A multi-technique data fusion approach applied to a case study on Tonda Gentile Trilobata from Piedmont, Italy

Results in Chemistry 17 (2025) 102532

Contents lists available at ScienceDirect

Results in Chemistry

journal homepage: www.sciencedirect.com/journal/results-in-chemistry

A multi-technique data fusion approach applied to a case study on Tonda Gentile Trilobata from Piedmont, Italy

Mattia Sozzi [a], Biancamaria Senizza [b], Leilei Zhang [b], Michele Remo Chierotti [c], Massimo Esposito [d], Roberto Gobetto [c,*], Luigi Lucini [b,*], Francesca Scandellari [d,*]

[a] Department of Applied Science and Technology, Polytechnic of Turin, Corso Duca degli Abruzzi 24, 10129, Turin, Italy
[b] Department for Sustainable Food Process, Universita Cattolica del Sacro Cuore, Via Emilia Parmense 84, 29122, Piacenza, Italy
[c] Department of Chemistry, University of Turin, Via P. Giuria 7, 10125, Turin, Italy
[d] WhiteLab/U-Series s.r.l., via Collamarini 21, 40138, Bologna, Italy

ARTICLE INFO

Keywords:
Food traceability
Metabolomics
Multi-omics
Stable isotopes
Food integrity

ABSTRACT

The increasing attention to food traceability calls for robust and sensitive analytical methodologies. Here we applied multivariate statistical analysis on data obtained from different analytical techniques typically used to determine food authenticity, namely nuclear magnetic resonance (^{1}H-NMR), liquid chromatography high-resolution mass spectrometry (LC-HRMS), and bulk stable isotope analysis (BSIA). The data were firstly analysed as independent data sets and then the compatible data were merged using a multi-omics data fusion approach. As a case study, Tonda Gentile Trilobata hazelnut from Piemonte (known as Nocciola Piemonte PGI) was used, considering different origins and cultivars collected for two consecutive years. A first exploration of data from each technique revealed a strong temporal component therefore the information related to origin and cultivar was evaluated on data from each year separately; the three techniques highlighted differences between origins, while only ^{1}H-NMR and LC-HRMS could discriminate cultivars. The ^{1}H-NMR and the LC-HRMS datasets were then merged using Data Integration Analysis for Biomarker discovery using Latent variable approaches for Omics studies (DIABLO) framework, a supervised multivariate method for multi-omics integration, designed to identify correlated variables across datasets and maximize class discrimination. The merged data were adequately classified for the geographical origin and cultivar with minimum error rate while the features across the two datasets recorded similar up or down-modulation. Data fusion also confirmed the hierarchically stronger effect of annual variability in agreement with the outcome of individual analytical approaches. With this work, we show that data fusion increases robustness and enhances the extracted information by leveraging the strengths of each analytical technique.

1. Introduction

Food safety is an increasingly hot topic in many countries worldwide, including the European Union, which has developed three schemes (the protected designation of origin PDO, the protected geographical indication PGI, and the traditional speciality guaranteed TSG) to promote and protect its agricultural products (Regulation (EU) No 1151/2012). Several analytical techniques were demonstrated to be suitable to verify compliance with product specifications as foreseen by the EU Regulation. These techniques include Nuclear Magnetic Resonance (^{1}H-NMR), Liquid Chromatography coupled with High-Resolution Mass Spectrometry (LC-HRMS), and Bulk Stable Isotope Analysis (BSIA), which detect variations in the composition of metabolites caused by factors such as geographical origin, cultivar or production year [1]. ^{1}H-NMR is one of the most employed techniques thanks to its non-destructivity, fastness, and simultaneous detection of all the major organic classes of compounds [2,3]. The high reproducibility of the ^{1}H-NMR, combined with the possibility to automatize the entire analytical process, even with a very high number of samples, makes it suitable for high-throughput analysis [4]. The discriminating ability of ^{1}H-NMR is due to its capacity to produce distinct ^{1}H-NMR profiles; even if not all the signals can be identified, the signal pattern in the spectrum can be regarded as a "fingerprint" of the specific sample, leading to ^{1}H-NMR-based discriminations. This technique was tested

* Corresponding authors.
E-mail addresses: mattia.sozzi@polito.it (M. Sozzi), biancamaria.senizza@unicatt.it (B. Senizza), leilei.zhang@unicatt.it (L. Zhang), michele.chierotti@unito.it (M.R. Chierotti), massimo@u-series.com (M. Esposito), roberto.gobetto@unito.it (R. Gobetto), luigi.lucini@unicatt.it (L. Lucini), francesca.scandellari@u-series.com (F. Scandellari).

https://doi.org/10.1016/j.rechem.2025.102532
Received 13 February 2025; Accepted 14 July 2025
Available online 28 July 2025

with many food products such as hazelnuts [5,6], extra virgin olive oils [7], and Parmesan cheese [8]. LC-HRMS has higher sensitivity, identification capabilities, and dynamic range reliability [9] compared to ^{1}H-NMR despite having lower reproducibility and being more limited in absolute quantification. Noteworthy, based on the respective compound coverage profiles, these two techniques are complementary analytical methods [10]. LC-HRMS has been applied to a broad range of diverse high-value food matrices such as hazelnuts [11], extra-virgin olive oils [12,13], and saffron [14]. BSIA is based on the analysis of stable isotopes [15]. Many elements in nature present stable isotopes and can virtually be used for this purpose, although hydrogen, oxygen, carbon, and nitrogen are the most widely used [16]. In plants, nitrogen, oxygen, and hydrogen mainly derive from root uptake, while carbon is assimilated during photosynthesis; the stable isotope values of these elements present a geographical pattern that is, to some extent, maintained during the metabolic pathways. Therefore, the stable isotope composition of the plant biomass reflects that of the surrounding environment [16]. In comparison to other techniques, stable isotope analysis is more flexible in terms of sample handling and it has been applied to products such as hazelnuts [17], potatoes [18], and apples [19]. Since each analytical technique has specific advantages and disadvantages, as well as its distinctive metabolite coverage, the recent literature has highlighted the benefits of using orthogonal multi-technique data fusion approaches [20] as they provide more robust results than those obtained independently from each dataset and with potentially complementary or synergic effects [21]. The integration of multi-omics data was initially developed to combine different omics datasets, such as genomics, transcriptomics, proteomics, metabolomics, and more, to gain a comprehensive understanding of biological systems. In the context of food protection and authenticity, this approach can extract information from different analytical techniques and provide a holistic view of the food's composition to ensure the safety, quality, and authenticity of different food matrices, including hazelnuts [22,23]. Recent uses of data fusion techniques merged data from ultraviolet–visible (UV–Vis) and infrared (IR) spectroscopy to differentiate saffron based on geographic origin [24], while the integration of ^{1}H-NMR spectroscopy with LC-HRMS distinguished black peppers according to their geographic origins and processing methods [25]. Because of its importance in the market [26], hazelnut represents an excellent case study to test the innovative potential of data fusion techniques. In particular, the Tonda Gentile Trilobata (TGT) produced in Piemonte, Italy, is a hazelnut cultivar registered as Protected Geographical Indication (PGI) (Regulation 1151/2012) under the name "Nocciola Piemonte". TGT is highly appreciated for its organoleptic properties and is considered a "gold standard" for product quality [27]: its richness in secondary metabolites translates into a well-recognized sensory quality that makes it an excellent ingredient in high-quality confectionery industry. TGT characteristics are strictly bound to its geographical origin and represent an exemplary application of the concept of *terroir* beyond the wine sector [28]. In this work, we applied supervised multivariate statistical analysis to data obtained from different orthogonal analytical techniques, namely ^{1}H-NMR, LC-HRMS, and BSIA. The data were firstly analysed as independent data sets and then the compatible data were merged into a single dataset using a multi-omics data fusion approach. As a case study, we used data obtained from hazelnuts of the TGT, which were compared with hazelnuts of different origins and cultivars. This work aimed to propose an improved discrimination model in the field of food authenticity.

2. Materials and methods

2.1. Samples collection and preparation

In this study, a total of 54 hazelnut samples were analysed. Of these, 44 samples were from different Italian regions, 4 samples from Turkey, 2 samples from Romania, 2 samples from Chile, 1 sample from Georgia, and 1 sample from Bulgaria; 29 were of the cultivar Tonda Gentile Trilobata, 6 Giffoni, 2 Nocchione, 1 Nostrale, and 4 of other minor cultivars. Samples were collected in 2020 (16) and 2021 (38). More detailed information about the origin and cultivar is provided in the supplementary materials (Table S1 and Figure S1). The sample preparation, handling and storage were the same for the three techniques. Coarsely ground samples were dispersed in an 80:20 methanol:water solution using an Ultra-Turrax. The samples were then centrifuged (15 min, 8.000 x g) and filtered (cellulose membrane, 0.22 μm) into vials for analysis [29]. The LC-HRMS and ^{1}H-NMR analyses were performed on the methanolic extract, while BSIA was performed on the centrifuge cake.

2.2. ^{1}H-NMR sample preparation, spectra acquisition and data pre-processing

For proton nuclear magnetic resonance analysis (^{1}H-NMR), the methanol solvent was removed by under-vacuum evaporation, and the metabolic mixtures were redissolved in deuterated methanol. Each ^{1}H-NMR tube was prepared at least in duplicate by adding 600 μL of the sample solution and 10 μL of trimethylsilylpropanoic-d4 (TSP-d4) acid standard solution 0.005 M. The ^{1}H-NMR analysis was performed on a Jeol ECZR 600 spectrometer (JEOL Ltd., Akishima, Tokyo, Japan) operating at 600.17 MHz for protons. The spectra were collected at a fixed temperature of 298 K by acquiring 32768 points and performing 256 scans for each sample, using a 30 s relaxation delay. A solvent suppression procedure (Dante pre-saturation) was applied to remove the water signal. The spectra were baseline- and phase-corrected using the DELTA processing tool offered by JEOL Ltd. The raw ^{1}H-NMR spectra were imported and processed under the MATLAB R2021b environment (Mathworks, Natick, MA, USA). For all the spectra, the ppm scale was referenced to the TSP peak (0.00 ppm) (Figure S2). To increase the comparability of spectra, the "Icoshift" tool [30,31] was applied to align the most important signals located inside specific manually selected intervals. The collected raw spectra contains 26214 variables. The spectra width was corrected to include only signals between −0.1 ppm and 9.3 ppm to remove unwanted and noisy areas. To avoid interferences, methanol signals (from 3.26 ppm to 3.38 ppm) and water residues (from 4.55 and 5.05 ppm) were removed from the spectra. To reduce the number of variables without affecting spectral information, and to make the dataset dimension comparable with the LC-HRMS dataset, the variables of the NMR dataset were reduced by picking one point every two. The reduced NMR dataset was composed by 10508 variables.

2.3. LC-HRMS untargeted profiling and data pre-processing

The hazelnut metabolite was also profiled through Ultra-High-Pressure Liquid Chromatography (UHPLC; 1290 series, Agilent Technologies, Santa Clara, CA, USA) coupled to a Quadrupole-Time-Of-Flight Mass Spectrometer (Q-TOF MS; 6550 iFunnel, Agilent Technologies), as reported previously (Senizza et al. 2023). The mass spectrometer worked in Full-Scan mode, with a positive ionization (ESI +), to acquire accurate masses in the 100–1200 m/z range. The chromatographic separation was performed on an Agilent Zorbax Eclipse plus C18 analytical column (50 × 2.1 mm, 1.8 μm), using water-acetonitrile gradient elution (from 6% to 94% organic in 34 min), with 6 μL of injection volume (three technical replicates). The features were aligned for mass (5 ppm accuracy) and retention time (0.05 min) and annotated according to the "find by-formula" algorithm using the Agilent Profinder B.07 software against the database FooDB. To this aim, the whole isotopic pattern of molecular features (accurate monoisotopic mass, isotope spacing, and ratio) was used as previously described [14]. The annotation approach corresponded to a Level 2 of confidence, with reference to the COSMOS initiative for standardization in metabolomics [32]. Data filtering and normalization were also carried out in Profinder B.07, retaining only the compounds identified

within 100% of replications of the treatment grouping. The Agilent Mass Profiler Professional B.12.06 software was finally used as post-acquisition pre-processing [33]. Therein, compounds were filtered by abundance, considering only those compounds with an area > 5000 counts, normalized at the 75th percentile, and baselined to their median in the dataset.

2.4. BSIA sample preparation and acquisition

Methanol-extracted samples were finely ground using a ceramic mortar and pestle and left at room temperature under a chemical hood for at least 4 h to let residual methanol evaporate. Stable isotope analysis was performed using a High-Temperature Conversion Elemental analyser coupled with a Delta V Advantage Isotope Ratio Mass Spectrometer (Thermo Fisher Scientific GmbH, Bremen, Germany). About 0.4 mg were weighted in silver cups for hydrogen ($^{2}H/^{1}H$) and oxygen ($^{18}O/^{16}O$) isotope ratio analysis and in tin cups for carbon ($^{13}C/^{12}C$) while about 1.5 mg were weighted in tin cups for nitrogen ($^{15}N/^{14}N$) isotope ratio. Every sample was analysed at least in duplicate. The isotope ratios are reported in the standard delta (δ) notation calculated with respect to the internationally recognized standard for each element (VSMOV for hydrogen and oxygen, VPDB for carbon, and Air for nitrogen). For this purpose, a three-point calibration was used to ensure the robustness of the measurements. The reference materials were all analysed in triplicate. A quality control was also added to each run. The δ values were multiplied by 1000 and expressed in ‰.

2.5. Multivariate data analysis to explore single techniques independently

The datasets obtained from the ^{1}H-NMR, LC-HRMS, and BSIA were pre-processed and treated separately. The datasets were first normalized to the total sum. In addition, a mean centre scaling was used to preprocess ^{1}H-NMR and LC-HRMS data, while autoscaling was used to preprocess BSIA data, according to the different characteristics of the three databases. Principal Component Analysis (PCA) was performed using a MATLAB toolbox specific for exploratory analysis [34]. PCA models were applied to each individual dataset to naively explore data patterns and search for similarities and differences among the samples and treatments. In addition, to visualize groupings and outliers, the samples in the scores' plots were coloured according to the different features inspected (harvest year, geographical origin, and hazelnut cultivar). The statistical significance of group separation was quantified using the Hotelling's T2 test using the ICSNP package for R [35] calculating the *p*-value using the F-distribution. A supervised sparse Partial Least Squares Discriminant Analysis, sPLS-DA [36] modelling approach was next used to investigate the ability to discriminate specific classes of samples (i.e., Piedmont hazelnuts and TGT cultivar). For each dataset, two models were created, one to observe how the samples with Piedmont geographical origin were separated from the others, and the second to evaluate the samples' distribution with respect to the hazelnuts of TGT cultivar. Stable isotope data were further explored by Soft Independent Modelling of Class Analogies (SIMCA) using the mdatools package for R [37]. The SIMCA model was built using only samples obtained from Piedmont for each year of harvest. To circumvent the small sample size, several models were produced, each excluding one sample from Piedmont. Each one of these models was validated using the Procrustes approach [38]. Finally, each model was used to predict the classification of the whole dataset for the year, including the sample from Piedmont previously left out of the training sets. A sample is considered correctly assigned when most models assigned it to the correct class (Piedmont or none). The sensitivity was calculated as the number of samples correctly assigned to the target class (true positive), the specificity was calculated as the number of samples correctly assigned to the alternate class (true negative), the accuracy was calculated as the number of samples correctly assigned to the respective class (true positive + true negative) respect to the total.

2.6. Multi-omics data fusion

The ^{1}H-NMR dataset and the LC-HRMS dataset were merged using Data Integration Analysis for Biomarker discovery using Latent variable approaches for Omics studies (DIABLO) framework implemented within the "mixOmics" R package (version 6.22) to evaluate the complementarity of the two techniques and to understand which variables better characterize the geographical origin and variety of hazelnuts. The BSIA dataset was not included in this analysis because of the large diversity in dataset dimensions (4 variables compared to the 2464 variables of LC-HRMS and the 10508 variables of ^{1}H-NMR). Moreover, before merging the data, the ^{1}H-NMR dataset was down-sampled by taking one data point every two to obtain half of the initial variables to be more comparable to the LC-HRMS dataset. The DIABLO model was optimized using the framework's tuning function, selecting the number of components (based on the overall error rate (BER) and the number of variables (defined by repeated and stratified cross-validation analysis (3-fold CV, repeated 50 times)). Specifically, two principal components were used to generate the integration model considering, for the LC-HRMS data, 8 and 40 variables, whereas for ^{1}H-NMR data, 100 and 100, for components 1 and 2, respectively. The classification error rate was performed in different steps, one component by one, to forecast the optimal number of variables with the strongest discrimination ability. In addition, repeated cross-validation was used to evaluate the performance of sPLS-DA models and estimate their ability to correctly classify each sample addition based on harvest year, geographical origin, and cultivar.

3. Results and discussion

3.1. Datasets exploration and patterns

The datasets obtained with ^{1}H-NMR LC-HRMS, and BSIA analysis were inspected individually by performing different PCA models (unsupervised) to search for groupings related to harvest year, geographical origin, and hazelnut cultivar. The information associated with each sample was used to visualize the different classes in the score plots obtained from the PCA models. Furthermore, to better investigate the information related to Piedmont's geographical origin and TGT cultivar, the datasets were also investigated using a supervised approach (PLS-DA and SIMCA).

3.1.1. ^{1}H-NMR dataset

The first exploration of the NMR dataset was done using unsupervised methods. A first PCA model was performed including all samples. Considering seven principal components (PC), a total explained variance of 81.31% was described. From the scores plot of the first two PCs (Fig. 1a), two major clusters associated with the harvest year were observed. In addition, by exploring the samples, a cluster related to the late-harvested hazelnuts was identified, later confirmed by LC-HRMS data. Since the largest part of the information in the dataset distinguished the harvest year (Fig. 1a), data from each year were modelled separately to evaluate information related to the origin and the cultivar. To evaluate the information related to the origin and the cultivar, data from each year were modelled separately; however, because of the higher sample number and better distribution in terms of geographical origins and cultivars, only data from 2021 were included in these exploratory statistical analyses. A PCA model was created to evaluate the information about the origin and the cultivar (Figure S3). Different clusters related to the geographical origin were found in the scores plot of the first two PCs respectively describing 27.46% and 19.97% of the total variance, which are reasonable values considering that we are working with NMR spectral full profile. In particular, three groups were identified for hazelnuts from Piedmont, Emilia, and from outside Italy. As previously obtained by Bachmann et al. [6], the ^{1}H-NMR analysis allowed us to distinguish Italian hazelnuts with respect

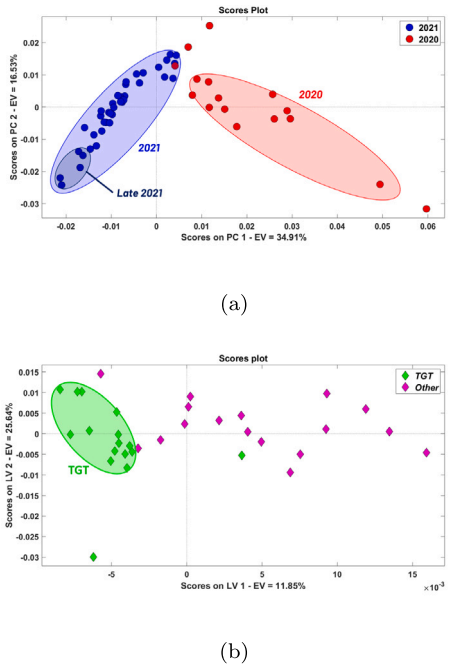


Fig. 1. ^{1}H-NMR dataset. (a) Scores plot of a PCA model grouped by harvest year. (b) Scores plot of a sPLS-DA model on 2021 samples grouped as TGT or other cultivars.

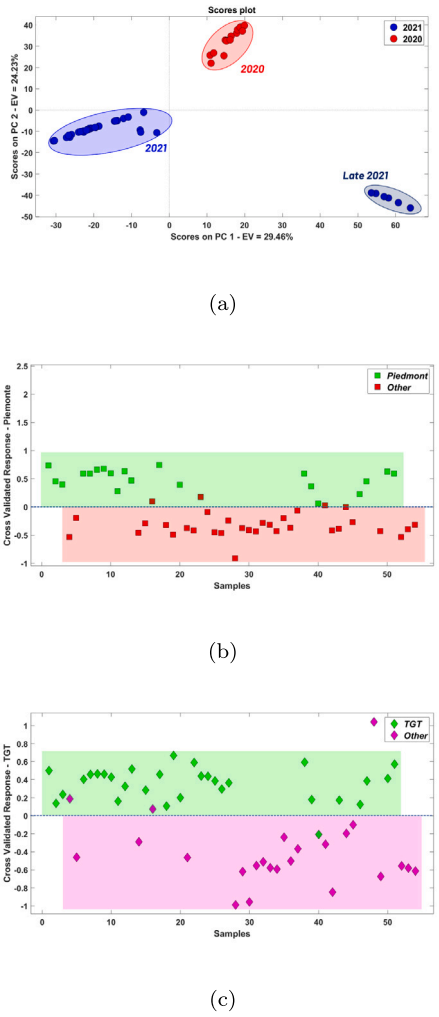


Fig. 2. LC-HRMS dataset. (a) Scores plot of a PCA model grouped by harvest year. (b) Plots of the model response in cross-validation for two sPLS-DA supervised models grouped as (b) from Piedmont or from other origin and (c) TGT or other cultivars.

to foreign samples. However, the variance of a single sample group is sometimes larger than the distance between two groups, especially for samples coming from the same country. From the scores plot of the first two PCs, it was possible to identify clusters related to TGT and Giffoni cultivars (Figure S4). Moreover, different from the results obtained by Caligiani et al. [5], a separation between Italian and foreign TGT hazelnuts was found during the exploration of the samples. Two distinct supervised models were also created to better evaluate the possibility of discriminating the TGT hazelnuts coming from Piedmont. A first PLS-DA model was developed to identify the TGT cultivar among the other cultivars from the 2021 samples. Three latent variables (LV) were selected, accounting for a Y-response explained variance of 61.45% and an X-dataset explained variance of 48.38%. By looking at the scores plot of the first two, a grouping related only to the TGT cultivar was spotted (Fig. 1b). The same approach was used to evaluate a possible separation between Piedmont hazelnuts and the others. For this PLS-DA model, three latent variables were selected, leading to a Y-response explained variance of 62.11% and an X-dataset explained variance of 50.68%. By exploring the scores plot of the first two LVs, a slight separation based on geographical origin was spotted (Figure S5). These two pieces of evidence confirmed the results previously highlighted by Bachmann et al. [6], which successfully differentiated hazelnut samples according to geographical origin and improved the results obtained by Caligiani et al. [5] about the TGT hazelnut characterization. This first exploration suggests that, for the ^{1}H-NMR spectroscopy, even if most of the information allows us to distinguish the samples according to the harvest year, it is possible to find information strictly related to the geographical origin and the cultivar of the samples. The results obtained from the two PLS-DA models highlight that, both for variety and geographical origin, half of the total explained variance (48.38% and 50.68% respectively) can be enough to successfully identified clusters related to TGT variety or Piedmont geographical origin. To better understand and evaluate the information found with the ^{1}H-NMR analyses, a tentative identification of the extracted compounds was performed on the ^{1}H-NMR spectra. The signals in the spectra were assigned by comparison with the literature [5,6,39] and with the Human Metabolome Database (HMDB), which contains metabolites also found in hazelnuts. The 28 identified metabolites are shown in Table 1, which also contains information related to the chemical shift and the multiplicity of the assigned metabolites. Other groups of signals were identified in the spectra but not assigned to any metabolites.

3.1.2. LC-HRMS dataset

The dataset obtained from LC-HRMS was explored through different statistical analyses, where the samples were labelled according to the harvest year, the geographical origin, and the cultivar. The first PCA model was performed considering all the samples. Four principal components was selected, corresponding to a total explained variance of 90.68%. Three major clusters were found, associated respectively with samples from 2020, samples from 2021 and samples late harvested in 2021. PC2 allows the separation of the sample from seasons 2020 and 2021, while PC1 discriminates the late samples from 2021 from the other 2021 hazelnuts (Fig. 2(a)). As for ^{1}H-NMR most of the information in the dataset clusters according to the harvest year, while no grouping related to other features has been found in the first PCA model. The PCA analysis revealed the presence of three large clusters, however, to build models with an acceptable amount of samples, the subsequent PLS-DA analyses were performed considering the samples from all the three clusters together. A first PLS-DA model was built considering two classes: samples from Piedmont and samples with different geographical origin. Six latent variables were selected, corresponding

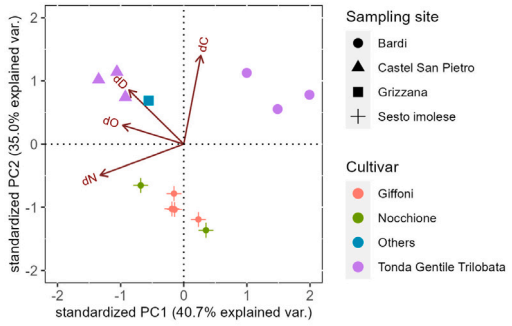
Table 1
List of the assigned metabolites with tentative names, chemical shifts (δ, ppm) and signal multiplicity.

Compound name	Chemical shift (δ, ppm)	Multiplicity	Reference
TSP	0	S	
Beta sitosterol	0.707	S	Caligiani et al. [5]
Isoleucine	0.87	t	Caligiani et al. [5]
	1.06	d	Bachmann et al. [6],Schmitt et al. [39]
Leucine	0.9	T	Caligiani et al. [5],Schmitt et al. [39]
Valine	1.015	d	Caligiani et al. [5]
	1.037	d	Bachmann et al. [6],Schmitt et al. [39]
Threonine	1.3	D	Bachmann et al. [6],Schmitt et al. [39]
Alanine	1.46	D	Caligiani et al. [5],Bachmann et al. [6],Schmitt et al. [39]
Arginine	1.72	M	Caligiani et al. [5],Schmitt et al. [39]
Acetic acid	1.932	S	Caligiani et al. [5],Bachmann et al. [6],Schmitt et al. [39]
Acetyl glutamate	2.015	s	[40]
	2.27	t	
Glutamic acid	2.16	S	Caligiani et al. [5],Schmitt et al. [39]
Malate	2.34	M	Bachmann et al. [6],Schmitt et al. [39]
Succinic acid	2.555	S	Caligiani et al. [5]
Malic acid	2.775	Dd	Caligiani et al. [5]
Asparagine	2.9	Dd	Caligiani et al. [5]
Choline derivate	3. 179	s	Caligiani et al. [5]
Choline	3. 201	s	Caligiani et al. [5],Bachmann et al. [6],Schmitt et al. [39]
Sucrose	3.44	dd	Caligiani et al. [5]
	3.615	m	Bachmann et al. [6]
	3.76	m	Schmitt et al. [39]
	4.01	t	
	4.08	s	
	4.11	s	
	5.377	d	
Glycerol	3.505	m	Caligiani et al. [5]
α-hydroxy acid	4.3	m	Caligiani et al. [5]
Ribose nucleotides	4.485	m	Caligiani et al. [5]
Oligosaccharides	5.399	m	Caligiani et al. [5]
	5.41	m	
	5.515	m	
Fumarate	6.65	s	Bachmann et al. [6]
Tyrosine	6.773	d	Caligiani et al. [5]
	7.128	m	Bachmann et al. [6]
Indole derivate	6.9	d	
	7.035	dt	
	7.283	dt	
	7.53	d	
Tryptophan	7.384	m	Caligiani et al. [5]
	7.69	d	
Trigonelline	8.043	dd	Caligiani et al. [5]
	8.841	d	
	8.889	d	
	9.178	s	
Adenosine	8.14	s	Caligiani et al. [5]
Formic acid	8.48	s	Caligiani et al. [5],Bachmann et al. [6],Schmitt et al. [39]

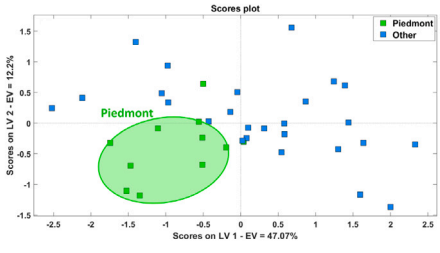
to a Y-response explained variance of 69.41% and to a X-dataset total explained variance of 93.07%. The model response plot reported in Fig. 2b confirm the possibility of discriminating hazelnuts samples from Piedmont according to their geographical origin. A second PLS-DA model was developed considering two variety-based classes: TGT variety and other cultivars. Seven LVs were selected, resulting in a Y-response explained variance of 61.28% and a X-dataset explained variance of 90.94%. The model response plot reported in Fig. 2c confirmed the possibility of discriminating TGT variety hazelnuts from other cultivars. These outcomes confirmed the possibility of finding information related to origin and cultivar using LC-HRMS analysis, even if the largest part of the information in the obtained dataset separates the samples according to the harvest year.

3.1.3. BSIA dataset

The overall stable isotope values (mean ±standard deviation) were −27.54 ±1.93, 1.45 ±1.53, 20.12 ±2.52, and −166.1 ±14.8 ‰ for $\delta^{13}C$, $\delta^{15}N$, $\delta^{18}O$ and $\delta^{2}H$, respectively. As for the other techniques, a first PCA model was performed with all samples. A clear effect of the year of sampling was observed when the individual sampling sites were considered ($p < 0.05$), while considering the whole dataset, the annual variability was hidden by the spatial variability. On the contrary, no difference was detected between early and late 2021 samples from Piedmont. The annual variability could be attributed to different climate, whose effect on stable isotopes is well known [41]. The cultivar had no clear effect on stable isotope signature, as indicated by the data collected in Emilia Romagna in 2021 that grouped according to the site of origin (Fig. 3(a)). Within Piedmont, all samples except one belonged to the cultivar TGT; therefore, it was not possible to determine the effect of this parameter. The cultivar effect on stable isotope values has been reported in a few cases [42], but only when considering small regions. Genetic variations are known to affect several metabolic pathways such as photosynthesis, which has an important role in determining the isotope composition of carbon and oxygen [43]. However, it is likely that at a wider spatial scale, the climatic and topological effect would outweigh the genetic effect. A supervised PLS-DA model was created on the 2021 samples, where a clustering of the Piedmont samples could be observed, with respect to other origins (Fig. 3(b)). However, because unbalanced samples could affect the power of the classification procedure [44], a Soft Independent Modelling of Class Analogy (SIMCA) was also applied. The results gave a good indication about the suitability of stable isotope data to confirm geographical origin, in this case, the Piedmont or other regions in Italy. In both



(a)



(b)

Fig. 3. BSIA dataset. (a) Score plot of a PCA model on the 2021 samples from the region Emilia Romagna, Italy, grouped by sampling site and cultivar. (b) Scores plot of a sPLS-DA model on 2021 samples grouped as from Piedmont or from other origin.

sampling years, 7 samples from Piedmont out of 11 were classified correctly, giving a sensitivity of 0.64. In 2020, 8 samples originated from outside Piedmont out of 11 were classified correctly. In 2021, 17 samples originated from outside Piedmont out of 27 were correctly classified. The specificity was, therefore, 0.727 in 2020 and 0.630 in 2021. The relatively low accuracy of class assignment, 0.65 on average, can be mainly ascribed to the small sample size, especially regarding the data from the area of interest. In fact, while the leave-one-out cross-validation decreases the probability of detecting outliers [45], a small sample size can affect the robustness of the SIMCA analysis [46].

3.2. Multi-omic data fusion and supervised sparse PLS-DA

Previous works suggested the great potential of data fusion coupled with chemometrics and multivariate statistical analysis in food authenticity to characterize the quality of different foodstuffs, including olive oil, honey, fish and meat [47–50]. In this study, the whole ^{1}H-NMR and LC-HRMS datasets, comprehensive of the two harvest years, were integrated using an sPLS-DA model (from DIABLO framework), and their variables contribution was underlined by the circle plots to visualize the correlations for geographical origin (Figure S6), cultivar (Figure S7), and harvest year (Figure S8). The sPLS-DA model adequately classified samples for the geographical origin and cultivar type with the minimum error rate (a one-sided t-test), using two and three components, respectively. Evident cluster points were outlined among the sparse components on the first two components, with positive correlations among the features from the two datasets. In particular, in the geographical origin model, the features resulted either positively or negatively correlated along component 1. In contrast, in component 2, some features were negatively correlated, while the correlation structure was ambiguous for others (Figure S6). The cultivar model outlined two separate clusters on the first component but not distinct on component 2 (Figure S7). Finally, in the harvest year model, the selected features generated two positive and negative clusters, separated by the first component (Figure S8).

3.2.1. Discrimination models by geographical origin

The first sPLS-DA model (Fig. 4) was constructed to discriminate samples from Piedmont with respect to other origins. As reported, there was a general agreement between LC-HRMS and ^{1}H-NMR outputs, in agreement with the good correlation coefficients between the datasets reported by the correlation circle plot, where features deriving from both datasets were characterized by similar clusterization potential (Fig. S6). This result was further confirmed by the arrow plot (Fig. 4) reporting the distribution of the samples belonging to the two datasets; the distances between the features are outlined by the arrow length that connects the same sample under the two datasets dimensions. Specifically, the features selected by the first two dimensions reported the capability to discriminate Piedmont hazelnuts from the "Others". Moreover, the hierarchical clustering analysis (Fig. 4), which was obtained from the modulation of all features selected across the two sparse components, shows that the selected features produced different clusters according to their origins, and the features across the two datasets recorded similar up or down-modulation in the sub-clustering groups. Finally, the loading plots of the first 8 (LC-HRMS) and 100 (^{1}H-NMR) selected variables for the sparse component 1 and 40 and 100 variables for sparse component 2 are represented in Fig. 4. All the features recorded an overall Pearson's correlation of r = 0.86 for component 1 and r = 0.85 for component 2. The features contributing to component 1 mainly belonged to the Piedmont hazelnuts group, characterized by p-coumaric acid ethyl ester, L-aspartic acid, moracin D, m-chlorobenzoic acid, while those contributing to component 2 were the "Others" group, including geranyl rhamnosyl-glucoside, kanokoside A and neoacrimarine K. In addition, the features differently related to the two datasets were associated with flavonoids, carbohydrates, and fatty acyls. These data agree with previous works highlighting that the chemical lipid composition, in particular, triacylglycerols but also phosphatidylcholines, phosphatidylethanolamines, diacylglycerols, and γ-tocopherol compositions [51], allow identifying hazelnuts according to their country of origin [52].

3.2.2. Discrimination by the cultivar

The second sPLS-DA model aimed to discriminate different hazelnut cultivars to distinguish TGT from the others (Fig. 5). The model separated, but did not group, the TGT samples from the "Others" according to their cultivars. In fact, these results were further confirmed by the arrow plot (Fig. 5a) representing the sample distribution, with no clear grouping according to different cultivars. However, the cultivar factor was clearly distinguished based on the contribution of the first three main components by selecting specific biomarkers that characterize the TGT and the "Others" cultivars. These features were used to build hierarchical clustering analysis, producing different clusterizations in accordance with their cultivar. The features across the two datasets recorded similar up or down-modulation in the sub-clustering groups (Fig. 5b). The loading plots reported in Figs. 5c-d show the first two main components contributing to the discrimination potential. Interestingly, these features recorded an overall Pearson's correlation of r = 0.89 for component 1, and r = 0.79 for component 2. The features contributing to component 1 mainly belonged to the "Others" hazelnuts group, while those contributing to components 2 was mainly related to TGT. The features related to the two datasets were mostly associated with flavonoids, amino acids, carbohydrates, fatty acyls, fatty acids and lipids. The main compounds characterizing "Others" (component 1) belonged to the class of lipids and included 3,4-dimethyl-5-pentyl-2-furanundecanoic acid, methyl-[12]-gingerdiol and 6,10,14-trimethyl-2-methylenepentadecanal while TGT cultivar (component 2) were characterized by p-coumaric acid

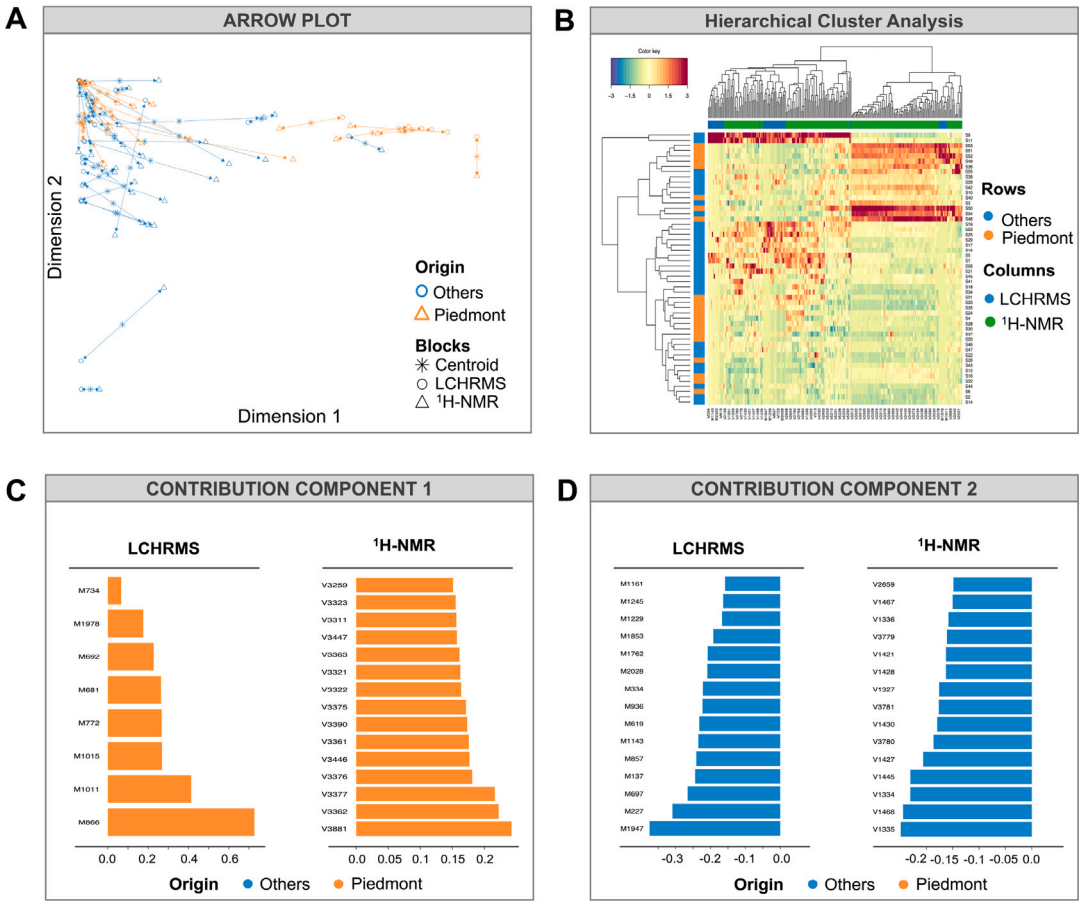


Fig. 4. DIABLO model performed based on hazelnuts' origin using ^{1}H-NMR and LC-HRMS datasets. (a) Arrow plot from integrated sPLS-DA projected into the space. The start of the arrow indicates the centroid between the datasets and the tip indicates the location of the same sample in each block. (b) Heatmap of the selected features by the sPLS-DA across the first two components; samples are in rows, and the datasets are in columns. (c) and (d) Loadings plot of the selected features by sPLS-DA on component 1, and component 2. Features are rated according to their loading weights and represented as bars; the colours indicate which class a particular compound is maximally present.

ethyl ester, aspartic acid, coumaric acid. Previous LC-HRMS-based studies identified several phenolic compounds differentially accumulated depending on the cultivars, including kaempferol, quercetin, rutin, myricetin, and catechin, and an alteration of the phenylpropanoids biosynthesis [11]. Kang and Suh [53], through metabolomics analysis, identified carbohydrates (glucose, fructose, sucrose, and xylose), amino acids (asparagine and glutamic acid), organic acids, fatty acids (oleic, α-linolenic and stearic acids) as markers associated with different Italian hazelnut genotypes. In particular, the authors reported that Tonda Gentile Trilobata from Piedmont showed the highest lipid and total polyphenols content, compared to Tonda Giffoni and Tonda Romana and compared to Chilean TGT. Noteworthy, no differences were found between TGT from Piedmont Chile at genetic level, strengthening the genotype x environment interaction as the main driver of chemical fingerprints [52].

3.2.3. Effect of the harvest year on the discrimination potential of the sPLS-DA model applied to fused data

It is well known that weather conditions can shape metabolite signatures in plants [52,54], and the multivariate analysis performed on all three individual datasets highlighted that the largest part of the information in the dataset is in fact related to the harvest year. After the success of the sPLS-DA model on fused data in highlighting the most important features discriminating geographical origins and cultivars, we decided to apply it to determine whether it is also possible to extrapolate those features most affected by annual variability. The third fusion model reported clear discrimination between the two harvest times for both LC-HRMS and ^{1}H-NMR data (Fig. 6a). This output was also confirmed by the heat map from the hierarchical clustering analysis (Fig. 6b), in which samples were grouped into clusters based on the harvest year. All features selected provided an overall Pearson's correlation of r = 0.80 for component 1 (Fig. 6c) and r = 0.85 for component 2 (Fig. 6d). Therefore, data fusion confirmed the hierarchically stronger effect of annual variability in agreement with the outcome of individual analytical approaches. Considering the features contributing to component 1, they belonged to the 2020 year, whereas those contributing to component 2 referred to the year 2021. By exploring the results, the features related to the two datasets were mostly associated with amino acids, carbohydrates, fatty acyls, polyunsaturated fatty acids, flavonoids, and lipids. It is clear that these results are not generalizable to the whole lifetime of a hazelnut grove. However, our results show that this approach can provide valuable information also to study inter-annual variability.

3.2.4. The advantages and disadvantages of DIABLO multi-technique data fusion for hazelnut authenticity

To discuss the potential benefits of the combined multi-omics metabolomic profiling of hazelnuts, it is important to point out how data were collected. In particular, a single preparation protocol was used for homogenization and extraction, starting from the same raw materials, without dedicated workflows. One of the disadvantages of using the data integration approach is the compatibility of data and analytical techniques, in which different approaches should be compatible and integrated. In this sense, the choice of analytical techniques used comes from the fact that ^{1}H-NMR and LC-HRMS are highly complementary, and their combination may increase the

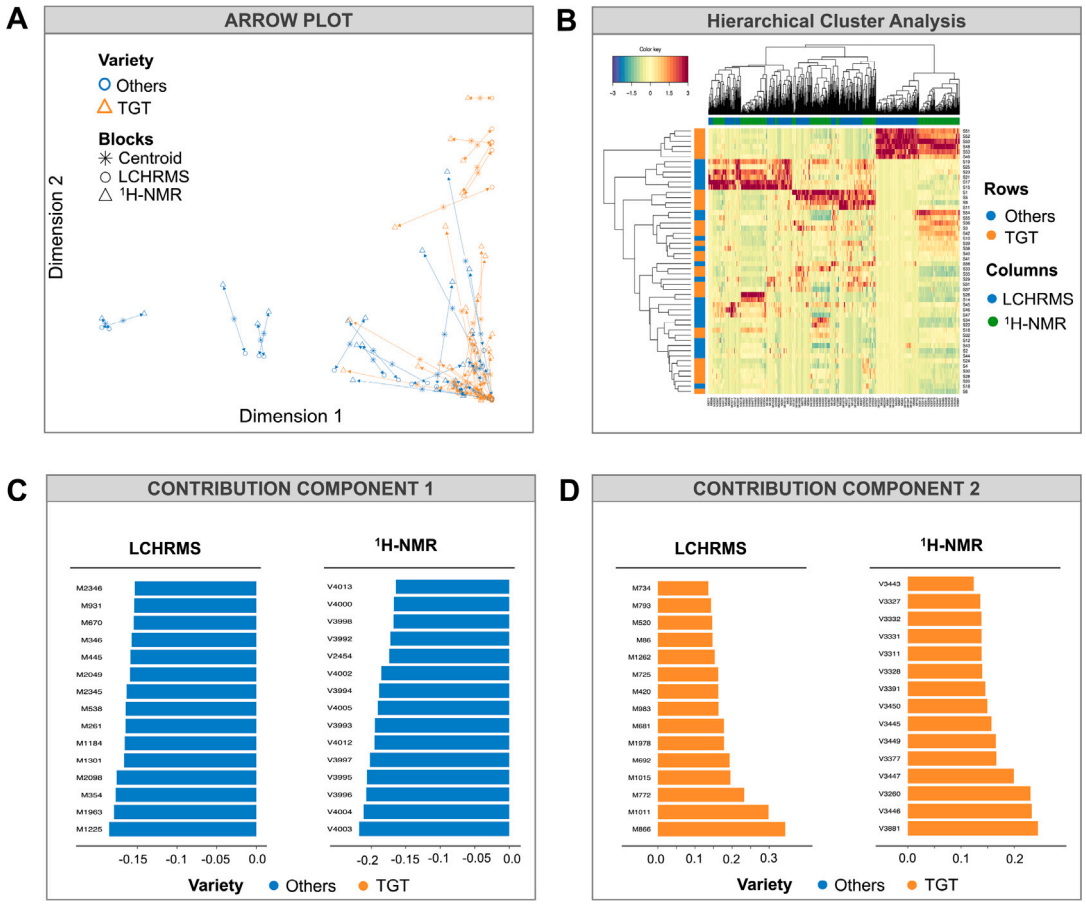


Fig. 5. DIABLO model based on hazelnuts' cultivar performed using ^{1}H-NMR and LC-HRMS datasets. (a) Arrow plot from integrated sPLS-DA projected into the space. The start of the arrow indicates the centroid between the datasets and the tip indicates the location of the same sample in each block. (b) Heatmap of the selected features by the sPLS-DA across the first two components; samples are in rows, and the datasets are in columns. (c) and (d) Loadings plot of the selected features by sPLS-DA on component 1, and component 2. Features are rated according to their loading weights and represented as bars; the colours indicate which class a particular compound is maximally present.

opportunities for improving compound coverage and delivering a more holistic metabolic profiling [55]. Indeed, ^{1}H-NMR requires minimal handling, is more quantitative, but typically detects the most abundant metabolites. Conversely, LC-HRMS has a broader compound coverage down to lower concentrations with a larger dynamic range but with challenging metabolite quantification. BSIA proved to be less suitable for data fusion with ^{1}H-NMR and LC-HRMS data. In a few cases the data fusion included also stable isotope results, but these works mostly included data sets with relatively similar dimensions, such as elemental analysis and volatile compounds concentration, in addition to stable isotopes [56,57]. In our work, the high number of features selected for data fusion prevented the inclusion of the stable isotope data. However, on the ground of the features identified in this work based on NMR and LC data, it could be possible to envision the selection of fewer, more specific features to reduce dimensionality to a more appropriate level [58]. In addition, complex biological samples can provide an overwhelming amount of data that challenges interpretations. Consequently, isotope tracer-based approaches like BSIA can significantly simplify traceability compared to metabolomics. Experimental barriers preventing data fusion could not be observed in our experiments, provided data management and chemometrics are also considered. Most works using metabolomics for traceability have been limited to a single analytical approach despite recent chemometrics tools that can accommodate data fusion from different techniques. On the one side, our findings confirmed the suitability of each separate technique while pointing out the importance of the experimental design. In this regard, it becomes fundamental to consider different seasons when the genotype *x* environment interaction is to be described for traceability purposes. Moreover, it was essential to use supervised statistics to investigate the contribution of different factors (cultivar, origin, and year of harvest, in our case) in the discrimination and identify specific vs. common marker compounds. Supervised statistics associated with an adequate sampling size have been shown to be highly valuable in the presence of confounding factors that may complicate interpretations in real situations. On the other side, our findings highlighted the excellent correlations between ^{1}H-NMR and LC-HRMS datasets, with cluster analysis demonstrating a limited overlapping and a substantial complementarity. Chen et al. [59], individually modelling ^{1}H-NMR and LC-HRMS data and then combining the scores from each analysis into a third score plot, obtained a better separation than the original ^{1}H-NMR or LC-HRMS scores alone. Notwithstanding, the sPLS-DA analysis showcases the robustness of the combined approach. Indeed, using appropriate tools for data fusion may represent a concrete advantage beyond a simple merge of data from different techniques because the latter does not consider the highly informative correlations between ^{1}H-NMR and LC-HRMS datasets. Despite looking like classical PLS models, multiblock methods enhance the sensitivity and specificity of the multiple analytical sources [60]. Accordingly, the dataset from each technique is modelled into separate "blocks", and their co-analysis is used to investigate within-block and between-block correlations [61]. Consequently, when the blocks share common patterns, such as in our study, between-block correlations will provide a better performance in terms of between-group discrimination ability. Thus, they can improve the sensitivity and specificity of biomarker discovery. Furthermore, integrating datasets coming from different techniques can reduce the noise and enhance the overall predictability of the model. However,

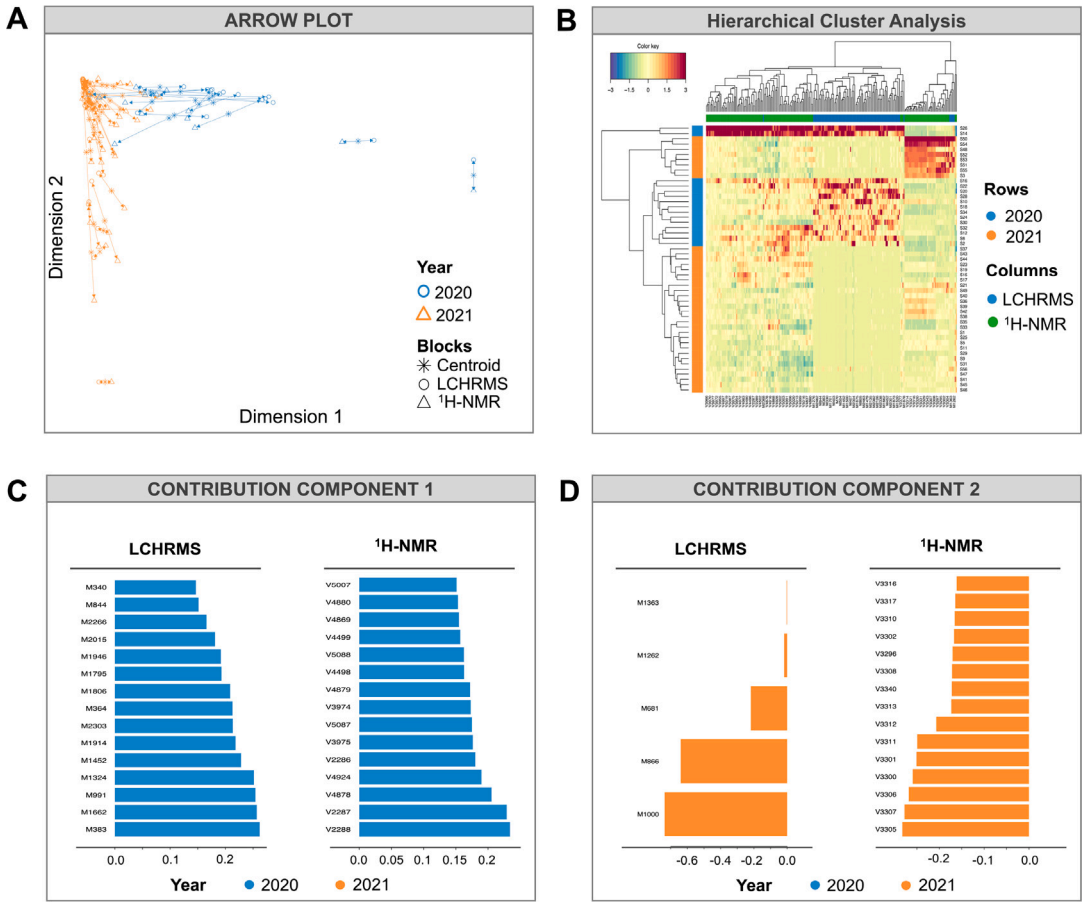


Fig. 6. DIABLO model based on year of production performed using ^{1}H-NMR and LC-HRMS datasets. (a) Arrow plot from integrated sPLS-DA projected into the space. The start of the arrow indicates the centroid between the datasets and the tip indicates the location of the same sample in each block. (b) Heatmap of the selected features by the sPLS-DA across the first two components; samples are in rows, and the datasets are in columns. (c) and (d) Loadings plot of the selected features by sPLS-DA on component 1, and component 2. Features are rated according to their loading weights and represented as bars; the colours indicate which class a particular compound is maximally present.

data overfitting could occur when a model represents the training data too well and captures noise that does not generalize to new data. To overcome this problem, we built the models after the tuning process by using k-fold cross-validation to ensure good performances when using the selected features [62].

4. Conclusions

A multi-technique unbiased platform was used, together with the intrinsic benefits of each individual analytical approach, to highlight the advantage of a multi-omics approach in ensuring the integrity of food products using hazelnuts as a case study. To the best of our knowledge, this approach has never been applied to hazelnut samples by combining ^{1}H-NMR and LC-HRMS techniques. Thus, the identification and characterization of hazelnuts biomarkers based on the metabolomics profile obtained from two different complementary techniques could be exploited to confirm the origins and the varieties of hazelnuts. In more detail, this study characterized hazelnuts of the Tonda Gentile Trilobata cultivar grown in Piedmont using three complementary analytical techniques and highlighted the differences in chemical signatures when the effects of cultivar, origin, and harvest year overlap. Although the selected features were primarily affected by the harvest year, the results within each season were comparable. This result, being based on only two subsequent harvests, is clearly not generalizable to the whole lifetime of a hazelnut grove, but it highlights that the multi-technique data-fusion approach can provide information even in the presence of such variability. The analysis of the individual databases pointed out similar conclusions. In particular, all three techniques highlighted differences between the 2020 and the 2021 harvests and, though with a different level of confidence, were able to distinguish hazelnuts grown in Piedmont from those harvested outside the region. In addition, LC-HRMS and ^{1}H-NMR could also differentiate the Tonda Gentile Trilobata from other cultivars. In the future, stable isotope analysis could be performed on specific compounds identified by the LC-HRMS and ^{1}H-NMR as indicative of origin and cultivar to incorporate into the data analysis information strictly bound to the site of origin of the produce. This approach is expected to facilitate data interpretation in the presence of confounding factors such as inter-annual and agronomic variability. Furthermore, the dataset could be further tested by adding samples from different origins and cultivars to test the robustness of data interpretation in the presence of an increasing variability of the sample. This information can be of high interest when aiming at exploiting these techniques for food security. These results indicate that data fusion increases the prediction and the classification ability with respect to individual results, despite the limitations in relation to robustness due to the small sample size. Notwithstanding, the chemometric outputs gained by the fusion of LC-HRMS and ^{1}H-NMR data allowed the identification of the most relevant features distinguishing the geographical origin and the cultivar, even considering different harvest years of hazelnut samples. In conclusion, despite the limited sample size, combining orthogonal techniques with complementary information associated with proper data fusion approaches can significantly advance the identification of the geographical and botanical (cultivar) origin of hazelnuts.

CRediT authorship contribution statement

Mattia Sozzi: Writing – original draft, Software, Formal analysis, Data curation. **Biancamaria Senizza:** Writing – original draft, Software, Formal analysis, Data curation. **Leilei Zhang:** Writing – original draft, Software, Formal analysis, Data curation. **Michele Remo Chierotti:** Writing – review & editing, Supervision. **Massimo Esposito:** Writing – review & editing, Funding acquisition. **Roberto Gobetto:** Writing – review & editing, Supervision, Conceptualization. **Luigi Lucini:** Writing – review & editing, Supervision, Conceptualization. **Francesca Scandellari:** Writing – review & editing, Writing – original draft, Supervision, Software, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Funding

This work was financially supported by Regione Piemonte, Italy (POR FESR Piemonte 2014–2020 Asse I Azione I.1b.1.2 PRISM-E; project n. J11F20000010009 CoryTeVa).

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.

Acknowledgements

The authors thank the "Romeo ed Enrica Invernizzi" foundation (Milan, Italy) for kindly supporting the metabolomics facility at Università Cattolica del Sacro Cuore. Leilei Zhang was the recipient of a PhD fellowship by the AgriSystem school at Università Cattolica del Sacro Cuore. The authors thank the team of U-Series s.r.l., who made possible the participation at the project. U-Series is accredited according to EN ISO 9001 and 17025. The authors would also like to thank the farmers and farmer associations for supplying the hazelnut samples.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.rechem.2025.102532.

Data availability

Data will be made available on request.

References

[1] G.P. Danezis, A.S. Tsagkaris, F. Camin, V. Brusic, C.A. Georgiou, Food authentication: Techniques, trends & emerging approaches, TRAC Trends Anal. Chem. 85 (2016) 123–132, http://dx.doi.org/10.1016/j.trac.2016.02.026, URL: https://linkinghub.elsevier.com/retrieve/pii/S0165993615301291.

[2] A.P. Sobolev, F. Thomas, J. Donarski, C. Ingallina, S. Circi, F. Cesare Marincola, D. Capitani, L. Mannina, Use of NMR applications to tackle future food fraud issues, Trends Food Sci. Technol. 91 (2019) 347–353, http://dx.doi.org/10.1016/j.tifs.2019.07.035, URL: https://linkinghub.elsevier.com/retrieve/pii/S092422441830147X.

[3] Q. Qu, L. Jin, Application of nuclear magnetic resonance in food analysis, Food Sci. Technol. 42 (2022) e43622, http://dx.doi.org/10.1590/fst.43622.

[4] A.-H. Emwas, R. Roy, R.T. McKay, L. Tenori, E. Saccenti, G.A.N. Gowda, D. Raftery, F. Alahmari, L. Jaremko, M. Jaremko, D.S. Wishart, NMR spectroscopy for metabolomics research, Metabolites 9 (7) (2019) 123, http://dx.doi.org/10.3390/metabo9070123.

[5] A. Caligiani, J.D. Coisson, F. Travaglia, D. Acquotti, G. Palla, L. Palla, M. Arlorio, Application of 1H NMR for the characterisation and authentication of "Tonda Gentile Trilobata" hazelnuts from Piedmont (Italy), Food Chem. 148 (2014) 77–85, http://dx.doi.org/10.1016/j.foodchem.2013.10.001, URL: https://linkinghub.elsevier.com/retrieve/pii/S0308814613014337.

[6] R. Bachmann, S. Klockmann, J. Haerdter, M. Fischer, T. Hackl, ^{1}H NMR spectroscopy for determination of the geographical origin of hazelnuts, J. Agricult. Food Chem. 66 (44) (2018) 11873–11879, http://dx.doi.org/10.1021/acs.jafc.8b03724, URL: http://pubs.acs.org/doi/10.1021/acs.jafc.8b03724.

[7] V. Maestrello, P. Solovyev, L. Bontempo, L. Mannina, F. Camin, Nuclear magnetic resonance spectroscopy in extra virgin olive oil authentication, Comp. Rev. Food Sci. Food Safe 21 (5) (2022) 4056–4075, http://dx.doi.org/10.1111/1541-4337.13005, URL: https://ift.onlinelibrary.wiley.com/doi/10.1111/1541-4337.13005.

[8] N. Cavallini, L. Strani, P. Becchi, V. Pizzamiglio, S. Michelini, F. Savorani, M. Cocchi, C. Durante, Tracing the identity of Parmigiano Reggiano "Prodotto di Montagna - Progetto Territorio" cheese using NMR spectroscopy and multivariate data analysis, Anal. Chim. Acta 1278 (2023) 341761, http://dx.doi.org/10.1016/j.aca.2023.341761, URL: https://linkinghub.elsevier.com/retrieve/pii/S0003267023009820.

[9] D. Schütz, E. Achten, M. Creydt, J. Riedl, M. Fischer, Non-targeted LC-MS metabolomics approach towards an authentication of the geographical origin of grain maize (*Zea mays* L.) samples, Foods 10 (9) (2021) 2160, http://dx.doi.org/10.3390/foods10092160, URL: https://www.mdpi.com/2304-8158/10/9/2160.

[10] R. Boiteau, D. Hoyt, C. Nicora, H. Kinmonth-Schultz, J. Ward, K. Bingol, Structure Elucidation of unknown metabolites in metabolomics by combined NMR and MS/MS prediction, Metabolites 8 (1) (2018) 8, http://dx.doi.org/10.3390/metabo8010008, URL: http://www.mdpi.com/2218-1989/8/1/8.

[11] V. Lelli, R. Molinari, N. Merendino, A.M. Timperio, Detection and comparison of bioactive compounds in different extracts of two hazelnut skin varieties, Tonda Gentile Romana and Tonda Di Giffoni, using a metabolomics approach, Metabolites 11 (5) (2021) 296, http://dx.doi.org/10.3390/metabo11050296, URL: https://www.mdpi.com/2218-1989/11/5/296.

[12] S. Ghisoni, L. Lucini, F. Angilletta, G. Rocchetti, D. Farinelli, S. Tombesi, M. Trevisan, Discrimination of extra-virgin-olive oils from different cultivars and geographical origins by untargeted metabolomics, Food Res. Int. 121 (2019) 746–753, http://dx.doi.org/10.1016/j.foodres.2018.12.052, URL: https://linkinghub.elsevier.com/retrieve/pii/S0963996918309992.

[13] M.B. Mohamed, G. Rocchetti, D. Montesano, S.B. Ali, F. Guasmi, N. Grati-Kamoun, L. Lucini, Discrimination of Tunisian and Italian extra-virgin olive oils according to their phenolic and sterolic fingerprints, Food Res. Int. 106 (2018) 920–927, http://dx.doi.org/10.1016/j.foodres.2018.02.010, URL: https://linkinghub.elsevier.com/retrieve/pii/S0963996918301030.

[14] B. Senizza, G. Rocchetti, S. Ghisoni, M. Busconi, M. De Los Mozos Pascual, J.A. Fernandez, L. Lucini, M. Trevisan, Identification of phenolic markers for saffron authenticity and origin: An untargeted metabolomics approach, Food Res. Int. 126 (2019) 108584, http://dx.doi.org/10.1016/j.foodres.2019.108584, URL: https://linkinghub.elsevier.com/retrieve/pii/S0963996919304624.

[15] A. Aguzzoni, F. Scandellari, The geographical origin of fresh horticultural products: analytical methods to prevent food frauds, Italus Hortus (24) (2017) 41–57, http://dx.doi.org/10.26353/j.itahort/2017.1.4157.

[16] J.F. Carter, L.A. Chesson (Eds.), Food Forensics: Stable Isotopes as a Guide to Authenticity and Origin, CRC Press, Taylor & Francis Group, 2017.

[17] B. Torres-Cobos, M. Rosell, A. Soler, M. Rovira, A. Romero, F. Guardiola, S. Vichi, A. Tres, Investigating isotopic markers for hazelnut geographical authentication: Promising variables and potential applications, Food Chem. 449 (2024) 139083, http://dx.doi.org/10.1016/j.foodchem.2024.139083, URL: https://linkinghub.elsevier.com/retrieve/pii/S0308814624007325.

[18] D.A. Magdas, A. Dehelean, I. Feher, S. Radu, Isotopic and multielemental fingerprinting of organically and conventionally grown potatoes, Isot. Environ. Heal. Stud. 53 (6) (2017) 610–619, http://dx.doi.org/10.1080/10256016.2017.1335722, URL: https://www.tandfonline.com/doi/full/10.1080/10256016.2017.1335722.

[19] A. Aguzzoni, M. Bassi, E. Pignotti, P. Robatscher, F. Scandellari, W. Tirler, M. Tagliavini, Sr isotope composition of Golden Delicious apples in Northern Italy reflects the soil ^{87}Sr/^{86}Sr ratio of the cultivation area, J. Sci. Food Agric. (2020) jsfa.10399, http://dx.doi.org/10.1002/jsfa.10399, URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jsfa.10399.

[20] J. Van De Steene, J. Ruyssinck, J.-A. Fernandez-Pierna, L. Vandermeersch, A. Maes, H. Van Langenhove, C. Walgraeve, K. Demeestere, B. De Meulenaer, L. Jacxsens, B. Miserez, Fingerprinting methods for origin and variety assessment of rice: development, validation and data fusion experiments, Food Control 151 (2023) 109780, http://dx.doi.org/10.1016/j.foodcont.2023.109780, URL: https://linkinghub.elsevier.com/retrieve/pii/S0956713523001809.

[21] M.P. Callao, I. Ruisánchez, An overview of multivariate qualitative methods for food fraud detection, Food Control 86 (2018) 283–293, http://dx.doi.org/10.1016/j.foodcont.2017.11.034, URL: https://linkinghub.elsevier.com/retrieve/pii/S0956713517305674.

[22] S. Hassani, U. Dackermann, M. Mousavi, J. Li, A systematic review of data fusion techniques for optimized structural health monitoring, Inf. Fusion 103 (2024) 102136, http://dx.doi.org/10.1016/j.inffus.2023.102136, URL: https://linkinghub.elsevier.com/retrieve/pii/S1566253523004529.

[23] X. Lin, S. Chao, D. Yan, L. Guo, Y. Liu, L. Li, Multi-sensor data fusion method based on self-attention mechanism, Appl. Sci. 13 (21) (2023) 11992, http://dx.doi.org/10.3390/app132111992, URL: https://www.mdpi.com/2076-3417/13/21/11992.

[24] A. Biancolillo, M. Foschi, A.A. D'Archivio, Geographical classification of Italian Saffron (*Crocus sativus* L.) by multi-block treatments of UV-Vis and IR spectroscopic data, Molecules 25 (10) (2020) http://dx.doi.org/10.3390/molecules25102332.

[25] A. Rivera-Pérez, R. Romero-González, A.G. Frenich, Application of an innovative metabolomics approach to discriminate geographical origin and processing of black pepper by untargeted UHPLC-Q-Orbitrap-HRMS analysis and mid-level data fusion, Food Res. Int. 150 (2021) 110722, http://dx.doi.org/10.1016/j.foodres.2021.110722, URL: https://www.sciencedirect.com/science/article/pii/S0963996921006219.

[26] INC International Nuts & Dried Fruit, Hazelnuts global statistical review, 2023, URL: https://inc.nutfruit.org/hazelnuts-global-statistical-review-2/.

[27] F. Ortega-Gavilán, S. Squara, C. Cordero, L. Cuadros-Rodríguez, M.G. Bagur-González, Application of chemometric tools combined with instrument-agnostic GC-fingerprinting for hazelnut quality assessment, J. Food Comp. Anal. 115 (2023) 104904, http://dx.doi.org/10.1016/j.jfca.2022.104904, URL: https://linkinghub.elsevier.com/retrieve/pii/S0889157522005221.

[28] L. Lucini, G. Rocchetti, M. Trevisan, Extending the concept of terroir from grapes to other agricultural commodities: an overview, Curr. Opin. Food Sci. 31 (2020) 88–95, http://dx.doi.org/10.1016/j.cofs.2020.03.007, URL: https://linkinghub.elsevier.com/retrieve/pii/S2214799320300266.

[29] S. Ghisoni, L. Lucini, G. Rocchetti, G. Chiodelli, D. Farinelli, S. Tombesi, M. Trevisan, Untargeted metabolomics with multivariate analysis to discriminate hazelnut (*Corylus avellana* L.) cultivars and their geographical origin, J. Sci. Food Agric. 100 (2) (2020) 500–508, http://dx.doi.org/10.1002/jsfa.9998, URL: https://onlinelibrary.wiley.com/doi/10.1002/jsfa.9998.

[30] F. Savorani, G. Tomasi, S. Engelsen, icoshift: A versatile tool for the rapid alignment of 1D NMR spectra, J. Magn. Reson. 202 (2010) 190–202, http://dx.doi.org/10.1016/j.jmr.2009.11.012.

[31] F. Savorani, G. Tomasi, S. Engelsen, Alignment of 1D NMR data using the iCoshift Tool: A tutorial, in: J. Van Duynhoven, P.S. Belton, G.A. Webb, H. Van As (Eds.), Special Publications, Royal Society of Chemistry, Cambridge, 2013, pp. 14–24, http://dx.doi.org/10.1039/9781849737531-00014, URL: http://ebook.rsc.org/?DOI=10.1039/9781849737531-00014.

[32] R.M. Salek, S. Neumann, D. Schober, J. Hummel, K. Billiau, J. Kopka, E. Correa, T. Reijmers, A. Rosato, L. Tenori, P. Turano, S. Marin, C. Deborde, D. Jacob, D. Rolin, B. Dartigues, P. Conesa, K. Haug, P. Rocca-Serra, S. O'Hagan, J. Hao, M. Van Vliet, M. Sysi-Aho, C. Ludwig, J. Bouwman, M. Cascante, T. Ebbels, J.L. Griffin, A. Moing, M. Nikolski, M. Oresic, S.-A. Sansone, M.R. Viant, R. Goodacre, U.L. Günther, T. Hankemeier, C. Luchinat, D. Walther, C. Steinbeck, COordination of Standards in MetabOlomicS (COSMOS): facilitating integrated metabolomics data access, Metabolomics 11 (6) (2015) 1587–1597, http://dx.doi.org/10.1007/s11306-015-0810-y, URL: http://link.springer.com/10.1007/s11306-015-0810-y.

[33] G. Rocchetti, G. Chiodelli, G. Giuberti, F. Masoero, M. Trevisan, L. Lucini, Evaluation of phenolic profile and antioxidant capacity in gluten-free flours, Food Chem. 228 (2017) 367–373, http://dx.doi.org/10.1016/j.foodchem.2017.01.142, URL: https://linkinghub.elsevier.com/retrieve/pii/S0308814617301541.

[34] D. Ballabio, A MATLAB toolbox for Principal Component Analysis and unsupervised exploration of data structure, Chemometr. Intell. Lab. Syst. 149 (2015) 1–9, http://dx.doi.org/10.1016/j.chemolab.2015.10.003, URL: https://linkinghub.elsevier.com/retrieve/pii/S0169743915002476.

[35] K. Nordhausen, S. Sirkia, H. Oja, D.E. Tyler, ICSNP: Tools for multivariate nonparametrics, 2018, URL: https://CRAN.R-project.org/package=ICSNP.

[36] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: linear models. PLS-DA, Anal. Methods 5 (16) (2013) 3790, http://dx.doi.org/10.1039/c3ay40582f, URL: http://xlink.rsc.org/?DOI=c3ay40582f.

[37] S. Kucheryavskiy, mdatools – R package for chemometrics, Chemometr. Intell. Lab. Syst. 198 (2020) 103937, http://dx.doi.org/10.1016/j.chemolab.2020.103937, URL: https://linkinghub.elsevier.com/retrieve/pii/S0169743919305672.

[38] S. Kucheryavskiy, O. Rodionova, A. Pomerantsev, Procrustes cross-validation of multivariate regression models, Anal. Chim. Acta 1255 (2023) 341096, http://dx.doi.org/10.1016/j.aca.2023.341096, URL: https://linkinghub.elsevier.com/retrieve/pii/S0003267023003173.

[39] C. Schmitt, T. Schneider, L. Rumask, M. Fischer, T. Hackl, Food Profiling: Determination of the geographical origin of walnuts by ^{1}H NMR spectroscopy using the polar extract, J. Agric. Food Chem. 68 (52) (2020) 15526–15534, http://dx.doi.org/10.1021/acs.jafc.0c05827, URL: https://pubs.acs.org/doi/10.1021/acs.jafc.0c05827.

[40] D.S. Wishart, A. Guo, E. Oler, F. Wang, A. Anjum, H. Peters, R. Dizon, Z. Sayeeda, S. Tian, B.L. Lee, M. Berjanskii, R. Mah, M. Yamamoto, J. Jovel, C. Torres-Calzada, M. Hiebert-Giesbrecht, V.W. Lui, D. Varshavi, D. Varshavi, D. Allen, D. Arndt, N. Khetarpal, A. Sivakumaran, K. Harford, S. Sanford, K. Yee, X. Cao, Z. Budinski, J. Liigand, L. Zhang, J. Zheng, R. Mandal, N. Karu, M. Dambrova, H.B. Schiöth, R. Greiner, V. Gautam, HMDB 5.0: the human metabolome database for 2022, Nucleic Acids Res. 50 (D1) (2022) D622–D631, http://dx.doi.org/10.1093/nar/gkab1062.

[41] P. Ghosh, W.A. Brand, Stable isotope ratio mass spectrometry in global climate change research, Int. J. Mass Spectrom. 228 (1) (2003) 1–33, http://dx.doi.org/10.1016/S1387-3806(03)00289-6, URL: https://linkinghub.elsevier.com/retrieve/pii/S1387380603002896.

[42] K.B. Bat, R. Vidrih, M. Neemer, B.M. Vodopivec, I. Muli, P. Kump, N. Ogrinc, Characterization of Slovenian apples with respect to their botanical and geographical origin and agricultural production practice, Food Technol. Biotechnol. 50 (1) (2012) 107–116.

[43] Y. Scheidegger, M. Saurer, M. Bahn, R. Siegwolf, Linking stable oxygen and carbon isotopes with stomatal conductance and photosynthetic capacity: a conceptual model, Oecologia 125 (3) (2000) 350–357, http://dx.doi.org/10.1007/s004420000466, URL: http://link.springer.com/10.1007/s004420000466.

[44] R.M. Alonso-Salces, J.M. Moreno-Rojas, M.V. Holland, F. Reniero, C. Guillou, K. Héberger, Virgin olive oil authentication by multivariate analyses of ^{1}H NMR fingerprints and δ^{13}C and δ^{2}H data, J. Agric. Food Chem. 58 (9) (2010) 5586–5596, http://dx.doi.org/10.1021/jf903989b, URL: https://pubs.acs.org/doi/10.1021/jf903989b.

[45] O.Y. Rodionova, A.V. Titova, A.L. Pomerantsev, Discriminant analysis is an inappropriate method of authentication, TRAC Trends Anal. Chem. 78 (2016) 17–22, http://dx.doi.org/10.1016/j.trac.2016.01.010, URL: https://linkinghub.elsevier.com/retrieve/pii/S0165993615302193.

[46] R. Vitale, M. Cocchi, A. Biancolillo, C. Ruckebusch, F. Marini, Class modelling by Soft Independent Modelling of Class Analogy: why, when, how? A tutorial, Anal. Chim. Acta 1270 (2023) 341304, http://dx.doi.org/10.1016/j.aca.2023.341304, URL: https://linkinghub.elsevier.com/retrieve/pii/S0003267023005251.

[47] M. Ottavian, L. Fasolato, L. Serva, P. Facco, M. Barolo, Data fusion for food authentication: Fresh/Frozen–Thawed discrimination in West African Goatfish (*Pseudupeneus prayensis*) fillets, Food Bioprocess Technol. 7 (4) (2014) 1025–1036, http://dx.doi.org/10.1007/s11947-013-1157-x, URL: http://link.springer.com/10.1007/s11947-013-1157-x.

[48] S. Schwolow, N. Gerhardt, S. Rohn, P. Weller, Data fusion of GC-IMS data and FT-MIR spectra for the authentication of olive oils and honeys—is it worth to go the extra mile? Anal. Bioanal. Chem. 411 (23) (2019) 6005–6019, http://dx.doi.org/10.1007/s00216-019-01978-w, URL: http://link.springer.com/10.1007/s00216-019-01978-w.

[49] C. Robert, W. Jessep, J.J. Sutton, T.M. Hicks, M. Loeffen, M. Farouk, J.F. Ward, W.E. Bain, C.R. Craigie, S.J. Fraser-Miller, K.C. Gordon, Evaluating low- mid- and high-level fusion strategies for combining Raman and infrared spectroscopy for quality assessment of red meat, Food Chem. 361 (2021) 130154, http://dx.doi.org/10.1016/j.foodchem.2021.130154, URL: https://linkinghub.elsevier.com/retrieve/pii/S0308814621011602.

[50] Y. Hong, N. Birse, B. Quinn, Y. Li, W. Jia, P. McCarron, D. Wu, G.R. Da Silva, L. Vanhaecke, S. Van Ruth, C.T. Elliott, Data fusion and multivariate analysis for food authenticity analysis, Nat. Commun. 14 (1) (2023) 3309, http://dx.doi.org/10.1038/s41467-023-38382-z, URL: https://www.nature.com/articles/s41467-023-38382-z.

[51] S. Klockmann, E. Reiner, R. Bachmann, T. Hackl, M. Fischer, Food fingerprinting: Metabolomic approaches for geographical origin discrimination of hazelnuts (*Corylus avellana*) by UPLC-QTOF-MS, J. Agric. Food Chem. 64 (48) (2016) 9253–9262, http://dx.doi.org/10.1021/acs.jafc.6b04433, URL: https://pubs.acs.org/doi/10.1021/acs.jafc.6b04433.

[52] M. Locatelli, J.D. Coïsson, F. Travaglia, E. Cereti, C. Garino, M. D'Andrea, A. Martelli, M. Arlorio, Chemotype and genotype chemometrical evaluation applied to authentication and traceability of "Tonda Gentile Trilobata" hazelnuts from Piedmont (Italy), Food Chem. 129 (4) (2011) 1865–1873, http://dx.doi.org/10.1016/j.foodchem.2011.05.134, URL: https://linkinghub.elsevier.com/retrieve/pii/S0308814611008375.

[53] M.J. Kang, J.H. Suh, Metabolomics as a tool to evaluate nut quality and safety, Trends Food Sci. Technol. 129 (2022) 528–543, http://dx.doi.org/10.1016/j.tifs.2022.11.002, URL: https://linkinghub.elsevier.com/retrieve/pii/S0924224422004381.

[54] M. Di Nunzio, Hazelnuts as source of bioactive compounds and health value underestimated food, Curr. Res. Nutr. Food Sci. 7 (1) (2019) 17–28, http://dx.doi.org/10.12944/CRNFSJ.7.1.03, URL: http://www.foodandnutritionjournal.org/volume7number1/hazelnuts-as-source-of-bioactive-compounds-and-a-health-value-underestimated-food/.

[55] Z. Pan, D. Raftery, Comparing and combining NMR spectroscopy and mass spectrometry in metabolomics, Anal. Bioanal. Chem. 387 (2) (2007) 525–527, http://dx.doi.org/10.1007/s00216-006-0687-8, URL: http://link.springer.com/10.1007/s00216-006-0687-8.

[56] G. Sammarco, M. Rossi, M. Suman, D. Cavanna, L. Viotto, P. Pettenà, C. Dall'Asta, P. Iacumin, Hazelnut products traceability through combined isotope ratio mass spectrometry and multi-elemental analysis, JSFA Rep. 3 (12) (2023) 633–645, http://dx.doi.org/10.1002/jsf2.171, URL: https://onlinelibrary.wiley.com/doi/10.1002/jsf2.171.

[57] K. Soni, R. Frew, B. Kebede, Multi-source data fusion for soybean origin traceability: Stable isotopes, elemental composition, & volatile organic compounds, Food Chem. 485 (2025) 144497, http://dx.doi.org/10.1016/j.foodchem.2025.144497, URL: https://linkinghub.elsevier.com/retrieve/pii/S0308814625017480.

[58] E. Borràs, J. Ferré, R. Boqué, M. Mestres, L. Aceña, O. Busto, Data fusion methodologies for food and beverage authentication and quality assessment – A review, Anal. Chim. Acta 891 (2015) 1–14, http://dx.doi.org/10.1016/j.aca.2015.04.042, URL: https://linkinghub.elsevier.com/retrieve/pii/S0003267015005528.

[59] H. Chen, Z. Pan, N. Talaty, D. Raftery, R.G. Cooks, Combining desorption electrospray ionization mass spectrometry and nuclear magnetic resonance for differential metabolomics without sample preparation, Rapid Comm. Mass Spectrom. 20 (10) (2006) 1577–1584, http://dx.doi.org/10.1002/rcm.2474, URL: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/10.1002/rcm.2474.

[60] J.A. Westerhuis, T. Kourti, J.F. MacGregor, Analysis of multiblock and hierarchical PCA and PLS models, J. Chemom. 12 (5) (1998) 301–321, http://dx.doi.org/10.1002/(SICI)1099-128X(199809/10)12:5<301::AID-CEM515>3.0.CO;2-S, URL: https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1099-128X(199809/10)12:5%3C301::AID-CEM515%3E3.0.CO;2-S.

[61] D.D. Marshall, R. Powers, Beyond the paradigm: Combining mass spectrometry and nuclear magnetic resonance for metabolomics, Prog. Nucl. Magn. Reson. Spectrosc. 100 (2017) 1–16, http://dx.doi.org/10.1016/j.pnmrs.2017.01.001, URL: https://linkinghub.elsevier.com/retrieve/pii/S0079656516300656.

[62] F. Rohart, B. Gautier, A. Singh, K.-A. Lê Cao, mixOmics: An R package for 'omics feature selection and multiple data integration, in: D. Schneidman (Ed.), PLoS Comput. Biol. 13 (11) (2017) e1005752, http://dx.doi.org/10.1371/journal.pcbi.1005752, URL: https://dx.plos.org/10.1371/journal.pcbi.1005752.