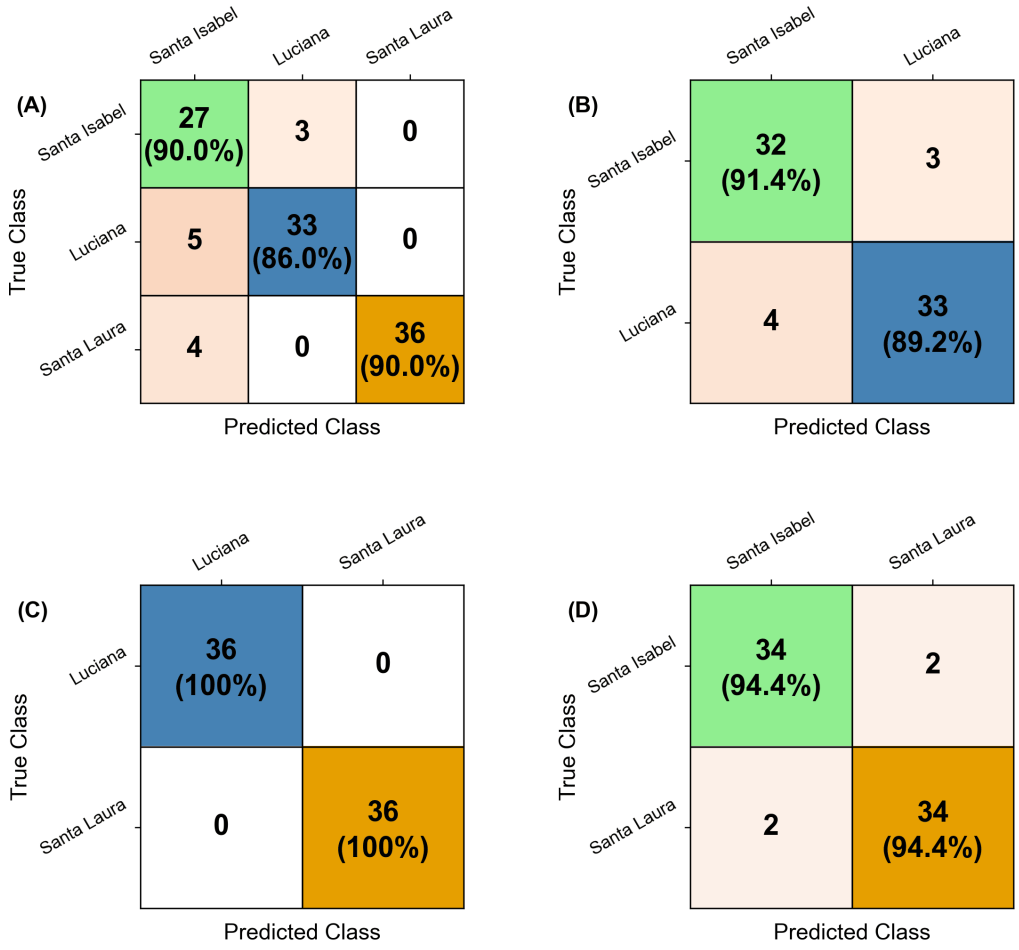

Figure 5.7.10: Results of the two and three-class PLS–DA classification models through $\mathbf{C}_{Features}$. **A)** Three-class confusion matrix for model A (Santa Isabel – Luciana – Santa Laura). **B)**, **C)**, and **D)**, confusion matrices for model B (Santa Isabel – Luciana), model C (Luciana – Santa Laura), and model D (Santa Isabel – Santa Laura).

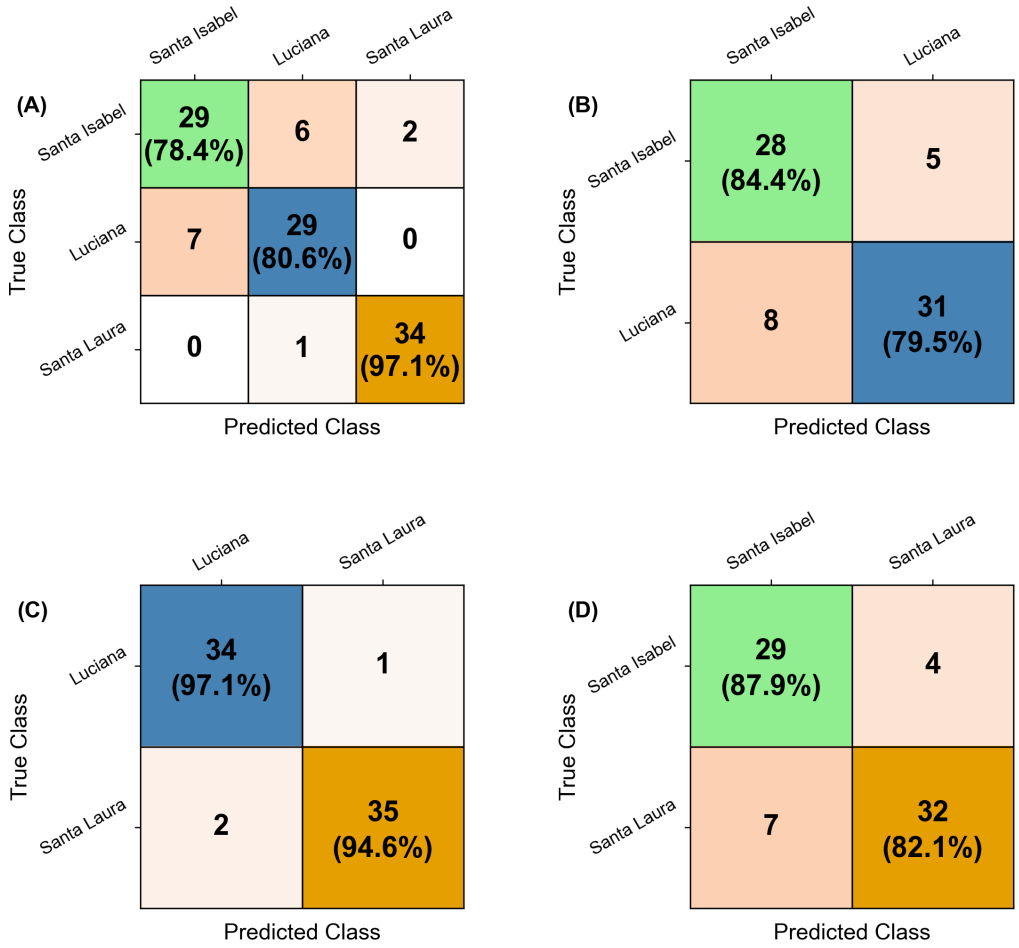

Figure 5.7.11: Results of the two and three-class PLS–DA classification models through $\mathbf{B}_{Spectra}$. **A)** Three-class confusion matrix for model A (Santa Isabel – Luciana – Santa Laura). **B)**, **C)**, and **D)**, confusion matrices for model B (Santa Isabel – Luciana), model C (Luciana – Santa Laura), and model D (Santa Isabel – Santa Laura).

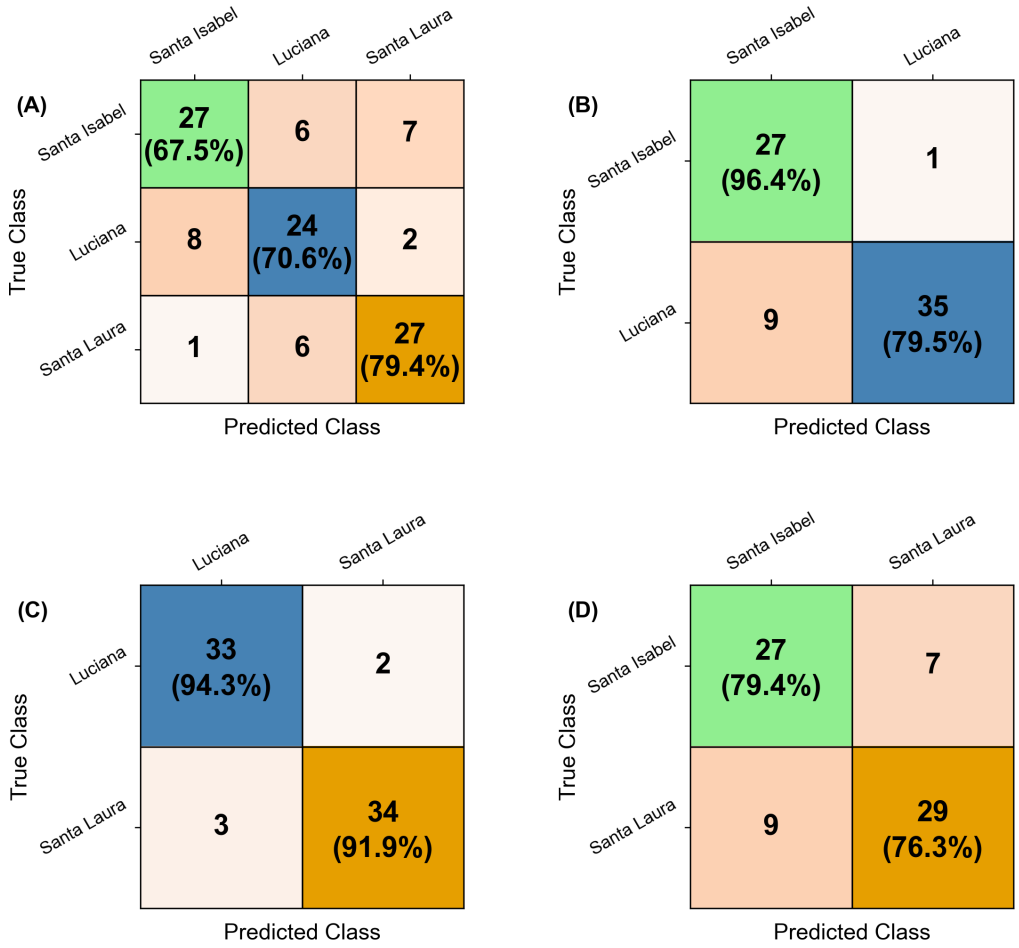

Figure 5.7.12: Results of the two and three-class PLS–DA classification models through $\mathbf{P}_{Spectra}$. **A)** Three-class confusion matrix for model A (Santa Isabel – Luciana – Santa Laura). **B)**, **C)**, and **D)**, confusion matrices for model B (Santa Isabel – Luciana), model C (Luciana – Santa Laura), and model D (Santa Isabel – Santa Laura).

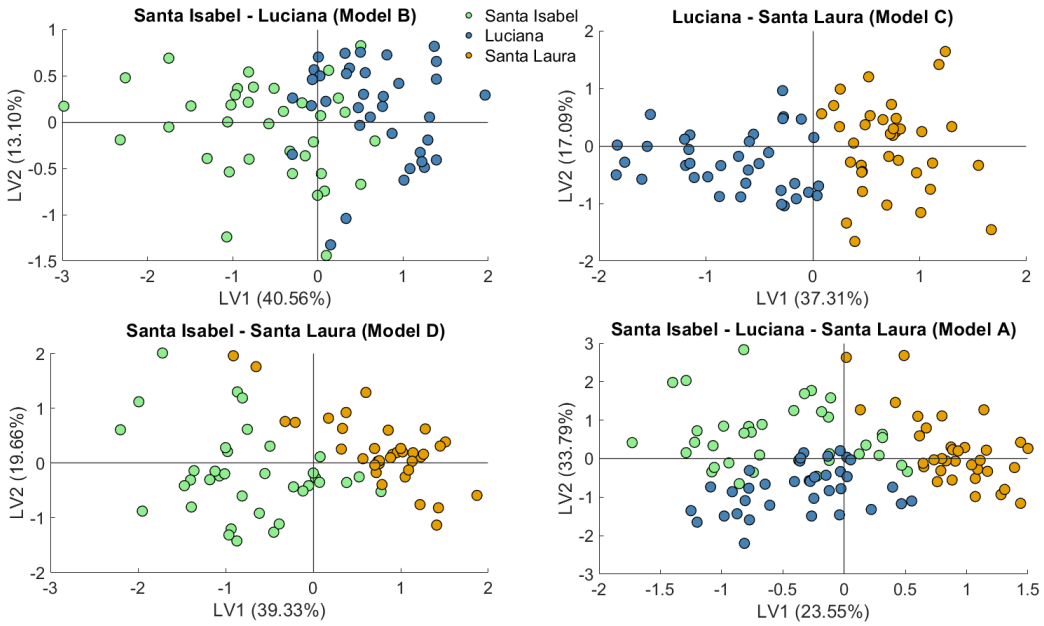

Figure 5.7.13: PLS–DA scores plot (LV1 vs LV2) of the two and three-class classification models via $\mathbf{C}_{Features}$.

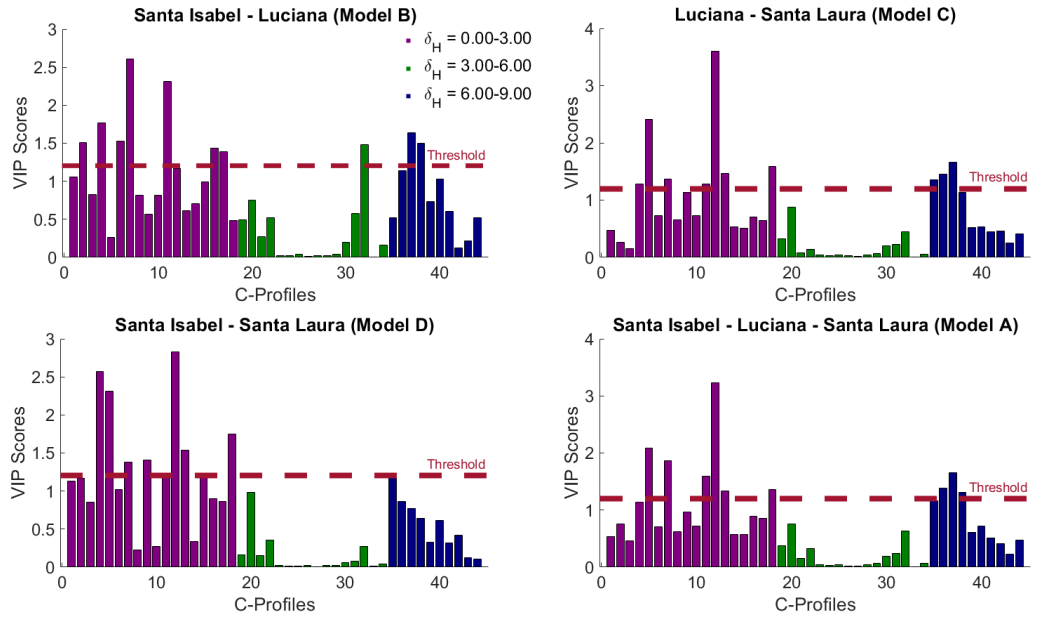

Figure 5.7.14: VIP plots obtained of the two and three-class PLS–DA classification models. The threshold value (1.2) is represented as a red line bar. VIP ≥ 1.2 were analyzed for statistical significance by a one-way ANOVA test with Bonferroni-Holm correction.

Table 5.7.1: Equations implemented in InRA for fit quality of the MCR and PCA models.

No.	Name	Equation
MCR		
1	Standard deviation of residuals	$\sigma = \sqrt{\frac{\sum_{i,j} e_{ij}^2}{n_{rows} \times n_{columns}}}$
2	Lack of fit (%)	$LOF\,(\%) = 100 \times \sqrt{\frac{\sum_{i,j} e_{ij}^2}{\sum_{i,j} d_{ij}^2}}$
PCA		
4	Root mean square error of calibration	$RMSEC = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$
5	Root mean square error of cross-validation	$RMSECV = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_{i,CV})}{n}}$

-$e_{ij} = d_{ij} - d_{ij}^*$

-d_{ij} = Element of the original matrix **D**.

-d_{ij}^* = Element of matrix **D** reproduced by MCR.

-$y_i - \hat{y}_i$ = Difference between the observed and predicted values in the calibration set.

-$y_i - \hat{y}_{i,CV}$ = Difference between the observed and predicted values in the test set during CV.

Table 5.7.2: List of sugar beet samples analyzed.

No.	ID	Field	MC[a](%)	No.	ID	Field	MC[a](%)	No.	ID	Field	MC[a](%)	No.	ID	Field	MC[a](%)
1	SI-1	Santa Isabel	74.77	**25**	SI-25	Santa Isabel	75.87	**49**	L-17	Luciana	73.48	**73**	SL-9	Santa Laura	77.56
2	SI-2	Santa Isabel	76.72	**26**	SI-26	Santa Isabel	77.75	**50**	L-18	Luciana	77.66	**74**	SL-10	Santa Laura	75.93
3	SI-3	Santa Isabel	74.67	**27**	SI-27	Santa Isabel	76.65	**51**	L-19	Luciana	77.12	**75**	SL-11	Santa Laura	76.67
4	SI-4	Santa Isabel	77.51	**28**	SI-28	Santa Isabel	74.65	**52**	L-20	Luciana	74.98	**76**	SL-12	Santa Laura	76.21
5	SI-5	Santa Isabel	76.27	**29**	SI-29	Santa Isabel	77.25	**53**	L-21	Luciana	76.61	**77**	SL-13	Santa Laura	77.72
6	SI-6	Santa Isabel	75.80	**30**	SI-30	Santa Isabel	76.65	**54**	L-22	Luciana	77.50	**78**	SL-14	Santa Laura	75.77
7	SI-7	Santa Isabel	74.97	**31**	SI-31	Santa Isabel	74.65	**55**	L-23	Luciana	73.82	**79**	SL-15	Santa Laura[b]	76.80
8	SI-8	Santa Isabel	73.64	**32**	SI-32	Santa Isabel	77.25	**56**	L-24	Luciana	76.35	**80**	SL-16	Santa Laura	77.86
9	SI-9	Santa Isabel	74.78	**33**	L-1	Luciana	77.04	**57**	L-25	Luciana	76.95	**81**	SL-17	Santa Laura	77.75
10	SI-10	Santa Isabel	77.89	**34**	L-2	Luciana	76.49	**58**	L-26	Luciana	74.14	**82**	SL-18	Santa Laura	76.88
11	SI-11	Santa Isabel	76.76	**35**	L-3	Luciana	76.13	**59**	L-27	Luciana	74.74	**83**	SL-19	Santa Laura	76.67
12	SI-12	Santa Isabel	77.740	**36**	L-4	Luciana	74.87	**60**	L-28	Luciana	77.75	**84**	SL-20	Santa Laura	76.40
13	SI-13	Santa Isabel	76.43	**37**	L-5	Luciana	77.29	**61**	L-29	Luciana	75.62	**85**	SL-21	Santa Laura	76.98
14	SI-14	Santa Isabel	74.21	**38**	L-6	Luciana	74.98	**62**	L-30	Luciana	75.02	**86**	SL-22	Santa Laura	77.54
15	SI-15	Santa Isabel	77.56	**39**	L-7	Luciana	76.26	**63**	L-31	Luciana	74.80	**87**	SL-23	Santa Laura	75.33
16	SI-16	Santa Isabel	74.81	**40**	L-8	Luciana	74.88	**64**	L-32	Luciana	77.22	**88**	SL-24	Santa Laura	75.75
17	SI-17	Santa Isabel	77.78	**41**	L-9	Luciana	74.62	**65**	SL-1	Santa Laura	77.43	**89**	SL-25	Santa Laura	76.23
18	SI-18	Santa Isabel	77.19	**42**	L-10	Luciana	76.90	**66**	SL-2	Santa Laura	77.87	**90**	SL-26	Santa Laura	77.86
19	SI-19	Santa Isabel	75.11	**43**	L-11	Luciana	77.40	**67**	SL-3	Santa Laura	76.12	**91**	SL-27	Santa Laura	75.54
20	SI-20	Santa Isabel	75.44	**44**	L-12	Luciana	75.77	**68**	SL-4	Santa Laura	77.50	**92**	SL-28	Santa Laura	76.44
21	SI-21	Santa Isabel[b]	77.86	**45**	L-13	Luciana	75.68	**69**	SL-5	Santa Laura	76.45	**93**	SL-29	Santa Laura	77.51
22	SI-22	Santa Isabel	74.89	**46**	L-14	Luciana[b]	74.80	**70**	SL-6	Santa Laura	75.84	**94**	SL-30	Santa Laura	77.42
23	SI-23	Santa Isabel	76.20	**47**	L-15	Luciana	73.89	**71**	SL-7	Santa Laura	76.61	**95**	SL-31	Santa Laura	76.39
24	SI-24	Santa Isabel	73.49	**48**	L-16	Luciana	77.33	**72**	SL-8	Santa Laura	77.10	**96**	SL-32	Santa Laura	77.04

[a]Moisture content. [b]Samples used to monitor the robustness and reproducibility of the extraction process.

Chapter 6

Multivariate Analysis for a Comparative Study of Pectin Production Methods using TD-NMR with Fast Fourier Transform (FFT)

This chapter presents and discusses the most recent results that have laid the groundwork for a third manuscript currently under preparation, entitled "Multivariate Analysis for a Comparative Study of Pectin Production Methods using TD-NMR with Fast Fourier Transform (FFT)".

This study, which complements the Ph.D. thesis, investigates the application of TD-NMR as an analytical technique for the study of sugar beet-derived pectin samples. As its principal innovation, the study introduces a multivariate strategy incorporating second-order models, specifically multivariate curve resolution with alternating least squares (MCR-ALS) and parallel factor analysis (PARAFAC), to allow the discrimination of pectin produced through distinct precipitation procedures.

Therefore, this manuscript is framed as a complementary contribution to the specific objective 2, where the implementation of advanced multivariate analytical approaches facilitated the development of supervised classification models with enhanced selectivity and discriminative power.

Multivariate Analysis for a Comparative Study of Pectin Production Methods using TD-NMR with Fast Fourier Transform (FFT)

Cristian A. Fuentes[a,b*], Esmanur Ilhan[c*], Maria Ivanova[d*], Aylin Özgür Göksu[e], Mecit Halil Öztop[b], Leonid Grunin[d], Rosario del P. Castillo[a,b**]

[a]Department of Food Engineering, Middle East Technical University, Ankara 06800, Türkiye

[b]Departamento de Análisis Instrumental, Facultad de Farmacia, Universidad de Concepción, Concepción 4070386, Chile

[c]Laboratorio de Bioespectroscopia y Quimiometría (BioSpeQ), Centro de Biotecnología, Universidad de Concepción, Concepción 4070386, Chile

[d]Resonance Systems GmbH, 73230 Kirchheim unter Teck, Germany

[e]Kayseri Sugar R&D Center, Kayseri Sugar Factory, Kayseri 38070, Türkiye

***Authors with equal contribution**

****Corresponding author:**

Rosario del P. Castillo

Tel: +56 41 220 7298

Email: rosariocastillo@udec.cl

Abstract

Pectin, widely used in the food industry for its gelling and stabilizing properties, exhibits functional characteristics that are strongly influenced by the extraction and precipitation methods used, such as ultrafiltration and maltodextrin addition. This study presents a novel methodology based on time-domain nuclear magnetic resonance (TD-NMR) relaxometry combined with multivariate analysis, aimed at differentiating and classifying pectin samples according to their production process. The relaxation curves transformed to spectra by FFT were decomposed using second-order algorithms via multivariate curve resolution alternating least squares (MCR-ALS) and parallel factor analysis (PARAFAC) to extract chemically meaningful components. Three groups of samples were analyzed using CPMG and SE-MSE sequences: pure pectins obtained by different production methods, one commercial pectin, and manually prepared binary mixtures. Classification performance was evaluated using k-nearest neighbors (k-NN), partial least squares-discriminant analysis (PLS-DA), and soft independent modeling of class analogy (SIMCA). The results showed that the combination of FFT with MCR-ALS achieved 100% classification accuracy in all models, including binary mixtures. PARAFAC-based scores also produced high classification performance, with accuracies between 96.5% and 100%. In contrast, models based solely on CPMG relaxation curves showed a reduced ability to resolve subtle compositional differences, particularly for mixtures. Overall, this integrative approach demonstrates the potential of TD-NMR combined with advanced multivariate decomposition techniques as a powerful, non-destructive tool for characterizing structurally similar samples and supporting quality control in complex food matrices.

Keywords – Pectin, low field nuclear magnetic resonance, chemometrics, parallel factor analysis, multivariate curve resolution

6.1 Introduction

Pectin is a complex polysaccharide that constitutes one of the principal structural components of the middle lamella and the primary cell wall in higher plants [1]. Its predominant molecular structure is composed of a linear backbone of α-D-galacturonic acid residues linked via (1→4) glycosidic bonds, interrupted by highly branched rahmnogalacturonan regions bearing neutral sugar side chains, such as arabinose and galactose [2, 3]. The latter structural complexity imparts pectin with unique functional properties, making it widely used as a gelling agent, emulsifier, thickener, and stabilizer in a wide range of food matrices [4, 5].

Industrial pectin production has traditionally relied on the extraction of citrus peels and apple pomace, primarily due to their high pectin content and seasonal availability [6]. However, in recent years, increasing attention has been directed towards more sustainable alternative sources, such as sugar beet pulp (*Beta vulgaris* L.), a by-product of the sugar industry that is generated in substantial quantities (estimated at more than 20 million tons annually in Europe) [7]. The residual biomass, which comprises a significant fraction of polysaccharides (22–24% cellulose, 24–32% hemicellulose and 15–25% pectin), has historically been used for animal feed, despite its high potential value as a raw material for the recovery of bio-active compounds of industrial interest [8].

Pectin extracted from sugar beet pulp exhibits physicochemical properties that are distinct from those of other conventional sources. These differences include a higher degree of acetylation in galacturonic acid residues, an increased presence of rhamnose, and the incorporation of phenolic esters such as ferulic acid into its side chains [9]. In addition, its notable emulsifying capacity has been attributed to a relatively high protein content, making it a functionally attractive biopolymer [10]. Extraction is typically performed through acid hydrolysis at elevated temperatures using mineral acids, e.g. hydrochloric or nitric acid, followed by purification and precipitation, commonly with ethanol [11, 12]. However, the efficiency of the extraction process and the quality of the final product are critically dependent on the extraction conditions, the precipitation method, and the intrinsic characteristics of the raw material.

Given the structural complexity of pectin, the use of robust, rapid, and non-destructive analytical techniques is essential for the efficient characterization of its physicochemical properties. In this context, time-domain nuclear magnetic resonance (TD-NMR) has emerged as a powerful tool for the analysis of complex molecular systems in food matrices [13]. TD-NMR does not require sample preparation or deuterated solvent, is non-invasive, offers high reproducibility, and allows measurements in solid, semi-solid or liquid states [14]. Based on measurement of longitudinal (T_1) and transverse (T_2) relaxation times, this technique provides valuable information on molecular mobility and intermolecular interactions within the system under investigation [14, 15]. Its application has been extended to the study of dairy products [16], meat [17], fruits [18], cheese [19] and starch [20], positioning TD-NMR as a versatile technique in food quality control.

The information provided by TD-NMR can be challenging for direct interpretation; nevertheless, the combination with multivariate analysis offers substantial potential for qualitative and quantitative analysis by allowing the extraction of meaningful insights from complex datasets [21]. Among the most widely used approaches that combine TD-NMR with multivariate analysis are principal component analysis (PCA), soft independent modeling of class analogy (SIMCA), k-nearest neighbors (k-NN), partial least squares discriminant analysis (PLS-DA), partial least squares regression (PLS-R) and support vector machine (SVM). These models have been successfully applied for exploratory analysis, classification, or quantification tasks, e.g., the selection of industrial tomatoes, the detection and quantification of formaldehyde in milk, and the monitoring of the diesel content in fuel [22, 23, 24]. Moreover, models such as multivariate curve resolution with alternating least squares (MCR-ALS) and parallel factor analysis (PARAFAC) offer powerful decomposition strategies. MCR-ALS resolves complex instrumental responses into individual contributions of each component in a mixture through constrained bilinear modeling [25]. PARAFAC decomposes multi-way trilinear arrays, allowing the identification and quantification of independent underlying signals, termed components [26]. Although both approaches have been extensively used in infrared and fluorescence spectroscopy, their application in TD-NMR remains relatively unexplored.

Therefore, the aim of this study was to evaluate the potential of TD-NMR relaxometry to differentiate pectin samples obtained through distinct production methods. For this purpose, a novel methodology was proposed in which relaxation curves were transformed into the frequency domain using fast Fourier transform (FFT) and subsequently decomposed using second-order techniques such as MCR-ALS and PARAFAC. In addition, classification models including k-NN, PLS-DA, and SIMCA were applied to assess the discriminative capacity of the proposed approach.

6.2 Study design

6.2.1 Pectin samples

Five pectin samples using sugar beet pulp as raw material plus an additional commercially purchased sample were provided by the Kayseri Sugar facility (Türkiye). The extraction process was carried out under acidic conditions (pH 2.0) and temperatures ranging from 80 to 90°C to promote pectin solubilization. Following extraction, the liquid phase containing dissolved pectin was mechanically separated from the residual solid pulp by means of a pressing technique. The precipitation of each pectin from the liquid extract was achieved through distinct methods: ultra-filtration, isopropyl alcohol, or a combination of both methods. In some formulations, varying concentrations of maltodextrin were incorporated to enhance the drying efficiency and the physicochemical properties of the resulting pectin. In addition, to evaluate the clustering performance of classification models against new sources of variation, a series of binary mixtures were prepared, including commercially purchased pectin derived from sugar beet pulp. Table 6.2.1 summarizes the full set of samples, the precipitation method, the concentration of maltodextrin, ID, and the formulated specific binary mixtures.

Table 6.2.1: Summary of pectin samples employed for evaluation.

Pectin ID	**Maltodextrin (%)**	**Precipitation Method**
SBPP	-	Commercially purchased
MD30	30	Isopropyl alcohol
UF	0	Ultra-filtration
UF IPA	0	Isopropyl alcohol + Ultra-filtration
UF IPA MD10	10	Isopropyl alcohol + Ultra-filtration
UF MD10	10	Ultra-filtration
Binary mixtures		**Ratio (%)**
MD30 - UF		20:80 ; 50:50 ; 80:20
SBPP - UF IPA		20:80 ; 50:50 ; 80:20
UF MD10 - UF IPA MD10		20:80 ; 50:50 ; 80:20

6.2.2 TD-NMR measurements and methodological workflow

To maximize sensitivity to structural heterogeneity arising from different pectin production methods, particularly those related to molecular mobility variations influenced by the degree of methoxylation and the presence of rigid domains, two complementary pulse sequences were employed: Carr-Purcell-Meiboom-Gill (CPMG) and Solid Echo–Magic Sandwich Echo (SE-MSE). Fig. 6.2.1 shows the overall workflow implemented in this study, which started with the spectral acquisition of pectin samples using both pulse sequences, followed by spectral processing, matrix organization according to pulse type with sample composition, and concluded with the application of two-way and three-way multivariate analysis to investigate compositional differences in the sample set related to the production method.

TD-NMR acquisitions were performed using two NMR analyzers operating at a ^{1}H frequency of 18 MHz from Resonance Systems GmbH (Germany): “Spin Track” for CPMG and “Spin Mate” for SE-MSE experiments (Fig. 6.2.1A). All measurements were performed by personnel from the Middle East Technical University (METU) and Resonance Systems. The experimental conditions were standardized for both pulse sequences, i.e, a probe temperature of 30 ± 0.005°C, a 90° RF pulse of 3.2 μs, a 180° RF pulse of 6.3 μs, a ringing time of 8 μs, a receiver gain of 500 kHz and a scan repetition delay of 1200 ms. SE-MSE measurements were acquired at Tau (τ) values of 13,000, 36,000, 120,000, and 286,000 ns to improve the resolution of proton dipolar interactions. For each sample (pure and binary mixtures), five independent spectral replicates (n = 5) were acquired.

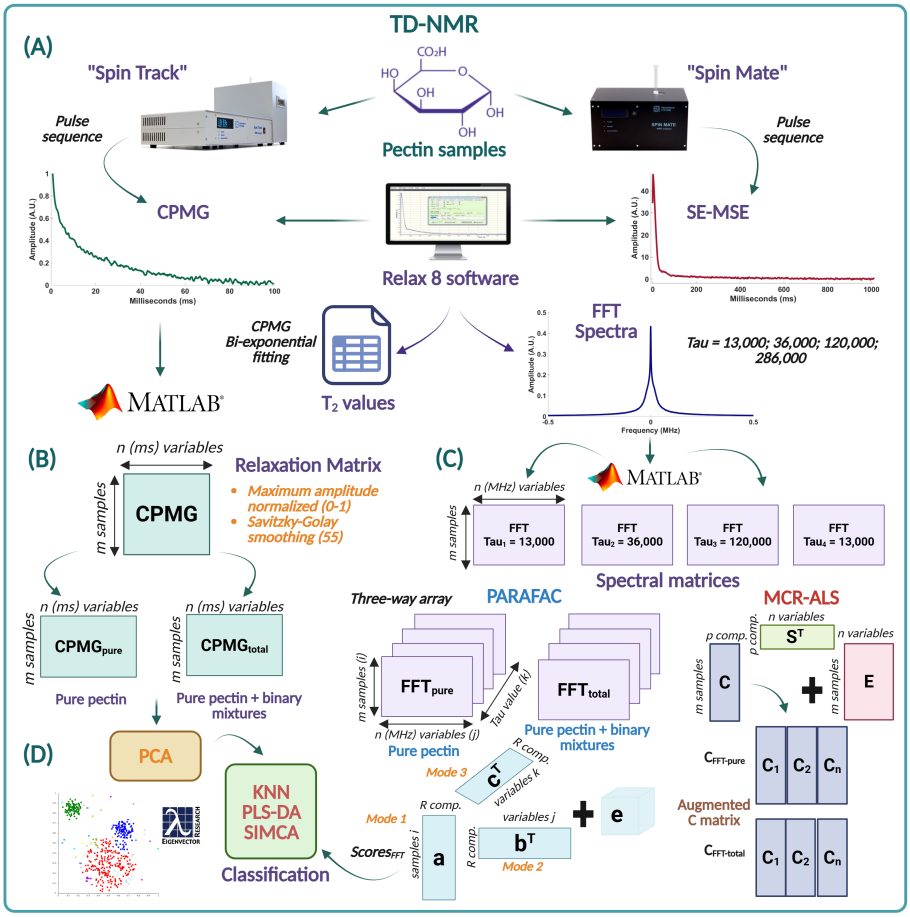

Figure 6.2.1: Proposed workflow for the comparison study of pectin samples using TD-NMR and multivariate analysis. **A)** CPMG and SE-MSE measurements. **B)** Two-way (MCR-ALS) and Three-way (PARAFAC) decomposition applied to the FFT spectra.

The CPMG relaxation curves of pure pectin samples were fitted to a bi-exponential model in Relax 8 software (Resonance Systems GmbH, Germany) to extract the characteristic transverse relaxation components T_{2a} and T_{2b}, respectively. Furthermore, all CPMG relaxation curves were imported into the MATLAB R2023b environment (The MathWorks Inc., Natick, MA, USA) as a spectral matrix. Each curve was normalized to its maximum amplitude (0-1) and smothered using a second-order Savitzky-Golay polynomial algorithm (window width = 55) using in-house MATLAB routines. The spectral matrix was partitioned into two independent matrices: a first matrix, referred to as $\mathbf{CPMG}_{pure}$, contained only the relaxation curves of the pure pectin samples, and a second matrix, designated $\mathbf{CPMG}_{total}$, comprised pure pectin samples and binary mixtures, as illustrated in Fig. 6.2.1B.

In parallel, the free induction decays (FIDs) obtained with the SE-MSE sequence at various values of τ were converted to the frequency domain by means of a fast Fourier transform (FFT) in Relax 8 software, with a frequency range of -0.5 to 0.5 MHz and a ringing time of 2 μs to minimize electronic artifacts and enhance spectral resolution. The resulting FFT spectra were imported into MATLAB 2023b as four independent matrices according to each τ value, as shown in Fig. 6.2.1C. To maintain consistency with the matrix structure applied to CPMG, the FFT matrices were integrated into two subsets: $\mathbf{FFT}_{pure}$ and $\mathbf{FFT}_{total}$.

6.2.2.1 Two-way decomposition of FFT spectra via MCR-ALS

The FFT spectra matrices ($\mathbf{FFT}_{pure}$ and $\mathbf{FFT}_{total}$) with m rows (samples) and n columns (frequency) according to their respective τ value can be described as a bilinear model based on the following equation:

$$\mathbf{D} = \mathbf{C}\mathbf{S}^T + \mathbf{E} \tag{6.2.1}$$

where $\mathbf{D}$ $(m \times n)$ is the input matrix, $\mathbf{C}$ $(m \times p)$ is the matrix of the concentration profiles of the compounds estimated; $\mathbf{S}^T$ $(p \times n)$ is the matrix of each component p related to their pure FFT spectra; and $\mathbf{E}$ $(m \times n)$ is the residual matrix with the variance of the unmodeled data.

To resolve the bilinear model described in Eq. 6.2.1, the $\mathbf{FFT}_{pure}$ and $\mathbf{FFT}_{total}$ spectra matrices were independently analyzed using the MCR-ALS GUI 2.0 Toolbox within the MATLAB R2023b environment [27]. To initialize the iterative procedure, the single value decomposition (SVD) was used to find the optimal number of components for the initial estimates of the purest $\mathbf{C}$ and $\mathbf{S}^T$. Due to the existence of rotational ambiguities in the solution, non-negativity constraints were applied to $\mathbf{C}$ and non-negativity with uni-modality to $\mathbf{S}^T$ to decrease the range of feasible solutions. The maximum number of iterations was set to 250. The resolved $\mathbf{S}^T$ profiles were normalized by the total sum criterion during the ALS optimization. The quality of the MCR-ALS models was determined by the percentage of lack of fit (%LOF) to evaluate the difference between the original $\mathbf{D}$ and the approximation made by the product $\mathbf{C}\mathbf{S}^T = \hat{\mathbf{D}}$ (Eq. 6.2.2), and the percentage of explained variance (%R^2) (Eq. 6.2.3) according to the following

expressions:

$$LOF\,(\%) = 100 \times \sqrt{\frac{\sum_{i,j} e_{ij}^2}{\sum_{i,j} d_{ij}^2}} \tag{6.2.2}$$

$$R^2(\%) = 100 \times \left(1 - \frac{\sum_{i,j} e_{ij}^2}{\sum_{i,j} d_{ij}^2}\right) \tag{6.2.3}$$

where e_{ij} are the residuals obtained in the ALS optimization and d_{ij} are the original values in $\mathbf{D}$. An augmented matrix ($\mathbf{C}_{FFT}$) was constructed by vertically concatenating the $\mathbf{C}$-profiles obtained from the FFT-modeled spectra according to the different values of τ. Two versions of the $\mathbf{C}_{FFT}$ matrix were organized depending on the sample set: $\mathbf{C}_{FFT-pure}$ and $\mathbf{C}_{FFT-total}$, as shown in Fig. 6.2.1C. $\mathbf{C}$-profiles were arranged such that those associated with shorter τ (i.e., 13,000 and 36,000 ns) appeared first, while those corresponding to longer τ (102,000 and 286,000 ns) were placed later. The latter structure preserved the spectral evolution as a function of increasing τ.

6.2.2.2 Three-way decomposition of FFT spectra via PARAFAC

The $\mathbf{FFT}_{pure}$ and $\mathbf{FFT}_{total}$ spectra matrices were stacked along a third mode to construct a three-way array (samples $\times$ frequency $\times$ τ) Each frontal slice of the cube corresponds to an individual FFT matrix in a specific τ, allowing the third mode to represent the evolution of spectral features across relaxation delays, as shown in Fig. 6.2.1C.

PARAFAC decomposes a three-way array into a set of trilinear terms by minimizing the sum of squares of the residuals e_{ijk} that contain all unexplained variations described by the following equation [26]:

$$x_{ijk} = \sum_{r=1}^{R} a_{if} b_{if} c_{kf} + e_{ijk} \tag{6.2.4}$$

where x_{ijk} can be described as an input in the three-way array of FFT spectra for the sample i, frequency j and value τ k; a_{if} are the scores proportional to the relative concentration of the R component in the i sample; b_{if} and c_{kf} are the loadings of the pure FFT spectra and the value τ for each component R.

The three-way FFT matrices were decomposed by PARAFAC (Eq. 6.2.4) using PLS_Toolbox 9.2 (Eigenvector Research Inc.) within the MATLAB R2023b environment. The PARAFAC models were built with non-negativity constraints applied to the scores (Mode 1) and non-negativity with uni-modality to the loadings of the pure FFT spectra (Mode 2). The maximum number of iterations was set to 500. To determine the optimal number of components R and evaluate the quality of the PARAFAC models, two criteria were considered: the percentage of core consistency (%CC) and the percentage of R^2 as expressed in the following equations:

$$CC\,(\%) = 100 \times \left(\frac{1 - \sum_{d=1}^{R} \sum_{e=1}^{R} \sum_{f=1}^{R} (g_{def} - t_{def})^2}{\sum_{d=1}^{R} \sum_{e=1}^{R} \sum_{f=1}^{R} t_{def}^2} \right) \tag{6.2.5}$$

$$R^2(\%) = 100 \times \left(1 - \frac{SSE}{SSX}\right) \tag{6.2.6}$$

where, in Eq. 6.2.5, g_{def} denotes the actual values of the core tensor obtained from the fitted PARAFAC model, and t_{def} corresponds to the ideal values of the core tensor expected under a perfectly trilinear structure. In line with Eq. 6.2.3, the percentage of R^2, as defined in Eq. 6.2.6, was calculated based on the sum of the squares of the residuals (SSE) and the sum of the squares of the elements of the systems (SSX). The scores obtained were combined by rows in a new matrix called $\mathbf{Scores}_{FFT}$ for supervised classification analyzes.

6.2.2.3 PCA and supervised classification methods

Independent PCA models were constructed for the $\mathbf{CPMG}_{pure}$, $\mathbf{CPMG}_{total}$, $\mathbf{C}_{FFT-pure}$ and $\mathbf{C}_{FFT-total}$ to assess the clustering potential of pure pectin samples according to their production method and the impact of incorporating binary mixtures. The resulting score distributions were compared with the biplots derived from PARAFAC models to contrast the differentiation capabilities of both previous approaches.

To evaluate the predictive performance of the different strategies (i.e, CPMG relaxation curves and FFT decomposition through MCR-ALS and PARAFAC), supervised classification methods were developed via k-NN, PLS-DA and SIMCA, as illustrated in Fig. 6.2.1D. Four classes were defined: **Class 1** (MD30 and SBPP); **Class 2** (UF and UF MD10); **Class 3** (UF IPA and UF IPA MD10), and **Class 4** encompassing the binary mixtures described in Table 6.2.1. In the case of SIMCA, only **classes 1-3** were used to construct the model, while **class 4** (three randomly selected samples of each binary mixture) was exclusively reserved for external prediction.

Due to the limited number of samples per class, a two-level cross-validation (CV) approach was implemented to prevent overfitting and evaluate the quality of the models. An external k-fold CV (k = 2) was applied. In each iteration, one part was used as a calibration set, while the second part was reserved as a prediction set. During calibration, the leave-one-out CV (LOOCV) method was applied as an interval CV. The optimal number of principal components (PCs and latent variables [LVs]) was selected based on the percentage of explained variance and the values of the root mean square error of cross-validation (RMSECV). For k-NN, the number of neighbors was set at 3 (k = 3). All models were conducted using PLS_Toolbox 9.5 via MATLAB R2023b. Discrimination capacity was evaluated on the basis of sensitivity (Sn), specificity (Sp), error rate (ER) and classification accuracy (CA). The latter figures of merit are presented in Eq. 6.2.7, 6.2.8, 6.2.9, 6.2.10, and obtained according to the following expressions:

$$\%\text{Sensitivity}\,(Sn) = \frac{TP}{(TP+FN)} \times 100 \tag{6.2.7}$$

$$\%\text{Specificity}\,(Sp) = \frac{TN}{(TN+FP)} \times 100 \tag{6.2.8}$$

$$\%\text{Error rate}\,(ER) = 1 - \left(\frac{Sn+Sp}{2}\right) \times 100 \tag{6.2.9}$$

$$\%\text{Classification accuracy}\,(CA) = \frac{TP+TN}{(TP+TN+FP+FN)} \times 100 \tag{6.2.10}$$

where *FN*, *FP*, *TP* and *TN* correspond to false negative, false positive, true positive and true negative, respectively.

6.3 Results and discussion

6.3.1 Relaxation behavior of CPMG curves and FFT spectra

The average of the maximum normalized (0-1) CPMG relaxation curves of pure and binary pectin mixtures is shown in Fig. 6.3.1A and 6.3.1B, respectively. In all cases, the relaxation curves exhibited a bi-exponential behavior, characteristic of a heterogeneous system comprising at least two proton environments with distinct transverse relaxation times (T_2) [28]. These components (T_{21} and T_{22}) reflect molecular dynamics with restricted and enhanced mobility.

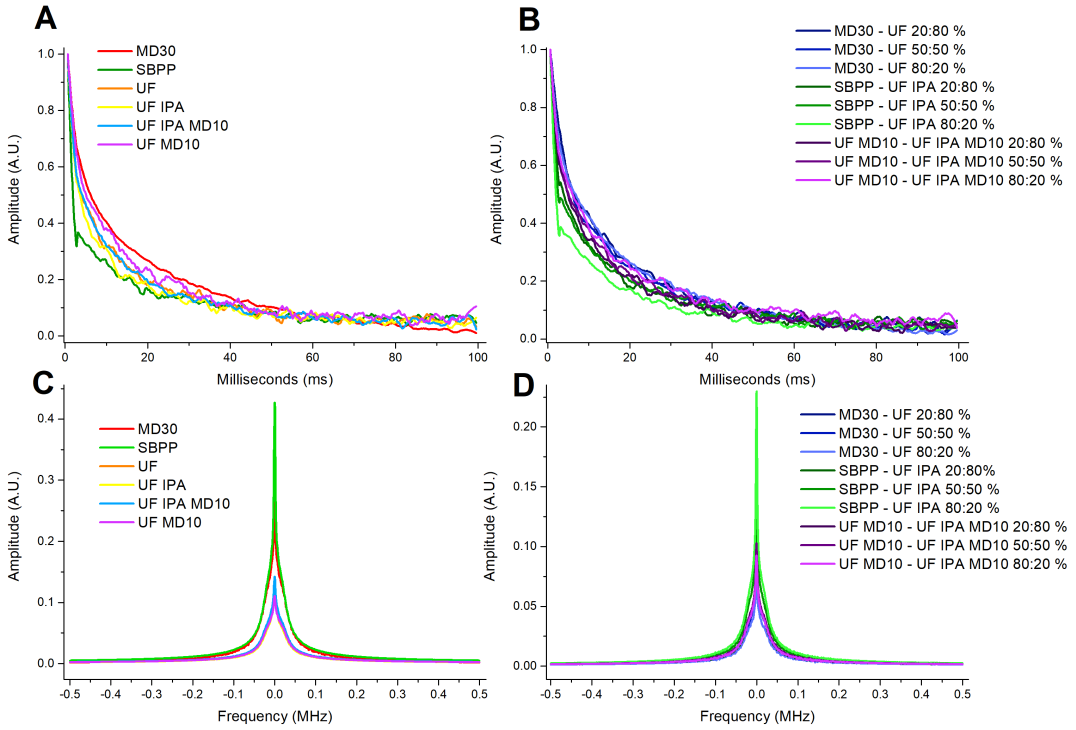

Figure 6.3.1: Maximum normalized (0–1) average CPMG relaxation curves **A, B)** and the corresponding FFT spectra **C, D)** for pure pectin and binary mixtures, respectively.

Table 6.3.1: T_2 relaxation times obtained through bi-exponential fitting of CPMG relaxation curves. Values expressed as mean ± standard deviation.

Pectin	**T_2** (ms)	
	T_{21}	**T_{22}**
MD30	2.583 ± 0.152	27.80 ± 1.42
SBPP	0.792 ± 0.054	18.03 ± 2.48
UF	1.592 ± 0.193	17.59 ± 2.33
UF IPA	1.621 ± 0.155	16.64 ± 1.63
UF IPA MD10	1.523 ± 0.245	16.16 ± 1.64
UF MD10	1.885 ± 0.235	19.51 ± 3.87

Table 6.3.1 summarizes the mean ± standard deviation of T_2 relaxation times obtained from bi-exponential fitting of the CPMG curves. The longest T_{22} component was observed in MD30 (27.80 ± 1.42 ms), suggesting a highly flexible pectin structure with extensive molecular motion and greater water interaction. In contrast, the commercial SBPP sample exhibited significantly lower T_{21} (0.792 ± 0.054 ms), suggesting a more rigid structure with reduced segmental mobility, probably influenced by differences in industrial-scale production. Among the

pectin subjected to ultra-filtration showed intermediate relaxation behavior. UF IPA MD10 and UF IPA samples presented T_{22} values of 16.16 ± 2.48 ms and 16.64 ± 1.63 ms, reflecting moderate rigidity and limited hydration. In particular, UF MD10 also showed a relatively short T_{21} of 1.885 ± 0.235 ms, compared to MD30, consistent with tighter molecular packing and reduced water mobility.

The FFT spectra shown in Fig. 6.3.1C and 6.3.1D provide a complement based on a frequency-domain perspective of the proton dynamics within each sample [29]. Although all spectra showed central peaks associated with water-associated protons, subtle differences in peak intensity and width revealed distinctions in the molecular environment of each pectin. The broader frequency distributions observed in pectin subjected to ultra-filtration are consistent with reduced molecular mobility and more restricted water environments, in agreement with their shorter T_2 values. Furthermore, binary mixtures exhibited intermediate relaxation profiles, combining the molecular signatures of their constituent pectin. Furthermore, the relaxation curves (Fig. 6.3.1B) and the FFT spectra (Fig. 6.3.1D) varied according to the mixture ratios, and UF pectin exerted a dominant influence on rigidity, particularly in samples containing a higher proportion of UF IPA or UF MD10. Although certain differences in T_2 relaxation times and spectral characteristics were observed in the frequency domain, these were insufficient to establish a clear and consistent differentiation between the samples analyzed. Therefore, the application of multivariate analysis was necessary to comprehensively interpret the spectral complexity and discover significant patterns in the set of samples.

6.3.2 Exploratory Analysis

6.3.2.1 CPMG-based relaxation

Initially, PCA was performed to the $\mathbf{CPMG}_{pure}$ and $\mathbf{CPMG}_{total}$ matrices to explore the feasibility of sample clustering using a straightforward approach based on relaxation curves. Fig. 6.3.2A presents the PCA score plot derived from the $\mathbf{CPMG}_{pure}$ matrix, where the first two PCs account for 98.52% of the total variance. A clear separation along PC1 (92.69%) was observed between MD30 with SBPP and the ultra-filtration pectin (UF, UF MD10, UF IPA, and UF IPA MD10). PC2 (5.93%) further distinguished between MD30 and SBPP, but did not contribute to the differentiation among the samples labeled UF, which clustered

closely. The latter suggests limited intrinsic variability between the pectin under ultra-filtration through relaxation curves. The observed patterns are consistent with the T_2 relaxation values summarized in Table 6.3.1 and the corresponding loading plot in Fig. 6.3.2B. The separation along PC2 is mainly attributed to the distinct decay constants of MD30 and SBPP, which exhibited markedly different long and short T_{21} components (2.583 ms and 0.792 ms, respectively).

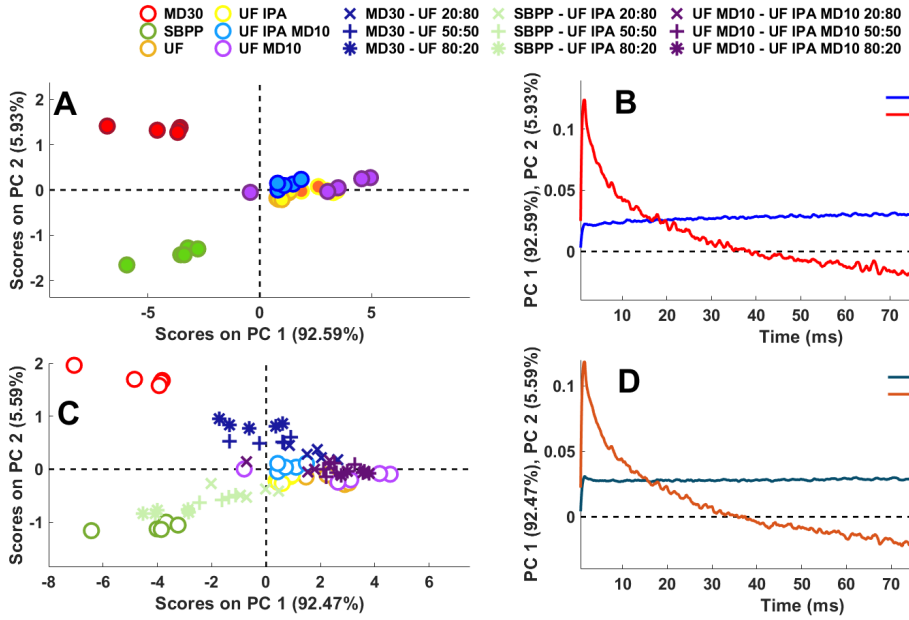

Figure 6.3.2: Comparison of PCA score and loading plots (PC1 vs PC2) for $\mathbf{CPMG}_{pure}$ and $\mathbf{CPMG}_{total}$ matrices. **A-B)** pure samples; **C-D)** pure + binary mixtures.

The inclusion of binary mixtures as additional sources of variance was evaluated to determine whether the previously identified clustering among pure pectin samples was preserved or modified due to the influence of the mixed compositions. The PCA score plot derived from the $\mathbf{CPMG}_{total}$ matrix 6.3.3C explained 98.06% of the total variance. PC1 (92. 47%) captures the main variability associated with differences in relaxation behavior between samples, and the binary mixtures exhibited minor clustering trends depending on their composition. The MD30 - UF mixtures showed a slight clustering along PC1 as a function of their ratio, suggesting additive behavior and the sensitivity of the CPMG relaxation to compositional changes. In contrast, UF MD10 - UF IPA MD10 showed a tighter grouping near the pure pectin samples, indicating less pronounced variation among its mixtures. PC2 (5.59%) captured clear differences between MD30 and SBPP; however, the UF samples were partially overlapped by the associated binary

mixtures. Moreover, the SBPP - UFA IPA system, although clustering near SBPP, further illustrates the dominant relaxation contribution of SBPP in mixtures due to specific relaxation domains (6.3.3)D.

These findings indicate that while the pronounced relaxation differences between MD30 and SBPP allowed their clear separation, the CPMG-based relaxation analysis was insufficient to distinguish the ultra-filtrated pectin samples or their binary mixtures. It is suggested that in the absence of strong relaxation contrast, the method does not effectively capture clustering trends associated with the production process.

6.3.2.2 MCR-ALS and PARAFAC-based methodology

To contrast the results obtained from CPMG, the clustering behavior was compared with those derived from the two FFT-based approaches: PCA applied to the **C**-profiles obtained through MCR-ALS, and the score biplot generated from the PARAFAC models.

The decomposition of the $\mathbf{FFT}_{pure}$ and $\mathbf{FFT}_{total}$ matrices via MCR-ALS (according to the specific value of τ) required the extraction of two components per matrix to achieve an appropriate fit. The latter resulted in a total of eight **C**-profiles. The complete set of **C**-profiles, obtained for pure pectin samples and binary mixtures, is provided in the Supplementary Fig. 6.6.1. In terms of model performance, an average LOF of 0.27% was achieved with an explained variance of 98.8%, indicating a high-quality representation of the original spectral information by the extracted profiles.

Figure 6.3.3A displays the PCA score plot based on the $\mathbf{C}_{FFT-pure}$ matrix, which represented 99.92% of the total variance. Compared to the PCA obtained by CPMG (Fig. 6.3.3A), the $\mathbf{C}_{FFT-pure}$ matrix produced a clearer separation into three distinct clusters. The primary source of variance (PC1, 99.19%) effectively differentiated the commercial SBPP sample from the remaining pectin produced at the Kayseri Sugar facility. PC2 (0.73%), although explaining a minor portion of the total variance, provided additional differentiation between SBPP, UF, UF IPA MD10, and UF MD10 versus MD30 and UF IPA. The corresponding loading plot is shown in Fig. 6.3.3B, where the contribution of each variable increases with higher values of τ, indicating a greater influence of longer relaxation components

in PC1 and, to a lesser extent, in PC2. This trend is further elucidated by the resolved $\mathbf{S}^T$ associated with each **C**-profile presented in Fig. 6.3.2C and 6.3.3D. In general, the resolved FFT spectra associated with the SBPP sample exhibited narrower and more intense peaks, indicating greater structural rigidity and lower proton mobility, in contrast to the slightly broader spectra associated with a higher molecular mobility.

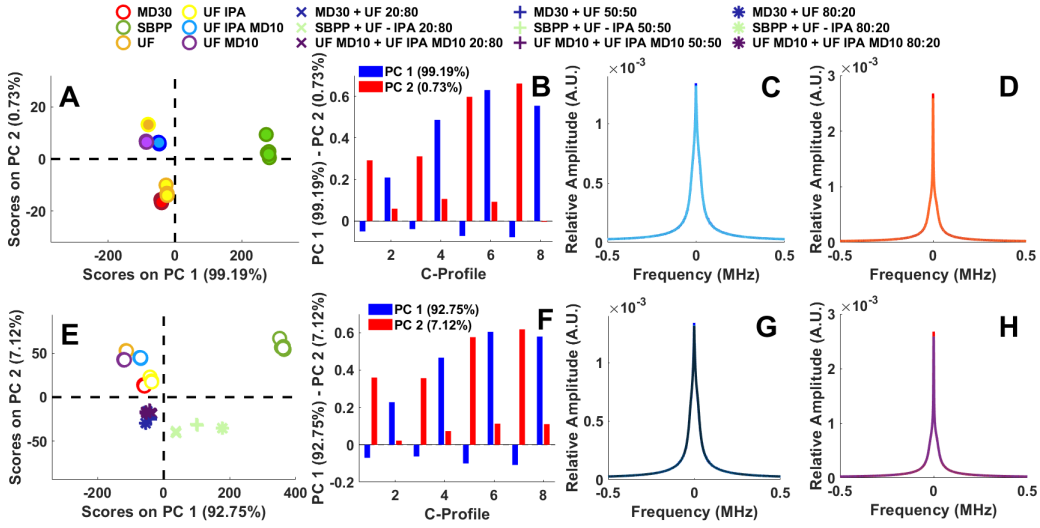

Figure 6.3.3: Comparison of PCA score (PC1 vs PC2), loading plots and resolved $\mathbf{S}^T$ profiles for the $\mathbf{FFT}_{pure}$ and $\mathbf{FFT}_{total}$ matrices. **A-B-C-D)** pure samples; **E-F-G-H)** pure + binary mixtures.

Regarding the $\mathbf{C}_{FFT}$ matrix, a higher total variance was explained (99.87%), and the separation between mixtures became more pronounced (Fig. 6.3.3)E. PC1 (92.75%) effectively distinguished samples with SBPP, while PC2 (7.12%) provided clear clustering between pure and mixed samples, especially between samples that had partially overlapped in the CPMG-based PCA. The MD30 - UF mixtures maintained their ratio-dependent spread, and the UF MD10 - UF IPA MD10 samples were clearly differentiated from pure pectin. Furthermore, the SBPP UF IPA mixtures remained grouped, consistent with the stronger relaxation characteristics of SBPP. As expected, the loading plots in Fig. 6.3.3)D indicated that longer values of τ were crucial for clustering. The resolved $\mathbf{S}^T$ supported these findings, with broader and narrower peaks suggesting rigid or mobile environments.

As part of the spectral decomposition strategy, the application of PARAFAC to three-way arrays constructed from $\mathbf{FFT}_{pure}$ and $\mathbf{FFT}_{total}$ required like MCR-ALS the extraction of two components per matrix to achieve an appropriate fit. For the model based on pure pectin samples, a CC of 97.4% and a residual error of 1.36 were obtained. In the case of the model incorporating pure pectin and binary mixtures, the CC increased to 99.7%, with a residual error of 0.476. The CC plots for both PARAFAC models are provided in the Supplementary Fig. 6.6.2 and 6.6.3, respectively.

In agreement with the trends observed with the $\mathbf{C}_{FFT-Pure}$ matrix, the PARAFAC scores plot (Fig. 6.3.4A reveals a clear separation between all pure pectin samples. To further investigate these differences, the relative concentration profiles of the two main PARAFAC components were evaluated (Fig. 6.3.4B. Our results highlight that the most substantial variations depend between SBPP and UF IPA MD10, while the remaining samples show only minor differences in their component contributions. These findings are supported by the resolved spectral components (Fig. 6.3.4C and 6.3.4D), and are consistent with the MCR-ALS results. The FFT spectra were characterized by distinct spectral features associated with heterogeneous proton environments.

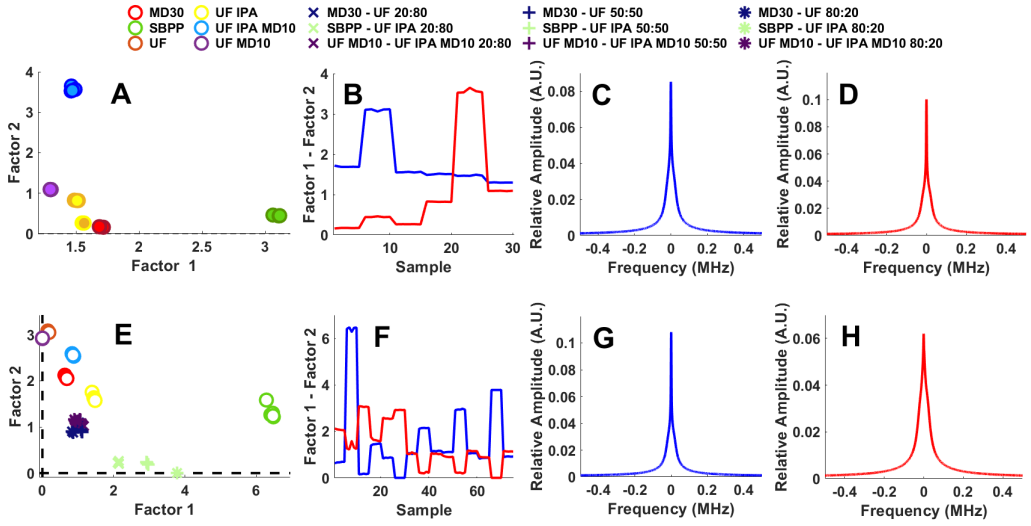

Figure 6.3.4: PARAFAC results for the $\mathbf{FFT}_{pure}$ and $\mathbf{FFT}_{total}$ matrices. **A-B-C-D)** pure samples; **E-F-G-H)** pure + binary mixtures.

The scores plot of the PARAFAC model of pure pectin samples and binary mixtures (Fig. 6.3.4G showed a similar clustering to the PCA obtained via $\mathbf{C}_{FFT-Total}$. The MD30 - UF and UF MD10 - UF IPA MD10 samples exhibited near clustering, consistent with additive behaviors linked to their mixing rations. The latter suggest the presence of intermediate relaxation properties, in agreement with their shared production method. In contrast, SBPP - UF IPA mixtures were located separately from the other samples, indicating a strong contribution of one of the resolved factors, probably associated with the distinct spectral features of SBPP. The factor profiles (Fig. 6.3.4H) illustrated the sample-specific contributions of each component, showing how PARAFAC was able to unravel overlapping spectral features that were not fully resolved in PCA. Unlike the results obtained from the PCA by CPMG, the MCR-ALS and PARAFAC strategies revealed a clear grouping between the pectin samples. Both multivariate approaches reduced intra-group variability by modeling noise independently in the residual matrix (Eq. 6.2.1 and 6.2.4). The latter capability improves sensitivity to subtle spectral differences, thus improving the differentiation of similar pectin samples (UF, UF IPA, UF IPA MD10 and UF MD10) compared to traditional CPMG-based PCA.

6.3.3 Classification of pectin samples

In the initial phase, the discriminative capability of k-NN and PLS-DA was evaluated for the classification of pure pectin samples produced by the different production methods under study. The classes were defined hierarchically according to the precipitation protocol, irrespective of maltodextrin grade: **Class 1** consisted of samples without an ultra-filtration; **Class 2** comprised pectin samples subjected solely to ultra-filtration; and **Class 3** included those processed by ultra-filtration and isopropyl alcohol. Table 6.3.2 summarizes the classification results for both models, based on their confusion matrices and merit figures (Sn, Sp, ER, and CA) for the input matrices CPMG, $\mathbf{C}_{FFT}$, and $\mathbf{Scores}_{FFT}$.

The classification models based on CPMG relaxation curves showed accuracies ranging from 70.0% to 96.7%, with PLS-DA producing the highest overall metric performance. Pectin samples from **Class 1** (MD30 and SBPP) were the most accurately classified by both models due to their distinctive relaxation profiles. Misclassifications occurred mainly between classes 2 and 3, with **Class 2** showing less performance in terms of sensitivity and specificity, particularly in the k-

NN model. These results may be attributed to the similar relaxation behavior exhibited by samples with ultra-filtration (regardless the addition of isopropyl alcohol), highlighting certain limitations of relaxation curves in distinguishing subtle composition differences related to the production method.

Table 6.3.2: Summary of confusion matrices and performance metrics for k-NN and PLS-DA applied to the different strategies.

Matrix	**Model**	**Figures of merit CV**						
		Act. Class	**Pred. Class**			**%Sn**	**%Sp**	**%ER**	**%CA**
			1	**2**	**3**				
CPMG	k-NN	**1**	10	0	0	100	100	0.00	100
		2	0	4	6	40.0	85.0	37.5	70.0
		3	0	3	7	70.0	70.0	30.0	70.0
	PLS-DA	**1**	9	0	1	90.0	100	5.00	96.7
		2	0	7	3	70.0	90.0	20.0	83.3
		3	0	2	8	80.0	80.0	20.0	80.0
$\mathbf{C}_{FFT}$	k-NN	**1**	10	0	0	100	100	0.00	100
		2	0	10	0	100	100	0.00	100
		3	0	0	10	100	100	0.00	100
	PLS-DA	**1**	10	0	0	100	100	0.00	100
		2	0	10	0	100	100	0.00	100
		3	0	0	10	100	100	0.00	100
$\mathbf{Scores}_{FFT}$	k-NN	**1**	10	0	0	100	100	0.00	100
		2	0	10	0	100	100	0.00	100
		3	0	0	10	100	100	0.00	100
	PLS-DA	**1**	10	0	0	100	95.0	2.50	96.7
		2	1	9	0	90.0	100	5.00	96.7
		3	0	0	10	100	100	0.00	100

Class 1: MD30 - SBPP; **Class 2**: UF - UF MD10;
Class 3: UF IPA - UF IPA MD10.

The use of the $\mathbf{C}_{FFT}$ and $\mathbf{Scores}_{FFT}$ matrices obtained through the MCR-ALS and PARAFAC decompositions proposed in this study led to a substantial improvement in classification performance. k-NN and PLS-DA models achieved accuracies of up to 100%, with maximum sensitivity and specificity in most classes, as shown in Table 6.3.2. The lowest classification rate observed was 96.7%, corresponding to the PLS-DA model using PARAFAC-derived scores. The enhancement in class discrimination can be attributed to the ability of both decomposition methods

Table 6.3.3: Summary of confusion matrices and performance metrics for SIMCA models applied to the different strategies.

Matrix	**PCs**	**Figures of merit CV**							
		Act. Class	**Pred. Class**				**%Sn**	**%Sp**	**%ER**	**%CA**
			1	**2**	**3**	**4**				
CPMG	2	**1**	10	0	0	0	100	100	0.00	100
	3	**2**	0	8	2	0	80.0	78.7	20.6	78.9
	3	**3**	0	2	8	0	80.0	74.5	22.8	75.4
	-	**4**	0	8	10	9	33.3	100	33.3	68.4
$\mathbf{C}_{FFT}$	1	**1**	10	0	0	0	100	100	0.00	100
	2	**2**	0	10	0	0	100	100	0.00	100
	2	**3**	0	0	10	0	100	100	0.00	100
	-	**4**	0	0	0	27	100	100	0.00	100
$\mathbf{Scores}_{FFT}$	1	**1**	10	0	0	0	100	100	0.00	100
	2	**2**	0	10	0	0	100	100	0.00	100
	2	**3**	0	0	10	0	100	95.7	2.13	96.5
	-	**4**	0	0	2	25	92.6	100	3.70	96.5

Class 1: MD30 - SBPP; **Class 2**: UF - UF MD10
Class 3: UF IPA - UF IPA MD10; **Class 4**: Binary mixtures

to extract meaningful spectral variance profiles that amplify subtle differences associated with the molecular mobility properties of the samples. The inclusion of multiple τ values allowed the capture of variation sources in closely related samples (i.e., classes 2 and 3), due to improved resolution of proton dipolar interactions in the FFT spectra.

Table 6.3.3 shows the classification results obtained using SIMCA across the different input matrices. Since SIMCA models each class independently, the inclusion of binary mixtures could compromise the internal homogeneity required for reliable class modeling. Therefore, binary mixtures (**Class 4**) were excluded from the model calibration and instead used as an external prediction set to assess the specificity and robustness of class discrimination.

For the CPMG relaxation curves, the classification results for classes 1, 2, and 3 were consistent with those obtained using k-NN and PLS-DA. However, the **Class 4** classification was markedly limited, with a precision of 68.4% and a sensitivity of 33.3%, highlighting the limited capacity of the model to identify mixed samples not included in the calibration. In contrast, the use of the $\mathbf{C}_{FFT}$ matrix resulted in perfect classification (100%) in all classes. Likewise, PARAFAC-derived scores

(**Scores**$_{FFT}$) yielded 100% accuracy for the modeled classes and 96.5% accuracy with 92.6% sensitivity for **Class 4**. Furthermore, the absence of assignment of samples belonging to **Class 4** to any of the modeled classes was interpreted as a favorable result, since it indicated that the SIMCA models via $\mathbf{C}_{FFT}$ and **Scores**$_{FFT}$ retained their discriminative ability without erroneously associating binary mixtures with a specific production method.

Overall, our results confirm that, although models based on CPMG relaxation curves can provide a preliminary approximation to class differentiation, they have limitations in resolving subtle differences between samples with similar relaxation profiles. In contrast, models built from MCR-ALS and PARAFAC derived matrices showed significantly better performance, evidencing greater discriminative power. These approaches not only allowed an accurate classification of pure pectin samples obtained without ultra-filtration (MD30 and SBPP) versus those subjected to ultra-filtration (UF-label), but were also able to correctly identify various binary mixtures, even under conditions where the relaxation signals were partially masked by predominant components, as in the case of SBPP-derived samples.

6.4 Conclusions

This study presents a novel methodology that combines TD-NMR relaxometry with second-order multivariate decomposition techniques, such as MCR-ALS and PARAFAC, in order to improve the differentiation of pectin samples produced by different methods. Although conventional analysis of CPMG relaxation curves allowed the identification of general trends in proton mobility and rigidity, their ability to discriminate between samples with subtle structural differences (i.e., those subjected to ultra-filtration or binary mixtures) proved to be limited.

The use of FFT spectra generated at different τ values as input for MCR-ALS and PARAFAC enabled the decomposition of relaxation signatures into chemically meaningful spectral components. This strategy effectively captured variations in molecular mobility and hydration states, resulting in clear separation between pure pectin samples and binary mixtures. The classification models built from $\mathbf{C}_{FFT}$ and **Scores**$_{FFT}$ matrices showed outstanding performance, with accuracies ranging from 96.5% to 100% for the k-NN, PLS-DA and SIMCA models, clearly outperforming those based solely on CPMG curves (68.4%– 96.7%).

Overall, the integration of MCR-ALS and PARAFAC with TD-NMR relaxometry data constitutes a robust, selective, and non-destructive analytical strategy for the discrimination of structurally similar pectins. This approach not only enhances the capabilities of traditional relaxometry, but also represents a valuable tool for addressing analytical challenges in complex food matrices, highlighting the potential of advanced multivariate decomposition in highly heterogeneous systems.

Conflict of interest

The authors declare that they do not have a conflict of interest.

Acknowledgments

This research was funded by the European Union's Horizon 2020 Research and Innovation program-MSCA RISE under grant agreement #101008228 and by ANID Chile: Beca Doctorado Nacional [21240532] and FONDECYT [1221287]. The authors express their gratitude to FENARE (Federación Nacional de Remolacheros) for their continued support and supply of the necessary samples.

6.5 References

[1] Potential role of cell wall pectin polysaccharides, water state, and cellular structure on twice "increase–decrease" texture changes during kohlrabi pickling process. *Food Research International*, 173(P1):113308, 2023.

[2] B. Westereng, T. E. Michaelsen, A. B. Samuelsen, and S. H. Knutsen. Effects of extraction conditions on the chemical structure and biological activity of white cabbage pectin. *Carbohydrate Polymers*, 72(1):32–42, 2008.

[3] Xin Huang, Dong Li, and Li-jun Wang. Characterization of pectin extracted from sugar beet pulp under different drying conditions. *Journal of Food Engineering*, 211:1–6, 2017.

[4] Nong xue QIU, Yu xia TIAN, Shu tao QIAO, and Hong DENG. Apple Pectin Behavior Separated by Ultrafiltration. *Agricultural Sciences in China*, 8(10):1193–1202, 2009.

[5] Hyun-wook Kim, Yong Jae Lee, and Yuan H Brad Kim. Effects of membrane-filtered soy hull pectin and pre-emulsified fiber / oil on chemical and technological properties of low fat and low salt meat emulsions. *Journal of Food Science and Technology*, 53(June):2580–2588, 2016.

[6] Florina Dranca and Silvia Mironeasa. Green Extraction of Pectin from Sugar Beet Flakes and Its Application in Hydrogels and Cryogels. *Gels*, 10(4), 2024.

[7] Berlowska Joanna, Binczarski Michal, Dziugan Piotr, Wilkowska Agnieszka, Kregiel Dorota, and Witonska Izabela. *Sugar Beet Pulp as a Source of Valuable Biotechnological Products*. Elsevier Inc., 2018.

[8] M. Hutnan, M. Drtil, and L. Mrafkova. Anaerobic biodegradation of sugar beet pulp. *Biodegradation*, 11(4):203–211, 2000.

[9] Analese Roman-Benn, Carolina A. Contador, Man Wah Li, Hon Ming Lam, Kong Ah-Hen, Pilar E. Ulloa, and María Cristina Ravanal. Pectin: An overview of sources, extraction and applications in food products, biomedical, pharmaceutical and environmental issues. *Food Chemistry Advances*, 2(September 2022):100192, 2023.

[10] Sen Ma, Shu Juan Yu, Xue Ling Zheng, Xiao Xi Wang, Qing Dan Bao, and

Xiao Ming Guo. Extraction, characterization and spontaneous emulsifying properties of pectin from sugar beet pulp. *Carbohydrate Polymers*, 98(1):750–753, 2013.

[11] Vinay Chandel, Deblina Biswas, Swarup Roy, Devina Vaidya, Anil Verma, and Anil Gupta. Current Advancements in Pectin: Extraction, Properties and Multifunctional Applications. *Foods*, 11(17):1–30, 2022.

[12] Zarifeh Raji, Faramarz Khodaiyan, Karamatollah Rezaei, Hossein Kiani, and Seyed Saeid Hosseini. Extraction optimization and physicochemical properties of pectin from melon peel. *International Journal of Biological Macromolecules*, 98:709–716, 2017.

[13] Ileana Menegazzo, Stefano Mammi, Paolo Sgarbossa, Alessandra Bartolozzi, Mirto Mozzon, Roberta Bertani, Michele Forzan, Raji Sundararajan, and Elisabetta Sieni. Time Domain Nuclear Magnetic Resonance (TD-NMR) to evaluate the effect of potato cell membrane electroporation. *Innovative Food Science and Emerging Technologies*, 65(July):102456, 2020.

[14] Fenfen Tang, Morgan Vasas, Emmanuel Hatzakis, and Apostolos Spyros. *Magnetic resonance applications in food analysis*, volume 98. Elsevier Ltd., 1 edition, 2019.

[15] J. van Duynhoven, A. Voda, M. Witek, and H. Van As. *Time-domain NMR applied to food products*, volume 69. Elsevier Ltd., 1 edition, 2010.

[16] Poliana M. Santos, Edenir R. Pereira-Filho, and Luiz A. Colnago. Detection and quantification of milk adulteration using time domain nuclear magnetic resonance (TD-NMR). *Microchemical Journal*, 124:15–19, 2016.

[17] Poliana M. Santos, Cátia C. Corrêa, Lucimara A. Forato, Rymer R. Tullio, Geraldo M. Cruz, and Luiz A. Colnago. A fast and non-destructive method to discriminate beef samples using TD-NMR. *Food Control*, 38(1):204–208, 2014.

[18] Malgorzata Nowacka, Luca Laghi, Katarzyna Rybak, M. Dalla Rosa, Dorota Witrowa-Rajchert, and Urszula Tylewicz. Water state and sugars in cranberry fruits subjected to combined treatments: Cutting, blanching and sonication. *Food Chemistry*, 299(January):125122, 2019.

[19] Yangyi Chen, William MacNaughtan, Paul Jones, Qian Yang, and Tim Foster. The state of water and fat during the maturation of Cheddar cheese. *Food Chemistry*, 303(August 2019):125390, 2020.

[20] R. Kovrlija, E. Goubin, and C. Rondeau-Mouro. TD-NMR studies of starches from different botanical origins: Hydrothermal and storage effects. *Food Chemistry*, 308(September 2019), 2020.

[21] R. Voccio, C. Malegori, P. Oliveri, M. Arimondi, G. Luciano, and M. Cettolin. Prediction of cross-linking properties in rubber-based materials by means of TD-NMR and chemometrics: data preprocessing, model building and chemical interpretation. *Analytica Chimica Acta*, 1361(April):344155, 2025.

[22] Karla R. Borba, Fernanda C.A. Oldoni, Tatiana Monaretto, Luiz A. Colnago, and Marcos D. Ferreira. Selection of industrial tomatoes using TD-NMR data and computational classification methods. *Microchemical Journal*, 164(December 2020):106048, 2021.

[23] Pablo T. Coimbra, Celso F. Bathazar, Jonas T. Guimarães, Nathalia M. Coutinho, Tatiana C. Pimentel, Roberto P.C. Neto, Erick A. Esmerino, Mônica Q. Freitas, Marcia C. Silva, Maria I.B. Tavares, and Adriano G. Cruz. Detection of formaldehyde in raw milk by time domain nuclear magnetic resonance and chemometrics. *Food Control*, 110(October 2019):107006, 2020.

[24] Poliana M. Santos, Renata S. Amais, Luiz A. Colnago, Åsmund Rinnan, and Marcos R. Monteiro. Time domain-NMR combined with chemometrics analysis: An alternative tool for monitoring diesel fuel quality. *Energy and Fuels*, 29(4):2299–2303, 2015.

[25] Anna De Juan, Joaquim Jaumot, and Romà Tauler. Multivariate Curve Resolution (MCR). Solving the mixture analysis problem. *Analytical Methods*, 6(14):4964–4976, 2014.

[26] Rasmus Bro. PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, 38(2):149–171, 1997.

[27] Joaquim Jaumot, Anna de Juan, and Romà Tauler. MCR-ALS GUI 2.0: New features and applications. *Chemometrics and Intelligent Laboratory Systems*, 140:1–12, 2015.

[28] Baris Ozel and Mecit H. Oztop. A quick look to the use of time domain nuclear magnetic resonance relaxometry and magnetic resonance imaging for food quality applications. *Current Opinion in Food Science*, 41:122–129, 2021.

[29] Leonid Grunin, Maria Ivanova, Veronika Schiraya, and Tatiana Grunina. Time-Domain NMR Techniques in Cellulose Structure Analysis. *Applied Magnetic Resonance*, 54(10):929–955, 2023.

6.6 Supplementary Material

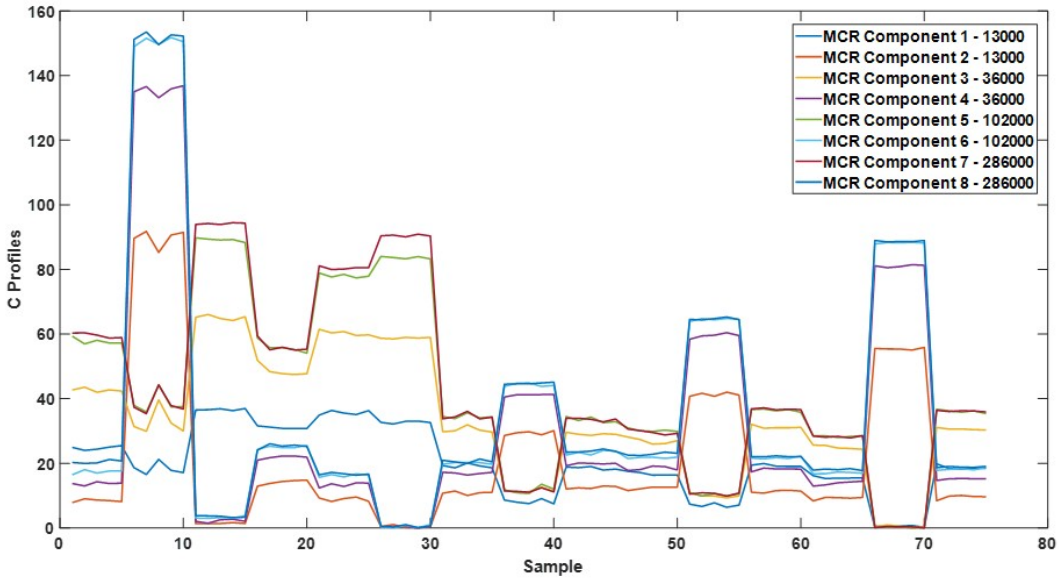

Figure 6.6.1: Sample-wise concentration profiles (**C**-profiles) obtained from MCR-ALS decomposition of FFT-spectra at different τ values, illustrating the contribution of eight resolved components across the analyzed pectin samples.

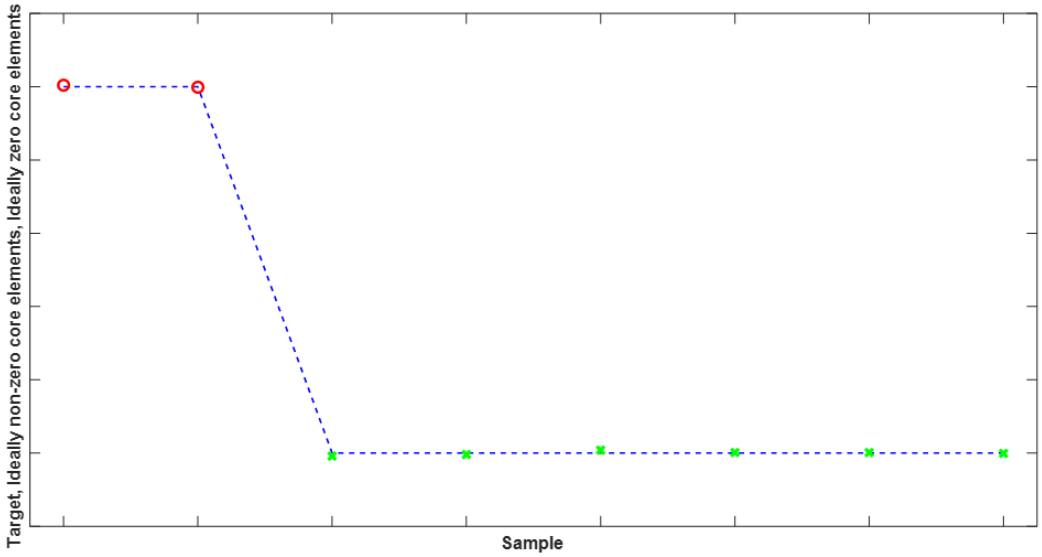

Figure 6.6.2: Core consistency diagnostic for PARAFAC modeling applied to pure pectin samples, indicating a reliable two-component solution.

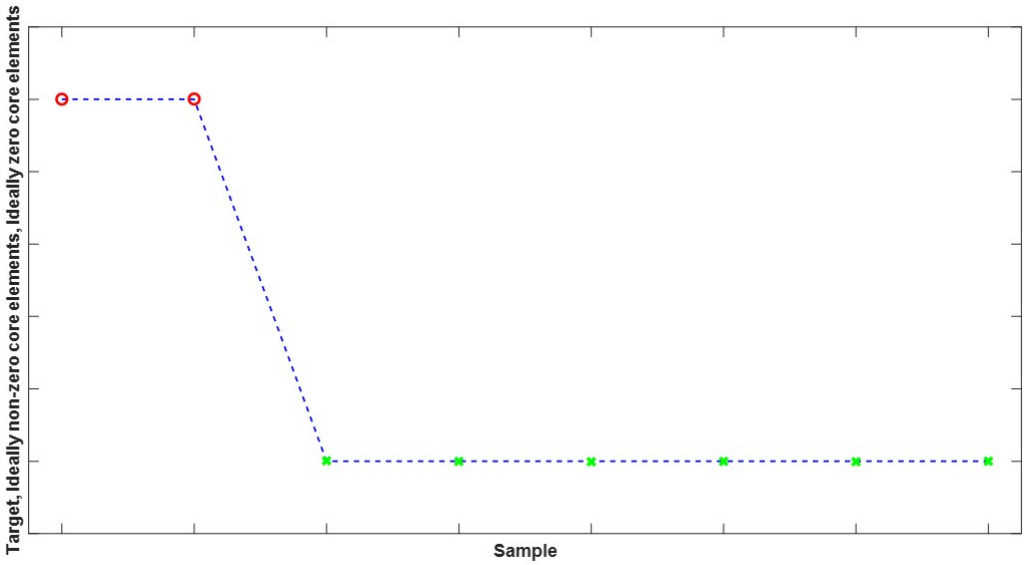

Figure 6.6.3: Core consistency diagnostic for PARAFAC modeling applied to pure pectin samples + binary mixtures, indicating a reliable two-component solution.

Chapter 7

Conclusions

This Ph.D. thesis established a selective analytical strategy based on ^{1}H NMR spectroscopy and multivariate analysis, aimed at product authentication and traceability in the sugar industry. Through the analysis of hydrophilic extracts obtained from sugar beet roots and commercial brown sugar, it was possible to characterize and identify relevant compounds associated with botanical origin, cultivation conditions, and production process.

The Interval Resonance Analysis (InRA) graphical user interface was developed to optimize and automate critical steps in the spectral workflow, including signal detection, relevant interval selection, and MCR-ALS decomposition to obtained resolved spectral and concentration profiles. This tool significantly improved the efficiency and reproducibility of the analytical process by reducing the dependence on manual decision-making and facilitating implementation by users with less experience in multivariate analysis applied to ^{1}H NMR spectroscopy.

Spectral decomposition using MCR-ALS applied to ^{1}H NMR spectra preserved key structural information and overcome the limitations inherent to the binning method. This enabled unsupervised (PCA) and supervised (PLS-DA) models to efficiently differentiate between brown sugar samples according to their botanical origin (cane, beet, and coconut), as well as between sugar beet roots from different cultivation fields.

Discriminant spectral variables were identified by VIP analysis and univariate statistical tests (ANOVA with *post-hoc* test), which were assigned to specific chemical compounds. These compounds were interpreted on the basis of their potential as quality and traceability markers, considering their relationship with agricultural practices, environmental conditions, and refining or adulteration processes.

Additionally, TD-NMR analysis was incorporated, allowing evaluation of properties related to molecular mobility and proton relaxation to differentiate pectin samples according to their production method. This information was processed using second-order multivariate models, such as MCR-ALS and PARAFAC, which demonstrated high discriminatory capacity when evaluated with classification techniques such as PLS-DA, SIMCA and KNN.

Overall, this research demonstrates that the integration of ^{1}H NMR spectroscopy with multivariate analysis provides a robust, reproducible, and highly selective alternative analytical strategy to address the challenges of authentication, quality control, and traceability in agri-food products. Furthermore, the versatility and adaptability of the proposed methodology allow its application beyond the sugar sector, potentially extending to various complex matrices, i.e., other foods, biological tissues, and materials of analytical interest, opening up new possibilities for its implementation in quality control and research laboratories across different sectors.

Appendix A

Productivity

A1 Scientific articles participation

A1.1 Ph.D. thesis

(Under preparation). Fuentes C, Ilhan E, Ivanova M, Göksu A, Öztop M, Grunin L, Castillo R. Multivariate Analysis for a Comparative Study of Pectin Production Methods using TD-NMR with Fast Fourier Transform (FFT).

May 2025 (Published). Fuentes C, Montoya D, Öztop M, Rojas-Rioseco M, Bravo M, González F, Castillo R. Interval Resonance Analysis (InRA): A versatile tool for automated untargeted ^{1}H NMR fingerprinting – A case study in sugar beet field authentication. *Analytica Chimica Acta*. https://doi.org/10.1016/j.aca.2025.344175.

July 2023 (Published). Fuentes C, Öztop M, Rojas-Rioseco M, Bravo M, Göksu A, Manley M, Castillo R. Application of segmented analysis via multivariate curve resolution with alternating least squares to ^{1}H-nuclear magnetic resonance spectroscopy to identify different sugar sources. *Food Chemistry*. https://doi.org/10.1016/j.foodchem.2023.136817.

A1.2 Collaborations

August 2025 (Published). Bravo M, Sanfuentes E, Amigo José, Hasbún R, Fuentes C, Navarro A, Sanhueza P, Castillo R. Early detection of *Fusarium circinatum* in Pinus radiata cutting using VIS-NIR hyperspectral imaging and multivariate

analysis. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*. https://doi.org/10.1016/j.saa.2025.126778.

May 2025 (Published). Rojas-Rioseco M, Öztop M, Fuentes C, Bravo M, Smajlovic I, Smajlovic M, Kolodziejski K, Kruk D, Muñoz V. Authentication of Molasses' Botanical Origin Using Infrared Spectroscopy: Accuracy and Greenness Evaluation of Spectral Techniques. *Current Research in Food Science*. https://doi.org/10.1016/j.crfs.2025.101096.

February 2025 (Under review). Ilhan E, Ivanova M, Fuentes C, Kormaz H, Göksu A, Öztop M, Grunin L, Castillo R. Impact of Extraction and Purification Methods on the Structural and Dynamic Properties of Pectin: A Time-Domain NMR and FTIR Study. *Food Analytical Methods*.

February 2025 (Under review). Shah F, Rojas-Rioseco M, Fuentes C, Castillo R, Montoya D, Kruk D. Insight into molecular properties of molasses by means of ^{1}H NMR relaxation. *Food Biophysics*.

October 2024 (Published). Bravo M, Sanfuentes E, Hasbún R, Smith M, Sandoval V, Fuentes C, Rojas-Rioseco M, Navarro A, Ulloa-Fuentes J, Castillo R. Light-driven incubation of *Fusarium* species and near-infrared spectroscopy for an early *in vitro* identification of *Fusarium circinatum*. *Microchemical Journal*. https://doi.org/10.1016/j.microc.2024.111168.

A2 Conferences participation

A2.1 International

September 2024. XIX Chemometrics in Analytical Chemistry. Santa Fe, Argentina. **Oral presentation**: Non-targeted ^{1}H NMR fingerprinting via Interval Resonance Analysis (InRA): Multivariate and data fusion approach for sugar beet crop field discrimination. Fuentes C, Montoya D, Öztop M, Rojas-Rioseco M, Bravo M, González F, Castillo R.

September 2024. XIX Chemometrics in Analytical Chemistry. Santa Fe, Argentina. **Poster**: Interval Resonance Analysis (InRA): An Alternative Approach for Automated Preprocessing of ^{1}H NMR Data in Multivariate Analysis. Fuentes C, Montoya D, Öztop M, Rojas-Rioseco M, Bravo M, Castillo R.

August 2022. XVIII Chemometrics in Analytical Chemistry. Rome, Italy. **Poster**: Application of a segmented analysis by MCR-ALS on ^{1}H NMR spectroscopy for the identification of adulterations in brown sugars. Fuentes C, Öztop M, Rojas-Rioseco M, Bravo M, Göksu A, Manley M, Castillo R.

A2.2 National

January 2025. XVI Encuentro de Química Analítica y Ambiental. Concepción, Chile. **Oral presentation**: Desarrollo de un flujo de trabajo analítico automatizado para la identificación selectiva de características espectrales en matrices complejas mediante espectroscopía de ^{1}H NMR no dirigida. Fuentes C, Montoya D, Öztop M, Rojas-Rioseco M, Bravo M, González F, Castillo R.

January 2025. XVI Encuentro de Química Analítica y Ambiental. Concepción, Chile. **Poster**: Multivariate Strategy Based on Time-Domain Nuclear Magnetic Resonance (TD-NMR) with Fast Fourier Transform (FFT) for Comparative Analysis of Pectin Production Methods from Sugar Beet. Fuentes C, Ilhan E, Ivanova M, Kormaz H, Göksu A, Öztop M, Grunin L, Castillo R.

January 2024. XXXIV Jornadas Chilenas de Química. Puerto Varas, Chile. **Poster**: Interval Resonance Analysis (InRA): A Versatile Approach for Automated Preprocessing in Non-Targeted ^{1}H NMR Fingerprinting for Multivariate Analysis. Fuentes C, Montoya D, Öztop M, Rojas-Rioseco M, Bravo M, González F, Castillo R.

April 2023. EXPOPOSTER. Concepción, Chile. Application of segmented analysis via multivariate curve resolution with alternating least squares to ^{1}H-nuclear magnetic resonance spectroscopy to identify different sugar sources. Fuentes C, Öztop M, Rojas-Rioseco M, Bravo M, Göksu A, Manley M, Castillo R.

October 2022. XV Encuentro de Química Analítica y Ambiental. Iquique, Chile. **Poster**: Application of a segmented analysis by MCR-ALS on ^{1}H NMR spectroscopy for the identification of adulterations in brown sugars. Fuentes C, Öztop M, Rojas-Rioseco M, Bravo M, Göksu A, Manley M, Castillo R.

A3 Research projects participation

Collaborative perssonel at the ANID FOVI 220172 "Green & Sustainable Analytical Technologies for Food Science: Biosensors and Time Domain Nuclear Magnetic Resonance (TD-NMR)"

Technical personnel at the ANID FONDECYT 1221387 "Selective early detection and tracing of the pathogenic fungi Fusarium circinatum in Pinus radiata seedlings based on fast. sensitive and environmentally-friendly spectroscopic and hyperspectral imaging techniques, using soft modelling chemometrics models"

Early Stage Researcher (ESR) at the #101008228 Horizon 2020 - MSCA-RISE (Marie Sklodowska-Curie Actions - Research an Innovation Staff Exchange) "Alternative Quality and Authenticity Methods for Sugar and Confectionary Industry" (SuChAQuality).

A4 Secondments

July 2024 - September 2024. Internship at the Department of Physics and Biophysics, University of Warmia and Mazury, Poland.

November 2023 - December 2023. Internship at SG Isotech, Belgrade, Serbia.

October 2023 - November 2023. Internship at the Department of Physics, Nottingham Trent University, Nottingham, United Kingdom.

November 2022 - February 2023. Internship at Resonance Systems GmbH, Kirchheim unter Teck, Germany.

October 2022 - November 2022. Internship at the Department of Food Engineering, Middle East Technical University, Ankara, Türkiye.

January 2022 - February 2022. Internship at the Department of Food Engineering, Middle East Technical University, Ankara, Türkiye.

A5 Obtained resources

February 2025. Complementary benefit "Extensión de Beca para Redacción de Tésis Doctoral". Agencia Nacional de Investigación y Desarrollo (ANID), Chile.

July 2024. Complementary benefit “Gastos Operacionales del Proyecto de Tésis Doctoral”. Agencia Nacional de Investigación y Desarrollo (ANID), Chile.

March 2024. Scholarship “Beca Doctorado Nacional” Folio [21240532]. Agencia Nacional de Investigación y Desarrollo (ANID), Chile.

August 2022. Benefit “Fondo Apoyo de Asistencia a Eventos en el Extranjero UCO 1866”. Universidad de Concepción y Ministerio de Educación, Chile.

Appendix B

Additional results

B1 NMR spectroscopy

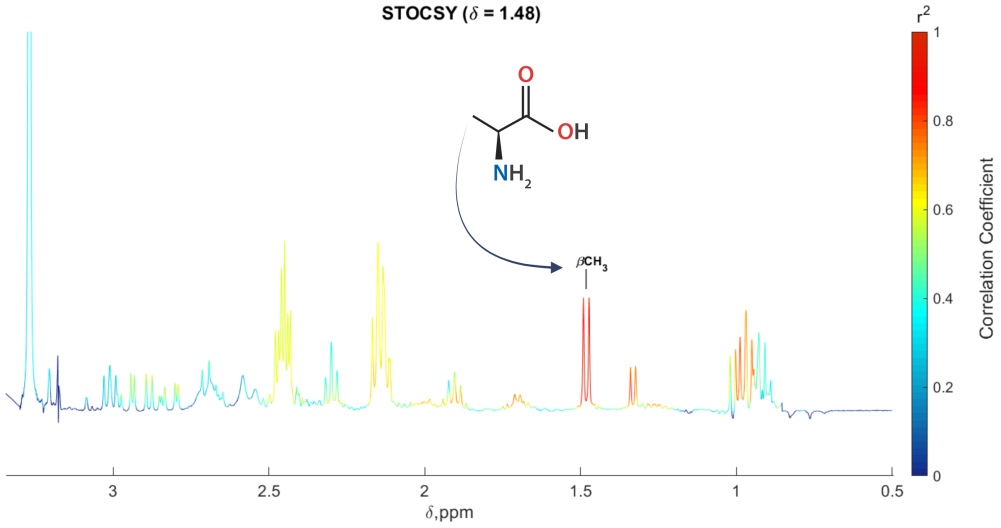

Figure B1.1: One-dimensional Statistical Total Correlation Spectroscopy (STOCSY) analysis performed for the selected "driver peak" (reference signal) at δ 1.48 ppm (Alanine). The resulting correlation profile across the spectrum is color-coded and projected on the spectrum in which the signal exhibits its maximum intensity.

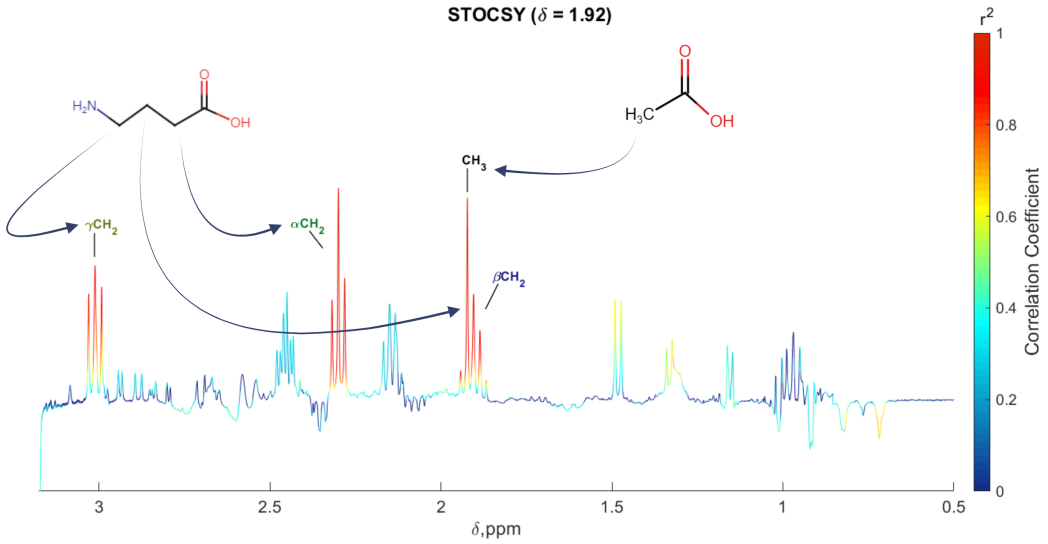

Figure B1.2: One-dimensional STOCSY analysis performed for the selected "driver peak" at δ 1.92 ppm (γ-aminobutyrate + acetate).

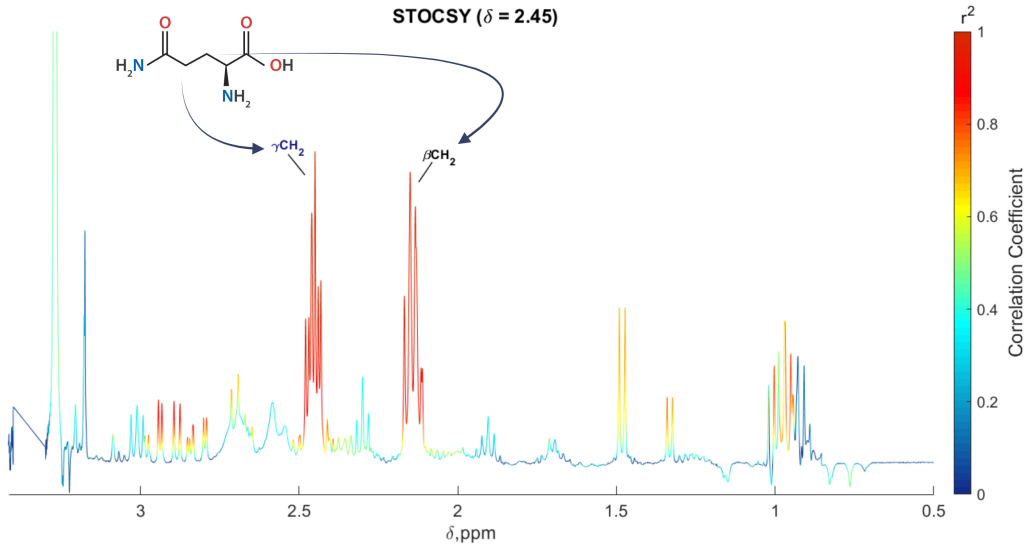

Figure B1.3: One-dimensional STOCSY analysis performed for the selected "driver peak" at δ 2.45 ppm (Glutamine).

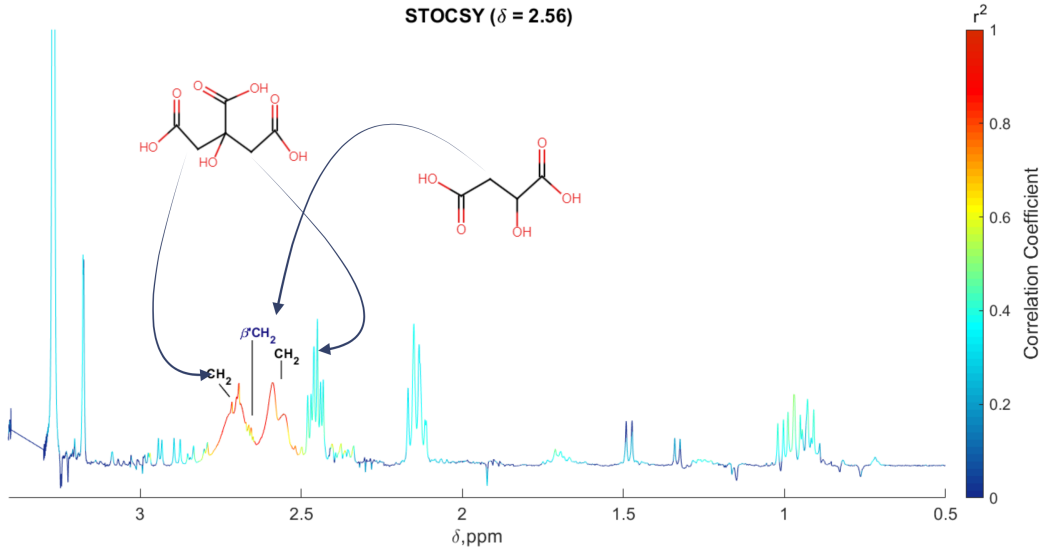

Figure B1.4: One-dimensional STOCSY analysis performed for the selected "driver peak" at δ 2.56 ppm (Citrate + Malate).

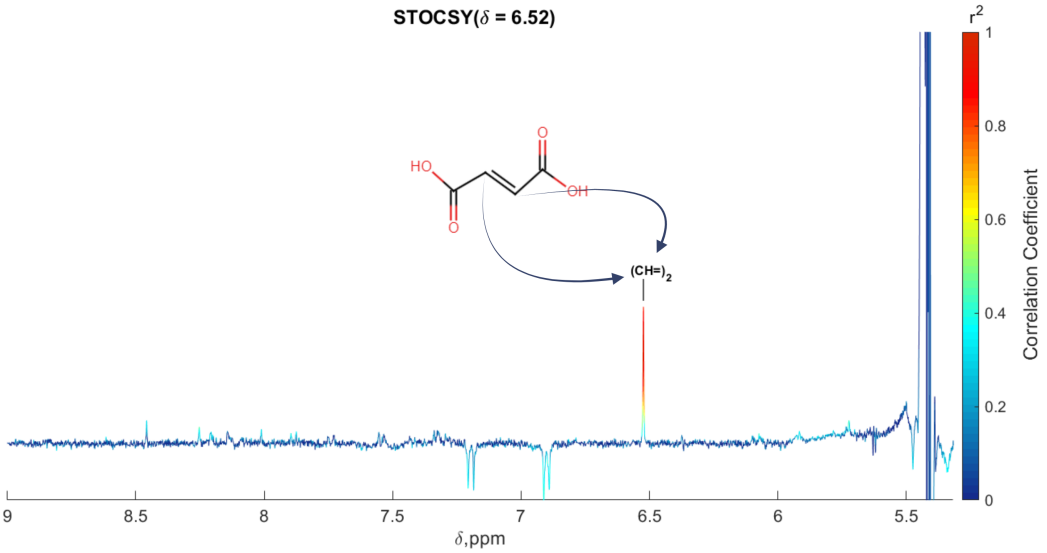

Figure B1.5: One-dimensional STOCSY analysis performed for the selected "driver peak" at δ 6.52 ppm (Fumarate).

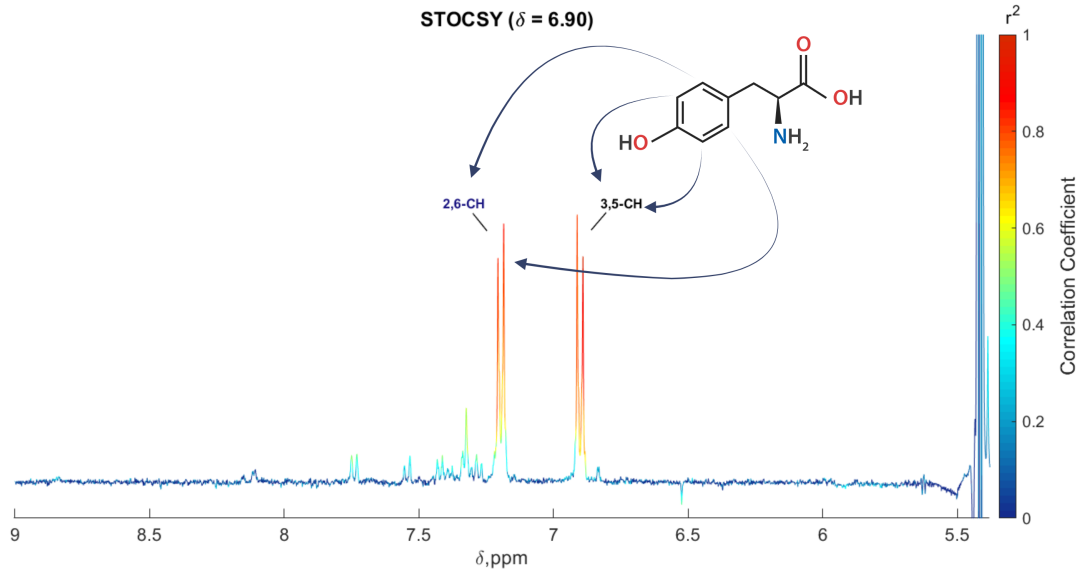

Figure B1.6: One-dimensional STOCSY analysis performed for the selected "driver peak" at δ 2.45 ppm (Tyrosine).

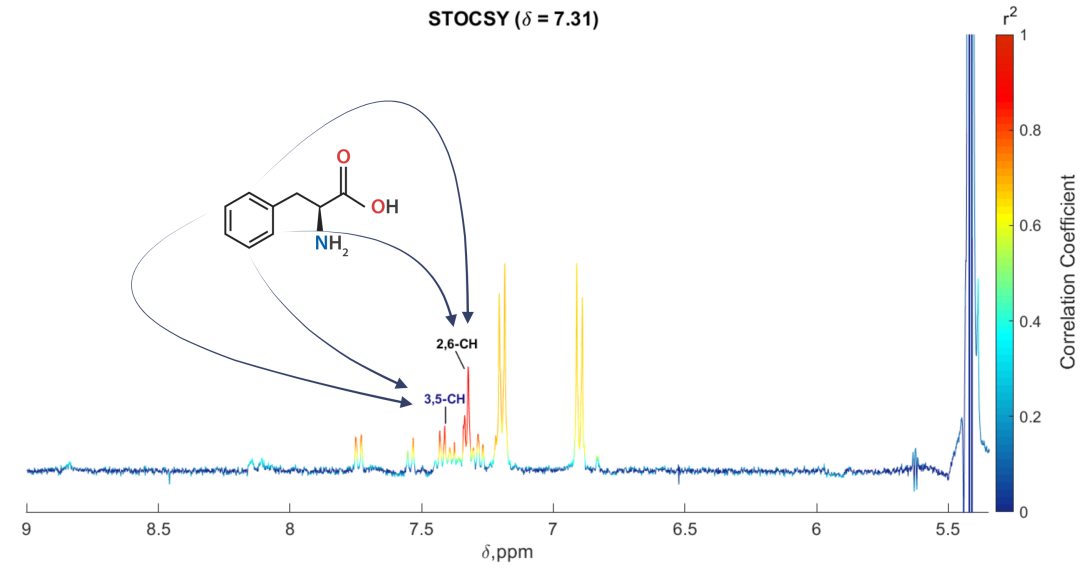

Figure B1.7: One-dimensional STOCSY analysis performed for the selected "driver peak" at δ 2.45 ppm (Phenylalanine).

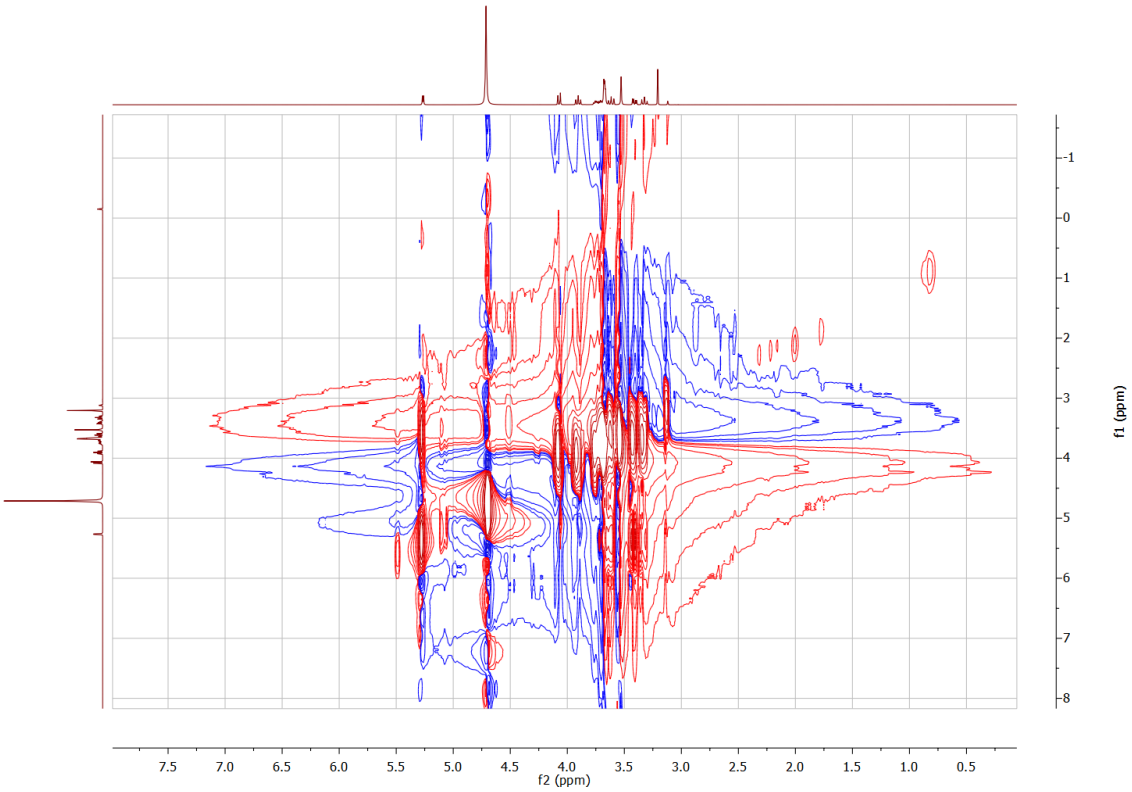

Figure B1.8: ^{1}H - ^{1}H TOCSY spectrum (D_2O with K_2HPO_4/KH_2PO_4 180 mM at pH 7.0, 400.13 MHz) of a representative hydrophilic sugar beet extract.

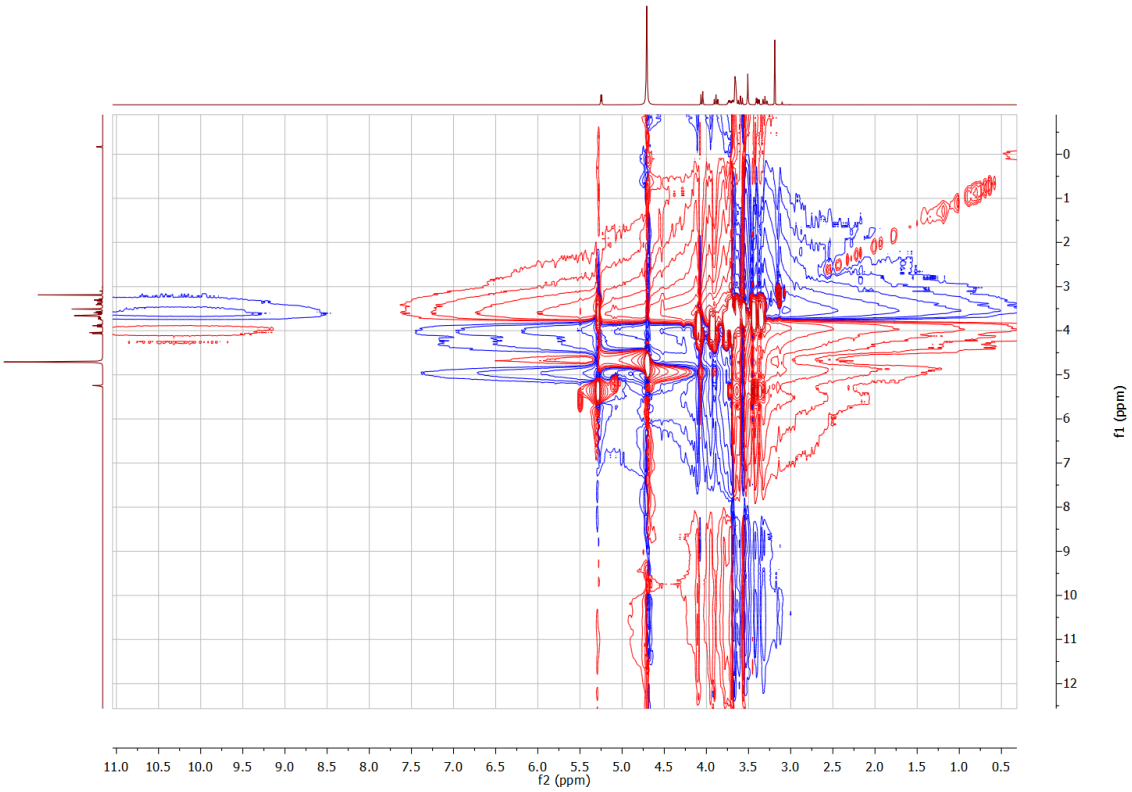

Figure B1.9: ^{1}H - ^{1}H TOCSY spectrum (D_2O with K_2HPO_4/KH_2PO_4 100 mM at pH 7.4, 400.13 MHz) of a representative brown commercial sugar sample.

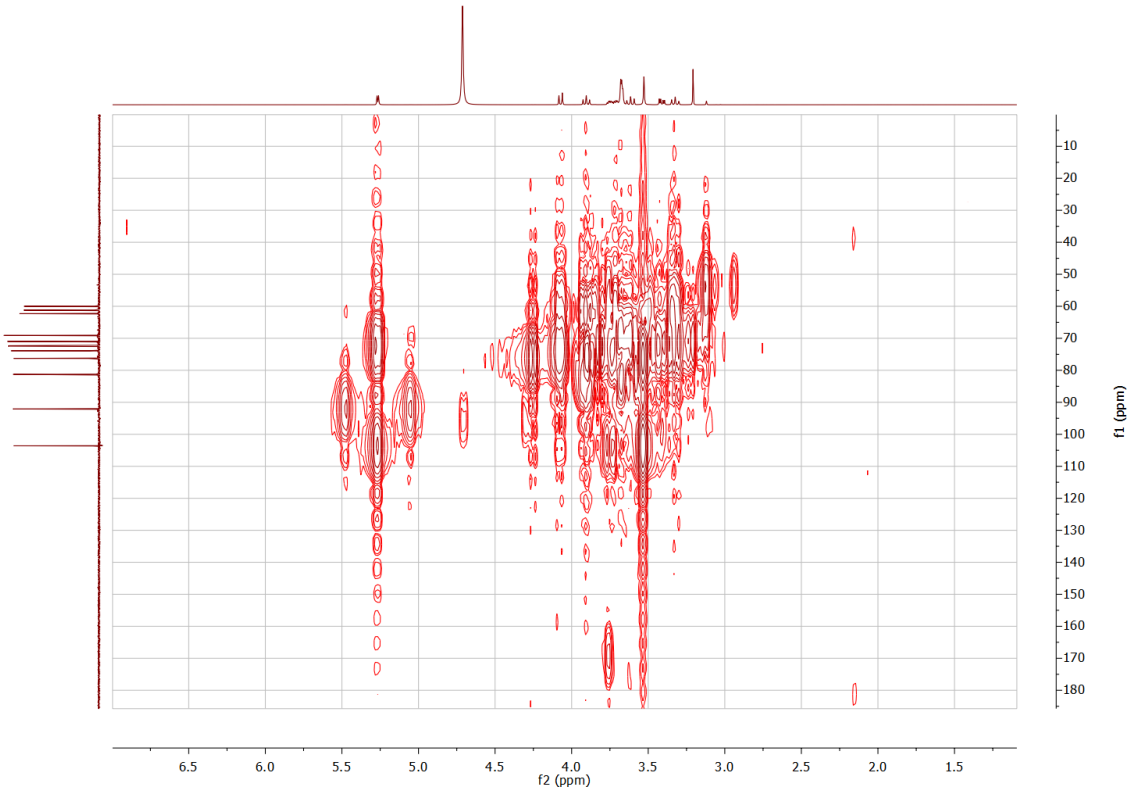

Figure B1.10: ^{1}H - ^{13}C HMBC spectrum (D_2O with K_2HPO_4/KH_2PO_4 180 mM at pH 7.0, 400.13 MHz) of a representative hydrophilic sugar beet extract.

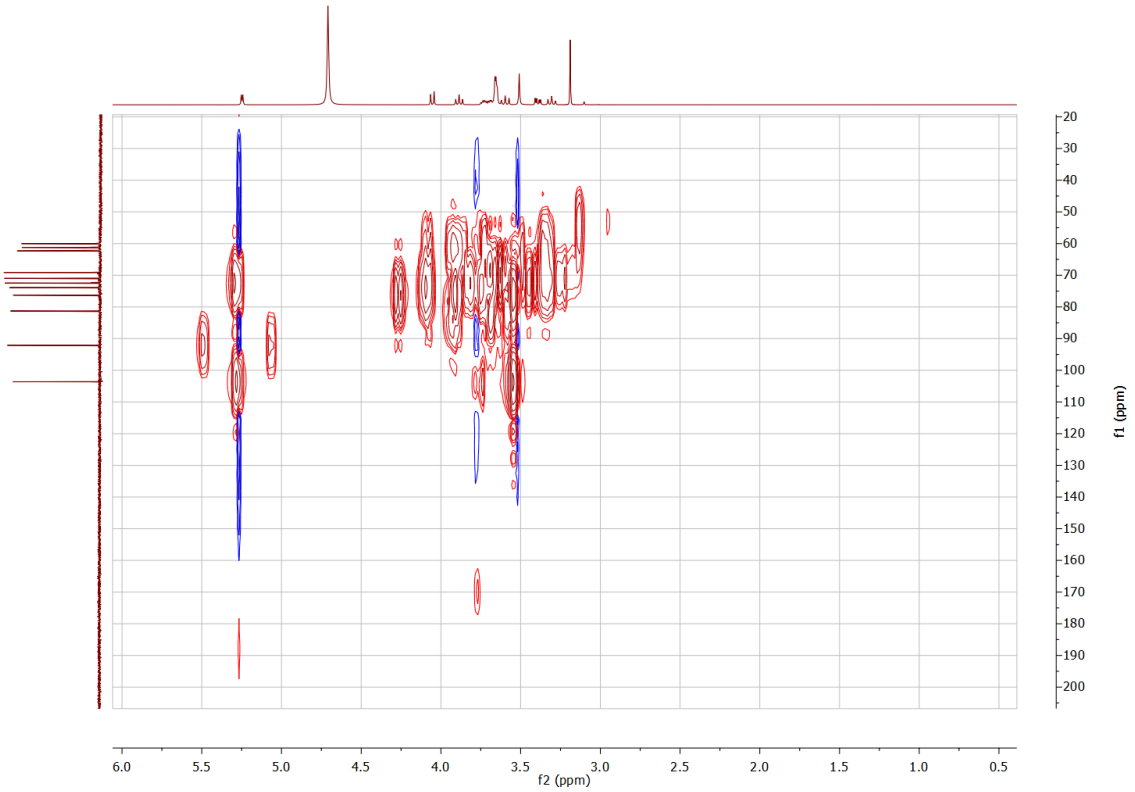

Figure B1.11: ^{1}H - ^{13}C HMBC spectrum (D_2O with K_2HPO_4/KH_2PO_4 100 mM at pH 7.4, 400.13 MHz) of a representative brown commercial sugar.

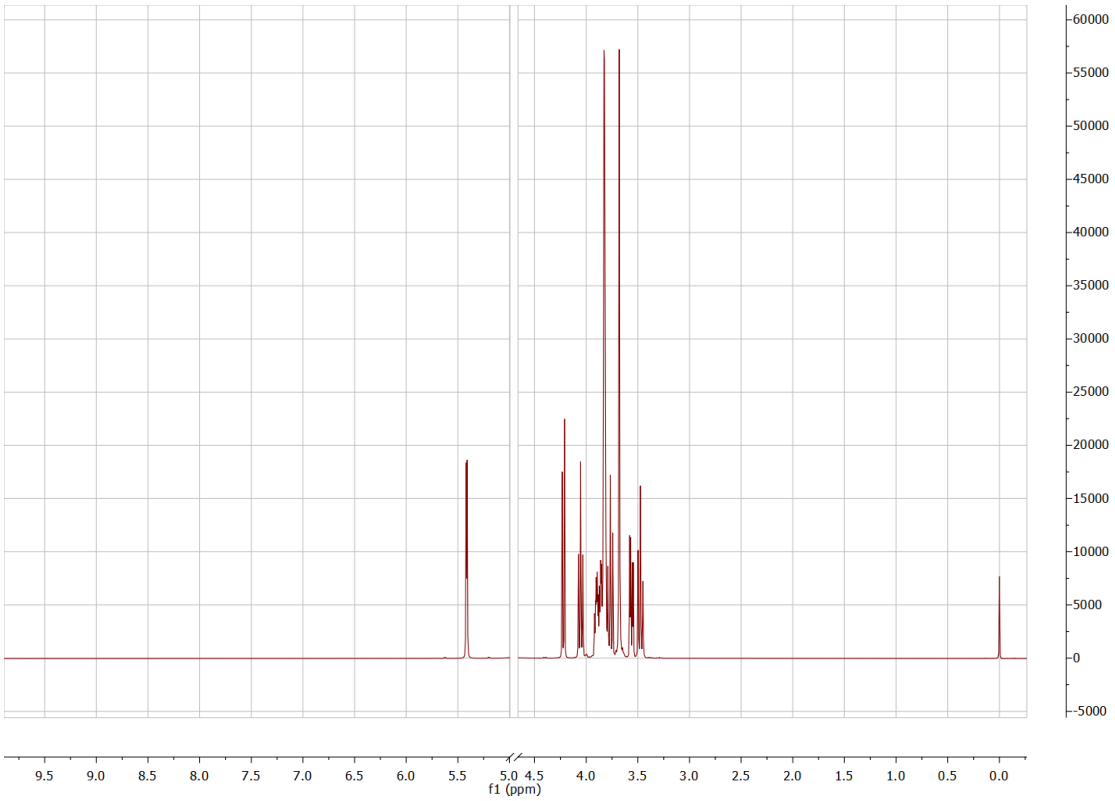

Figure B1.12: Representative ^{1}H NMR spectrum (D_2O with K_2HPO_4/KH_2PO_4 180 mM at pH 7.0, 400.13 MHz) of brown granulated sugar (BGS).

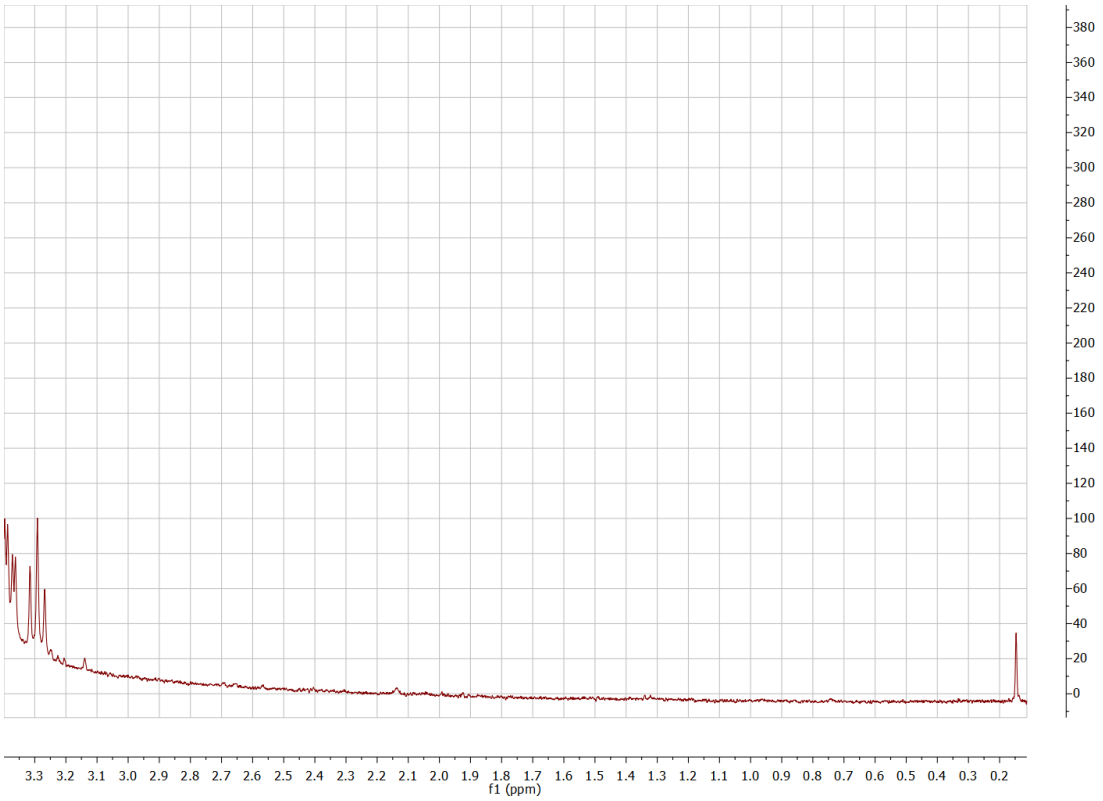

Figure B1.13: Expanded spectral region ($\delta_H = 0.20 - 3.33$) of a representative BGS sample.

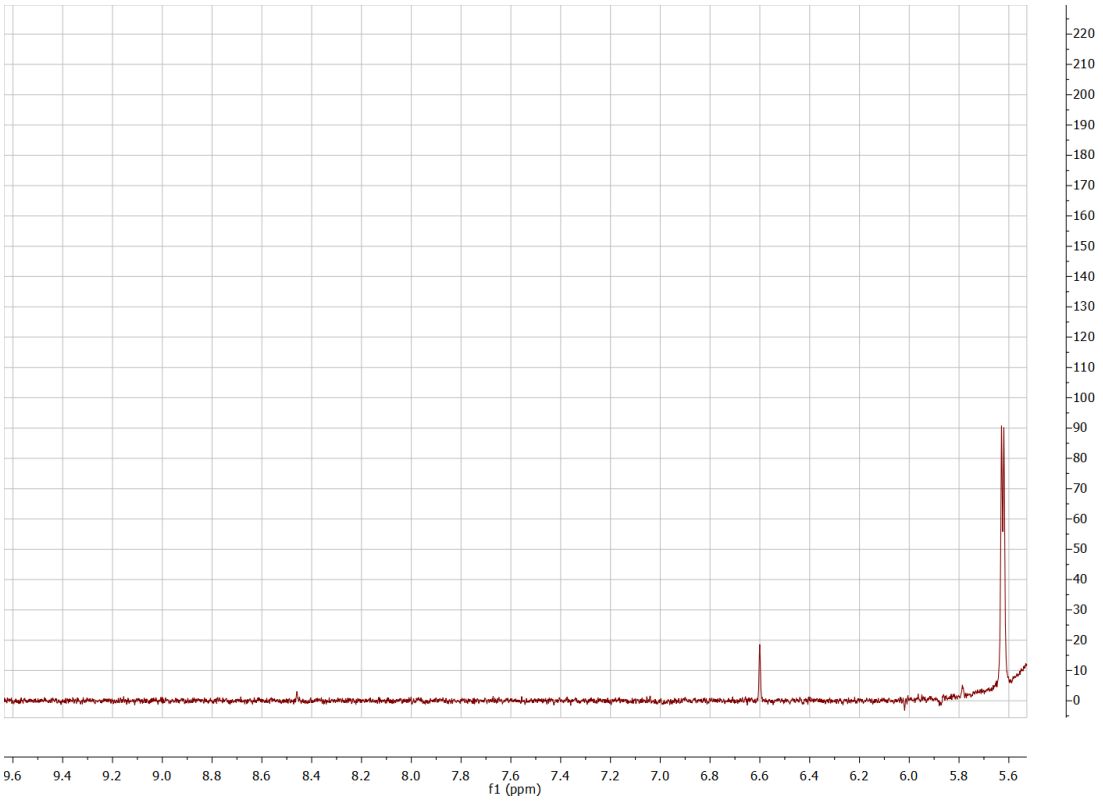

Figure B1.14: Expanded spectral region (δ_H $= 5.60 - 9.60$) of a representative BGS sample.

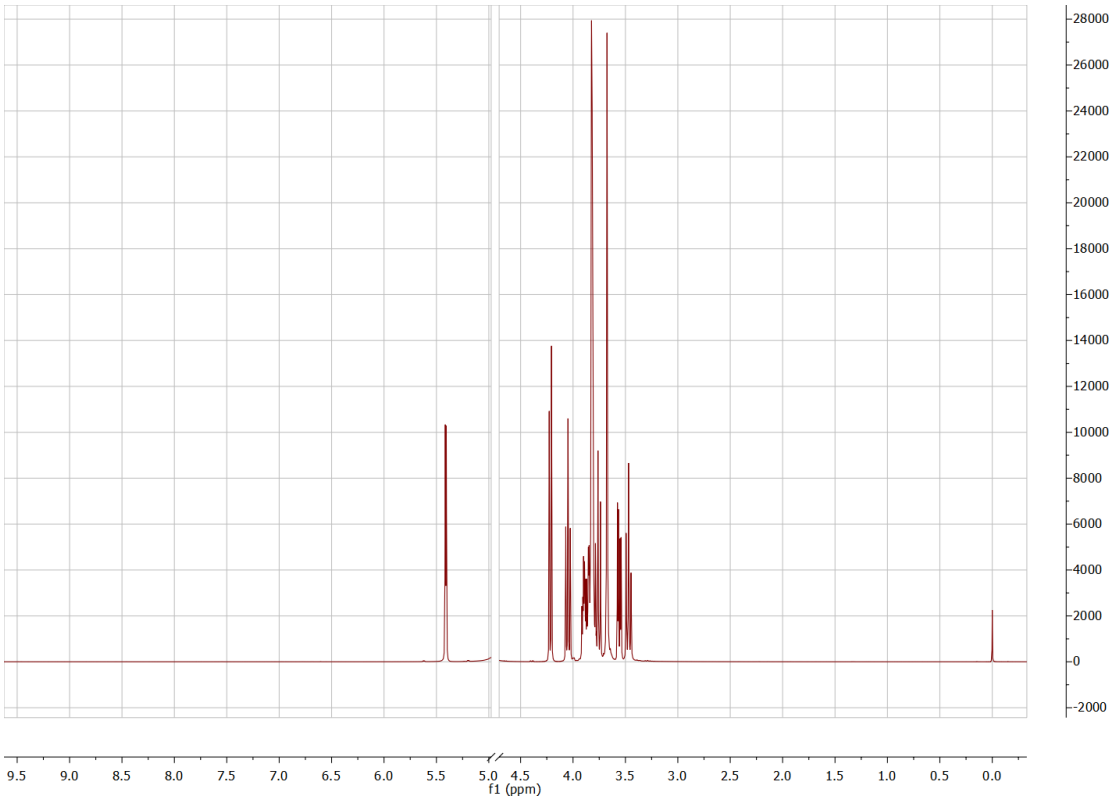

Figure B1.15: Representative ^{1}H NMR spectrum (D_2O with K_2HPO_4/KH_2PO_4 180 mM at pH 7.0, 400.13 MHz) of non-centrifugal sugar (NCS).

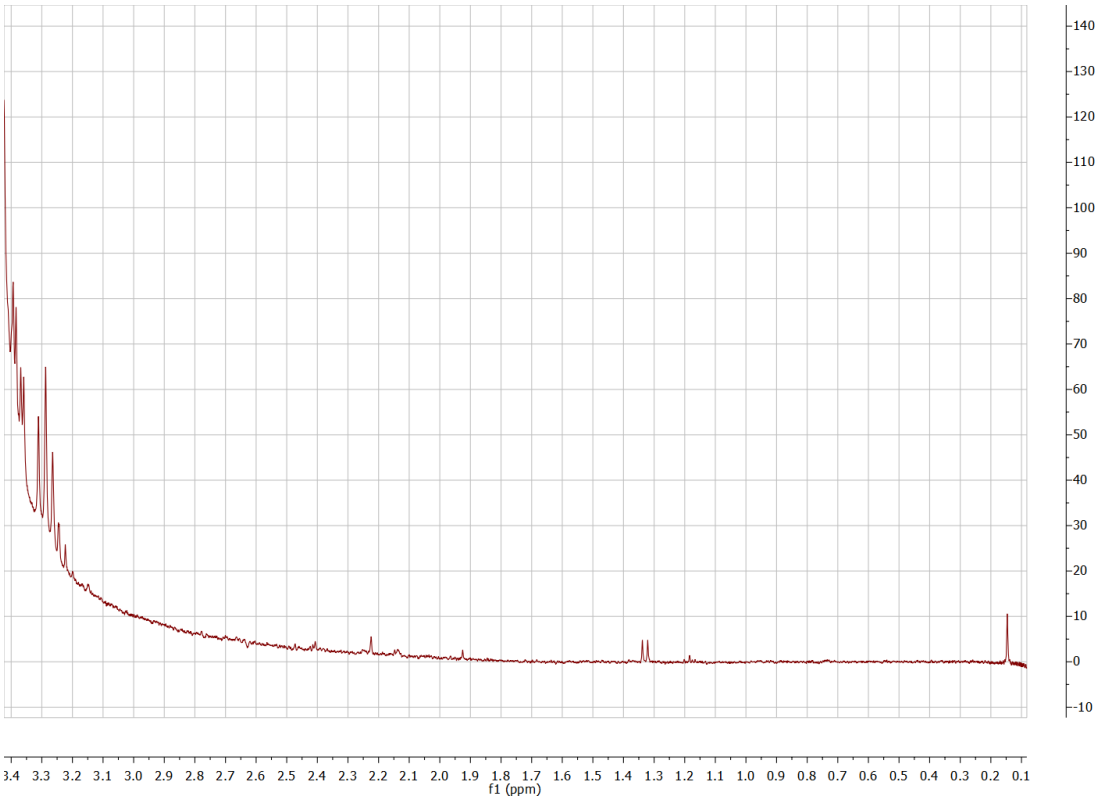

Figure B1.16: Expanded spectral region ($\delta_H = 0.10 - 3.40$) of a representative NCS sample.

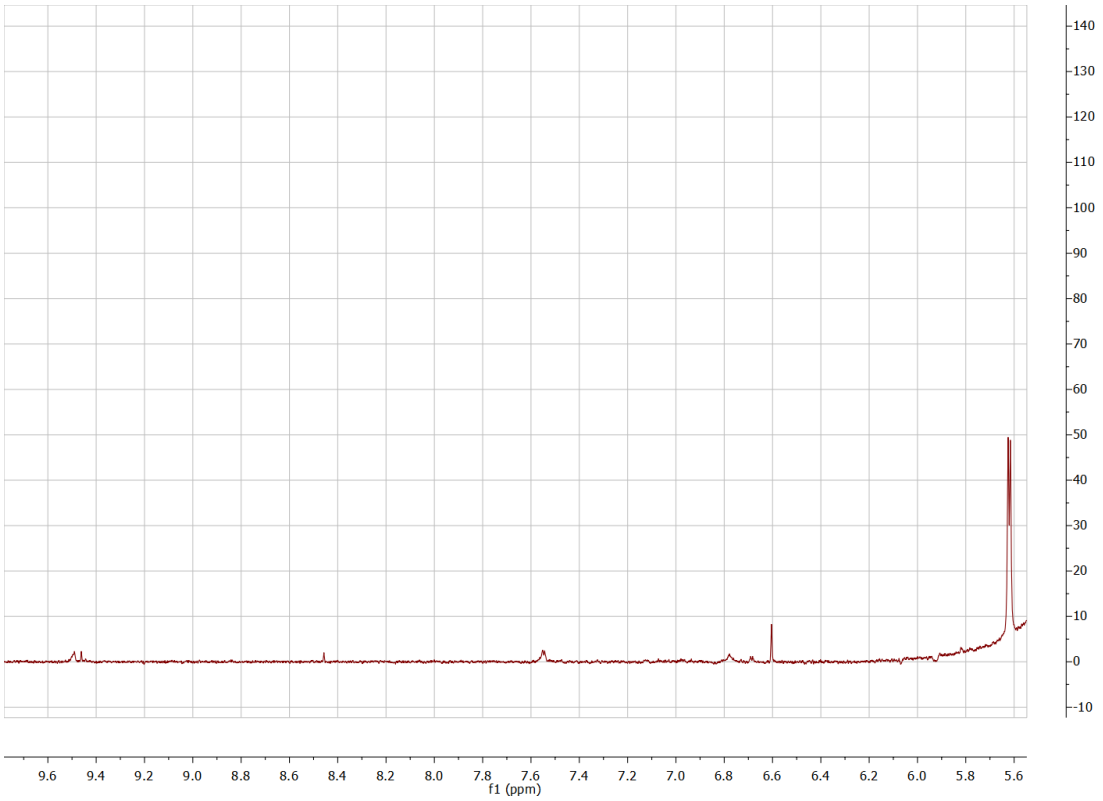

Figure B1.17: Expanded spectral region (δ_H $= 5.60 - 9.60$) of a representative NCS sample.

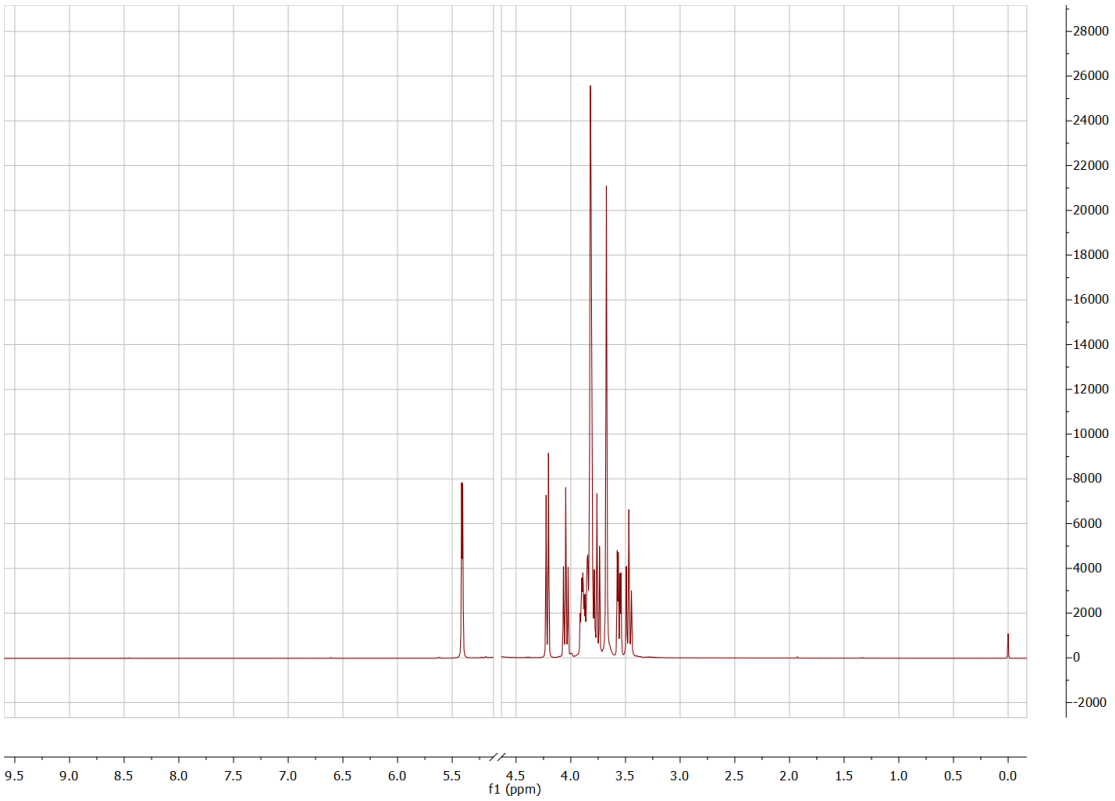

Figure B1.18: Representative ^{1}H NMR spectrum (D_2O with K_2HPO_4/KH_2PO_4 180 mM at pH 7.0, 400.13 MHz) of muscovado sugar (MS).

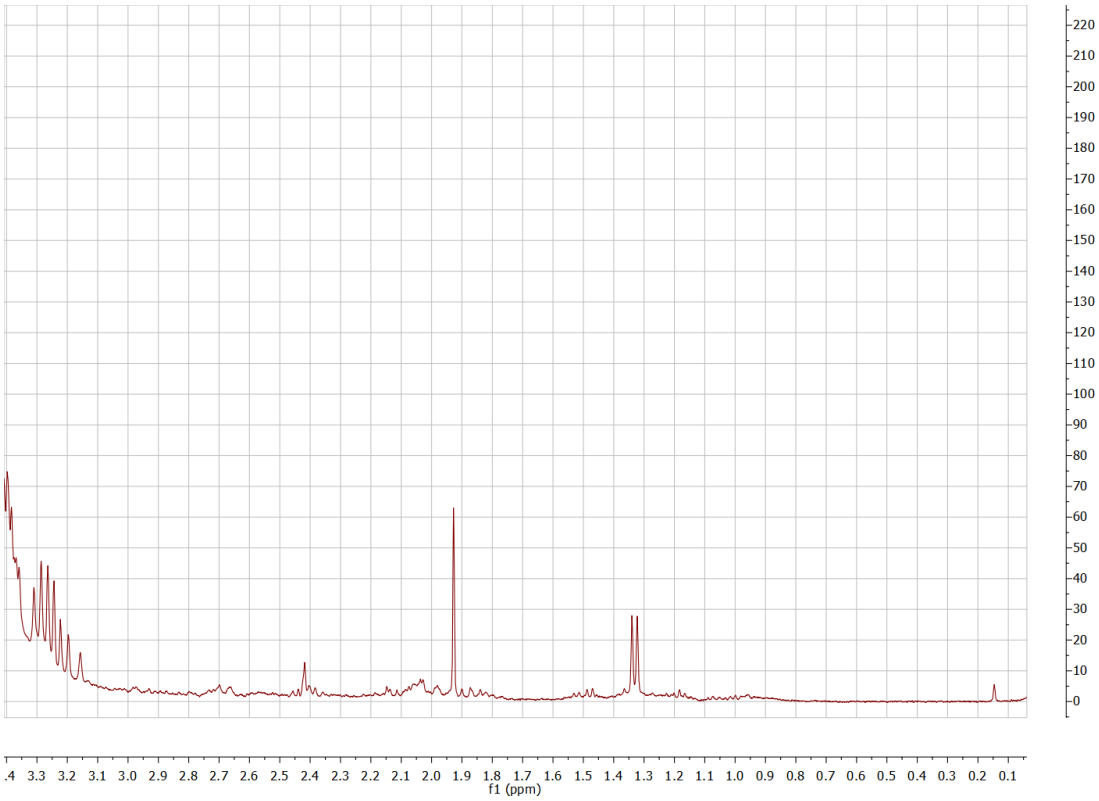

Figure B1.19: Expanded spectral region (δ_H $= 0.10 - 3.30$) of a representative MS sample.

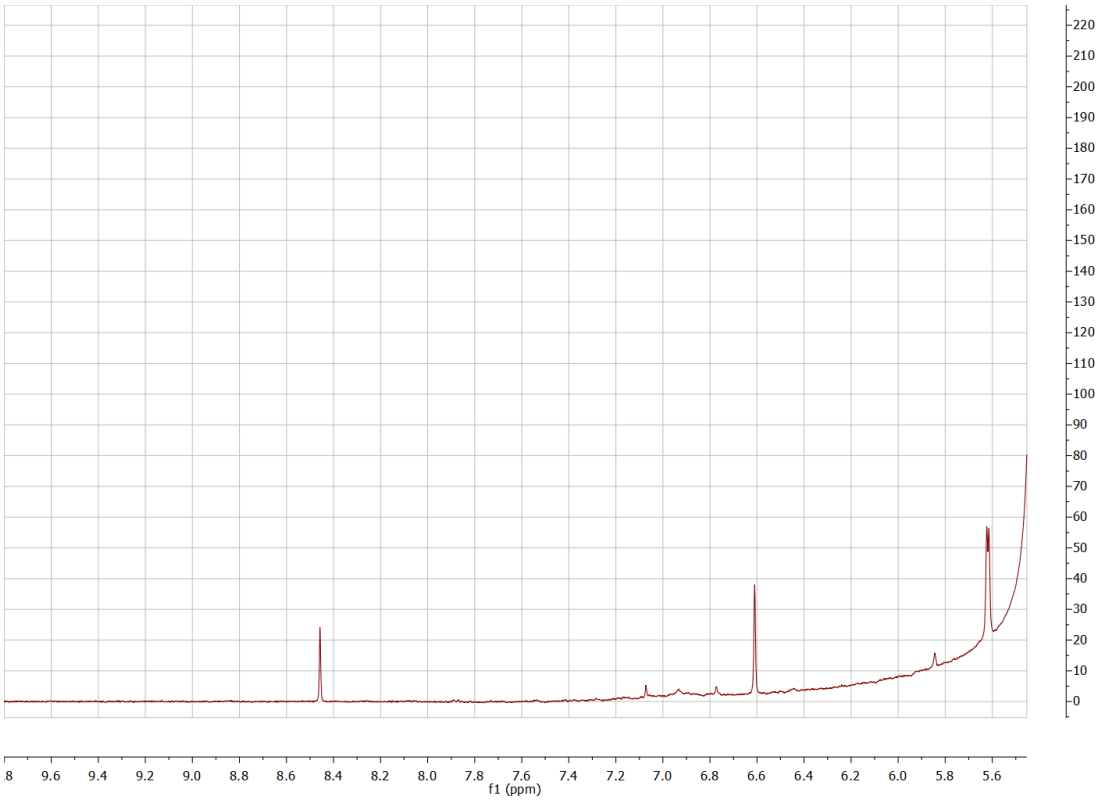

Figure B1.20: Expanded spectral region (δ_H $= 5.60 - 9.80$) of a representative MS sample.

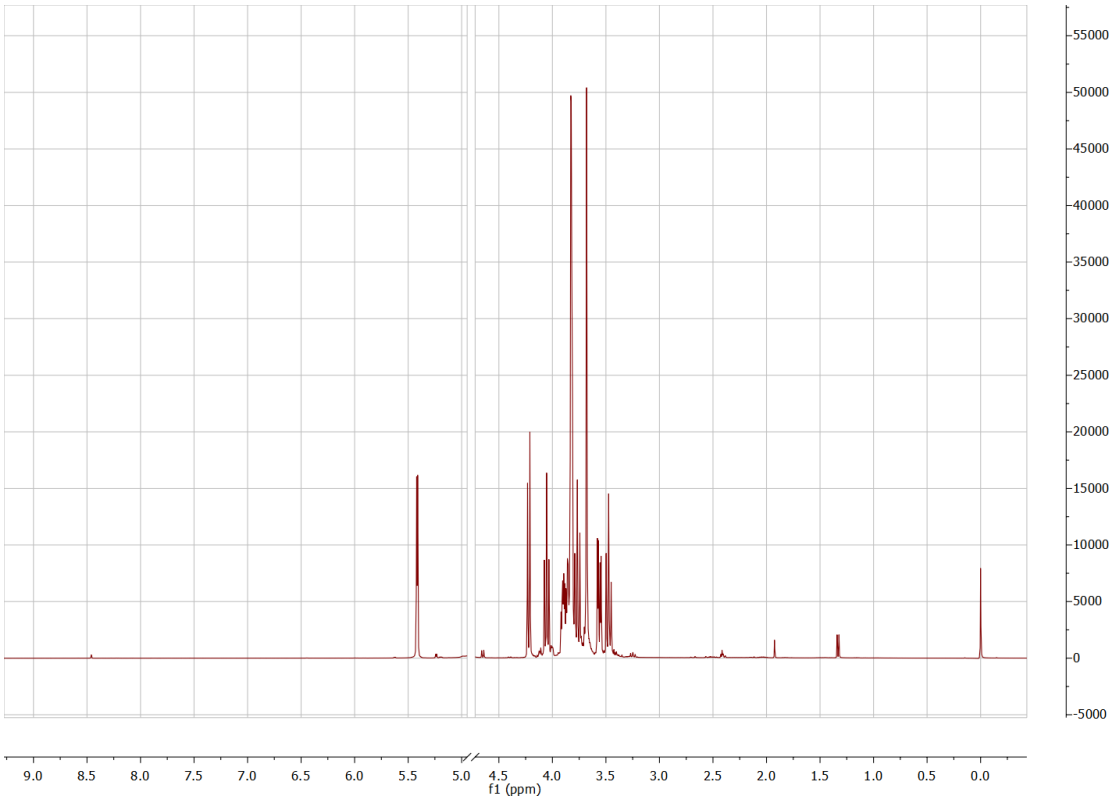

Figure B1.21: Representative ^{1}H NMR spectrum (D_2O with K_2HPO_4/KH_2PO_4 180 mM at pH 7.0, 400.13 MHz) of coconut sugar.

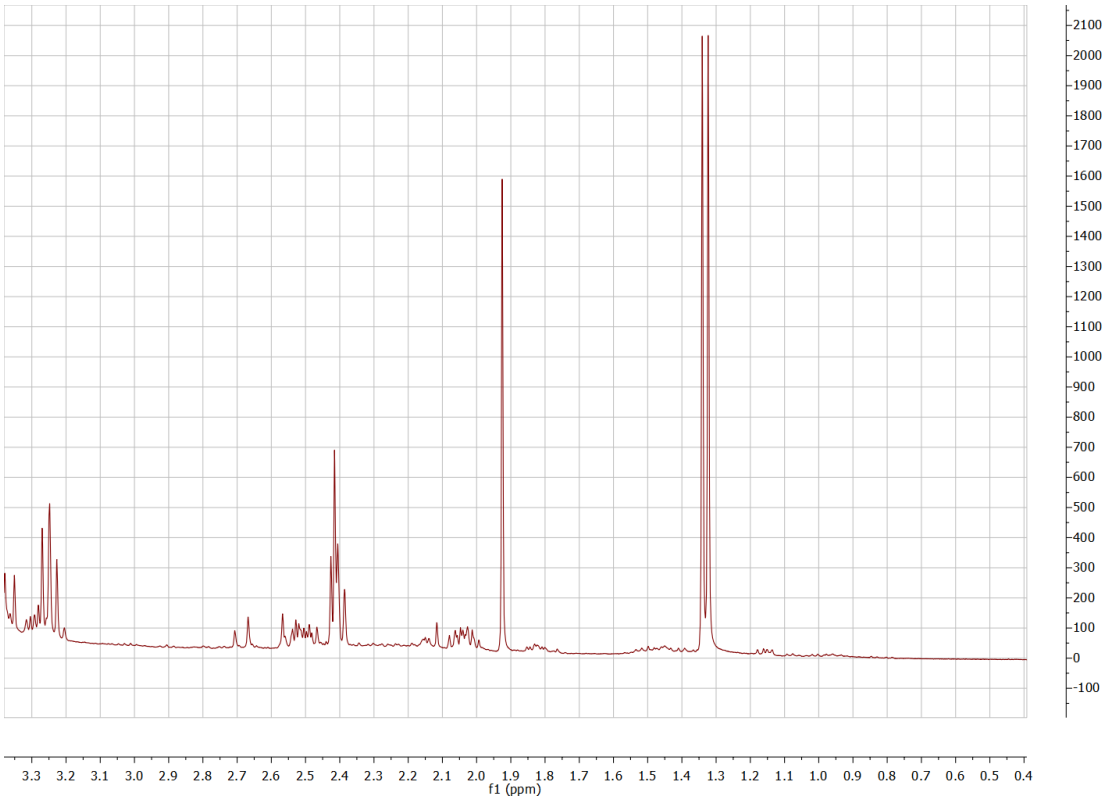

Figure B1.22: Expanded spectral region (δ_H $= 0.40 - 3.30$) of a representative coconut sugar sample.

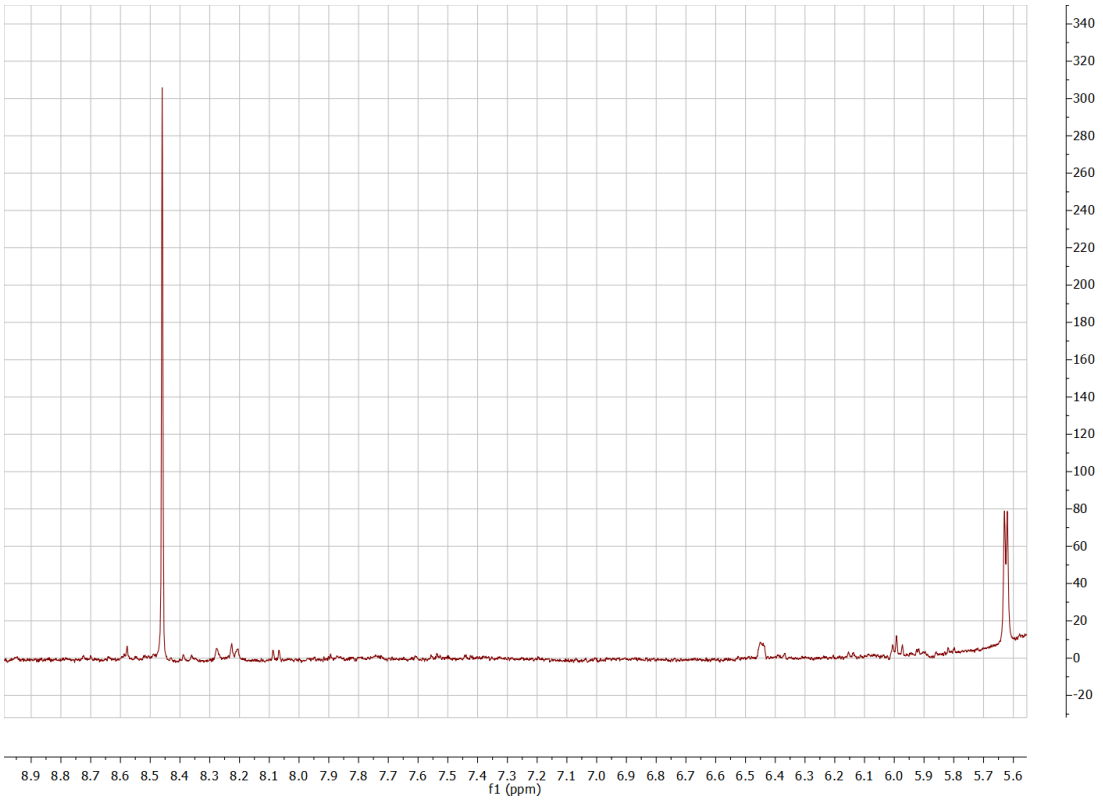

Figure B1.23: Expanded spectral region (δ_H $= 5.60 - 8.90$) of a representative coconut sugar sample.

B2 Fast Field Cyling (FFC) Relaxometry

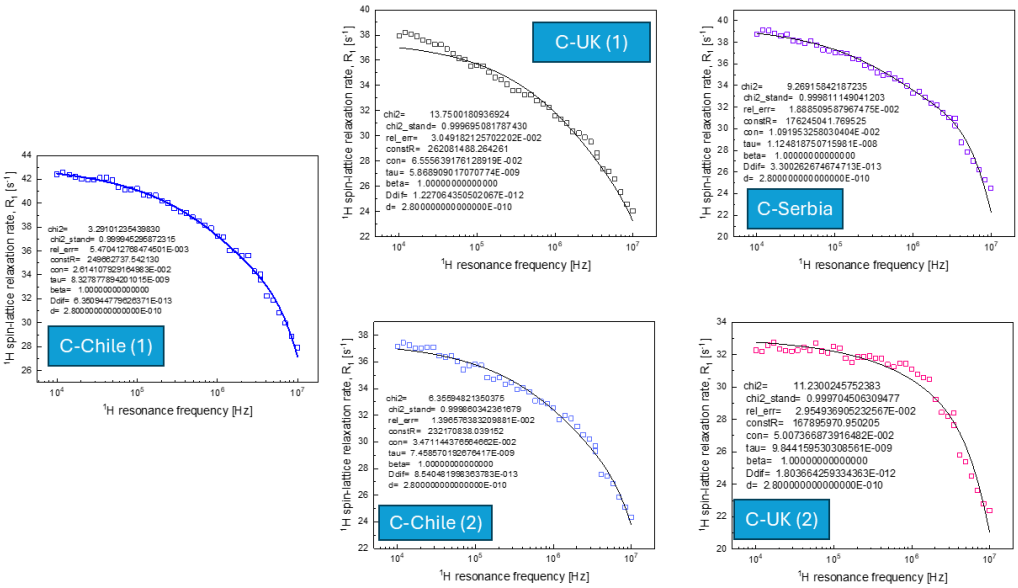

Figure B2.1: ^{1}H spin-lattice relaxation for coconut brown sugar samples according by country. Fast Field Cycling (FFC) NMR relaxation were acquired using a SMARtracer NMR Relaxometer (Stelar s.r.l., Mede, Italy). The measurements were performed for hydrated samples (2.0 g sample in 1 g of water [200% w/v]) with a frequency range of 10 kHz to 10 MHz. The temperature was set to 25°C with an accuracy of 1°C.

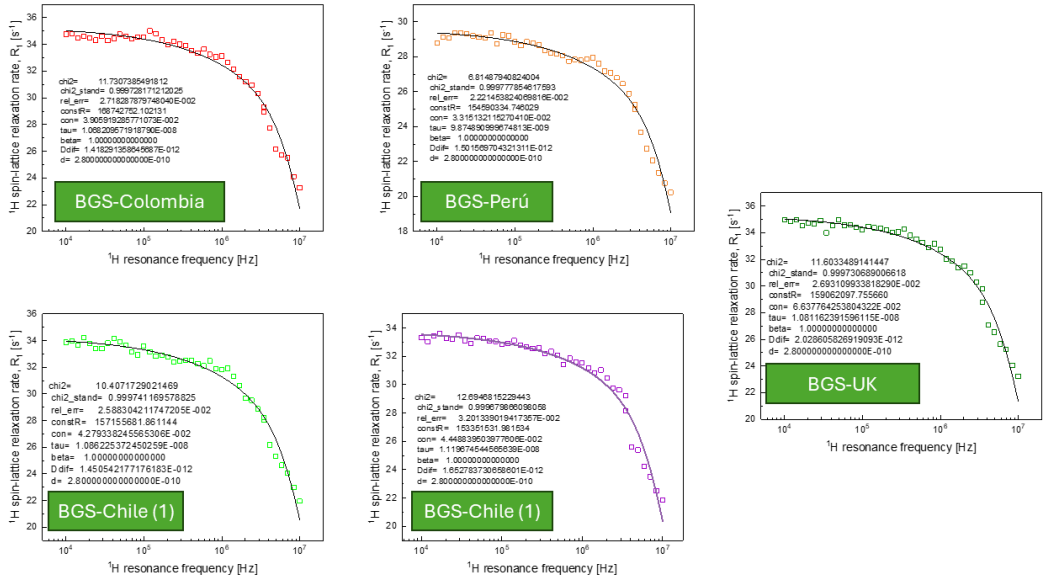

Figure B2.2: ^{1}H spin-lattice relaxation for Brown Granulated Sugar (BGS) samples according by country.

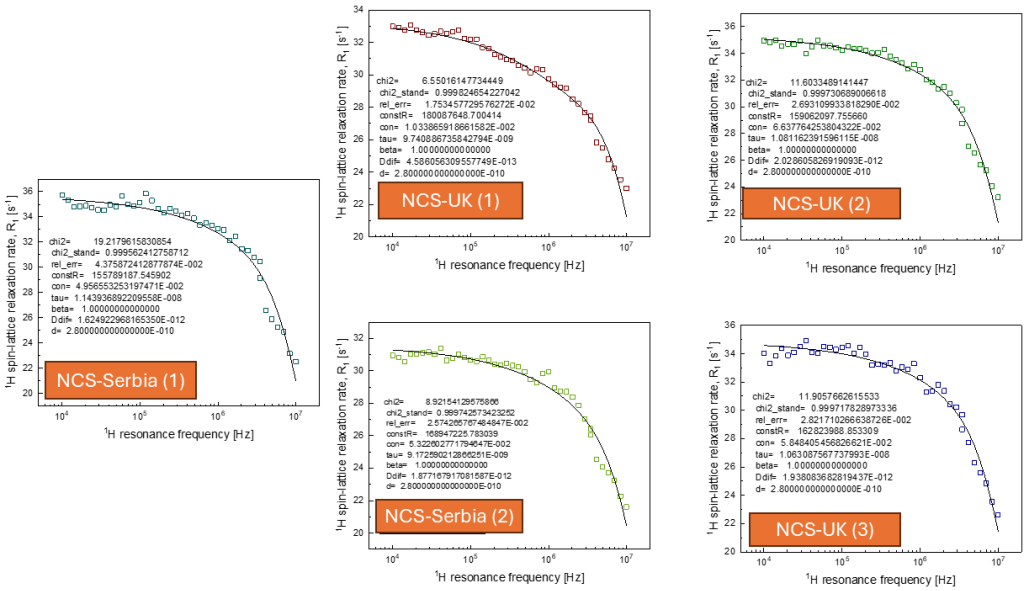

Figure B2.3: ^{1}H spin-lattice relaxation for Non-Centrifugal Sugar (NCS) samples according by country.

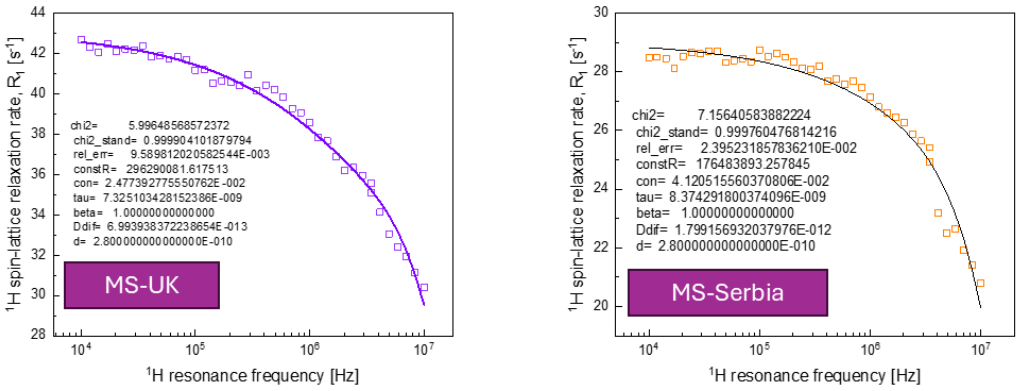

Figure B2.4: ^{1}H spin-lattice relaxation for Muscovado Sugar (MS) samples according by country.

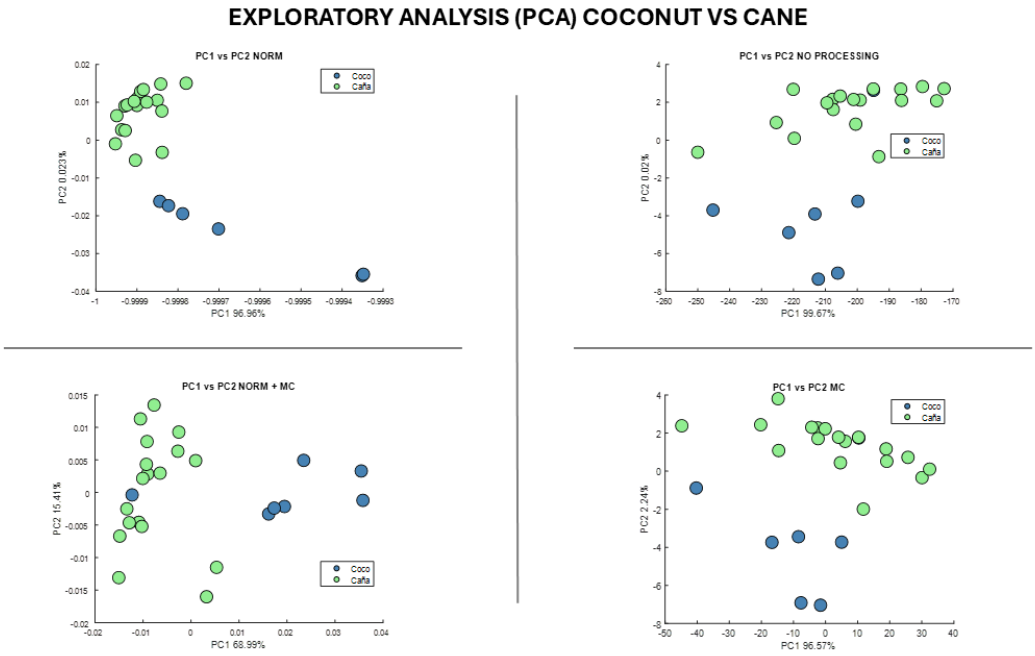

Figure B2.5: PCA score plot (PC1 vs PC2) for coconut vs cane obtained via ^{1}H spin-lattice relaxation curves with different pre-treatments. Coco: Coconut brown sugar; Caña: Brown sugarcane.

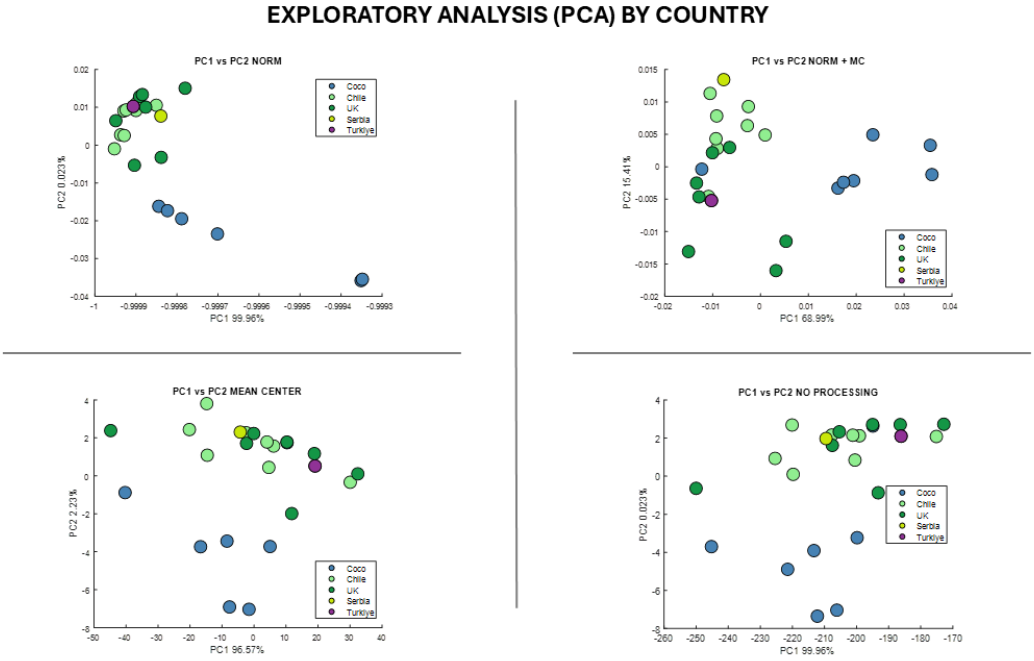

Figure B2.6: PCA score plot (PC1 vs PC2) by country (only for sugarcane-derived samples) obtained via ^{1}H spin-lattice relaxation curves with different pre-treatments.Coco: Coconut brown sugar.

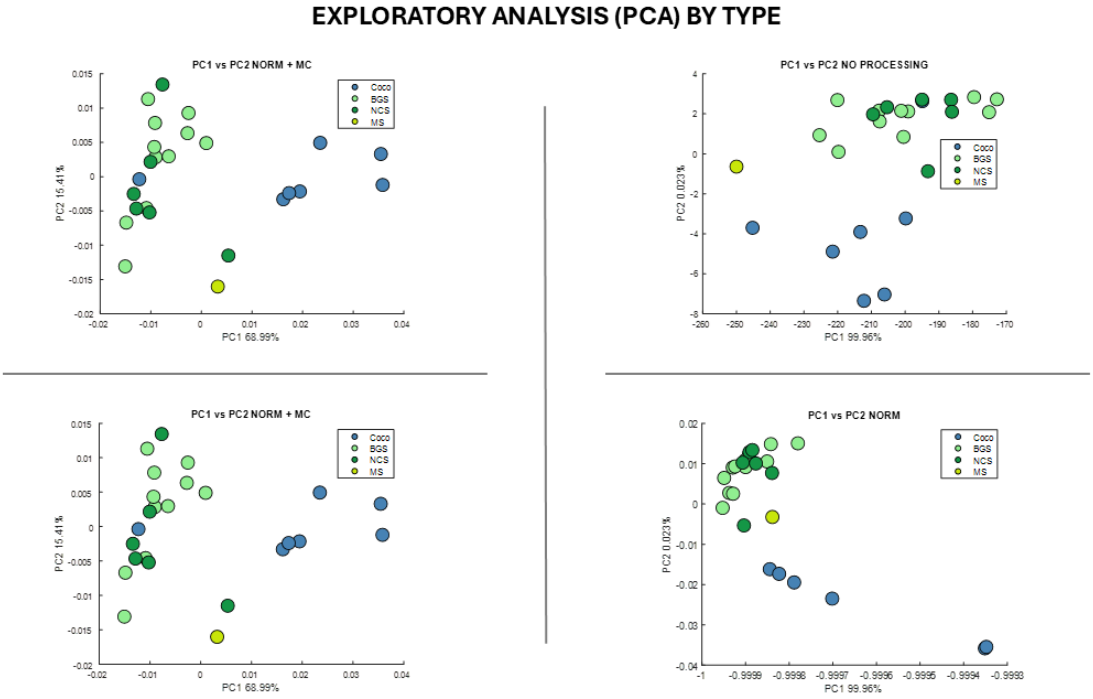

Figure B2.7: PCA score plot (PC1 vs PC2) by brown sugar type obtained via ^{1}H spin-lattice relaxation curves with different pre-treatments. Coco: Coconut brown sugar; BGS: Brown granulated sugar; NCS: Non-centrifugal sugar, MS: Muscovado sugar.

B3 Infrared spectroscopy

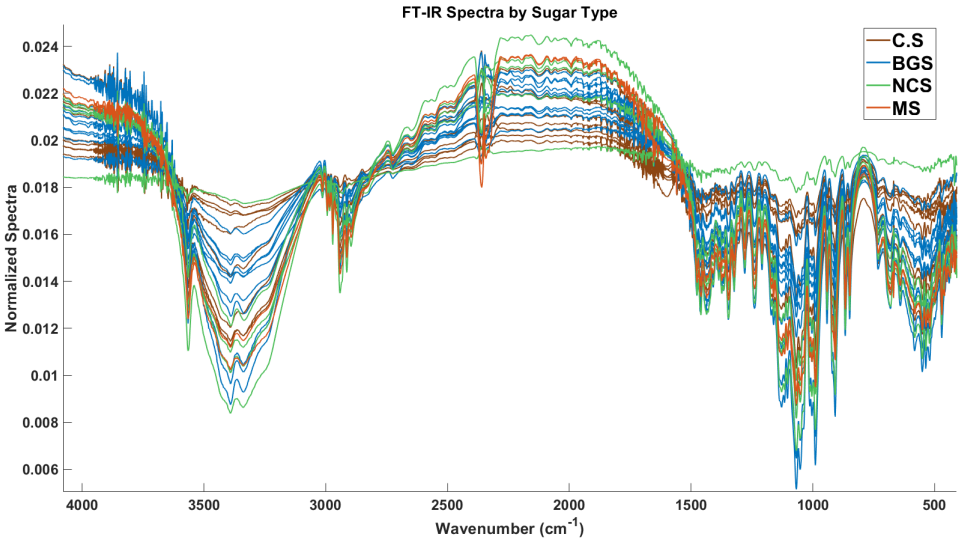

Figure B3.1: FT-IR spectra of brown commercial sugars were acquired using an IR Spirit Shimadzu FT-IR spectrometer. Spectral acquisition was carried out in transmittance mode with an spectral range of 4000 to 400 cm^{-1}, resolution of 4 cm^{-1}, slit of 2 cm^{-1}, and 32 scans. C.S: Coconut sugar; BGS: Brown granulated sugar; NCS: Non-centrifugal sugar; MS: Muscovado sugar.

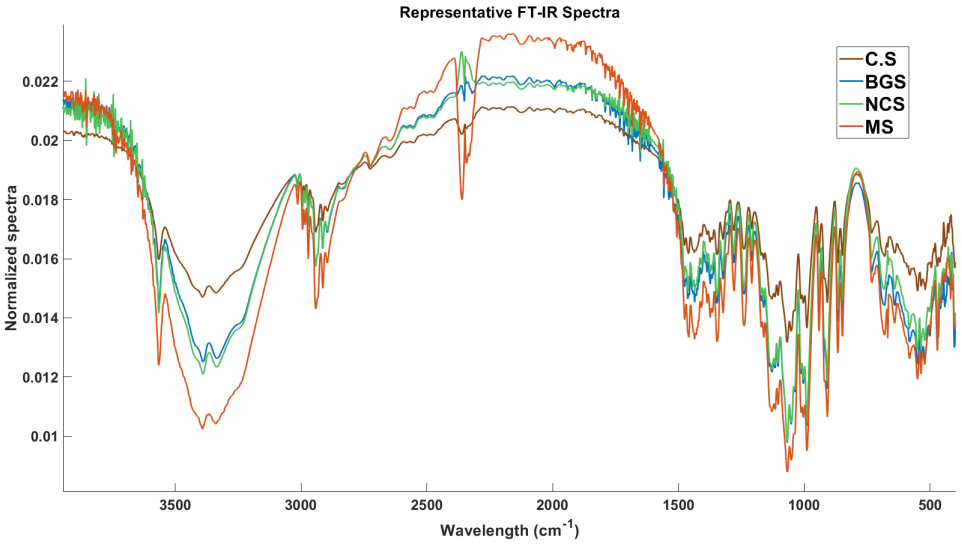

Figure B3.2: FT-IR spectra of representative brown commercial sugar samples. C.S: Coconut sugar; BGS: Brown granulated sugar; NCS: Non-centrifugal sugar; MS: Muscovado sugar.

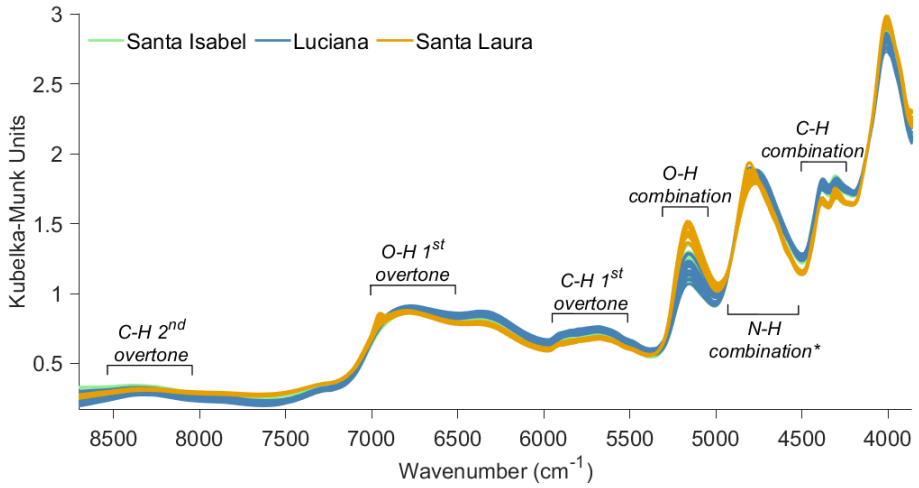

Figure B3.3: Near infrared (NIR) spectra for sugar beet roots were acquired using a Multi-Purpose Analyzer (MPA) Bruker FT-NIR spectrometer (Bruker Optics Inc., Germany). Spectral acquisition was carried out in reflectance diffuse mode with an spectral range of 12,500 to 3600 cm^{-1} (800 - 2777 mm) in Kubelka-Munk units.

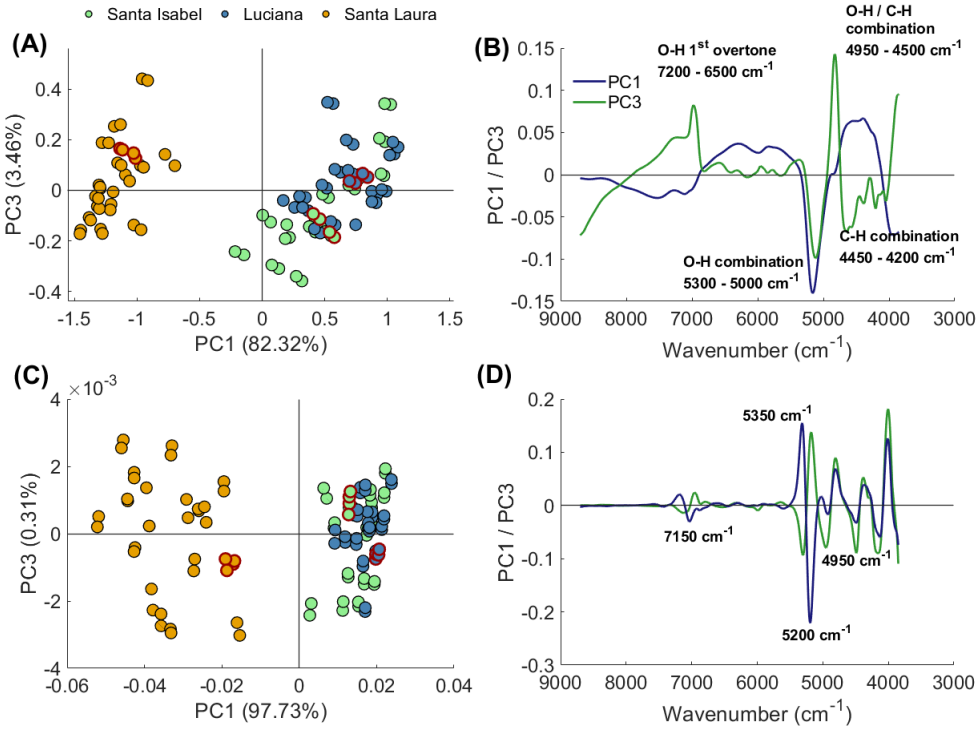

Figure B3.4: PCA (PC1 vs PC3) results obtained by FT-NIR. **A)** Score plot with smooth 15-point window + MSC + mean center. **B)** Score plot with smooth 15-point window + second derivative + mean center.