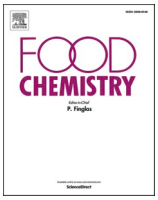
Food Chemistry 441 (2024) 138294

Contents lists available at ScienceDirect

Food Chemistry

journal homepage: www.elsevier.com/locate/foodchem



Prospective exploration of hazelnut's unsaponifiable fraction for geographical and varietal authentication: A comparative study of advanced fingerprinting and untargeted profiling techniques

B. Torres-Cobos [a,b], B. Quintanilla-Casas [c], M. Rovira [d], A. Romero [d], F. Guardiola [a,b], S. Vichi [a,b,*], A. Tres [a,b]

[a] University of Barcelona, Department of Nutrition, Food Sciences and Gastronomy, Prat de la Riba 171, Santa Coloma de Gramenet 08921, Spain
[b] University of Barcelona, Institute of Research on Food Nutrition and Safety (INSA-UB), Prat de la Riba 171, Santa Coloma de Gramenet 08921, Spain
[c] University of Copenhagen, Department of Food Science, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark
[d] Institute of Agrifood Research and Technology (IRTA), Ctra. de Reus – El Morell Km 3.8, Constantí 43120, Spain

ARTICLE INFO

Keywords:
Hazelnut
Geographical and varietal authentication
Unsaponifiable fraction
Fingerprinting
Untargeted profiling
PLS-DA

ABSTRACT

This study compares two data processing techniques (fingerprinting and untargeted profiling) to authenticate hazelnut cultivar and provenance based on its unsaponifiable fraction by GC–MS. PLS-DA classification models were developed on a selected sample set (n = 176). As test cases, cultivar models were developed for "Tonda di Giffoni" vs other cultivars, whereas provenance models were developed for three origins (Chile, Italy or Spain). Both fingerprinting and untargeted profiling successfully classified hazelnuts by cultivar or provenance, revealing the potential of the unsaponifiable fraction. External validation provided over 90 % correct classification, with fingerprinting slightly outperforming. Analysing PLS-DA models' regression coefficients and tentatively identifying compounds corresponding to highly relevant variables showed consistent agreement in key discriminant compounds across both approaches. However, fingerprinting in selected ion mode extracted slightly more information from chromatographic data, including minor discriminant species. Conversely, untargeted profiling acquired in full scan mode, provided pure spectra, facilitating chemical interpretability.

1. Introduction

Hazelnuts are widely used raw, roasted and as a key ingredient in food and confectionery products, adding flavour and texture to various sweet and savoury products. They rank third in the global nut market with a production volume higher than one million tons per year (FAOstat, 2021). The main hazelnut producing countries are Turkey (63.5 %), Italy (7.9 %), United States (6.5 %), Azerbaijan (6.3 %), Georgia (4.3 %) and Chile (3.3 %), followed by China, Iran, France and Spain (<3% each) (FAOstat, 2021). Among the most prominent cultivars are 'Tombul', 'Palaz', 'Çakildak' (Turkey); 'Tonda di Giffoni', 'Tonda Romana', 'Tonda Gentile delle Langhe' (Italy); 'Negret' and 'Pauetet' (Spain) (Król & Gantner, 2020). Sensory and qualitative attributes of hazelnuts are strongly influenced by varietal and geographical factors (Amaral et al., 2006; Król & Gantner, 2020; Parcerisa, Richardson, Rafecas, Codony, & Boatella, 1998). Their prices can also vary greatly based on their cultivar and geographical origin (FAOstat, 2021), being higher for hazelnuts with special geographical indications such as Protected Designation of Origin (PDO) or Protected Geographical Indication registered in the European Union. The great value of hazelnuts makes them susceptible to economically motivated fraud, which is further aggravated by the growth of emerging nut producing countries, the expansion of markets and the lack of effective fraud detection methods. All these factors contribute to a growing vulnerability that counterfeiters can exploit. Hence, having suitable tools to verify the cultivar and origin of hazelnuts is crucial to guarantee their authenticity and to protect the consumer.

In this regard, phenotypic observations based on physical characteristics are currently used for this purpose. However, the fact that they are susceptible to external influences and can only be used on whole kernels (Ciarmiello et al., 2014; Król & Gantner, 2020) limits their efficiency. For this reason, several studies to explore more suitable tools for hazelnut authentication have been carried out in the last decade.

* Corresponding author at: Departament de Nutrició, Ciències de l'Alimentació i Gastronomia, Campus De l'Alimentació Torribera, Facultat de Farmàcia i Ciències de l'Alimentació, Universitat de Barcelona, Av Prat de la Riba, 171, 08921 Santa Coloma de Gramenet, Spain.
E-mail address: stefaniavichi@ub.edu (S. Vichi).

https://doi.org/10.1016/j.foodchem.2023.138294
Received 30 June 2023; Received in revised form 22 December 2023; Accepted 26 December 2023
Available online 4 January 2024

DNA-based methods offer high accuracy (Lang et al., 2021), but their complex and expensive procedures make them unsuitable for routine analysis. Furthermore, both phenotypic and DNA-based methods are limited to varietal authentication only. In contrast, methods for geographical authentication of nuts often involve analysing the mineral composition (Inaudi et al., 2020; Oddone, Aceto, Baldizzone, Musso, & Osella, 2009) or isotope ratios of light or heavy elements (Krauß, Vieweg, & Vetter, 2019; Zannella et al., 2017) associated with the growing area. Unfortunately, these methods are not suitable for determining the cultivar. Similarly, models based on near infrared spectrometry (Biancolillo et al., 2018; Sammarco & Dall'Asta, & Suman, 2023) have been suggested to verify the geographical origin of some Italian PDOs. Alternatively, methods based on the analysis of hazelnut metabolites have been proposed to verify their varietal or geographical origin, given that metabolomics is a state-of-the-art approach for food authentication. These approaches analyse protein/peptide compounds, phenolic profiles and components of the lipid fraction, measured by chromatographic techniques, such as gas (Parcerisa et al., 1998; Tüfekci and Karataş, 2018) and liquid chromatography (Ciarmiello et al., 2014; Ghisoni et al., 2020; Klockmann, Reiner, Bachmann, Hackl, & Fischer, 2016) as well as proton nuclear magnetic resonance (^{1}H NMR) (Bachmann, Klockmann, Haerdter, Fischer, & Hackl, 2018). Nevertheless, some of these markers such as phenols may be unstable under certain conditions (light, temperature, time) and none of the methods developed so far have been tested for their suitability in verifying both the hazelnut's cultivar and origin. This underscores the need reliable methods that can fulfil this objective.

In the pursuit of appropriate candidates for hazelnut cultivar and geographical markers, the unsaponifiable lipid fraction stands out for presenting relatively stable metabolites under storage conditions, which are known to be influenced by both genetic (Amaral et al., 2006; Matthäus & Özcan, 2012; Parcerisa et al., 1998) and environmental factors (Benitez-Sanchez, León-Camacho, & Aparicio, 2003; Ghisoni et al., 2020; Matthäus & Özcan, 2012). This rich fraction contains several families of secondary metabolites such as linear and terpene alcohols and hydrocarbons, sterols, methylsterols and dimethylsterols, among others (Benitez-Sanchez et al., 2003; Goriainov et al., 2021). This makes the unsaponifiable lipid fraction a promising candidate for the geographical and varietal authentication of hazelnuts. The most appropriate technique for its analysis is gas chromatography-mass spectrometry (GC–MS) (Goriainov et al., 2021; Phillips, Ruggio, & Ashraf-Khorassani, 2005), as it provides comprehensive molecular-level information through three-way data (an array sized of intensity × retention time × *m/z,* for each sample), together with a high sensitivity and widespread availability in routine labs.

In addition to selecting the appropriate authentication markers and analytical technique, an essential aspect of the authentication strategy is the data analysis approach, especially when dealing with complex chromatograms such as those from nut's unsaponifiable lipid fraction. Untargeted approaches are an advantageous alternative to conventional targeted methods, as they provide more information and overcome the difficulties of identifying and quantifying analytical compounds in complex chromatograms (Quintanilla-Casas et al., 2020a). In fact, untargeted methods coupled to chemometric pattern recognition techniques, such as partial least squares discriminant analysis (PLS-DA), proved to be efficient tools for authentication purposes (Quintanilla-Casas et al., 2020a,b; Riedl, Esslinger, & Fauhl-Hassek, 2015; Torres-Cobos et al., 2021).

Among untargeted methods, fingerprinting operates on high dimensional data (i.e. two-way or three-way data) such as spectra or chromatograms and consists on finding specific patterns, known as fingerprints, which are unique to a specific characteristic of the food sample, such as cultivar or geographical origin (Ballin & Laursen, 2019; Bosque-Sendra, Cuadros-Rodríguez, Ruiz-Samblás, & de la Mata, 2012). Fingerprinting methods have been widely tested for food authentication, proving to be successful (Quintanilla-Casas et al., 2020a,b; Torres-Cobos et al., 2021). Fingerprinting of three-way data, such as GC–MS data (intensity × retention time × *m/z*, for each sample), typically entails complex multi-way chemometric algorithms, but a recently introduced approach simplifies the process by transforming the data into a manageable two-way format (retention time × intensity, for each ion and sample). This process involves the creation of models using the unfolded matrix of extracted chromatograms of specific ions and has proven successful for authentication purposes (Quintanilla-Casas et al., 2020a,b; Torres-Cobos et al., 2021, 2023). Another alternative for analysing three-way GC–MS data is using advanced untargeted profiling techniques, such as powerful deconvolution tools, to extract the maximum information from the samples (Rinnan, Amigo, & Skov, 2014). Among them, a deconvolution and identification tool called PARADISe has emerged for GC–MS, which is based on PARAllel FACtor analysis 2 (PARAFAC2). PARAFAC2 models provide estimates for each mode – relative concentration, elution profile and pure mass spectra – for each analyte, while handling common issues in chromatographic data, including co-elution, baseline variations and retention time shifts (Johnsen, Skou, Khakimov, & Bro, 2017; Baccolo, Quintanilla-Casas, Vichi, Augustijn, & Bro, 2021). This user-friendly software allows an efficient untargeted analysis of large GC–MS datasets, while minimizing inter-user variability. Several studies have reported its usefulness for chromatographic datasets in different fields (Johnsen et al., 2017; Baccolo et al., 2021; Ríos-Reina, Aparicio-Ruiz, Morales, & García-Gonzalez, 2023; Sales, Portolés, Johnsen, Danielsen, & Beltran, 2019). Baccolo et al. (2021) evidenced its advantages in time-saving, comprehensiveness of the chromatographic results and tentative identification over manual profiling. To our knowledge, no comparisons have been made to evaluate the efficiency of fingerprinting and untargeted profiling approaches, along with other deconvolution methods, in extracting information from chromatographic data for authentication purposes.

This study aims to explore the potential of the unsaponifiable fraction for the cultivar and geographical authentication of hazelnuts and to compare two different data processing techniques (fingerprinting and untargeted profiling) to determine the best method for developing efficient authentication models. For this purpose, PLS-DA classification models based on GC–MS data of hazelnut unsaponifiable fraction were developed using both approaches. These were applied to the same test cases: a cultivar model to distinguish 'Tonda di Giffoni' (TG), one of the most widespread cultivars in the world, from other hazelnut cultivars (non-TG); a provenance model to discriminate three different countries of origin (Spain, Chile and Italy). Finally, the regression coefficients of the models have been explored with the sole purpose of gaining a deeper understanding of the models and their chemical significance, to ensure that models are based on genuine chemical information rather than arbitrary randomness.

2. Material and methods

2.1. Material and reagents

Diethyl ether stabilized with 7 mg/L of BHT, anhydrous sodium sulphate and anhydrous pyridine 99.5 % were purchased from Scharlau (Sentmenat, Spain). Methanol for gas chromatography ECD and FID SupraSolv® and Horning's silylating mixture II (N,O-bis(trimethylsilyl) acetamide/chlorotrimethylsilane/1-(trimethylsilyl)imidazole, 3:2:3, v/v/v) were purchased from Merck (Darmstadt, Germany). Potassium hydroxide 85 % for analysis in pellets form was purchased from Thermo Scientific (Waltham, Massachusetts, USA) and amberlite IRN78 OH hydroxide form from Supelco (Bellefonte, Pennsylvania, USA).

2.2. Sampling

The sample set consisted of 176 traceable hazelnuts collected over two consecutive harvest years, 2019 and 2020 (Supplementary material, Table S1). They were obtained in the framework of the TRACENUTS

project (PID2020-117701RB100), directly from producers. Out of these samples, 110 were of the TG cultivar from Chile (n = 40), Italy (n = 24) and Spain (n = 46), while 66 were from different cultivars (non-TG) produced in Spain. Samples were stored vacuum-packed at 4 °C until analysis.

2.2.1. Sample preparation

Around 30 g of hazelnuts were grinded and their lipid fraction was extracted using 50 mL of diethyl ether. The mixture was centrifuged at 1220 *g* for 10 min, the liquid phase was taken and the organic solvent was evaporated with a rotatory evaporator until only the hazelnut oil was left. Then, an aliquot of 1 g of the hazelnut oil was saponified by adding 4 mL of 2 M methanolic potassium hydroxide solution and heated for 30 min at 70 °C in a water bath. The reaction was quenched with ice for 10 min and 10 mL of water were added. Once the sample reached room temperature, the unsaponifiable fraction was extracted with 3 x 10 mL of diethyl ether, centrifuging each time (1220 *g*; 10 min) to separate the organic phase from the aqueous phase. The organic extracts were subsequently pooled and washed with 10 mL of distilled water. Following this, 2 g of amberlite adsorbent were added to remove the excess of dissolved free fatty acids. After removing the adsorbent, the organic phase was washed again with 10 mL of water and anhydrous sodium sulphate was added to remove any remaining moisture. Once the extract was cleaned, purified and any residual water was removed, the solvent was evaporated using a rotatory evaporator until the volume was reduced to approximately 1 mL. The resulting solution was transferred into a silylation tube, and the remaining solvent was evaporated to dryness by applying a stream of N_2. The dry unsaponifiable fraction was reconstituted with 50 μL of pyridine. Finally, 100 μL of silylating reagent were added and allowed to react for 20 min at room temperature prior to injection.

2.3. Gas chromatography-mass spectrometry (GC–MS)

The samples were analysed by an Agilent 6890 N Network GC system equipped with a Combi-pal autosampler (CTC Analytics, Zwingen, Switzerland) and coupled to an Agilent 5975C Inert MSD quadrupolar mass selective analyser (Agilent Technologies, Santa Clara, California, USA). Helium was the carrier gas at a flow rate of 1.5 mL/min. Analytes were separated on a ZB-5 ms capillary column (60 m × 0.25 mm i.d., 0.25 μm film thickness) from Phenomenex (Torrance, California, USA). Column temperature was initially held at 150 °C for 2 min, then increased to 260 °C at a rate of 10 °C/min, held for 2 min and then increased to 320 °C at 2 °C/min, holding the last temperature for 13 min. The ion source and the transfer line were set at 230 and 300 °C, respectively. Mass spectra were acquired at 1.9 scan/s and the electron energy was set at 70 eV. For the untargeted profiling approach, data acquisition was performed in the full scan mode within the 50–500 *m/z* range. For the fingerprinting approach, data were acquired using selected ion monitoring (SIM) of 15 ions that resulted characteristic of several compound families of the unsaponifiable fraction: *m/z* 57 (linear hydrocarbons); *m/z* 69, 81, 93 (terpene alcohols and hydrocarbons); *m/z* 73 (silylated compounds, i.e. any compound with a hydroxy group); *m/z* 75, 103 (linear alcohols); *m/z* 83, 117 (fatty acids); *m/z* 117, 129, 189, 199, 204, 218, 393 (sterols) and *m/z* 189, 204, 218 (4-methylsterols and 4,4-dimethylsterols) (Li, Beveridge, & Drover, 2007; Xu et al., 2018). Ions with *m/z* 75, 81, 83, 93, 103 and 117 were acquired from 14.85 to 42.5 min; *m/z* 129, 57, 69 and 73 were acquired from 14.85 to 58 min; *m/z* 189, 199, 204, 218 and 393 were acquired from 42.5 to 58 min.

2.4. Chemometrics

2.4.1. SIM fingerprinting approach

A fingerprinting approach was followed using the extracted ion chromatograms (EIC) of the 15 selected ions. The intensities of the scans from minute 14.85 to 58 (4903 scans per selected ion over 43 min) were considered for ions *m/z* 129, 57, 69 and 73; from 14.85 to 42.5 min for ions *m/z* 75, 81, 83, 93, 103 and 117; and from 42.5 to 58 min for ions *m/z* 189, 199, 204, 218 and 393 (4903 scans × 4 ions + 3171 scans × 6 ions + 1785 scans × 5 ions = 47563 variables per sample). A data matrix was built for each ion, with the scan intensities of each EIC (columns) for all samples (rows). Then, the EICs of each ion matrix were aligned using the Correlation Optimized Warping (COW) algorithm in Matlab® (Nielsen, Carstensen, & Smedsgaard, 1998) to correct the retention time shifts among samples. Finally, the 15 aligned EIC matrices were concatenated conforming a two-way unfolded matrix (176 samples × 47563 variables).

2.4.2. Untargeted profiling approach using PARADISe

The GC–MS raw dataset, acquired in full scan mode from 14.85 to 58 min, was imported in PARADISe and aligned to solve peak shifts using the automatic alignment tool that applies icoshift (Larsen, Van den Berg, & Engelsen, 2006) and COW algorithms (Tomasi, Van den Berg, & Andersson, 2004). Even if raw data alignment is not required, it eases the subsequent interval selection for each of the peaks. The optimal number of components of PARAFAC2 models for each interval was determined based on the combination of the following parameters: high number of true peaks based on the deep learning tool, high model fitting, high core consistency indicating better model adequacy, as well as low and random model residuals (Quintanilla-Casas, Bro, Hinrich, & Davie-Martin, 2023). After excluding baseline noise and other interferences, components corresponding to actual chemical compounds (Fig. 1) were selected and exported to a peak table (Supplementary material, Table S2), listing the intervals (Supplementary material, Table S3) and the peak areas (relative concentration) for all exported peaks in each sample, together with the resolved mass spectra. The data matrix used for further analysis only contained the peak areas for all exported peaks (155 columns) for each sample (176 rows).

2.4.3. Partial least squares discriminant analysis (PLS-DA) models

Principal Component Analysis (PCA) was performed on each dataset [fingerprinting (47563 columns × 176 rows) and untargeted profiling through PARADISe (155 columns × 176 rows)] to explore the data and to identify any potential outliers according to the Hotelling's T^2 range and model residuals.

Then, each data matrix (from fingerprinting and from untargeted profiling through PARADISe) was used to calibrate and validate independent PLS-DA classification models with SIMCA v13.0© (Umetrics AB, Sweden) to discriminate between: i) cultivars of Spanish hazelnuts; and ii) geographical origins of TG hazelnuts.

For the cultivar models, a binary PLS-DA model was applied to discriminate TG samples from those of other cultivars (non-TG). To eliminate any potential influences from other factors, only samples from the same origin (Spain) were considered (n = 112; 46 Spanish TG hazelnuts and 66 Spanish non-TG hazelnuts).

For the classification according to the geographical origin of TG hazelnuts, two discrimination approaches were developed, using hazelnut samples from the same cultivar (TG) obtained from different geographical origins (n = 110). The first geographical approach aimed to discriminate between 'European' (EU) (n = 70) and 'non-European' (non-EU: Chile) (n = 40) TG samples; whereas the second model aimed to classify the 70 TG EU samples into their specific country of origin: 'Spain (n = 46) or 'Italy' (n = 24).

For each type of authentication model, each sample set was randomly divided into training (80 % of the samples, TG/non-TG model n = 90, EU/non-EU model n = 88, Spain/Italy model n = 56) and validation set (20 % of the samples, TG/non-TG model n = 22, EU/non-EU model n = 22, Spain/Italy model n = 14). This splitting was run seven times (7 iteration for authentication model type) to evaluate the effect of the sample set composition and to increase the robustness of the external validation.

In PLS-DA binary models, classes are expressed as PLS dummy

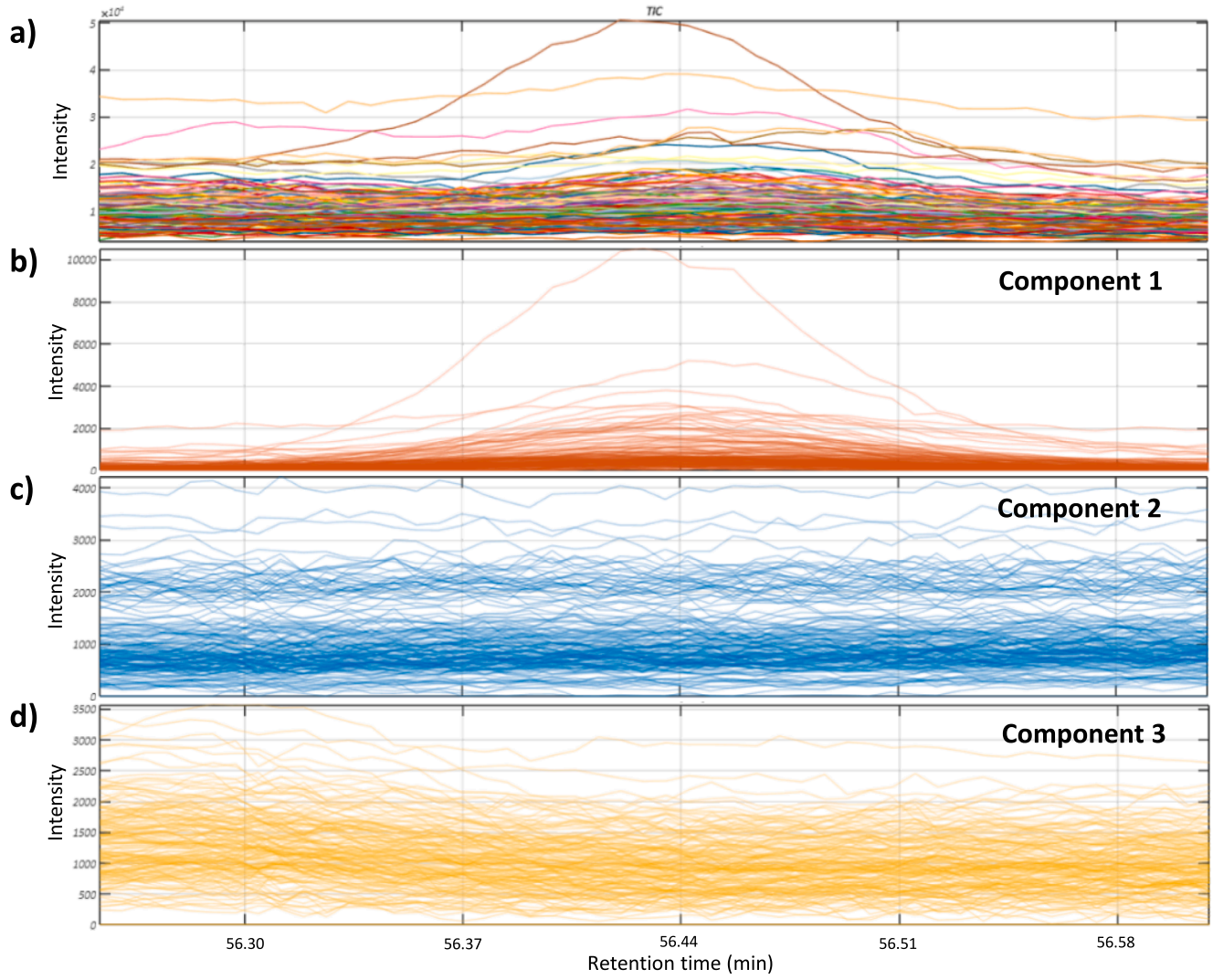


Fig. 1. Plot of the TIC interval between 56.2 and 56.6 min against the PARADISe extracted component plots. a) TIC plot, b) component 1 (orange) corresponding to chemical compound 31, c) component 2 (blue) is baseline noise, d) component 3 (yellow) is baseline noise. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

variables (here, 1 for 'non-TG', 'non-EU' and 'Italy' classes and 0 for 'TG', 'EU' and 'Spain' classes). Then, the PLS predicted value of each sample is used for its classification into one class or the other according to a classification threshold (predicted value = 0.5). In each iteration, models were first internally validated using the training set of samples through leave 10 %-out cross-validation, and the optimal number of latent variables (LV) was selected according to the lowest Root Mean Squared Error of Cross Validation (RMSEcv) criteria. The optimal pre-processing, according to the criteria below, for all the models was mean centring and scaling to the unit of variance. Permutation test (n = 20 permutations) and ANOVA on the cross-validated predictive residuals (p-value) were carried out to assess the models' overfitting (Supplementary material, Table S4). Then, the models were externally validated by predicting the class of the samples in the corresponding validation set, which had not been used to build the models. The suitability of each PLS-DA model was evaluated by the Q^2 values and efficiency, which was expressed as the percentage of correct classification of each class, and the sensitivity (true positives/ [true positives + false negatives]) and specificity (true negatives/ [true negatives + false positives]) values, positive samples being the non-TG, non-EU and Italian samples for the corresponding models. The performance of models from each data processing approach (fingerprint and untargeted profiling through PARADISe) was compared to determine the most suitable one for authentication.

2.4.4. Evaluation of PLS-DA regression coefficients

The regression coefficients of the PLS-DA models developed with all samples in the corresponding sample sets (cultivar n = 112, or origin n = 110) with both the fingerprinting and untargeted profiling approaches were compared to tentatively identify the key variables that contributed to the discrimination between classes. This comparison aimed to reveal the variables that were relevant for both approaches or for only one of the approaches. The jack-knife standard error of cross-validation (SEcv) was used to evaluate the significance of the regression coefficients, considering significant those with values higher than their corresponding SEcv (Torres-Cobos et al., 2021). Out of the significant variables, only the ones with the highest absolute values (25 % higher than the coefficient media) were considered and the corresponding compounds were tentatively identified based on their mass spectra and elution order.

Table 1
External validation of PLS-DA models ('Tonda di Giffoni vs non-Tonda di Giffoni; 'European' vs 'non-European' (Chilean samples) and 'Spanish' vs 'Italian') developed on the fingerprinting and untargeted profiling through PARADISe. Results are mean values (±standard deviation) obtained from seven iterations.

Cultivar model: TG/non-TG						
Fingerprinting (LVs = 7, $Q^2 > 0.64$, RMSEcv < 0.30) [a]						
	n	non-TG (n)	TG (n)	Correct classification (%)	Sensitivity	Specificity
non-TG	13	11.9 ± 0.7	1.1 ± 0.7	91.2 ± 5.3	0.91 ± 0.05	
TG	9	1.0 ± 1.0	8.0 ± 1.0	88.9 ± 11.1		0.89 ± 0.11
Total	22			**90.3 ± 6.7**		
PARADISe (LVs = 6–7, $Q^2 > 0.40$, RMSEcv < 0.36) [a]						
	n	**non-TG (n)**	**TG (n)**	**Correct classification (%)**	**Sensitivity**	**Specificity**
non-TG	13	10.7 ± 1.1	2.3 ± 1.1	82.4 ± 8.6	0.82 ± 0.09	
TG	9	0.9 ± 0.9	8.1 ± 0.9	90.5 ± 10.0		0.91 ± 0.10
Total	22			**85.7 ± 4.9**		
Geographical origin model: EU/non-EU						
Fingerprinting (LVs = 6–7, $Q^2 > 0.72$, RMSEcv < 0.29) [a]						
	n	**non-EU (n)**	**EU (n)**	**Correct classification (%)**	**Sensitivity**	**Specificity**
non-EU	8	8.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0	1.00 ± 0.00	
EU	14	0.1 ± 0.4	13.9 ± 0.4	99.0 ± 2.7		0.99 ± 0.03
Total	22			**99.4 ± 1.7**		
PARADISe (LVs = 5–6, $Q^2 > 0.60$, RMSEcv < 0.29) [a]						
	n	**non-EU (n)**	**non-EU (n)**	**Correct classification (%)**	**Sensitivity**	**Specificity**
non-EU	8	7.6 ± 0.8	0.4 ± 0.8	94.6 ± 9.8	0.95 ± 0.10	
EU	14	0.1 ± 0.4	13.9 ± 0.4	99.0 ± 2.7		0.99 ± 0.03
Total	22			**97.4 ± 3.6**		
Geographical origin model: Spanish/Italian						
Fingerprinting (LVs = 5–6, $Q^2 > 0.62$, RMSEcv < 0.31) [a]						
	n	**ITA (n)**	**ESP (n)**	**Correct classification (%)**	**Sensitivity**	**Specificity**
ITA	5	4.6 ± 0.5	0.4 ± 0.5	91.4 ± 10.7	0.91 ± 0.11	
ESP	9	0.3 ± 0.5	8.7 ± 0.5	96.8 ± 5.4		0.97 ± 0.05
Total	14			**94.9 ± 3.5**		
PARADISe (LVs = 4–5, $Q^2 > 0.67$, RMSEcv < 0.28) [a]						
	n	**ITA (n)**	**ESP (n)**	**Correct classification (%)**	**Sensitivity**	**Specificity**
ITA	5	4.9 ± 0.4	0.1 ± 0.4	97.1 ± 7.6	0.97 ± 0.08	
ESP	9	0.3 ± 0.5	8.7 ± 0.5	96.8 ± 5.4		0.97 ± 0.05
Total	14			**96.9 ± 3.8**		

For all models, ANOVA p-value < 0.05. [a] Model parameters: mean values obtained with the training sets from 7 iterations. TG: 'Tonda di Giffoni'; non-TG: other cultivars; EU: European (Spanish and Italian); non-EU: non-European (Chilean); ESP: Spanish hazelnuts; ITA: Italian hazelnuts.

3. Results

3.1. Performance of PLS-DA classification models: fingerprinting vs. untargeted profiling data

All models built on training sets (7 iterations per authentication model type) from both approaches achieved 100 % of correct classification in leave 10 %-out cross validation, which corresponds to the maximum value of sensitivity and specificity (sensitivity and specificity = 1) (Supplementary material, Figure S1) (Supplementary material, Table S4). Subsequently, the PLS-DA models developed on each approach were used to predict the class of the samples conforming the corresponding validation sets. Table 1 presents the mean values of the seven iterations obtained from the external validation of each authentication model type (Cultivar: TG/non-TG; Geographical origin: EU/non-EU and Spain/Italy) developed on fingerprinting and untargeted profiling approaches. No outliers were detected according to the Hotelling's T^2 range and model residuals.

In the case of cultivar authentication, the fingerprinting model outperformed the untargeted profiling one with higher sensitivity (0.91 vs 0.82) and total correct classification percentage (90.3 % vs 85.7 %). Although the specificity of the untargeted profiling model was slightly higher (0.91 vs 0.89), the fingerprinting model performed better overall.

Fingerprinting models were also more efficient in distinguishing between EU and non-EU samples, because even if both approaches achieved the same specificity (0.99), the fingerprinting model exhibited a higher sensitivity (1 vs 0.95) and overall correct classification (99.4 % vs 97.4 %).

Finally, in terms of classification by EU country, both approaches presented the same specificity (0.97) but untargeted profiling models discriminated better the Italian samples, with a higher sensitivity (0.97 vs 0.91, by arbitrarily considering Italian hazelnuts as the positive samples) and overall correct classification (96.9 % vs 94.9 %).

Regarding the standard deviation of mean external validation results calculated on 7 iterations, it ranged from 1.7 to 6.7 % and from 3.6 to 4.9 % in global correct classification by fingerprinting and untargeted profiling models, respectively, being the TG/non-TG and the EU/non-EU the models with the highest and lowest standard deviations, respectively, in both approaches.

3.2. Exploring models through regression coefficient analysis

To study and compare the most informative variables in PLS-DA models developed on the fingerprinting or untargeted profiling (PARADISe) data, we assessed the corresponding regression coefficients. The most relevant coefficients of PLS-DA models based on both approaches corresponded to variables distributed throughout the entire chromatogram. For the fingerprinting approach, they were present in the EICs of all the ions considered (Fig. 2) and several of them corresponded to minor components (Fig. 3).

Although a targeted approach was not the aim of the present study, we tentatively identified some of the most discriminant compounds

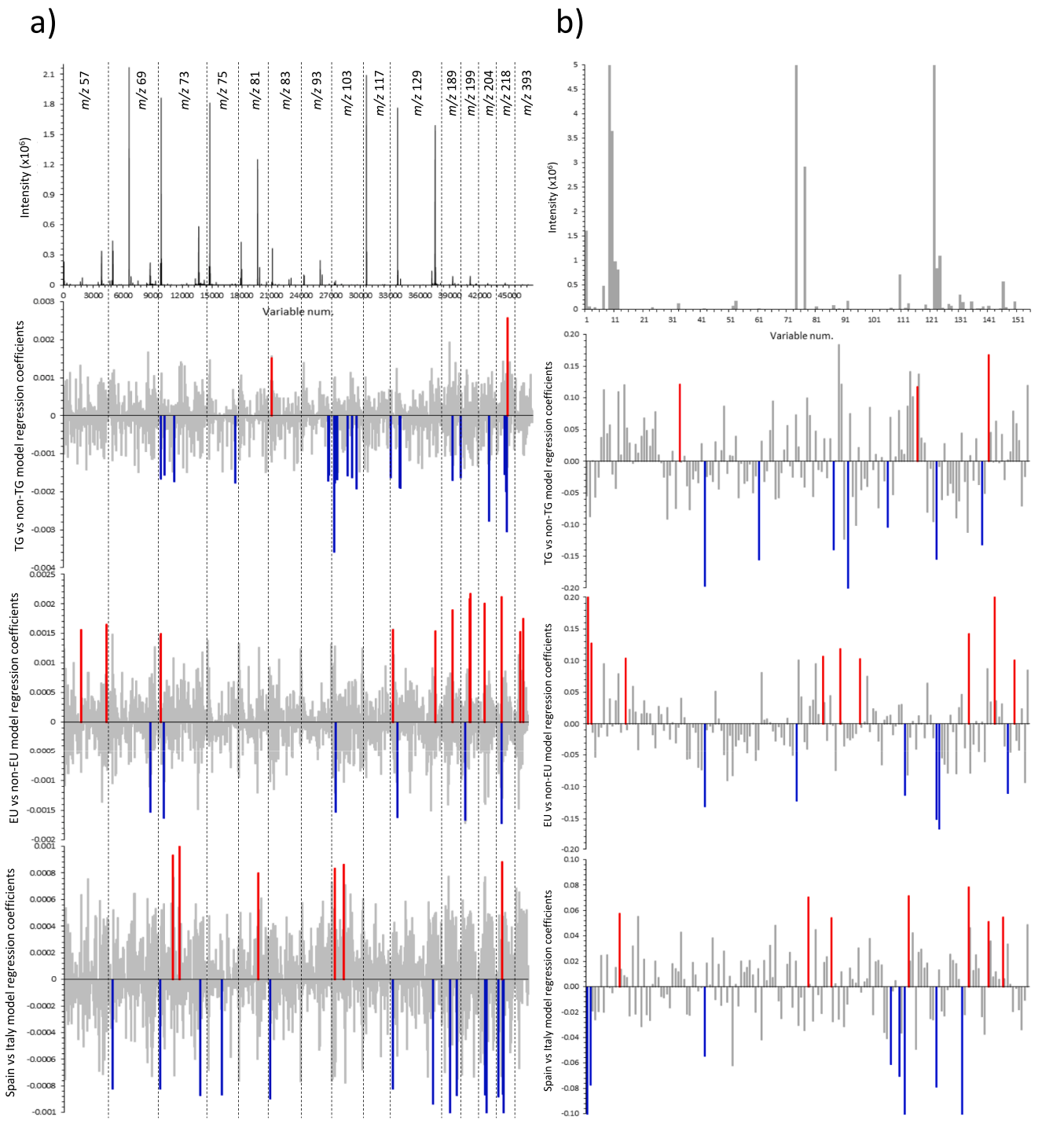
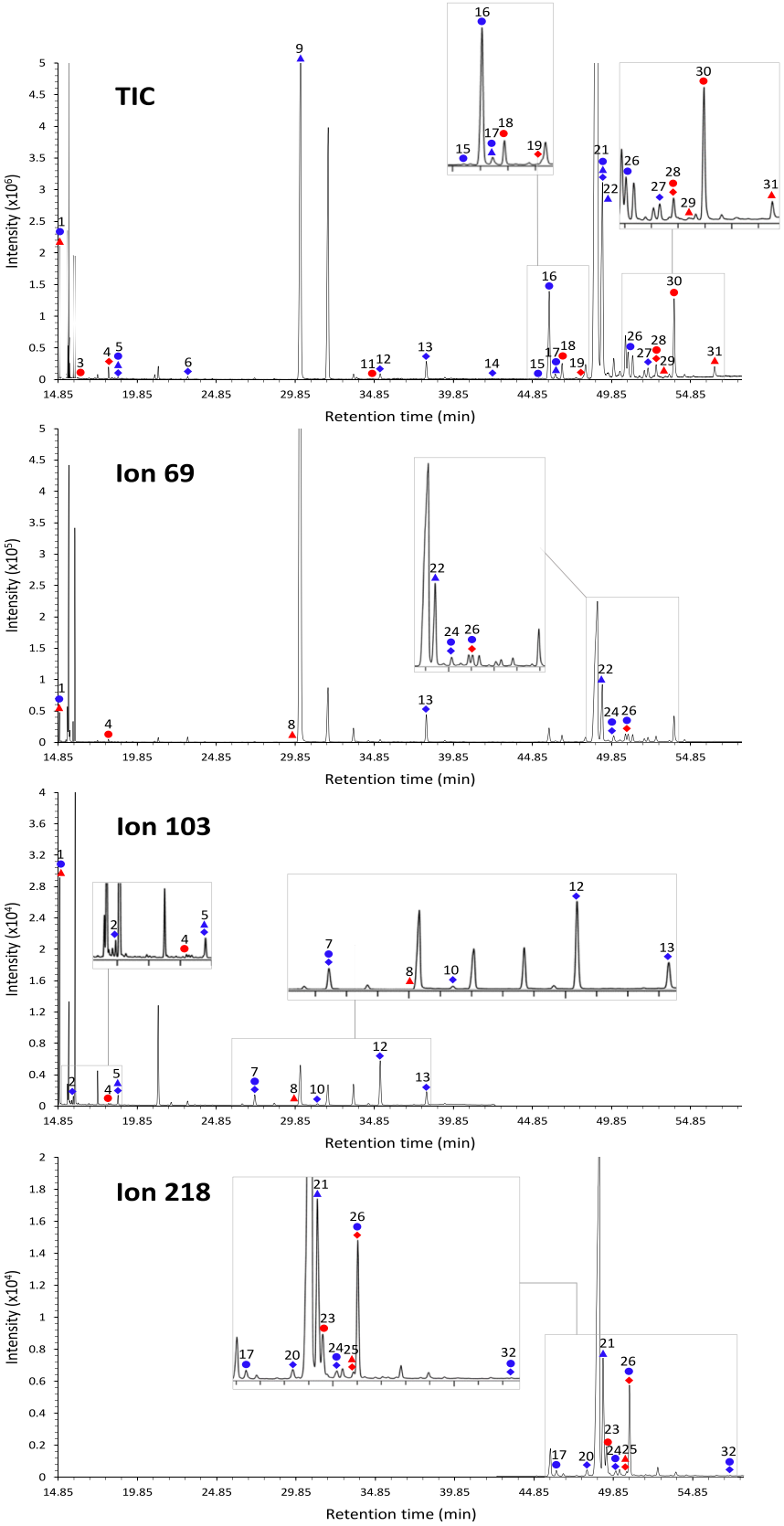


Fig. 2. Regression coefficients of the PLS-DA models ('Tonda di Giffoni -TG vs 'other cultivars' - non-TG; 'European' - EU vs 'non-European' – non-EU and 'Spain' – ESP vs 'Italy' - ITA) developed by a) fingerprinting, plotted against the variables (acquisition points) of concatenated EICs of a Tonda di Giffoni Spanish sample and b) PARADISe, plotted against the variables (detected compounds) of the TIC of the same sample. For each model, the most relevant coefficients for the prediction of the TG, EU and ESP classes are highlighted in blue (negative coefficients) and those relevant for non-TG, non-EU and ITA in red (positive coefficients). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

according to their elution order and the MS spectra (identification levels 2a and 3 according to Schymanski et al., 2014) obtained from full scan MS data (Li et al., 2007; Xu et al., 2018). Specifically, we tentatively identified the variables that were more relevant in classifying the samples within each category of the three authentication models (TG/non-TGs; EU/non-EU; Spain/Italy) (Table 2). The mass spectra of these compounds matched with those of a fatty acid (*FA*), linear (*LA*) and terpene alcohols (*TA*), sterols (*S*), methylsterols (*MS*) and dimethylsterols (*DMS*) previously described in hazelnut oil (Table 2). Additionally, some relevant coefficients corresponded to compounds that could not be linked to any specific structure or chemical family with sufficient confidence and are reported as unknown compounds.

3.2.1. *Cultivar model: TG/non-TG*

Concerning the cultivar model, the regression coefficients with higher absolute value for the TG class in fingerprinting model belonged



(caption on next page)

Fig. 3. Gas chromatograms of the MS response of the unsaponifiable fraction of hazelnuts: Total Ion Chromatogram (TIC) and three representative extracted ions (69, 113, 218). 1) phytol (*TA1*), 2) lineal alcohol C17 (*LA1*), 3) terpene alcohol (*TA2*), 4) unknown (*UNK1*), 5) unknown (*UNK2*), 6) terpene alcohol (*TA3*), 7) lineal alcohol C24 (*LA2*), 8) unknown (*UNK4*), 9) squalene (*TT1*), 10) lineal alcohol C25 (*LA3*), 11) terpene alcohol (*TA4*), 12) lineal alcohol C26 (*LA4*), 13) *cis*-farnesol (*TA5*), 14) lineal alcohol C28 (*LA5*), 15) unknown (*UNK6*), 16) campesterol (*S1*), 17) campestanol (*S2*), 18) stigmasterol (*S3*), 19) sterol (*S4*), 20) unknown (*UNK7*), 21) sitostanol (*S5*), 22) Δ5-avenasterol (*S6*), 23) unknown (*UNK8*), 24) dimethylsterol (*DMS1*), 25) Δ7-stigmastenol (*S7*), 26) lupeol (*DMS2*), 27) 28-methylobtusifoliol (*MS1*), 28) 24-methylenecycloartanol (*DMS3*), 29) dimethyl sterol (*DMS4*), 30) citrostadienol (*MS2*), 31) sterol (*S8*), 32) unknown (*UNK9*). Diamond blue: TG, red: non-TG; Triangle blue: EU, red: non-EU; Circle blue: SPA, red: ITA. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

to the EIC of *m/z* 103, 204 and 218, with the first exhibiting the greatest number of relevant coefficients within this category (Fig. 2). These coefficients may correspond to linear alcohols (*m/z* 103: C17, *LA1*; C25, *LA3;* C26, *LA4*), a terpene alcohol (*m/z* 69, 103: *cis*-farnesol, *TA5*), a not identified dimethylsterol at 50.0 min (*m/z* 218: dimethylsterol 1, *DMS1*) and three unknown compounds eluting at 18.6 min (*m/z* 73, 103 and 129: Unknown 2, *UNK2*), 48.4 min (*m/z* 218: Unknown 7, *UNK7*), and 57.2 min (*m/z* 189: Unknown 9, *UNK9*), respectively (Blue diamonds in EICs from Fig. 3; Table 2). The non-TG class showed the most relevant coefficients in the EIC of *m/z* 83 and 218 (Fig. 2), which could correspond to a fatty acid (*m/z* 83: an oleic acid isomer, *FA1*), a sterol (*m/z* 218; Δ7-stigmastenol, *S7*) and a 4,4-dimethylsterol (*m/z* 218: lupeol, *DMS2*) (Red diamonds in EICs from Fig. 3; Table 2).

In the case of untargeted profiling cultivar model, pure spectra revealed that the most significant coefficients for the TG class corresponded to: the same unknown compound eluting at 18.6 min (Unknown 2, *UNK2*), terpene alcohols (a not identified one eluting at 23.1 min: terpene alcohol 3, *TA3*) and *cis*-farnesol *TA5*), linear alcohols (C26, *LA4*; C28, *LA5*), a sterol (sitostanol, *S5*) and a 4-methylsterol (28-methylobtusifoliol, *MS1*) (Blue diamonds in TIC from Fig. 3, Table 2). Some of these compounds presented *m/z* that had also shown to be relevant for the fingerprint model. For the non-TG class, the most relevant coefficients corresponded to an unknown compound eluting at 18.1 min (Unknown 1, *UNK1*), a not identified sterol eluting at 47.7 min (Sterol, 4 *S4*) and a 4,4-dimethylsterol (24-methylenecycloartanol, *DMS3*) (Red diamonds in TIC from Fig. 3, Table 2).

3.2.2. Geographical origin model: EU/non-EU

For the EU/non-EU model based on fingerprinting data, the most relevant regression coefficients of for the EU category were in the EIC of *m/z* 69, 73, 103, 129, 199 and 218 (Fig. 2), and corresponded to the above-mentioned unknown compound eluting at 18.6 min (*m/z* 73, 103, 129: Unknown 2, *UNK2*), and to sterols (*m/z* 218: sitostanol, *S5*; *m/z* 69, 199: Δ5-avenasterol, *S6*)(Blue triangles in EICs from Fig. 3, Table 2). The most relevant regression coefficients for the non-EU class were distributed in several EIC (Fig. 2). Some of these discriminant variables corresponded to: a terpene alcohol (*m/z* 69, 103, 129: phytol, *TA1*), an unknown specie eluting at 30.0 min (*m/z*: 57: Unknown 4, *UNK4*) a sterol (*m/z* 393: Δ7-stigmastenol, *S7*) and a not identified 4,4-dimethylsterol eluting at 53.3 min (*m/z* 57, 129, 189, 199, 393: dimethylsterol 4; *DMS4*) (Red triangles in EICs from Fig. 3, Table 2).

For the untargeted profiling model, the relevant coefficients for the EU class corresponded to the previously mentioned unknown compound eluting at 18.6 min (Unknown 2, *UNK2*), a triterpene hydrocarbon (squalene, *TT1*) and three sterols (campestanol, *S2*; sitostanol, *S5;* Δ5-avenasterol, *S6*) (Blue triangles in the TIC of Fig. 3, Table 2). The important coefficients for the non-EU class, in this case, corresponded to: a terpene alcohol (phytol, *TA1*), the above-mentioned not identified dimethylsterol 4 (*DMS4*) and a not identified sterol eluting at 56.4 min (Sterol 8, *S8*) (Red triangles in the TIC of Fig. 3, Table 2).

3.2.3. Geographical origin model for TG samples: Spain/Italy

Regarding the discrimination of TG hazelnuts by their origin from Spain or Italy, for the model based on fingerprinting approach the most significant coefficients for the Spain class were detected in EIC of *m/z* 69, 73, 75, 83, 129, 189, 204 and 218 (Fig. 2). They could correspond to: a terpene alcohol (*m/z* 69, 73, 83, 103, 129: phytol, *TA1*), a lineal alcohol (*m/z* 75, 103: C24, *LA2*), a sterol (*m/z* 218: campestanol, *S2*), dimethylsterols (*m/z* 218: not identified dimethylsterol 1, *DMS1*; *m/z* 73, 129, 189, 204, 218: lupeol *DMS2*,) and the previously mentioned unknown compound eluting at 57.2 min (*m/z* 189: Unknown 9, *UNK9*) (Blue circles in the EICs of Fig. 3, Table 2). The most important coefficients for the Italian class corresponded to unidentified species eluting at 18.1 min (*m/z* 103: Unknown 1, *UNK1*), 25.9 min (*m/z* 73: Unknown 3, *UNK3*), 32 min (*m/z* 73, 81: Unknown 5, *UNK5*) and 49.5 min (*m/z* 218: Unknown 8, *UNK8*) (Red circles in the EICs of Fig. 3, Table 2).

Finally, the relevant coefficients for the Spain class in the untargeted profiling model related to: a terpene alcohol (phytol, *TA1*), two unknown compounds, eluting at 18.6 min (Unknown 2, *UNK2*) and at 45.5 min (Unknown 6, *UNK6*), three sterols (campesterol, *S1*; campestanol, *S2*; sitostanol, *S5*) and a 4,4-dimethylsterol (lupeol, *DMS2*) (Blue circles in the TIC of Fig. 3, Table 2). The Italian relevant coefficients corresponded to two not identified terpene alcohols eluting at 15.9 min (Terpene alcohol 2, *TA2*) and 34.8 min (Terpene alcohol 4, *TA4*), a sterol (stigmasterol, *S3*), a 4,4-dimethylsterol (24-methylenecycloartanol, *DMS3*) and a 4-methylsterol (citrostadienol, *MS2*) (Fig. 3, Table 2) (Red circles in the TIC of Fig. 3, Table 2).

4. Discussion

Both untargeted profiling and fingerprinting approaches successfully classified samples according to their cultivar or geographical origin depending on the variable selected for supervising the analysis, achieving percentages of correct classification in external validation higher than 90 % in almost all cases. The results confirm our hypothesis that the unsaponifiable fraction's secondary metabolites, which depend on genetic and environmental factors, have great potential for hazelnut varietal and geographical authentication. Although specificity values were similar for both approaches, fingerprinting outperformed untargeted profiling in two of the three models, providing higher sensitivity and overall correct classification for cultivar and provenance from EU or non-EU areas. This agreed with a previous study on spectroscopic data (Quintanilla-Casas et al., 2022) reporting slightly better prediction results using fingerprinting compared to untargeted profiling approach. On the other hand, the untargeted profiling model demonstrated higher sensitivity in classifying hazelnuts based on their country of origin (Spain or Italy, by arbitrarily considering Italian hazelnuts as the positive samples). In view of these results, we can affirm that both untargeted approaches applied to hazelnut unsaponifiable GC–MS data proved to be highly effective in extracting valuable sample information for the development of efficient authentication models, with the fingerprinting approach achieving slightly higher classification efficiency than untargeted profiling approach.

The standard deviation of the external validation results obtained from the randomly selected sample sets (7 iterations) can provide valuable insights into the dependence of the models on the sample set composition. This metric can be considered as an indicator of the robustness of the models and can help describe their performance in various scenarios. In this regard, the models generated by both approaches exhibited a remarkable low standard deviation, which implies that both showed low dependency on the composition of the validation sample set, indicating a high degree of reliability. Nevertheless, it should be considered that this study was designed to compare these two

Table 2
Tentative identification of compounds (based on the MS spectra from full scan acquisition at the same retention time for the fingerprinting approach and on the MS spectra of compounds extracted by PARADISe from full scan chromatograms for the untargeted profiling) corresponding to the variables with the highest regression coefficients for each class in binary PLS-DA models developed by fingerprinting and PARADISe approaches. The compounds that were relevant in the models developed by both approaches are evidenced in bold.

			TG/non-TG				EU/non-EU				Spanish/Italian			
			Fingerprinting		PARADISe		Fingerprinting		PARADISe		Fingerprinting		PARADISe	
N[a]	Chemical family	Tentative identification and level of annotation[b]	TG	non-TG	TG	non-TG	EU	non-EU	EU	non-EU	ESP	ITA	ESP	ITA
-[c]	Fatty acid 1	NI FA, 3	–	FA1	–	–	–	–	–	–	–	–	–	–
2	Linear alcohol 1	C17, 2a	LA1	–	–	–	–	–	–	–	–	–	–	–
7	Linear alcohol 2	C24, 2a	–	–	–	–	–	–	–	–	LA2	–	–	–
10	Linear alcohol 3	C25, 2a	LA3	–	–	–	–	–	–	–	–	–	–	–
12	Linear alcohol 4	C26, 2a	**LA4**	–	**LA4**	–	–	–	–	–	–	–	–	–
14	Linear alcohol 5	C28, 2a	–	–	LA5	–	–	–	–	–	–	–	–	–
1	Terpene alcohol 1	phytol, 2a	–	**–**	**–**	**–**	**–**	**TA1**	–	**TA1**	**TA1**	–	**TA1**	–
3	Terpene alcohol 2	NI TA (15.9 min), 3	–	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	TA2
6	Terpene alcohol 3	NI TA (23.1 min), 3	–	–	TA3	–	–	–	–	–	–	–	–	–
11	Terpene alcohol 4	NI TA (34.8 min) 3	–	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	TA4
13	Terpene alcohol 5	*cis*-farnesol, 2a	**TA5**	–	**TA5**	–	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**
9	Triterpenoid 1	squalene, 2a	–	–	–	–	–	–	TT1	–	–	–	–	–
16	Sterol 1	campesterol, 2a	–	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	S1	–
17	Sterol 2	campestanol, 2a	–	**–**	**–**	**–**	**–**	**–**	S2	–	**S2**	–	**S2**	–
18	Sterol 3	stigmasterol, 2a	–	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	S3
19	Sterol 4	NI S (47.7 min), 3	–	**–**	**–**	S4	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**
21	Sterol 5	sitostanol, 2a	–	**–**	S5	**–**	**S5**	**–**	**S5**	–	**–**	**–**	S5	–
22	Sterol 6	Δ5-avenasterol, 2a	–	**–**	**–**	**–**	**S6**	–	**S6**	–	**–**	**–**	**–**	**–**
25	Sterol 7	Δ7-stigmastenol, 2a	–	S7	–	–	–	S7	–	**–**	**–**	**–**	**–**	**–**
31	Sterol 8	NI S (56.4 min), 3	–	**–**	**–**	**–**	**–**	**–**	**–**	S8	**–**	**–**	**–**	**–**
27	4-methylsterol 1	28-methylobtusifoliol, 2a	–	**–**	MS1	–	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**
30	4-methylsterol 2	citrostadienol, 2a	–	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	MS2
24	4,4-dimethylsterol 1	NI DMS (50.0 min), 3	DMS1	–	–	–	–	–	–	–	DMS1	–	–	–
26	4,4-dimethylsterol 2	lupeol, 2a	–	DMS2	–	**–**	**–**	**–**	**–**	**–**	**DMS2**	–	**DMS2**	–
28	4,4-dimethylsterol 3	24-methylenecycloartanol, 2a	–	**–**	**–**	DMS3	**–**	**–**	**–**	**–**	**–**	**–**	**–**	DMS3
29	4,4-dimethylsterol 4	NI DMS (53.3 min), 3	–	**–**	**–**	**–**	**–**	**DMS4**	–	**DMS4**	**–**	**–**	**–**	**–**
4	Unknown 1	UNK 18.1 min	**–**	**–**	**–**	UNK1	**–**	**–**	**–**	**–**	**–**	UNK1	–	–
5	Unknown 2	UNK 18.6 min	**UNK2**	–	**UNK2**	–	**UNK2**	–	**UNK2**	–	**–**	**–**	UNK2	–
-[c]	Unknown 3	UNK 25.9 min	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	UNK3	–	–
8	Unknown 4	UNK 30.0 min	**–**	**–**	**–**	**–**	**–**	UNK4	**–**	**–**	**–**	**–**	–	–
-[c]	Unknown 5	UNK 32.0 min	–	–	–	–	–	–	–	–	–	UNK5	–	–
15	Unknown 6	UNK 45.5 min	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	UNK6	–
20	Unknown 7	UNK 48.4 min	UNK7	–	–	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**	**–**
23	Unknown 8	UNK 49.5 min	–	–	–	–	–	–	–	–	–	UNK8	–	–
32	Unknown 9	UNK 57.2 min	UNK9	–	–	**–**	**–**	**–**	**–**	**–**	UNK9	–	–	**–**

[a] : Compound code according to Fig. 3. TG: Tonda di Giffoni class; non-TG: other cultivars class; EU: European class (Spanish and Italian hazelnuts); non-EU: non-EU class (Chilean hazelnuts); ESP: Spanish hazelnuts class; ITA: Italian hazelnuts class; NI: not identified; UNK: unknown compound.

[b] : tentative molecular structure identification and level of annotation according to Schymanski et al., 2014 (2a: probable structure by library spectrum match; 3: tentative candidate, evidence exists for possible structure, but insufficient information for one exact structure only).

[c] : Compounds not shown in Fig. 3.

approaches, and to preliminarily evaluate the usefulness of the unsaponifiable fraction for these authentication purposes. Therefore, the sampling set included a limited number of regions and cultivars, which implies that these models are not representative of the real hazelnut production, and therefore their applicability is limited to this specific purpose. Further sample collecting including a wider natural variability (i.e. other main producing countries) is needed to develop models that can be applied in a real scenario.

Examining the regression coefficients of models generated by both fingerprinting and untargeted profiling it was evident that the distribution of most relevant compounds for the classification was throughout the entire chromatogram. This, combined with the fact that several of these compounds were present in low concentrations (Fig. 3), highlights the necessity for methods that enable comprehensive utilization of sample information such as the untargeted approaches evaluated in the present study. The tentative identification of the compounds corresponding to the most relevant variables, provided insights into the chemical families that played a crucial role in the classification process. This analysis also enabled to determine whether there were any differences in the key discriminant compounds according to the untargeted approach applied. It is worth clarifying that our intention was not to provide an exhaustive exploration of the discriminant variables, but to focus on some the most relevant variables to acquire an understanding of the type of compounds on which the classification was based on. Regarding the chemical families that mainly drove the classification in models obtained from both untargeted approaches, steroid compounds tentatively identified as sterols, 4-methylsterols, and 4,4-dimethylsterols; linear and terpene alcohols; and some unknown compounds were found to be the key discriminators, with the steroid compounds playing a crucial role in classification (Table 2). Previous studies demonstrated the influence of both genetic (Parcerisa et al., 1998; Amaral et al., 2006; Matthäus & Özcan, 2012) and environmental factors (Benitez-Sanchez et al., 2003; Matthäus & Özcan, 2012; Ghisoni et al., 2020) on the steroid fraction of hazelnuts, which supports the present findings.

The comparison of the key discriminant compounds between fingerprinting and untargeted profiling authentication models revealed partial agreement in relevant variables (Table 2). Compounds tentatively identified as unknown compound 2 (*UNK2*), C26 linear alcohol (*LA4*) and *cis*-farnesol (*TA5*) were significant to classify TG samples for both approaches. Likewise, unknown compound 2 (*UNK2*), sitostanol (*S5*) and Δ5-avenasterol (*S6*) were found to be relevant in discriminating EU samples, while phytol (*TA1*) and the not identified dimethylsterol 4 (*DMS4*) were characteristic of the non-EU samples. In addition, phytol (*TA1*), along with campestanol (*S2*) and lupeol (*DMS2*), were useful in discriminating Spanish from the Italian samples in both approaches.

However, in addition to the matching discriminant markers, it is worth noting that each approach identified distinct relevant variables in each of the authentication models (Table 2). This can be attributed to the fact that the information provided by the unfolded matrix-based fingerprinting and untargeted profiling approach varies in terms of both quantity and type, due to their differing mode of operation. On the one hand, the higher sensitivity of SIM acquisition in unfolded matrix-based fingerprinting can detect even minor compounds that may significantly contribute to sample categorization, that may be overlooked by untargeted profiling's full scan acquisition. For instance, minor compounds like the not identified minor fatty acid (*FA1*), linear alcohols C24 (*LA2*) and C25 (*LA3*), the not identified dimethylsterol 1 (*DMS1*) and Δ7-stigmastenol (*S7*) were found to be significant for classification in fingerprinting models but were not detected as chromatographic peaks by full scan untargeted profiling.

On the other hand, the selection of specific ions for acquisition in SIM mode might hinder the detection of other significant compounds characterized by different ions, which can, in turn, be detected by untargeted profiling in full scan mode. Nevertheless, in this case, compounds found as relevant only in untargeted profiling models (Table 2), such as those tentatively identified as linear alcohol C28 (*LA5*), terpene alcohols *TA2*, *TA3* and *TA4*, campesterol (*S1*), stigmasterol (*S3*), sterols *S4* and *S8,* 28-methylobtusifoliol (*MS1*), 24-methylenecycloartanol (*DMS3*) and citrostadienol (*MS2*), as well as some unknown compounds (*UNK6*), were also detected by SIM and thus, they were included in the unfolded matrix. However, they resulted to be less relevant for the classification in fingerprinting model compared to other minor compounds.

Therefore, if the representative ions of the chemical families being analysed are selected properly, the information obtained from the fingerprinting method using SIM acquisition appears to be greater than the information contained in the untargeted profiling matrix based on full scan acquisition. However, this assumption requires a general prior knowledge of the chemical families of compounds present in the samples, which is satisfied in the case of the unsaponifiable fraction of hazelnut but may present a challenge in other authentication scenarios. In this sense, one of the main advantages of untargeted profiling is its ability to provide chemically interpretable results, making it suitable for analysing samples with unknown compositions and allowing for easy identification of the markers of interest. It represents a straightforward way to identify the most relevant compounds as the pure mass spectra are provided, unlike SIM fingerprinting that does not allow for clear identification and requires further full scan analysis to properly assess compounds' mass spectra and chemical structure.

One final consideration that should be addressed concerns the applicability, ease of implementation and level of prior knowledge required by the user, and transferability for each of the untargeted approaches compared. PARADISe is a user-friendly interface to utilize PARAFAC2, but it does require a certain level of know-how for interval selection and optimization of PARAFAC2 models, which is not necessary for building the unfolded matrix in the fingerprinting approach. This issue may be resolved in future versions of PARADISe by enabling automatic interval selection, but at present, the fingerprinting unfolded matrix approach is easier to use and apply. On the other hand, transferring untargeted analytical methods to other laboratories or instruments can be a challenging task, especially for fingerprinting methods. In fact, while conventional strategies for target methods can be adapted to assess the performance of untargeted profiling results, thereby enabling easy in-house and inter-laboratory validations, a lack of precise guidelines regarding the validation procedure for fingerprinting methods make it even more challenging to transfer these methods, despite ongoing efforts to establish them (Quintanilla-Casas et al., 2020b).

5. Conclusions

In conclusion, the unsaponifiable fraction of the hazelnut oil has proven to be a promising tool for their geographical and varietal authentication. Even if it is not a fast-screening technique, the study has proved that GC–MS coupled with untargeted methods such as fingerprinting and advanced profiling techniques like untargeted profiling can provide high-dimensional molecular-level information for hazelnut authentication. Both untargeted profiling and fingerprinting proved to be successful in the authentication of hazelnuts, although fingerprinting provided slightly better prediction results. As revealed by the examination of the regression coefficients of the PLS-DA models, this may be due to the greater information extracted by the fingerprinting method from chromatographic data, which enabled considering even very minor discriminant species. However, untargeted profiling enables easier chemical interpretability than fingerprinting based on SIM data, providing the pure spectra of the relevant compounds. It is remarkable that these results were obtained in a challenging scenario in which the origin was discriminated between samples of the same cultivar, and in turn, the cultivar was discriminated between samples from the same origin. This positions the analytical strategy as a suitable candidate to verify challenging samples as a support to fast-screening tools. Nevertheless, optimal models should be further developed and evaluated using a large-scale dataset, that would include the natural heterogeneity

of the samples, the main producing regions and their principal cultivars in addition to several harvest years.

Funding

This work was developed in the context of the project TRACENUTS, PID2020-117701RB100 financed by MCIN/AEI/https://doi.org/10.13039/501100011033. B. Torres-Cobos thanks the Spanish Ministry of Universities predoctoral fellowships FPU20/014540. B. Quintanilla-Casas thanks the Fundación Alfonso Martín Escudero for the research grant for universities and centers abroad 2022. A. Tres received a Ramon y Cajal grant (RYC-2017-23601) funded by MCIN/AEI/https://doi.org/10.13039/501100011033 and by "ESF Investing in your future".

CRediT authorship contribution statement

B. Torres-Cobos: Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft. **B. Quintanilla-Casas:** Data curation, Investigation, Methodology, Validation, Writing – review & editing. **M. Rovira:** Conceptualization, Resources, Writing – review & editing. **A. Romero:** Conceptualization, Resources, Writing – review & editing. **F. Guardiola:** Supervision, Writing – review & editing. **S. Vichi:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing. **A. Tres:** 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.

Data availability

Data will be made available on request.

Acknowledgements

INSA-UB Maria de Maeztu Unit of Excellence (Grant CEX2021-001234-M) funded by MICIN/AEI/FEDER, UE. The authors would like to express their gratitude to Ferrero Hazelnut Company and Tuscia University (Department of Agriculture and Forest Science) for providing the hazelnut samples from Chile and Italy, respectively.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodchem.2023.138294.

References

Amaral, J. S., Casal, S., Citova, I., Santos, A., Seabra, R. M., & Oliveira, B. P. P. (2006). Characterization of several hazelnut (Corylus avellana L.) cultivars based in chemical, fatty acid and sterol composition. *European Food Research and Technology, 222*, 274–280. https://doi-org.sire.ub.edu/10.1007/s00217-005-0068-0.

Baccolo, G., Quintanilla-Casas, B., Vichi, S., Augustijn, D., & Bro, R. (2021). From untargeted chemical profiling to peak tables – A fully automated AI driven approach to untargeted GC-MS. *TrAC Trends in Analytical Chemistry, 145*, Article 116451. https://doi.org/10.1016/j.trac.2021.116451

Bachmann, R., Klockmann, S., Haerdter, J., Fischer, M., & Hackl, T. (2018). 1H NMR Spectroscopy for Determination of the Geographical Origin of Hazelnuts. *Journal of Agricultural and Food Chemistry, 66*, 11873–11879. https://doi.org/10.1021/acs.jafc.8b03724

Ballin, N. Z., & Laursen, K. H. (2019). To target or not to target? Definitions and nomenclature for targeted versus non-targeted analytical food authentication. *Trends in Food Science & Technology, 86*, 537–543. https://doi.org/10.1016/j.tifs.2018.09.025

Benitez-Sanchez, P. L., León-Camacho, M. L., & Aparicio, R. (2003). A comprehensive study of hazelnut oil composition with comparisons to other vegetable oils, particularly olive oil. *European Food Research and Technology, 218*, 13–19. https://doi.org/10.1007/s00217-003-0766-4

Biancolillo, A., De Luca, S., Bassi, S., Roudier, L., Bucci, R., Magrì, A. D., & Marini, F. (2018). Authentication of an Italian PDO hazelnut ("Nocciola Romana") by NIR spectroscopy. *Environmental Science and Pollution Research, 25*, 28780–28786. https://doi.org/10.1007/s11356-018-1755-2

Bosque-Sendra, J. M., Cuadros-Rodríguez, L., Ruiz-Samblás, C., & de la Mata, A. P. (2012). Combining chromatography and chemometrics for the characterization and authentication of fats and oils from triacylglycerol compositional data—A review. *Analytica Chimica Acta, 724*, 1–11. https://doi.org/10.1016/j.aca.2012.02.041

Ciarmiello, L. F., Mazzeo, M. F., Minasi, P., Peluso, A., De Luca, A., Piccirillo, P., Siciliano, R. A., & Carbone, V. (2014). Analysis of Different European Hazelnut (Corylus avellana L.) Cultivars: Authentication, Phenotypic Features, and Phenolic Profiles. *Journal of Agricultural and Food Chemistry, 62*, 6236–6246. https://doi.org/10.1021/jf5018324

FAOstat Food and Agriculture Organization of the United Nations, agricultural data 2021; https://www.fao.org/faostat/es/#data/PP (accessed Oct 15, 2023).

Ghisoni, S., Lucini, L., Rocchetti, G., Chiodelli, G., Farinelli, D., Tombesi, S., & Trevisan, M. (2020). Untargeted metabolomics with multivariate analysis to discriminate hazelnut (Corylus avellana L.) cultivars and their geographical origin. *Journal of the Science of Food and Agriculture, 100*, 500–508. https://doi.org/10.1002/jsfa.9998

Goriainov, S. V., Esparza, C. A., Borisova, A. R., Orlova, S. V., Vandyshev, V. V., Hajjar, F., Platonov, E. A., Chromchenkova, E. P., Novikov, O. O., Borisov, R. S., & Kalabin, G. A. (2021). Phytochemical Study of the Composition of the Unsaponifiable Fraction of Various Vegetable Oils by Gas Chromatography-Mass Spectrometry. *Journal of Analytical Chemistry, 76*, 1635–1644. https://doi.org/10.1134/S1061934821140045

Inaudi, P., Giacomino, A., Malandrino, M., La Gioia, C., Conca, E., Karak, T., & Abollino, O. (2020). The Inorganic Component as a Possible Marker for Quality and for Authentication of the Hazelnut's Origin. *International Journal of Environmental Research and Public Health, 17*, 447. https://doi.org/10.3390/ijerph17020447

Johnsen, L. G., Skou, P. B., Khakimov, B., & Bro, R. (2017). Gas chromatography – mass spectrometry data processing made easy. *Journal of Chromatography A, 1503*, 57–64. https://doi.org/10.1016/j.chroma.2017.04.052

Klockmann, S., Reiner, E., Bachmann, R., Hackl, T., & Fischer, M. (2016). Food Fingerprinting: Metabolomic Approaches for Geographical Origin Discrimination of Hazelnuts (Corylus avellana) by UPLC-QTOF-MS. *Journal of Agricultural and Food Chemistry, 64*, 9253–9262. https://doi.org/10.1021/acs.jafc.6b04433

Krauß, S., Vieweg, A., & Vetter, W. (2019). Stable isotope signatures (δ2H-, δ13C-, δ15N-values) of walnuts (Juglans regia L.) from different regions in Germany. *Journal of the Science of Food and Agriculture, 100*, 1625–1634. https://doi.org/10.1002/jsfa.10174

Król, K., & Gantner, M. (2020). Morphological Traits and Chemical Composition of Hazelnut from Different Geographical Origins: A Review. *Agriculture, 10*, 375. https://doi.org/10.3390/agriculture10090375

Lang, C., Weber, N., Möller, M., Schramm, L., Schelm, S., Kohlbacher, O., & Fischer, M. (2021). Genetic authentication: Differentiation of hazelnut cultivars using polymorphic sites of the chloroplast genome. *Food Control, 130*, Article 108344. https://doi.org/10.1016/j.foodcont.2021.108344

Larsen, F. H., Van den Berg, F., & Engelsen, S. B. (2006). An exploratory chemometric study of 1H NMR spectra of table wines. *Journal of Chemometrics, 20*, 198–208. https://doi.org/10.1002/cem.991

Li, T. S. C., Beveridge, T. H. J., & Drover, J. C. G. (2007). Phytosterol content of sea buckthorn (Hippophae rhamnoides L.) seed oil: Extraction and identification. *Food Chemistry, 101*, 1633–1639. https://doi.org/10.1016/j.foodchem.2006.04.033

Matthäus, B., & Özcan, M. M. (2012). The comparison of properties of the oil and kernels of various hazelnuts from Germany and Turkey. *European Journal of Lipid Science and Technology, 114*, 801–806. https://doi.org/10.1002/ejlt.201100299

Nielsen, N. P. V., Carstensen, J. M., & Smedsgaard, J. (1998). Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A, 805*, 17–35. https://doi.org/10.1016/S0021-9673(98)00021-1

Oddone, M., Aceto, M., Baldizzone, M., Musso, D., & Osella, D. (2009). Authentication and Traceability Study of Hazelnuts from Piedmont, Italy. *Journal of Agricultural and Food Chemistry, 57*, 3404–3408. https://doi.org/10.1021/jf900312p

Parcerisa, J., Richardson, D. G., Rafecas, M., Codony, R., & Boatella, J. (1998). Fatty acid, tocopherol and sterol content of some hazelnut varieties (Corylus avellana L.) harvested in Oregon (USA). *Journal of Chromatography A, 805*, 259–268. https://doi.org/10.1016/S0021-9673(98)00049-1

Phillips, K. M., Ruggio, D. M., & Ashraf-Khorassani, M. (2005). Phytosterol Composition of Nuts and Seeds Commonly Consumed in the United States. *Journal of Agricultural and Food Chemistry, 53*, 9436–9445. https://doi.org/10.1021/jf051505h

Quintanilla-Casas, B., Bro, R., Hinrich, J. L., Davie-Martin, C. L. (2023). Tutorial on PARADISe: PARAFAC2-based Deconvolution and Identification System for processing GC–MS data, PROTOCOL (Version 1), Protocol Exchange.

Quintanilla-Casas, B., Bertin, S., Leik, K., Bustamante, J., Guardiola, F., Valli, E., … Vichi, S. (2020a). Profiling versus fingerprinting analysis of sesquiterpene hydrocarbons for the geographical authentication of extra virgin olive oils. *Food Chemistry, 307*, Article 125556. https://doi.org/10.1016/j.foodchem.2019.125556

Quintanilla-Casas, B., Marin, M., Guardiola, F., García-González, D. L., Barbieri, S., Bendini, A., … Tres, A. (2020b). Supporting the Sensory Panel to Grade Virgin Olive Oils: An In-House-Validated Screening Tool by Volatile Fingerprinting and Chemometrics. *Foods, 9*, 1509. https://doi.org/10.3390/foods9101509

Quintanilla-Casas, B., Rinnan, Å., Romero, A., Guardiola, F., Tres, A., Vichi, S., & Bro, R. (2022). Using fluorescence excitation-emission matrices to predict bitterness and

pungency of virgin olive oil: A feasibility study. *Food Chemistry, 395*, Article 133602. https://doi.org/10.1016/j.foodchem.2022.133602

Riedl, J., Esslinger, S., & Fauhl-Hassek, C. (2015). Review of validation and reporting of non-targeted fingerprinting approaches for food authentication. *Analytica Chimica Acta, 885*, 17–32. https://doi.org/10.1016/j.aca.2015.06.003

Rinnan, A., Amigo, J. M., Skov, T. (2014). Multiway methods in food science, in: D. Granato, G. Ares (Eds.), Mathematical and Statistical Methods in Food Science and Technology, IFT Press/Wiley Blackwell, Chichester 143-174.

Ríos-Reina, R., Aparicio-Ruiz, R., Morales, M. T., & García-Gonzalez, D. L. (2023). Contribution of specific volatile markers to green and ripe fruity attributes in extra virgin olive oils studied with three analytical methods. *Food Chemistry, 399*, Article 133942. https://doi.org/10.1016/j.foodchem.2022.133942

Sales, C., Portolés, T., Johnsen, L. G., Danielsen, M., & Beltran, J. (2019). Olive oil quality classification and measurement of its organoleptic attributes by untargeted GC–MS and multivariate statistical-based approach. *Food Chemistry, 271*, 488–496. https://doi.org/10.1016/j.foodchem.2018.07.200

Sammarco, G., Dall'Asta, C., & Suman, M. (2023). Near infrared spectroscopy and multivariate statistical analysis as rapid tools for the geographical origin assessment of Italian hazelnuts. *Vibrational Spectroscopy, 126*, Article 103531. https://doi.org/10.1016/j.vibspec.2023.103531

Schymanski, E. L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H. P., & Hollender, J. (2014). Identifying small molecules via high resolution mass spectrometry: Communicating confidence. *Environmental Science & Technology, 48*, 2097–2098. https://doi.org/10.1021/es5002105

Tomasi, G., Van den Berg, F., & Andersson, C. (2004). Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics, 18*, 231–241. https://doi.org/10.1002/cem.859

Torres-Cobos, B., Quintanilla-Casas, B., Romero, A., Ninot, A., Alonso-Salces, R. M., Toschi, T. G., Bendini, A., Guardiola, F., Tres, A., & Vichi, S. (2021). Varietal authentication of virgin olive oil: Proving the efficiency of sesquiterpene fingerprinting for Mediterranean Arbequina oils. *Food Control, 128*, Article 108200. https://doi.org/10.1016/j.foodcont.2021.108200

Torres-Cobos, B., Quintanilla-Casas, B., Vicario, G., Guardiola, F., Tres, A., & Vichi, S. (2023). Revealing adulterated olive oils by triacylglycerol screening methods: Beyond the official method. *Food Chemistry, 409*, Article 135256. https://doi.org/10.1016/j.foodchem.2022.135256

Tüfekci, F., & Karataş, Ş. (2018). Determination of geographical origin Turkish hazelnuts according to fatty acid composition. *Food Science & Nutrition, 6*, 557–562. https://doi.org/10.1002/fsn3.595

Xu, B., Zhang, L., Ma, F., Zhang, W., Wang, X., Zhang, Q., Luo, D., Ma, H., & Li, P. (2018). Determination of free steroidal compounds in vegetable oils by comprehensive two-dimensional gas chromatography coupled to time-of-flight mass spectrometry. *Food Chemistry, 245*, 415–425. https://doi.org/10.1016/j.foodchem.2017.10.114

Zannella, C., Carucci, F., Aversano, R., Prohaska, T., Vingiani, S., Carputo, D., & Adamo, P. (2017). Genetic and geochemical signatures to prevent frauds and counterfeit of high-quality asparagus and pistachio. *Food Chemistry, 237*, 545–552. https://doi.org/10.1016/j.foodchem.2017.05.158