Food Research International 194 (2024) 114873

Contents lists available at ScienceDirect

Food Research International

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

Artificial intelligence decision making tools in food metabolomics: Data fusion unravels synergies within the hazelnut (*Corylus avellana* L.) metabolome and improves quality prediction

Simone Squara[a], Andrea Caratti[a], Angelica Fina[a], Erica Liberto[a], Nemanja Koljančić[a,b], Ivan Špánik[b], Giuseppe Genova[c], Giuseppe Castello[c], Carlo Bicchi[a], André de Villiers[d,*], Chiara Cordero[a,*]

[a] *Dipartimento di Scienza e Tecnologia del Farmaco, Università degli Studi di Torino, Via Pietro Giuria 9, Torino 10125, Italy*
[b] *Institute of Analytical Chemistry, Slovak University of Technology, Radlinského 9, Bratislava 812 37, Slovakia*
[c] *Soremartec Italia Srl, Piazzale Ferrero 1, Alba, Cuneo 12051, Italy*
[d] *Department of Chemistry and Polymer Science, Stellenbosch University, Matieland, Stellenbosch, Western Cape 7602, South Africa*

ARTICLE INFO

Keywords:
Corylus avellana L.
LC-HRMS and GC×GC-TOF-MS data
Data fusion
Volatilome
Metabolome
Food quality prediction

ABSTRACT

This study investigates the metabolome of high-quality hazelnuts (*Corylus avellana* L.) by applying untargeted and targeted metabolome profiling techniques to predict industrial quality. Utilizing comprehensive two-dimensional gas chromatography and liquid chromatography coupled with high-resolution mass spectrometry, the research characterizes the non-volatile (primary and specialized metabolites) and volatile metabolomes. Data fusion techniques, including low-level (LLDF) and mid-level (MLDF), are applied to enhance classification performance. Principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA) reveal that geographical origin and postharvest practices significantly impact the specialized metabolome, while storage conditions and duration influence the volatilome. The study demonstrates that MLDF approaches, particularly supervised MLDF, outperform single-fraction analyses in predictive accuracy. Key findings include the identification of metabolites patterns causally correlated to hazelnut's quality attributes, of them aldehydes, alcohols, terpenes, and phenolic compounds as most informative. The integration of multiple analytical platforms and data fusion methods shows promise in refining quality assessments and optimizing storage and processing conditions for the food industry.

1. Introduction

The emergence of metabolomics marks a significant advancement in biological exploration, offering deep insights into the complexity of biological systems (Collins, Koo, Peters, Smith, & Patterson, 2021). Metabolomics, a subset of systems biology, examines a vast array of over 40,000 metabolites with diverse properties, presenting both opportunities and challenges. The metabolome's diversity is particularly valuable in agricultural sciences, providing predictive markers for crop quality (Balkir, Kemahlioglu, & Yucel, 2021; Li et al., 2021; Pedrosa et al., 2021). As consumer demand for high-quality, nutrients-dense food increases, metabolomics becomes crucial for enhancing crop breeding, optimizing cultivation, reducing post-harvest losses, and improving storage (Schmid et al., 2021). To fully harness the metabolome's potential, integrating various analytical platforms is essential for deriving reliable markers that predict food quality at a molecular level and monitor it over time (Caratti et al., 2024; Jacobs, van den Berg, & Hall, 2021; Mack et al., 2017; Romo-Pérez et al., 2020; Ulaszewska et al., 2019).

Data fusion techniques play a pivotal role in metabolomics,

Abbreviations: LLDF, Low-level data fusion; MLDF, Mid-level data fusion; HLDF, High-level data fusion; UMLDF, Unsupervised mid-level data fusion; SMLDF, Supervised mid-level data fusion; GC×GC, Comprehensive 2D gas chromatography; MOX, O-methylhydroxylamine hydrochloride; BSTFA, N, O-bis(trimethylsilyl) trifluoroacetamide; TGT, Tonda Gentile Trilobata; TGR, Tonda Gentile Romana; GIR, Giresun; AKC, Akçakoca; UV, Under vacuum; SA, Standard atmosphere; PLS-DA, Partial least square – discriminant analysis; PCA, Principal component analysis; LV, Latent variable.

* Corresponding authors.
E-mail addresses: ajdevill@sun.ac.za (A. de Villiers), chiara.cordero@unito.it (C. Cordero).

https://doi.org/10.1016/j.foodres.2024.114873
Received 29 April 2024; Received in revised form 22 July 2024; Accepted 5 August 2024
Available online 14 August 2024

combining information from diverse sources to make more accurate and robust inferences. Similar to how the human brain integrates sensory information, data fusion consolidates inputs from various analytical methods to capture a system's characteristics, aiding reliable decision-making (Westerhuis, Kloet, & Smilde, 2019). This approach results in more precise insights, improved classifications, and reduced prediction errors compared to single techniques. However, integrating data from different platforms is challenging due to their varied nature and dimensionality (Azcarate, Ríos-Reina, Amigo, & Goicoechea, 2021; Wang et al., 2023). Effective data fusion requires a thorough understanding of the data structure and appropriate merging strategies, categorized into low-level, mid-level, and high-level fusion based on how data or features are combined (Hassani, Dackermann, Mousavi, & Li, 2024; Smolinska, Engel, Szymanska, Buydens, & Blanchet, 2019).

The study focuses on analyzing the metabolome of high-quality hazelnuts used in confectionery to predict industrial quality. Using comprehensive two-dimensional gas chromatography (GC×GC) and liquid chromatography coupled with high-resolution mass spectrometry (LC-HRMS), both non-volatile and volatile metabolites are mapped and analyzed. The volatile metabolome, crucial for sensory quality, is assessed using quantitative metabolomics. Data fusion techniques like low-level data fusion (LLDF) and mid-level data fusion (MLDF) are employed to enhance classification performance and information capacity. The study aims to predict industrial quality factors, including harvest region, post-harvest practices, storage conditions, and sensory quality, thereby providing a comprehensive approach to optimizing hazelnut quality for industrial use.

2. Materials and methods

2.1. Reagents and chemicals

Pure standards of *n*-alkanes (from n-C9 to n-C27) for system evaluation and linear retention index (I^T) determination as well as α/β-tujone for volatiles internal standardization (IS) were obtained from Merck (Milan, Italy). The mixture of *n*-alkanes for the I^T solution was prepared in cyclohexane at a concentration of 100 mg/L; internal standard (IS) α-thujone solution was prepared in diethyl phthalate (Sigma Aldrich 99 % of purity) at a concentration of 100 mg/L

Pure standards for identity confirmation of pyruvic acid, lactic acid, malonic acid, succinic acid, glyceric acid, fumaric acid, malic acid, citric acid, alanine-Ala, asparagine-Asn, aspartic acid-Asp, cysteine-Cys, glutamic acid-Glu, glycine-Gln, isoleucine-Ile, leucine-Leu, lysine-Lys, methionine-Met, ornithine-Orn, phenylalanine-Phe, proline-Pro, serine-Ser, threonine-Thr, tryptophan-Trp, tyrosine-Tyr, valine-Val, glycerol, xylitol, mannitol, myo-inositol, fructose, glucose, saccharose, catechin, epicatechin, procyanidin B2, ellagic acid, and the internal standards (ISs), 4-chlorophenylalanine (quality control – QC for derivatization) and 1,4-dibromobenzene (QC for GC normalization), were purchased from Merck.

Derivatization reagents and LC grade solvents: O-methylhydroxylamine hydrochloride (MOX), (N, O-bis(trimethylsilyl)trifluoroacetamide (BSTFA), methanol, acetone, pyridine, *n*-hexane, dichloromethane, toluene and formic acid (FA) were from Merck, while LC-MS grade acetonitrile (ACN) was obtained from ROMIL (Waterbeach Cambridge, England).

2.2. Hazelnut samples

Raw hazelnuts, sourced from various geographical and botanical origins, were supplied by Soremartec Italia Srl located in Alba, Cuneo, Italy. The raw materials, which were uniformly ground into *granella* (i.e., grain), were promptly stored at −80 °C until analysis. Samples were from four diverse cultivars/geographical origins: two Italian samples, Tonda Gentile Trilobata (TGT), and Tonda Gentile Romana (TGR), and two Turkish blends from Akçakoca (AKC) and Giresun (GIR) areas. Moreover, the GIR sample was also simulated to undergo improper postharvest handling, and thus simulated to be sun dried at 38–40 °C in a high-humidity environment. All samples were stored up to 12 months, with samples collected at 0, 4-, 6-, 9-, and 12-months intervals. Two storage conditions were tested: under vacuum (UV) after modified atmosphere with 1 % O_2 and 99 % N_2, and standard atmosphere (SA). Both conditions were maintained at 5 °C and kept below 65 % equilibrium relative humidity.

2.2.1. QC and batch handling

QC samples, according to accepted practices, were appositely created to monitor the analytical platforms performances, in particular retention times (t_R) stability and response fluctuations of the detector (Dudzik, Barbas-Bernardos, García, & Barbas, 2018). For the primary and specialized metabolites, which were all analyzed in a single analytical batch, the QC sample was created by mixing aliquots of each biological replicate. QCs were then analyzed once every 8 (randomized) samples. Regarding the volatile fraction, where analyses were acquired punctually at every time point, QC samples were created by mixing the biologically different T0 samples and stored at −80 °C; aliquots were used at each time point to check for system stability. In the latter case, given the larger fluctuations in the detector given by the fact that samples were acquired during the course of one year, the response was normalized on the IS (α/β-Thujone mixture), which was pre-loaded onto the SPME device before exposing the fiber to the sample.

2.3. Volatilome analysis by GC×GC-qMS/FID

Sampling consisted of SPME on 100 mg of hazelnut raw grain at 50 °C for 50 min under constant agitation with a 2 cm Divinylbenzene/Carboxen/Polydimethylsiloxane (DVB/CAR/PDMS) df 50/30 μm fiber. The fibers utilized were sourced from Merck, and prior to their application, they underwent the manufacturer's recommended conditioning process; this involved thermal desorption at 270 °C for a duration of 60 min in a split/splitless injector, maintained under a continuous carrier gas flow. The analyses were conducted using an Agilent 7890B unit coupled with an Agilent 5977B high efficiency source (HES) MS (Agilent Technologies, Little Falls, DE, USA). The MS operated in EI mode at 70 eV, with a scan range from 40 to 250 *m/z* (mass-to-charge ratio) at 10,000 amu/s resulting in a 28 Hz acquisition frequency. The MS source temperature was set at 230 °C, and the transfer line temperature was maintained at 280 °C. Modulation was achieved through a reverse-inject differential-flow modulator (Agilent Technologies, Little Falls, DE, USA). Additionally, parallel detection was accomplished using a flame ionization detector (FID) set at 300 °C, with an H_2 flow of 40 mL/min, an air flow of 450 mL/min, and a sampling frequency of 200 Hz. The split-splitless injector, operating in pulsed-split mode at 250 kPa until 2.5 min with a 1:5 split ratio, was set at 250 °C. Helium was used as the carrier gas, with a flow rate of 0.4 mL min^{-1} and 10 mL min^{-1} in the first dimension (^{1}D) and the second dimension (^{2}D), respectively. The modulation period (P_M) was set at 2.5 s with a sampling time of 0.25 s. The column configuration consisted of a ^{1}D DB-HeavyWax™ column (100 % polyethylene glycol – PEG; 20 m × 0.18 mm d_c × 0.18 μm d_f) coupled with a ^{2}D DB17 column ((50 %-phenyl)-methylpolysiloxane; 1.8 m × 0.18 mm d_c × 0.18 μm d_f), both from Agilent Technologies (Wilmington, DE, USA). After the ^{2}D column, the flow was split using a three-way unpurged capillary microfluidic splitter (G3181B, Agilent, Little Falls, DE, USA). The connections toward the MS and FID consisted of deactivated silica capillaries (Agilent Technologies, Wilmington, DE, USA) of dimensions 0.5 m × 0.1 mm d_c and 1.1 m x 0.18 mm d_c, respectively, resulting in a split ratio of 70:30 FID/MS. The bleeding capillary, consisting of deactivated silica with dimensions of 5.81 m × 0.1 mm d_c, was dimensioned using a validated calculator to obtain a minimal flow increase of 10 % (Giardina et al., 2018). The oven temperature program was as follows: 40 °C (2 min) to 130 °C (0 min) @ 4 °C min^{-1}, and to 260 °C (10 min) @ 8 °C min^{-1}. Data were acquired by MassHunter version

10.0 (Agilent Technologies, Wilmington, DE, USA), and processed using GC Image ver. 2022r1 (GC Image LCC, Lincoln, Nebraska, USA).

Analytes putative identification was by comparing the experimental EI 70 eV spectrum (option peak spectrum from the highest modulation) with those collected in commercial and in-house databases. The NIST similarity search algorithm was used and the direct match factor (DMF) threshold was set at 950; the ^{1}D I^{T} tolerance was at ± 10.

2.4. Primary metabolome analysis by GC×GC-TOF MS

The extraction/derivatization protocol applied in this study was optimized from previous research (Cialiè Rosso et al., 2020; Cialiè Rosso, Stilo, Bicchi, et al., 2021). Specifically, 0.5 g of hazelnuts *granella* were defatted with *n*-hexane (5 mL x 7 times) in an ultrasonic bath for 15 min at ambient temperature. 100 mg of the defatted grain was further extracted with 5.0 mL of H_2O/CH_3OH (98:2 *V/V*) in an ultrasonic bath for 15 min. After centrifugation for 5 min at 5,000 RPM at 4 °C, the supernatant was collected and filtered with Nylon HPLC filters with 20 µm pores. One mL of extract was then freeze-dried overnight after the addition of 15 µL of 4-chlorophenylalanine solution (4 mg/mL in methanol) (process internal standard – IS). To prevent the formation of multiple derivatives from sugar enols during the silylation steps, 45 µL of MOX (20 g/L in pyridine) was added and the mixture was allowed to react for two hours at 60 °C under agitation. Then, 60 µL of BSTFA were introduced, and the solution was maintained at 60 °C for 1 h. To serve as analytical IS, 20 µL of 1,4-dibromobenzene in dichloromethane at a concentration of 1 g/L were added and further diluted with 75 µL of dichloromethane, reaching a final volume of 200 µL. The prepared samples were promptly stored at −18 °C and analyzed within 24 h after the derivatization process.

GC×GC analyses were conducted using an Agilent 7890B coupled with a Markes BenchTOF Select™ mass spectrometer featuring Tandem Ionization™ (Markes International, Llantrisant, UK). The system incorporated a two-stage KT 2004 loop-type thermal modulator (Zoex Corporation, Houston, TX), cooled with liquid nitrogen and controlled by Optimode v2.0 (SRA Instruments, Cernusco sul Naviglio, Milan, Italy).

The column setup and operative conditions were as follows: ^{1}D DB5 (95 % polydimethylsiloxane, 5 % phenyl; 60 m × 0.25 mm d_c × 0.25 µm d_f), ^{2}D DB17MS (equivalent to (50 %-phenyl)-methylpolysiloxane; 2.0 m × 0.1 mm d_c × 0.10 µm d_f) from J&W (Agilent, Little Falls, DE, USA). The first 0.80 m of the ^{2}D column, connected in series to the ^{1}D column by a silTite µ-union (Trajan Scientific and Medical, Ringwood, Victoria, Australia), was utilized as a loop-capillary for cryogenic modulation. The carrier gas, helium, was maintained at a flow rate of 1.3 mL min^{-1} in constant flow mode. The P_M was 3.0 s, operating in multi-step mode from 0 to 15 min. The hot jet pulse time duration was 250 ms during the first 15 min, and 350 ms during the period 15–63 min. The cold jet flow was programmed for a linear decrease from 35 % of the Mass Flow Controller (MFC) maximum flow (40 L min^{-1}) to 5 % at the end of the run. The injector temperature was held at 280 °C, operating in split mode with a split ratio of 1:20. The oven temperature ramp was: 60 °C (2 min) to 120 °C at 10 °C min^{-1}, then to 300 °C (10 min) at 4 °C min^{-1}; the injection volume was 1 µL. TOF MS acquisition parameters included tandem ionization™ at 70 and 12 eV, with an acquisition rate of 50 Hz per channel within the mass range 45–650 *m/z*; the filament voltage was set at 1.8 V. The ion source and transfer line temperatures were both set at 290 °C. Data were processed using GC Image ver. 2022r1 (GC Image LCC, Lincoln, Nebraska, USA).

Analytes putative identification was by comparing the experimental EI 70 eV spectrum (option *peak spectrum* from the highest modulation) with those collected in commercial and in-house data bases. The NIST similarity search algorithm was used and the direct match factor (DMF) threshold was set at 950; the ^{1}D I^{T} tolerance was at ± 10. Where available, reference standards were also used to confirm the analytes identity.

2.5. Specialized metabolome analysis via LC-MS

Sample preparation followed the protocol reported by Ghirardello *et al.* (Ghirardello et al., 2014), with slight modifications. 1 ± 0.005 g of *granella* were weighted in a 15 mL vial and extracted with 10 mL of an acetone: water: formic acid mixture (75:24.5:0.5, *V/V/V*) in an ultrasonic bath for two hours at room temperature. Subsequently, samples were centrifuged at 6,000 RPM for 10 min and the supernatant collected. The extracts were washed with *n*-hexane (3 x 5 mL), the organic phase was removed using a separation funnel, and the aqueous phase recovered in a 10 mL vial. The acetone contained in the aqueous phase was evaporated under a gentle nitrogen stream, while water was subsequently removed using a VirTis BenchTop Pro freeze dryer (SP Scientific, NY USA) overnight. Freeze-dried extracts were stored at −18 °C away from UV exposure, and prior to analysis dissolved in 200 µL of water:methanol:formic acid mixture (66.6:33.3:0.1, *V/V/V*), filtered (0.22 µm), and immediately analyzed.

Analyses were carried out using an Acquity ultra-high-pressure liquid chromatography (UHPLC) system hyphenated to a photodiode array (PDA) detector (500 nL flow cell, 10 mm path length) and a Synapt-G2 quadrupole time-of-flight (q-TOF) mass spectrometer equipped with an electrospray ionization (ESI) source operating in negative ionization mode (Waters, Milford, MA, USA). The PDA acquisition ranged from 200 to 500 nm at a scan rate of 20 Hz. The MS scan ranged from 120 to 1500 amu (40 to 1500 for the high collision energy data) at a scan time of 0.2 s, with a capillary voltage of – 3.0 kV, a cone voltage of 20 V, a source temperature of 120 °C, and an extraction cone voltage of 4.0 V. The desolvation gas was nitrogen with a flow rate of 650 L h^{-1} at a temperature of 275 °C, and the cone gas flow (N_2) 50 L h^{-1}. Low and high collision energy data were acquired alternately in a single analysis in MS^{E} mode, using collision energies of 6 eV and a ramp of 15 – 65 eV, respectively. Accurate mass calibration was carried out via a sodium formate solution, while leucine enkephalin (*m/z* = 554.2615) was used as mass calibrant. The separation was achieved on a Kinetex (RP-18, 150 mm × 2.1 mm, 1.7 µm) superficially porous column (Phenomenex, Torrance, USA) using as mobile phase a combination of A (0.1 % FA in H_20, *V/V*) and B (0.1 % FA in ACN, *V/V*) at a constant flow rate of 0.3 mL min^{-1}; the elution gradient was as follows: 1 % B held for 3 min (0–3 min), a linear gradient from 1 % to 8 % B (3–16 min), from 8 % to 14 % B (16–22 min), from 14 % to 25 % B (22–38 min), from 25 % to 95 % B (38–44 min), held at 95 % B for 2 min (44 – 46 min), and followed by a re-equilibration step of 10 min at the initial conditions. The injection volume was 5 µL with the injector loop set in partial fill mode. Data were acquired using MassLynx (v4.2). Raw data files were converted to *.abf format using Reifycs Abf converter, and further processed using MS-DIAL (Tsugawa et al., 2015; Tsugawa et al., 2016, 2020), MS-Finder (Lai et al., 2018; Tsugawa et al., 2019), the Global Natural Product Social Molecular Networking (GNPS) website, and MassLynx (v4.2) (Waters).

Analytes putative identifications were based on relative retention, UV spectra and low- and high collision energy HR-MS data (Confidence level 3) (Schrimpe-Rutledge, Codreanu, Sherrod, & McLean, 2016). Exact mass error was below 5 ppm. This approach is aligned to the recently introduced scoring system for non-targeted screenings by LC-HRMS by Alygizakis et al. (Alygizakis et al., 2023).

2.6. Data analysis software and tools

Data fusion and chemometrics were performed using Matlab R2021a (The MathWorks, Inc., Natick, Massachusetts, United States) with the following packages: PCA toolbox (v1.5) (Ballabio, 2015) and Classification toolbox (v6.0) (Ballabio & Consonni, 2013).

LLDF involved merging data matrices from various techniques into a single matrix after scaling and centering each original block. For unsupervised MLDF (UMLDF), principal component analysis (PCA) was conducted on each scaled and centered data matrix. The first principal components that collectively explained at least 80 % of the total variance were then merged into a new matrix. In contrast, supervised MLDF (SMLDF) used partial least-squares discriminant analysis (PLS-DA) on each data matrix, and the first three latent variables (LVs) from each model were merged into a new matrix.

The resultant data matrices, each comprising 130 samples, had dimensions of 442, 674, and 44 variables for volatilome, primary metabolome, and non-volatile metabolome, respectively. To assess classification performance, Monte Carlo cross-validation was performed with an 80–20 % dataset split over 1,000 iterations, using a maximum assignation criterion for all three fusion levels. Quality metrics for classification models include classification accuracy, R^2 and Q^2.

3. Results and discussion

3.1. *Metabolite fractions informative potential: Targeted features*

3.1.1. *Primary metabolome and volatilome*

The hazelnut primary metabolome and volatilome GC×GC chromatograms were processed using the Untargeted – Targeted (UT) fingerprinting approach, as outlined in previous studies (Cordero et al., 2019; Stilo, Liberto, Reichenbach, et al., 2021), with established processing parameters for constructing an UT template (Squara, Manig, et al., 2023) comprehensively covering untargeted (unknown) and targeted (known) components. Optimized parameters for UT fingerprinting include a signal-to-noise (S/N) threshold of 50 data points (dp) to include a peak or peak-region into a template, a distance threshold of 10 data points in the Mass Spectrometry (MS) channel as search space in the retention times domain, along with a direct match factor (DMF) threshold of 700 to reliably align UT features across all chromatograms (Squara, Manig, et al., 2023; Stilo et al., 2019) and to limit false positive matches for features with inconsistent MS spectral signatures *vs* a template reference. Given that the volatilome fraction was analyzed on a dual detector instrumentation (*i.e.*, GC×GC–MS/FID), the FID response was used to generate the features data matrix; it was chosen to minimize the bias caused by the detector performance fluctuations over one year of sample analysis and acquisitions. The MS trace, in its turn, was exploited for identification/identity confirmation purposes and to guide

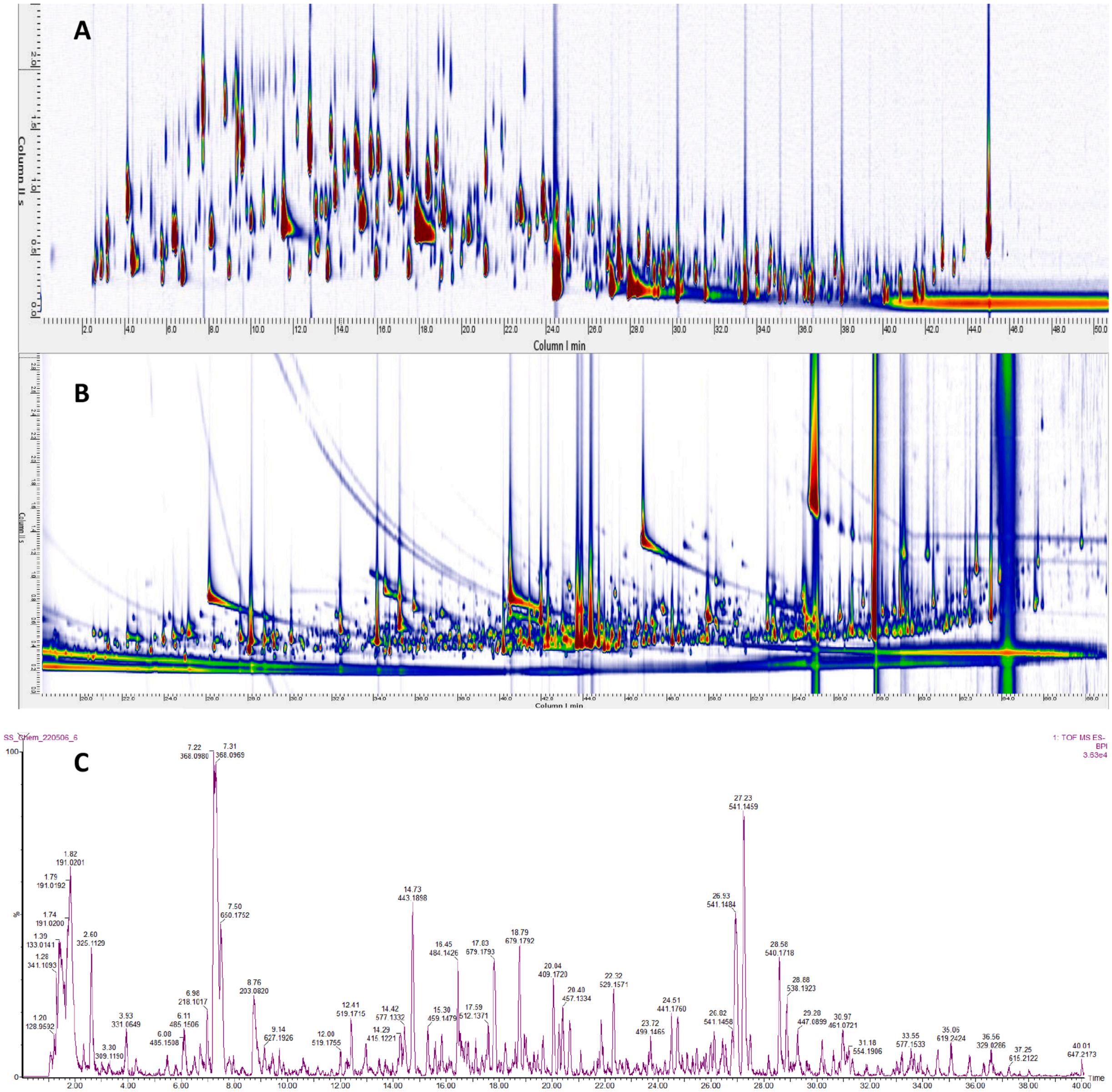

Fig. 1. (A) GC×GC-FID chromatogram illustrating the volatilome fraction of a TGT sample. (B) GC×GC–MS chromatogram illustrating the primary metabolome fraction of a TGT sample. (C) LC-MS chromatogram illustrating the specialized metabolome fraction of a TGT sample.

template realignment between the analyzed batches. The primary and specialized metabolome datasets were analyzed in a single batch, thus without relevant detector fluctuation bias. Fig. 1 shows the chromatograms of the three fractions of one exemplary sample, while Table 1 reports the list of targeted compounds, together with retention times and retentions indices (experimental and tabulated).

Within the volatilome many different chemical classes were detected. This fraction is correlated to hazelnut aroma quality due to the presence of key-aroma compounds in specific amounts; these odorants evoke the peculiar aroma profile and qualify aroma identity of the product (Caratti et al., 2023; Dunkel et al., 2014; Kiefl, Pollner, & Schieberle, 2013; Squara, Stilo, Cialiè Rosso, Liberto, Spigolon, et al., 2022). Moreover, the volatilome brings information about fat quality since it includes secondary products of lipid oxidation such as linear saturated aldehydes. These analytes in suitable amounts and relative ratios can be considered decision-makers for storage conditions and time (Ortega-Gavilán, Squara, Cordero, Cuadros-Rodríguez, & Bagur-González, 2023; Squara, Caratti, et al., 2022; Squara, Caratti, et al. 2023; Squara, Stilo, Cialiè Rosso, Liberto, Bicchi, et al., 2022). In addition, microbial development and related metabolism, enzymatic activation of the seeds in non-optimal post-harvest and storage, contribute to nuts spoilage and loss of industrial quality. These phenomena are detectable through the clear chemical signature of short chain fatty acids, lactones and primary alcohols (Stilo, Liberto, Spigolon, et al., 2021). Within the selected samples, covering two harvest countries (Italy and Turkey) and many cultivars/blends [Tonda Gentile Trilobata (TGT), Tonda Gentile Romana (TGR), Akçakoca (AKC), and Giresun (GIR)], optimal and sub-optimal post-harvest storage conditions [5 °C under vacuum (UV) or in standard atmosphere (SA)], it was expected to find information about all these quality traits, on the basis of which industry plans supplies and strategies, accompanied by the required sensory profile with perceivable positive attributes (nutty, malty, fresh, sweet floral, and fruity) and the absence of defects (rancid, mushroom like, metallic, stale, and solvent) (Kiefl et al., 2013; Squara, Stilo, Cialiè Rosso, Liberto, Spigolon, et al., 2022). The impact of post-harvest drying and storage conditions was extensively studied in previous research, interested readers can refer to available literature (Alasalvar, Shahidi, Ohshima, et al., 2003; Cialiè Rosso et al., 2018; Squara, Caratti, et al., 2023; Squara, Stilo, Cialiè Rosso, Liberto, Spigolon, et al., 2022).

About the volatilome known with its direct impact on aroma quality, the ketones class, and in particular 5-methyl-(*E*)-2-hepten-4-one, commonly known as filbertone, emerged as the most impactful odorant, characterized by a typical nutty and hazelnut-like aroma. Moreover, various aldehydes, including 2-methylpropanal, 2- and 3-methylbutanal, were also detected; they are associated with fruity, malty, nutty, and chocolate-like odors. Additionally, linear saturated and monounsaturated aldehydes were linked to perceptions of green, fatty, sweet floral, and fruity notes. The detection of five alcohols in raw hazelnuts further enriched the understanding of hazelnut aroma, with correlations between specific alcohols and distinctive sensory attributes such as dark chocolate, pungent, sweet, rancid, burnt, and wine-like notes. Aromatic hydrocarbons, although more abundant in roasted kernels, were also detected in raw hazelnuts, showcasing the method's sensitivity and improvement in mapping raw hazelnut volatilome (Alasalvar, Shahidi, & Cadwallader, 2003; Cialiè Rosso et al., 2018; Pedrotti et al., 2021; Squara, Stilo, Cialiè Rosso, Liberto, Spigolon, et al., 2022; Stilo et al., 2022).

The chemical dimensionality of primary metabolites includes a set of chemical classes, such as mono- and disaccharides, amino acids, low-molecular weight acids, and amines. Targeted analytes are reported in Table 1 in the form of derivatives together with retention times and $I^T s$. Among the primary metabolites identified in hazelnuts are glucose, fructose, sucrose, glutamine, alanine, and citric acid, alongside various amines. These constituents play pivotal roles in the nutritional composition and metabolic pathways of hazelnuts, contributing to their overall profile and potential health benefits. Moreover, this fraction is crucial in understanding the odorant formation during roasting (Cialiè Rosso, Stilo, Bicchi, et al., 2021; Squara, Stilo, Cialiè Rosso, Liberto, Spigolon, et al., 2022; Stilo et al., 2020) or informing about the stability of kernels against germination or bad post-harvest practices that do not properly or efficiently reduce moisture (Cialiè Rosso et al., 2020; Squara, Stilo, Cialiè Rosso, Liberto, Bicchi, et al., 2022; Squara, Stilo, Cialiè Rosso, Liberto, Spigolon, et al., 2022). Notably, while filbertone emerges as the primary odorant in roasted hazelnuts, the precursor(s) and formation pathway(s) remain elusive. Nonetheless, literature highlights various reactions occurring during thermal treatment, yielding potent odorants from primary metabolites, as a consequence of Maillard's reaction between reducing sugars and amino acids – especially lysine and arginine. A positive correlation between primary metabolites and odorous components developed during roasting was demonstrated by Cialiè Rosso *et al.* (Cialiè Rosso et al., 2020), which proposed the concept of aroma potential for hazelnut quality assessment.

3.1.2. Specialized non-volatile metabolome

The bioactive specialized metabolites found in hazelnut kernels and related plant portions were putatively identified (confidence level 3) based on relative retention, UV spectra and low- and high collision energy HR-MS data (Table 1).. Phenolic compounds emerged as the predominant class of metabolites. Within this category, phenolic acids – i.e., phenols featuring one carboxylic acid functional group, such as hydroxybenzoic and hydroxycinnamic acids as subcategories – emerged as the more abundant. In hazelnut kernels, several phenolic acids including mono-, di- and tri-hydroxybenzoic acids and their methoxylated and glycosylated derivatives were identified. This class of compounds has been demonstrated to possess antimicrobial activity, inhibiting the growth of pathogenic bacteria and fungi by disrupting their cell membranes and metabolic processes, thus inhibiting the growth of spoilage microorganisms and prolonging the product shelf life (Alasalvar & Bolling, 2015; Bottone et al., 2019; Jakopic et al., 2011; Pelvan, Olgun, Karadağ, & Alasalvar, 2018; Shahidi, Alasalvar, & Liyana-Pathirana, 2007). Moreover, it was found to correlate with geographical origin of samples likely because of the local pedo-climatic conditions impacting on the phenotype/chemotype expression (Ghisoni et al., 2020; Pelvan et al., 2018).

Flavonoids, polyphenolic compounds found abundantly in nature, exhibit diverse forms including aglycones, glycosides, and methylated derivatives. In the context of *C. avellana*, small quantities of aglycones are commonly present and occasionally contribute to the overall flavonoid content within the plant. The flavonols quercetin and myricetin were detected in hazelnut kernels and shells, while myricetin was also present in shells and skins. Flavonoid glycosides observed in hazelnuts are primarily *O*-glycosides, characterized by a sugar portion comprising one or two units. In hazelnuts, tannins are particularly abundant. Condensed tannins exist as oligomers or polymers classified into procyanidins and propelargonidins based on the flavan-3-ol unit. Specifically, hazelnut kernels were found to contain dimeric procyanidins with B- and A-type linkages, along with a series of B-type trimers. These compounds are present in the kernels and skins and possess antioxidant activity in addition to their antimicrobial activity. They help protect the kernels from oxidative damage caused by environmental stressors such as ultraviolet radiation and reactive oxygen species (ROS). By scavenging free radicals, flavonoids maintain the integrity of cellular structures and preserve the quality of the kernels during growth and maturation (Bottone et al., 2019; Fanali et al., 2018; Ghisoni et al., 2020).

Lastly, diarylheptanoids, compounds characterized by the 1,7-diphenylheptane skeleton, are also commonly found in hazelnuts. Cyclic diarylheptanoids, called giffonins, were isolated from various parts of hazelnut plants including leaves, leaf covers, flowers, and shells (Bottone et al., 2019; Ngouta et al., 2021; Singldinger et al., 2018).

Table 1
Identified metabolites in the volatile metabolome, primary and specialized non-volatile metabolites listed together with their average retention times (^1t_R min, ^2t_R sec), experimental and tabulated I^T. For specialized metabolites, retention times (^1t_R min), accurate mass, molecular formula, mass accuracy (ppm) and the main MS^E fragment ions are listed.

Volatile metabolome (GC×GC–MS/FID and differential-flow modulation)				
Targeted features	**^1t_R min**	**^2t_R min**	**Experimenal I^T**	**Literature I^T**
Aldehydes and Ketones				
Propanal, 2-methyl	4.53	0.83	817	808
Butanal	5.42	0.74	881	875
2-Butanone	5.96	0.75	901	905
Butanal, 2-methyl-	6.25	0.84	913	915
Butanal, 3-methyl-	6.38	0.80	918	922
Pentanal	8.00	0.81	980	978
2,3-Butandione	8.21	0.64	988	982
2-Pentanone, 4-methyl	8.92	0.89	1011	1010
2-Pentanone, 3-methyl	9.25	0.95	1020	1016
3-Hexanone	10.42	0.98	1053	1051
3-Pentanone, 2,4-dimethyl-	11.00	1.09	1069	–
Hexanal	11.50	0.93	1083	1080
4-Heptanone	13.04	1.05	1124	1131
3-Penten-2-one	13.04	0.76	1124	1131
2-Hexanone, 5-methyl-	13.67	0.97	1141	1142
4-heptanone, 3-methyl	13.96	1.17	1149	1178
2-Heptanone	15.13	0.97	1180	1175
Heptanal	15.25	0.98	1183	1186
Hexanal, 2-ethyl-	15.33	1.14	1185	1197
4-Heptanone, 2,6-dimethyl	15.92	1.02	1201	1189
2-Heptanone, 6-methyl-	17.21	1.02	1237	1228
2-Octanone	18.75	1.12	1279	1281
Acetoin	19.00	0.55	1286	1283
Octanal	19.08	1.01	1288	1289
5-methyl-(*E*)-2-hepten-4-one	19.1	1.12	1290	1290
2-Hexenal, 2-ethyl-	19.42	1.01	1301	1310
(*E*)-2-Heptenal	20.29	0.85	1322	1332
2-Nonanone	22.54	1.01	1387	1390
Nonanal	22.75	1.03	1393	1390
2-Decanone	25.92	0.76	1491	1495
Decanal	26.13	0.91	1497	1496
Benzaldehyde	26.92	0.64	1527	1530
(*E*)-2-Nonenal	27.17	0.77	1538	1543
(*E*)-2-Decenal	29.75	0.66	1632	1630
Tetradecanal	34.75	0.68	1920	1919
Alcohols				
2-Butanol	9.38	0.58	1024	1022
1-Propanol	9.83	0.55	1037	1037
1-Propanol, 2-methyl-	11.75	0.56	1090	1086
2-Pentanol	12.71	0.59	1115	1115
1-Butanol	13.58	0.56	1139	1146
2-Pentanol, 3-methyl-	15.54	0.65	1196	1202
1-Butanol, 2-methyl-	15.88	0.57	1201	1208
2-Hexanol	16.38	0.63	1214	1211
1-Pentanol	17.42	0.58	1242	1244
4-Heptanol	18.67	0.68	1277	1281
2-Heptanol	19.92	0.65	1312	1318
1-Hexanol	21.08	0.60	1345	1340
1-Hexanol, 2-ethyl-	25.67	0.61	1483	1484
4-Heptanol, 2,6-dimethyl	26.25	0.67	1502	1509
2,3-Butanediol	27.00	0.43	1530	1545
Ethanol, 2-butoxy-	22.92	0.62	1397	1405
2,3-Epoxyhexanol	23.96	0.74	1430	1428
(*E*)-*p*-2-Menthen-1-ol	27.38	0.64	1544	1563
1-Octanol	27.50	0.55	1549	1546
2,3-Butanediol	27.92	0.42	1565	1570
Menthol	29.54	0.58	1632	1626
Ethanol, 2-(2-butoxyethoxy)-	32.58	0.50	1786	1786
2-Phenylethanol	34.54	0.45	1906	1915
Carboxylic acids				
Acetic acid	24.50	0.41	1447	1449
Formic acid	26.29	0.45	1503	1510
Propanoic acid	27.13	0.41	1535	1534
Butanoic acid	29.29	0.38	1621	1637
Butanoic acid, 3-methyl-	30.21	0.38	1663	1666
Pentanoic acid	31.54	0.39	1729	1733
Hexanoic acid	33.42	0.38	1836	1839
Heptanoic acid	35.08	0.38	1943	1946
Octanoic acid	36.58	0.36	2048	2046
Nonanoic acid	38.00	0.38	2154	2144

(*continued on next page*)

Table 1 (*continued*)

Volatile metabolome (GC×GC–MS/FID and differential-flow modulation)				
Decanoic acid	39.29	0.35	2258	2265
Esters				
Allyl acetate	4.33	0.62	839	
Butyl acetate	11.25	1.04	1076	1075
Butanoic acid, butyl ester	16.54	1.18	1218	1221
Butyrolactone	29.50	0.51	1630	1626
2-Octynoic acid, methyl ester (ISTD)	29.92	0.64	1650	
γ-Hexalactone	31.08	0.55	1704	1703
Butyl benzoate	33.92	0.59	1866	1856
γ-Octalactone	34.79	0.51	1923	1916
γ-Nonalactone	36.21	0.42	2021	2028
Hexanedioic acid, bis(2-methylpropyl) ester	37.63	0.56	2125	2119
Terpenes				
α-Pinene	9.29	1.67	1022	1022
α-Thujene	9.54	1.55	1028	1010
α-Pinene	9.29	1.67	1022	1022
α-Thujene	9.54	1.55	1028	1010
Camphene	10.71	1.66	1061	1066
β-Pinene	12.17	1.69	1101	1110
Sabinene	12.75	1.52	1117	1120
δ-Carene	13.79	1.61	1144	1147
β-Myrcene	14.42	1.32	1161	1159
α-Terpinene	15.04	1.37	1178	1178
Limonene	15.63	1.43	1193	1199
β-Phellandrene	16.00	1.47	1203	1205
γ-Terpinene	17.38	1.45	1241	1243
p-Cymene	18.33	1.23	1268	1268
α-Terpinolene	18.75	1.46	1279	1280
cis-Sabinene hydrate	24.96	0.77	1461	1451
Terpinen-4-ol	28.67	0.65	1593	1600
Others				
Hexane	2.67	0.66	600	600
Octane	4.08	1.14	800	800
1-Octene	4.75	1.14	855	837
Cyclohexane, ethyl-	5.54	1.32	885	885
Nonane	5.92	1.57	900	900
Furan, 2,5-dimethyl-	7.33	0.86	955	952
n-Butyl ether	7.67	1.49	967	974
Decane	8.71	1.91	1000	1000
Furan, 2-ethyl-5-methyl-	9.75	1.04	1034	1028
Toluene	9.96	0.96	1040	1036
1-Decene	10.08	1.68	1043	1052
Furan, 2,3,5-trimethyl-	10.54	1.00	1056	1051
Undecane	12.00	2.11	1100	1100
Ethylbenzene	13.54	1.05	1138	1134
p-Xylene	13.54	1.05	1138	1134
Undecane, 5-methyl-	13.96	1.97	1149	1157
Dodecane	15.75	2.12	1200	1200
Furan, 2-pentyl-	16.96	1.12	1230	1230
Styrene	17.83	0.99	1254	1251
Benzene, 1,2,4-trimethyl-	18.75	1.12	1279	1275
Hexanenitrile	19.42	0.88	1297	1303
Hexane, 1,1′-oxybis-	21.63	1.68	1360	1367
α-Thujone (ISTD)	23.79	1.06	1424	1429
β-Thujone (ISTD)	24.46	1.04	1445	1451
Primary metabolome (GC×GC-TOF MS and loop-type thermal modulation)				
Targeted features	**^1t_R min**	**^2t_R min**	**Experimenal I^T**	**Literature I^T**
Amino acids				
Alanine 2TMS	10.08	1.45	1103	1110
Valine TMS	10.33	1.59	1111	1105
Glycine 2TMS	11.10	1.63	1136	1149
Isoleucine TMS	13.25	1.89	1201	1189
Valine 2TMS	14.17	2.31	1231	1221
Leucine 2TMS	15.95	1.85	1288	1299
Serine 2TMS	16.04	1.95	1275	1267
Proline 2TMS	16.42	1.79	1300	1308
Isoleucine 2TMS	16.75	1.75	1311	1302
Glycine 3TMS	16.92	1.75	1316	1317
Serine 3TMS	17.33	2.21	1328	1342
Methionine TMS	20.47	2.19	1431	1420
Methionine 2TMS	23.42	1.95	1526	1534
Phenylalanine TMS	24.25	2.35	1554	1554
Cysteine 3TMS	24.87	1.93	1575	1568
Ornithine 3TMS	26.08	1.90	1617	1624
Phenylalanine 2TMS	27.00	2.07	1649	1649
Ornithine 4TMS	31.92	1.85	1828	1814

(*continued on next page*)

Table 1 (*continued*)

Volatile metabolome (GC×GC–MS/FID and differential-flow modulation)				
Tyrosine 2TMS	33.83	3.48	1903	1901
Tryptophan 2TMS	41.08	1.77	2236	2236
Sugars				
Glycerol 3TMS	15.83	1.61	1282	1284
Threitol 4TMS	22.92	1.63	1509	1502
Erythritol 4TMS	23.17	1.63	1517	1508
Arabinose 4TMS ether ethyl oxime	27.50	2.27	1666	1662
1-Tridecanol TMS	27.92	1.62	1673	1661
Ribose 4TMS	27.92	1.78	1680	1668
Xylitol 5TMS	29.17	1.73	1723	1710
Fructofuranoside, methyl 1,3,4,6tetrakis-*O*-TMS	30.95	1.81	1786	1799
Fructofuranose, pentakis(trimethylsilyl) ether (a)	31.33	1.71	1797	1800
Fructofuranose, pentakis(trimethylsilyl) ether (b)	31.58	1.71	1816	1836
Fructose 5TMS (*anti*)	32.83	1.79	1862	1867
Fructose 5TMS (*syn*)	33.33	1.77	1881	1875
Glucose 5TMS	33.75	1.79	1900	1902
Mannitol 6TMS	34.81	1.73	1941	1928
Glucopyranose 5TMS derivative	35.63	1.74	1978	1971
Scyllo-Inositol 6TMS	36.99	1.77	2076	2090
Myo-Inositol 6TMS	38.25	1.87	2086	2096
Sucrose 8TMS	50.15	1.91	2620	2610
Maltose 8TMS	50.97	1.87	2686	2693
Organic acids				
Lactic acid 2TMS	9.17	1.55	1071	1068
Glycolic acid 2TMS	9.96	1.73	1099	1085
Pyruvic acid 2TMS	10.00	1.60	1100	1108
Oxalic acid 2TMS	11.42	1.97	1145	1150
Pyruvic acid oxime 2TMS	11.50	1.79	1147	1149
Succinic acid 2TMS	17.03	1.98	1321	1314
Fumaric acid 2TMS	18.50	1.92	1368	1358
Lactic acid dimer 2TMS	19.67	2.07	1405	1394
Glutaric acid 2TMS	20.25	1.97	1424	1413
Malic acid 2TMS	22.05	1.77	1481	1478
Malic acid 3TMS	22.05	1.99	1481	1478
Pyroglutamic acid TMS	22.58	3.67	1504	1511
Adipic acid 2TMS	23.58	1.99	1531	1522
2-Hydroxyglutaric acid 3TMS	25.33	1.95	1591	1589
2-Hydroxyadipic acid 3TMS	25.38	2.01	1593	1589
Glutamic acid 3TMS	25.67	1.81	1603	1612
3-Hydroxy-3-methylglutarate	26.08	1.91	1617	1619
Tartaric acid 4TMS	27.00	1.93	1656	1665
Pantothenic acid 3TMS	35.85	2.33	1987	1985
Galactonic acid 6TMS	35.87	1.85	1988	1981
Gluconic acid 6TMS	35.92	2.39	1984	1997
Galactaric acid 6TMS	36.25	1.95	2003	2014
Galacturonic acid 5TMS	36.42	1.93	2011	2010
Specialized non-volatile metabolome (LC-(HR)TOF MS)				
Targeted features	**t_R min**	**[MH]$^-$**	**Formula (ppm)**	**Fragments**
Phenolic acids				
Trihydroxybenzoic acid-*O*-hexoside (a)	3.93	331.0657	$C_{13}H_{15}O_{10}$ (−2.4)	169.0138, 125.234
Hydroxybenzoic acid-*O*-hexoside	5.82	299.0755	$C_{13}H_{15}O_8$ (−4.0)	137.0248, 93.0344
Dihydroxybenzoic acid-*O*-hexoside (a)	6.77	315.0715	$C_{13}H_{15}O_9$ (−0.3)	152.0119, 108.0201
Trihydroxybenzoic acid-*O*-hexoside (b)	7.76	331.0653	$C_{13}H_{15}O_{10}$ (−3.6)	169.0140, 125.239
Hydroxytyrosol-*O*-hexoside	8.98	315.1083	$C_{14}H_{19}O_8$ (1.0)	153.0552
Hydroxy-methoxybenzoic acid-*O*-hexoside (a)	9.18	329.0875	$C_{14}H_{17}O_9$ (0.6)	123.0442
Hydroxy-methoxybenzoic acid-*O*-hexoside (b)	10.67	329.0865	$C_{14}H_{17}O_9$ (−2.4)	165.0538
Hydroxy-methoxycinnamic acid-*O*-hexoside	15.57	355.1016	$C_{16}H_{19}O_9$ (−3.7)	193.0515,134.0360
Hydroxymethoxybenzoic acid-*O*-pentosyl hexoside	20.66	461.1299	$C_{19}H_{25}O_{13}$ (0.9)	167.0349,123.0451
Flavan-3-ols				
Procyanidin trimer (a)	8.37	865.1974	$C_{45}H_{37}O_{18}$ (−0.6)	575.1229,407.0750,287.0522,125.0240
Catechin	14.42	289.0705	$C_{15}H_{13}O_6$ (−2.4)	125.0241
Procyanidin dimer (a)	14.44	577.1323	$C_{30}H_{25}O_{12}$ (−4.0)	425.0859,407.0740,289.0718,125.0238
(Epi)gallocatechin	16.54	305.0681	$C_{15}H_{13}O_7$ (6.6)	125.0237
Procyanidin trimer (b)	17.38	865.1992	$C_{45}H_{37}O_{18}$ (1.4)	125.0242
Procyanidin dimer (b)	17.68	577.1334	$C_{30}H_{25}O_{12}$ (−2.1)	425.0891,407.0757,289.0735,125.0242
Procyanidin trimer (c)	18.42	865.1986	$C_{45}H_{37}O_{18}$ (0.7)	125.0244
Epicatechin	19.27	289.0717	$C_{15}H_{13}O_6$ (1.7)	125.0236
Procyanidin trimer (d)	19.38	865.1978	$C_{45}H_{37}O_{18}$ (−0.2)	407.0710,287.0535,125.0245
Procyanidin trimer (e)	19.72	865.1993	$C_{45}H_{37}O_{18}$ (1.5)	577.1387,407.0734,289.0702,125.0236
Procyanidin trimer (f)	22.17	865.1965	$C_{45}H_{37}O_{18}$ (−1.7)	407.0810,289.0726,125.0245
Procyanidin dimer (c)	22.62	577.1351	$C_{30}H_{25}O_{12}$ (0.9)	425.0996,407.0798,289.0717,125.0239
(Epi)catechin gallate	24.98	441.0845	$C_{22}H_{17}O_{10}$ (5.2)	289.0692,169.0155,125.0240
Procyanidin trimer (a)	8.37	865.1974	$C_{45}H_{37}O_{18}$ (−0.6)	575.1229,407.0750,287.0522,125.0240
Flavonols				
Myricetin-*O*-deoxyhexoside	25.64	463.0877	$C_{21}H_{19}O_{12}$ (1.9)	316.0923
Quercetin di-*O*-glucoside	26.63	625.1415	$C_{27}H_{29}O_{17}$ (1.6)	463.0882

(*continued on next page*)

Table 1 (*continued*)

Volatile metabolome (GC×GC–MS/FID and differential-flow modulation)				
Chalcones				
Phloretin-*O*-hexoside	30.20	435.1289	$C_{21}H_{23}O_{10}$ (−0.5)	273.0760
Diarylheptanoids				
Giffonin T	19.18	505.1721	$C_{25}H_{29}O_{11}$ (2.2)	343.1198
Giffonin P (a)	21.25	361.1281	$C_{19}H_{21}O_{7}$ (−1.7)	–
Giffonin P (b)	25.54	361.1294	$C_{19}H_{21}O_{7}$ (1.9)	–
Giffonin O	27.48	343.1174	$C_{19}H_{19}O_{6}$ (−2.3)	–
Giffonin S	33.67	321.1119	$C_{20}H_{17}O_{4}$ (−2.5)	306.0860
Others				
Dioxindole-3-acetic acid-*O*-hexoside	7.35	368.0982	$C_{16}H_{18}NO_{9}$ (0.0)	144.0452
Tryptophan	8.76	203.0821	$C_{11}H_{11}N_{2}O_{2}$ (0.0)	116.0499
Dioxindole-3-acetic acid-*O*-hexoside-2-(2-oxoindolinyl)acetate (a)	26.80	541.1467	$C_{26}H_{25}N_{2}O_{11}$ (1.7)	292.0818,190.0502,146.0604
Dioxindole-3-acetic acid-*O*-hexoside-2-(2-oxoindolinyl)acetate (b)	26.95	541.1475	$C_{26}H_{25}N_{2}O_{11}$ (3.1)	292.0822,190.0500,146.0609
Dioxindole-3-acetic acid-*O*-hexoside-2-(2-oxoindolinyl)acetate (c)	27.25	541.1456	$C_{26}H_{25}N_{2}O_{11}$ (−0.4)	292.0821,190.0512,146.0609
Hazelnutin D	28.59	540.1714	$C_{24}H_{30}NO_{13}$ (−0.6)	488.1216,189.0764,144.0447
1,7-bis(hydroxyphenyl)heptane-3,5-diol-*O*-hexoside	30.31	477.2122	$C_{25}H_{33}O_{9}$ (−0.6)	315.1616
Hydroxy-1,7-bis(hydroxyphenyl)heptan-3-one-*O*-hexoside	31.11	475.1971	$C_{25}H_{31}O_{9}$ (0.6)	315.1616
1,7-bis(hydroxyphenyl)heptane-3,5-diol (a)	33.82	315.1604	$C_{19}H_{23}O_{4}$ (2.5)	149.0587
1,7-bis(hydroxyphenyl)heptane-3,5-diol (b)	34.63	315.1596	$C_{19}H_{23}O_{4}$ (0.0)	149.0606
Dimethoxy coumaryl alcohol-*O*-pentosyl hexoside	35.76	503.1774	$C_{22}H_{31}O_{13}$ (1.8)	209.0819
Hydroxy-1,7-bis(hydroxyphenyl)heptan-3-one	36.29	313.1442	$C_{19}H_{21}O_{4}$ (0.6)	149.0609
Dimethoxy-ellagic acid	36.56	329.0300	$C_{16}H_{9}O_{8}$ (0.9)	314.0078,298.9829,270.9879

3.2. Untargeted and targeted features signatures in the metabolome: Information potential of single fractions

Results were firstly explored by independently analyzing each metabolome fraction. Data matrices had the following dimensions: 130 samples × 442 variables (volatilome), 130 samples × 674 variables (primary metabolites), and 130 samples × 44 variables (specialized metabolites).

PCA was firstly performed on each dataset independently. Results illustrated as score plots are reported in Electronic Supplementary Material – **Supplementary Figure 1 SF1**. The score plots for each dataset with samples colored according to biological variables impacting the dataset are reported to have a comprehensive understanding. **Supplementary Figure 1 –SF1 A, B, C** illustrate the showcase samples colored according to their geographical origin. Within the three, the specialized metabolome fraction mainly represented by phenolic compounds (Fig. 2**A**) is the one with a better defined clustering. The five clusters that are reported in the plot represent the different cultivars/blends examined, and the two sub-clusters in the top-left quadrant are the Giresun samples (Turkey) with a different postharvest treatment. On the other hand, the primary metabolome (**Supplementary Figure 1 –SF1B**) shows the lowest diversity among the samples. Postharvest practices are

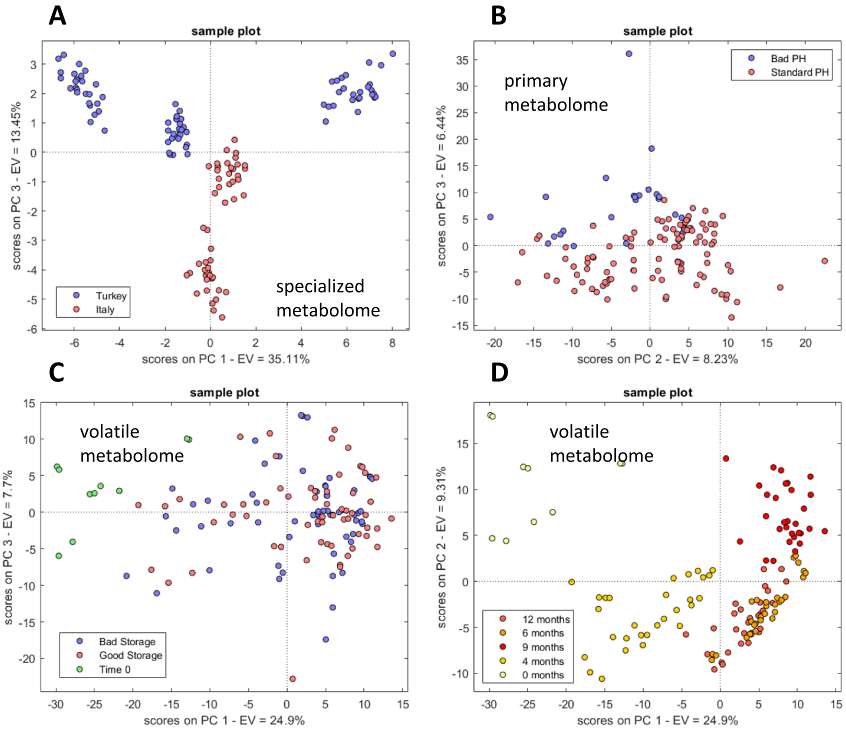

Fig. 2. PCA score plot illustrating natural samples clusters according to the detectable metabolome fractions and functional variables. **2A** refers to specialized metabolome comprising 130 samples and 44 variables describing harvest region; **2B** refers to the primary metabolome comprising 130 samples and 442 variables as a function of post-harvest practices (bad PH and good PH); **2C** and 2D refer to the volatilome comprising 130 samples and 674 variables as a function of storage conditions (**2C**) and storage time (**2D**). The data is presented after Z-score normalization.

visualized in **Supplementary Figure 1 –SF1 D, E, F**. Similar to the geographical origin, the phenolic fraction offers the clearer discrimination, but satisfactory results are also achieved with the other two datasets with primary metabolome showing a fairly differentiation of samples along PC3 as shown in Fig. 2**B**. Storage conditions and storage months are highlighted in **Supplementary Figure 1 –SF1 G, H, I** and **J, K, L**, respectively; in contrast to geographical origin and postharvest practices, the phenolic fraction (**Supplementary Figure 1 –SF1 I**) is not influenced by the industrial shelf life. On the other hand, the most varying fraction according to the storage conditions (Fig. 2**C**) and time is the volatilome, which has a clear trend on PC1 reflecting the storage months as shown in Fig. 2**D.**

PLS-DA classification models with three LVs were created and cross-validated to better understand the prospective of each of these fractions to be used as a predictive tool. The choice of using no more than three LVs was made to avoid overfitting and to draw fallacious conclusions, since a different dataset for blind validation was not available. Monte Carlo (80–20 % dataset split) with 1,000 iterations and maximum assignation criteria was used as the cross-validation strategy. Table 2 reports the confusion matrices of each classification model with their overall accuracy, expressed as percentage ratio of the correctly assigned predictions and the total number of predictions. As previously mentioned, the predictive capabilities are in line with the PCA observations. Geographical origin and postharvest practices present greater/higher accuracy percentage over all the investigated fractions, with the specialized metabolome being the one with optimal prediction, and the primary metabolome being the worst performing (≈ 88 % overall accuracy) with regard to geographical origin. Overall, satisfactory results were achieved. Regarding the prediction of storage time and conditions, the general accuracy achieves lower results. The volatilome is the best performing fraction among the three for storage conditions and months with ≈ 94 % and ≈88 % of overall accuracy, respectively. The specialized metabolome is the fraction that is less influenced by the latter external variables: its prediction accuracy is the worst among the three, indeed achieving only ≈ 57 % and 20 % of overall accuracy for storage conditions and duration, respectively.

VIP scores were calculated for each model; in PLS-DA, the VIP score for each variable quantifies its contribution to the model's ability to discriminate between classes. Higher VIP scores indicate greater importance in explaining the differences between classes. The variables with the average 15 highest VIPs for each fraction are reported in the Electronic Supplementary Material – **Supplementary Figure 2 SF2.**

Table 2

Performance evaluation, expressed as confusion matrices, % overall accuracy, R2, and Q2, of the classification models created after separately processing volatilome, primary metabolome, and specialized metabolome datasets.

	Real/ predicted	Turkey	Italy	Overall accuracy	R^2	Q^2
Volatile metabolome	**Turkey**	15229	341	**98.35 %**	0.890	0.715
	Italy	89	10341			
Primary metabolome	**Turkey**	13981	1570	**87.73 %**	0.610	0.474
	Italy	1620	8829			
Specialized metabolome	**Turkey**	15517	0	**100.00 %**	0.967	0.744
	Italy	0	10483			

	Real/ predicted	Bad PH	Standard PH	Overall accuracy	R^2	Q^2
Volatile metabolome	**Bad PH**	5232	0	**99.35 %**	0.842	0.727
	Standard PH	170	20598			
Primary metabolome	**Bad PH**	4789	356	**97.36 %**	0.574	0.393
	Standard PH	330	20525			
Specialized metabolome	**Bad PH**	5271	0	**100.00 %**	0.899	0.808
	Standard PH	0	20729			

	Real/ predicted	Bad Storage	Good Storage	Time 0	Overall accuracy	R^2	Q^2
Volatile metabolome	**Bad Storage**	11175	855	6	**93.53 %**	0.842	0.727
	Good Storage	790	11135	22			
	Time 0	10	0	2007			
Primary metabolome	**Bad Storage**	5513	5174	1281	**53.44 %**	0.396	0.294
	Good Storage	2862	7524	1712			
	Time 0	431	645	858			
Specialized metabolome	**Bad Storage**	8253	2596	1112	**56.77 %**	0.422	0.226
	Good Storage	3836	6223	1995			
	Time 0	273	1429	283			

	Real/ predicted	12 months	9 months	6 months	4 months	0 months	Overall accuracy	R^2	Q^2
Volatile metabolome	**12 months**	5244	501	0	224	0	**88.19 %**	0.669	0.586
	9 months	413	4739	292	22	0			
	6 months	132	267	6018	74	0			
	4 months	21	38	126	5581	556			
	0 months	0	0	0	404	1348			
Primary metabolome	**12 months**	3417	1262	314	0	1009	**48.37 %**	0.209	0.153
	9 months	3055	1910	286	83	661			
	6 months	330	315	3202	2185	0			
	4 months	30	14	2621	3313	31			
	0 months	803	347	1	76	735			
Specialized metabolome	**12 months**	1265	1615	957	950	1144	**19.85 %**	0.309	0.024
	9 months	1120	1682	1298	1367	615			
	6 months	978	1648	1230	1368	714			
	4 months	1012	1894	1530	720	898			
	0 months	688	337	482	224	264			

Within the volatilome fraction, analytes that on average perform better in discriminating the external/functional variables are aldehydes, alcohols, and terpenes; within them, aldehydes such as hexanal, heptanal, and octanal have higher VIPs with the storage months model, terpenes such as δ-carene and α-terpinolene are relevant with the storage conditions model, alcohols such as 1-pentanol, 2-pentanol, and 4-heptanol have higher scores when it comes to discriminating geographical origin, and last but not least ketones such as 5-methyl-2-heptanone have higher scores when discriminating post-harvest practices.

Within the primary metabolome, carbohydrates are those more capable to differentiate both geographical origin and post-harvest practices, the former due to the different cultivars and pedoclimatic conditions of the investigated samples, and the latter likely because of an enzymatic activation caused by the higher moisture levels in the early stages after harvest (Cialiè Rosso et al., 2020; Cialiè Rosso, Stilo, Bicchi, et al., 2021; Cialiè Rosso, Stilo, Mascrez, et al., 2021). Since the discrimination capabilities of the remaining models were not successful, the VIPs on the remaining models will not be further explored. Lastly, within the specialized metabolome, diaryl heptanoids (i.e., giffonins) show the strongest contributions to the geographical origin and post-harvest models.

The distribution of the mentioned analytes in the different classes is illustrated in Fig. 3. Regarding storage times and conditions, a general increase in the saturated aldehydes is due to the lipid oxidation process that naturally occurs on the hazelnut lipid fraction, while an increase in the amount of terpenoids in the unproperly stored samples is likely associated to a direct expression of the plant phenotype/chemotype, informing for the presence of bacteria and molds (Squara, Caratti, et al.,

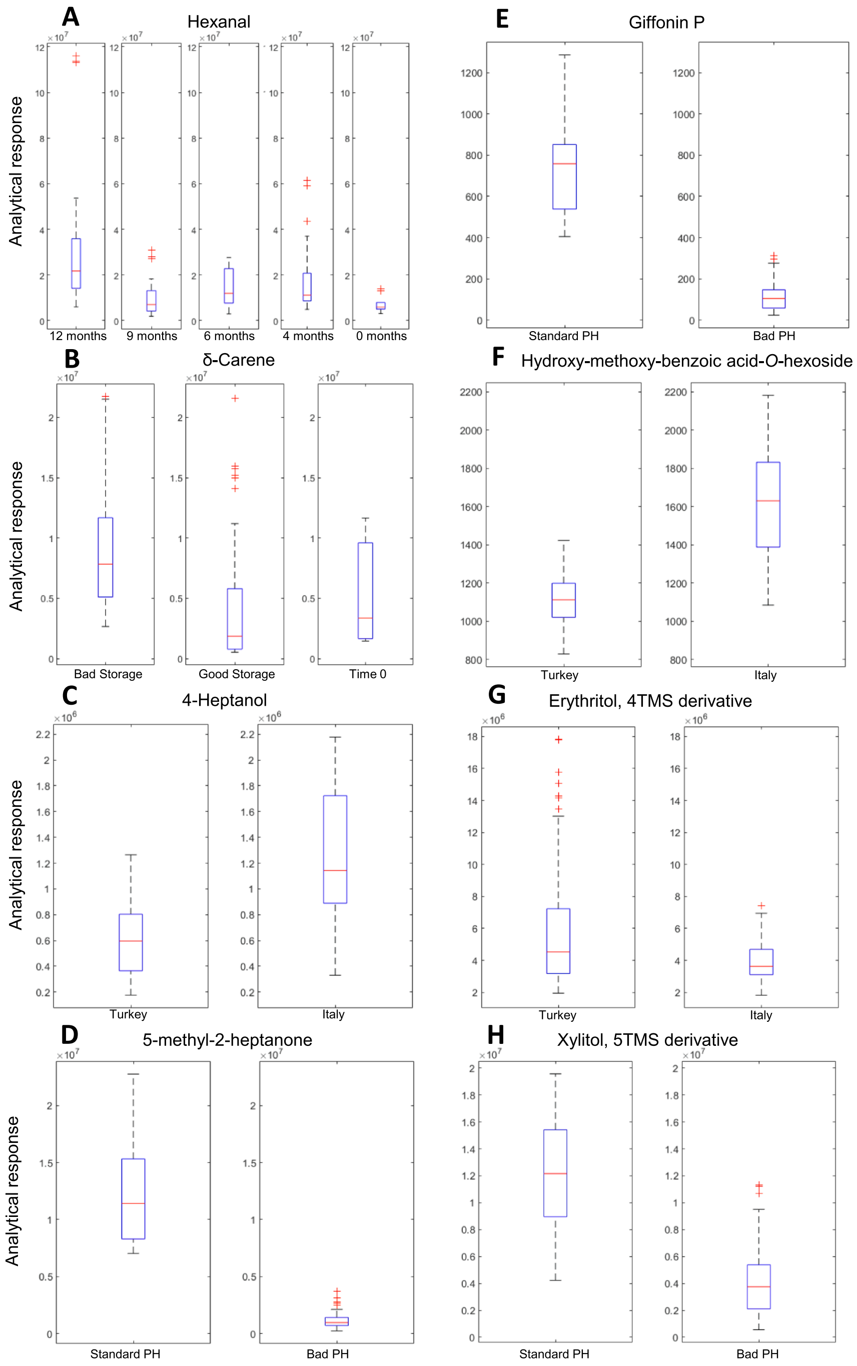

Fig. 3. Distribution of target analytes belonging to the different fractions: Volatilome (A, B, C, D), specialized metabolome (E, F), and primary metabolome (G, H).

2023; Stilo, Liberto, Spigolon, et al., 2021). When analyzing the specialized metabolome, giffonins are more abundant in samples that underwent a standard post-harvest procedure; these analytes serve as natural antioxidants and are likely being depleted in badly-dried samples, as well as carbohydrates (primary metabolome) that are probably consumed by the plant due to their higher metabolic activity induced by the higher moisture levels.

3.3. Data fusion improves quality prediction

LLDF can be achieved by placing datasets next to each other, studying different variables of the same set of samples, or coupling them across variables, measuring the same variables on samples from different batches. Often referred to as data augmentation or multi-block analysis, LLDF retains all original information from diverse sources but may introduce noise and redundant information, impacting modeling precision (Dankowska & Kowalewski, 2019; de Juan & Tauler, 2019). Different classical variable normalization methods, such as autoscaling or Pareto scaling, needs to be applied to each data matrix preemptively to equalize variance while preserving the variance ratio between variables within a block. Other normalization methods can be also chosen, such as mean-centering, root square scaling, and log scaling. This preliminary operation avoids that one matrix block prevails among the others (Azcarate et al., 2021). Moreover, given that the number of variables is much higher than the number of observations, the risk of creating models that are prone to overfitting exist. To try to minimize such risk, one of the key aspects is choosing the appropriate classification algorithm. Different classification algorithms were tested in the past with similar data matrix(Ortega-Gavilán et al., 2023), where, for instance, SIMCA, PLSDA, and SVM were tested. Between the three, SIMCA was the one performing the worse, while SVM was overfitting; PLS-DA, on the other hand, achieved satisfactory results but at the same

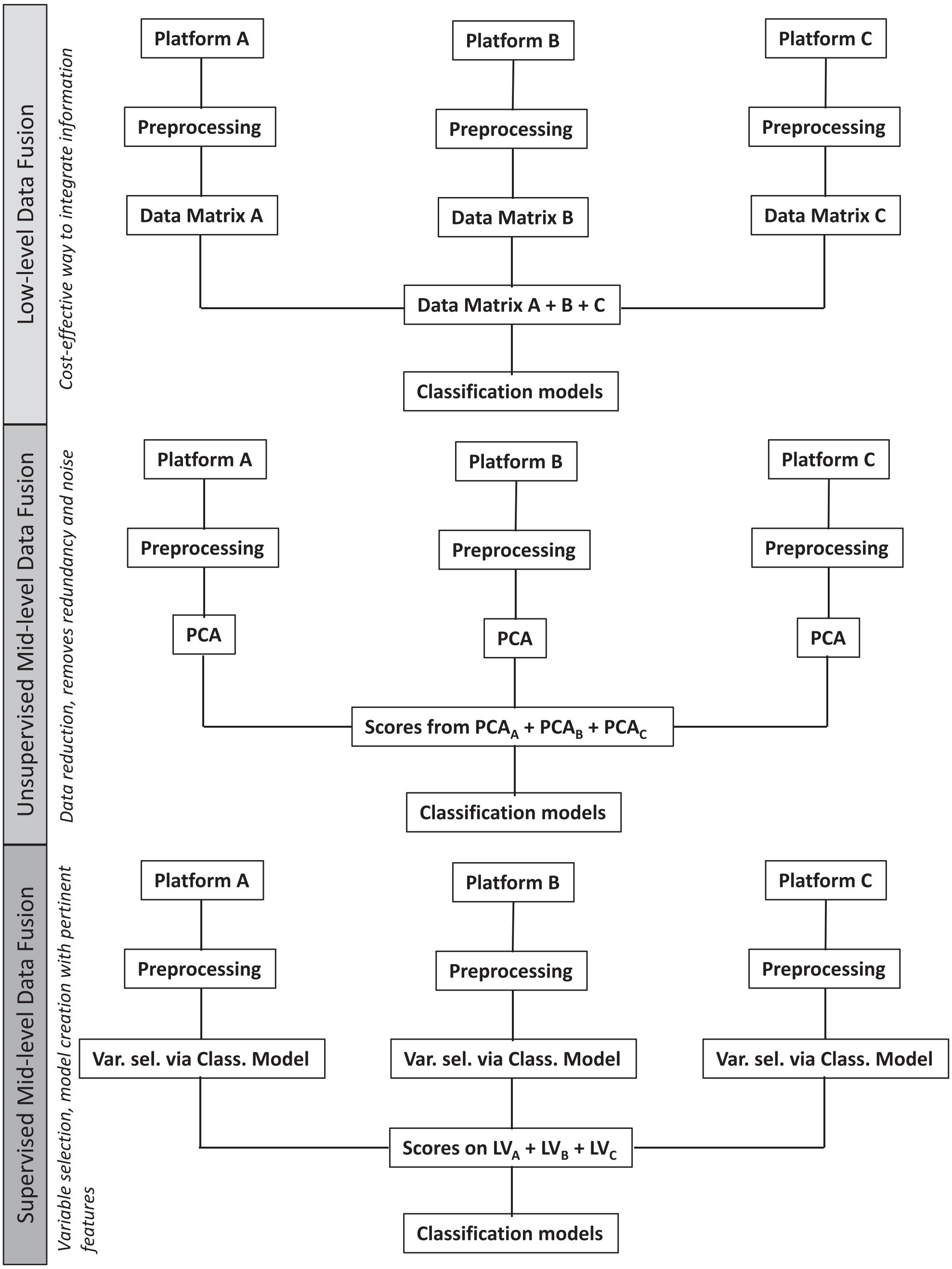

Fig. 4. Schematic flowchart representing low-level, unsupervised and supervised mid-level data fusion steps.

time was less prone to overfitting given that the model was created by using only three latent variables. Moreover, to be aware of possible overfitting, the similarity between R^2 and Q^2 was monitored.

MLDF is a two-step methodology: initially, pertinent features from individual data sources are extracted. Secondly, these extracted features are concatenated to generate a unified matrix for further processing. This approach necessitates a thorough evaluation of results in terms of raw variables, determining the connection between each feature's salience in the final model and its corresponding pattern in the original variables. The first step of MLDF involves calculating LVs or selecting features independently obtained from each analytical data matrix. This could be performed either via unsupervised or supervised algorithms, such as PCA and PLS-DA, respectively. Compared to LLDF, MLDF often results in improved classification performances due to the feature reduction step, which accounts for non-informative variance. For explanatory purposes the MLDF steps with a supervised algorithm are presented: PLS is employed on each matrix block for dimension reduction, and the scores corresponding to selected LVs are concatenated. Subsequently, another supervised technique (e.g., PLS-DA, SVM, SIMCA etc.) is applied to the concatenated score matrix to derive the final model (Alamar, Caramês, Poppi, & Pallone, 2020; Silvestri et al., 2014).

High-level data fusion (HLDF) operates at the decisions level. The classification of samples is conducted on each matrix block independently. Predictions from these independent models are then combined using different strategies. The consensus strategies applied to the predictions obtained from the single-block models include majority voting, which involves directly merging the predictions of single models, and Bayesian consensus with discrete probability distributions, that estimates the probability that samples belong to a specific class for each information source and combines these preliminary identity declarations to provide a fused probability (Ballabio et al., 2018; Fernández et al., 2012). HLDF approaches are less prevalent in the field of analytical chemistry, likely due to their higher complexity compared to LLDF and MLDF approaches. Many studies (Di Anibal, Callao, & Ruisánchez, 2011; Márquez, López, Ruisánchez, & Callao, 2016; Rodionova & Pomerantsev, 2023) highlight that this level primarily contributes to enhancing predictive accuracy compared to those obtained from individual models, but it does not provide pertinent information about biological variables. The three approaches are visually summarized in flowcharts in Fig. 4.

To achieve better classification performance, different data fusion levels were tested: low-level data fusion, mid-level unsupervised data fusion, and mid-level supervised data fusion. With LLDF, data matrices from the different techniques were merged in a single data matrix after each original block was centered and scaled. With unsupervised MLDF (UMLDF), a PCA was performed on each original data matrix after

Table 3
Performance evaluation, expressed as confusion matrices, % overall accuracy, R^2, and Q^2, expressed as confusion matrices, of the classification models created after different data fusion techniques combining information from all metabolome fractions.

	Real/predicted	Turkey	Italy	Overall accuracy	R^2	Q^2			
LLDF	**Turkey**	15534	98	**99.61 %**	0.888	0.674			
	Italy	3	10365						
UMLDF	**Turkey**	15642	2	**99.99 %**	0.923	0.670			
	Italy	1	10355						
SMLDF	**Turkey**	15546	0	**100.00 %**	0.938	0.774			
	Italy	1	10453						
	Real/predicted	**Bad PH**	**Standard PH**	**Overall accuracy**	**R^2**	**Q^2**			
LLDF	**Bad PH**	5149	0	**100.00 %**	0.934	0.778			
	Standard PH	0	20851						
UMLDF	**Bad PH**	15642	2	**99.99 %**	0.932	0.817			
	Standard PH	1	10355						
SMLDF	**Bad PH**	5206	0	**100.00 %**	0.949	0.763			
	Standard PH	0	20794						
	Real/predicted	**Bad Storage**	**Good Storage**	**Time 0**	**Overall accuracy**	**R^2**	**Q^2**		
LLDF	**Bad Storage**	10032	1920	13	**83.40 %**	0.557	0.315		
	Good Storage	2162	9763	68					
	Time 0	26	126	1890					
UMLDF	**Bad Storage**	11094	695	17	**94.15 %**	0.835	0.495		
	Good Storage	709	11433	0					
	Time 0	32	67	1953					
SMLDF	**Bad Storage**	11965	0	0	**99.91 %**	0.865	0.651		
	Good Storage	0	11972	0					
	Time 0	1	23	2039					
	Real/predicted	**12 months**	**9 months**	**6 months**	**4 months**	**0 months**	**Overall accuracy**	**R^2**	**Q^2**
LLDF	**12 months**	5286	367	219	116	9	**83.47 %**	0.497	0.279
	9 months	307	5301	434	14	1			
	6 months	651	621	4704	0	0			
	4 months	348	74	13	5335	219			
	0 months	50	116	0	738	1077			
UMLDF	**12 months**	5687	10	0	0	284	**93.12 %**	0.687	0.416
	9 months	5	5495	0	11	455			
	6 months	2	1	5917	132	3			
	4 months	67	1	3	5863	38			
	0 months	199	555	0	23	1249			
SMLDF	**12 months**	5817	57	0	10	0	**97.78 %**	0.851	0.614
	9 months	190	5835	1	15	0			
	6 months	4	16	5935	55	0			
	4 months	14	33	148	5817	0			
	0 months	1	32	2	0	2018			

scaling and centering, and the first n Principal Components that explained at least 80 % of the total variance were merged into a separate matrix. Lastly, supervised MLDF (SMLDF) was achieved by performing a PLS-DA on each data matrix, and the first 20 LVs from each model were merged into a new data matrix. The resulting data matrices consisted of 130 samples each x 1160 variables, 28 variables, and 60 variables for LLDF, UMLDF, and SMLDF, respectively.

PLS-DA classification performances expressed as confusion matrices after cross validation via Monte Carlo (80–20 % dataset split) with 1,000 iterations and maximum assignation criteria for the three levels are presented in Table 3. Interestingly, the average performance of LLDF over the four impacting variables (*i.e.*, harvest region, post-harvest conditions, storage time and temperature) is lower than the volatilome fraction on its own, 91.62 % *vs* 94.85 %, while both UMLDF and SMLDF classification models outperformed the single fractions on their own. The relatively poor results obtained with LLDF are attributable to the noisiness of the data from the primary and specialized metabolome with the storage times and conditions classification models. The resulting data matrix contained 10 times more variables than the number of samples. A model of such nature is prone to overfitting due to its excessive flexibility relative to the extent of the training dataset. As the model's flexibility increases, exemplified by the inclusion of additional variables in a regression model, and the number of samples remains the same, the likelihood increases that the model will accommodate random fluctuations within the training data that fail to adequately represent the genuine underlying distribution. The purpose of reduction algorithms is to mitigate the challenges related to dimensionality by simplifying data complexity, thereby enhancing data quality. Historically, PCA has been the prevailing method for dimensionality reduction. In this study, principal components were employed as an unsupervised technique for compressing data dimensions when integrating the three matrices. The selection of the number of components was based on evaluating the cumulative explained variance, which threshold was arbitrarily set to 80 %. With this approach, a general increase in the classification performances was achieved (average accuracy ≈ 97 %); as expected, the lower scores were achieved with storage conditions and duration classifications, despite achieving satisfactory results (≈ 93 % and ≈ 94 %, respectively). The last improvement was achieved through SMLDF, averaging 99.42 % classification accuracy, with the lowest performance being ≈ 98 % accuracy for storage months.

4. Conclusions

This research demonstrated the potential of using combined untargeted and targeted metabolomics with GC×GC–MS and UHPLC-HRMS to investigate metabolic signatures among *C. avellana* samples from different geographical origins, post-harvest processing methods, and one year of industrial shelf life under two storage conditions. Unlike modeling data individually, MLDF showed significant improvements in data analysis and classification accuracy. This approach has proven that adopting a suitable multivariate analysis strategy for authenticity testing yields highly reliable results. These findings offer promising prospects for enhancing the detection of mislabeling and increasing the reliability of defining the authenticity of plant-based products in the food industry, such as hazelnuts.

The remarkable capability of data fusion to integrate diverse sources of information and refine analytical insights underscores its crucial role in identifying chemical markers for selective monitoring and improving predictive capacity. The combination of these research outcomes and the description of the aroma blueprint serves as a valuable decision-making tool to guide and align strategic investments and value chains across the industry.

CRediT authorship contribution statement

Simone Squara: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Andrea Caratti:** Writing – review & editing, Formal analysis, Data curation. **Angelica Fina:** Writing – review & editing, Data curation. **Erica Liberto:** Writing – review & editing, Supervision. **Nemanja Koljančić:** Writing – review & editing, Formal analysis, Data curation. **Ivan Špánik:** Writing – review & editing, Supervision. **Giuseppe Genova:** Writing – review & editing, Project administration. **Giuseppe Castello:** Writing – review & editing, Project administration. **Carlo Bicchi:** Writing – review & editing, Supervision. **André de Villiers:** Writing – review & editing, Supervision, Conceptualization. **Chiara Cordero:** Conceptualization, Supervision, Project administration, Funding acquisition, Writing – review & editing.

Funding

I. Š. and N. K. are thankful for the funding provided by the Grant Agency of the Slovak Republic VEGA (project No. 1/0298/23).

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodres.2024.114873.

References

Alamar, P. D., Caramês, E. T. S., Poppi, R. J., & Pallone, J. A. L. (2020). Detection of Fruit Pulp Adulteration Using Multivariate Analysis: Comparison of NIR, MIR and Data Fusion Performance. *Food Analytical Methods, 13*(6), 1357–1365. https://doi.org/10.1007/s12161-020-01755-x

Alasalvar, C., & Bolling, B. W. (2015). Review of nut phytochemicals, fat-soluble bioactives, antioxidant components and health effects. *British Journal of Nutrition, 113*(S2), S68–S78. https://doi.org/10.1017/S0007114514003729

Alasalvar, C., Shahidi, F., & Cadwallader, K. R. (2003). Comparison of natural and roasted Turkish Tombul hazelnut (Corylus avellana L.) volatiles and flavor by DHA/GC/MS and descriptive sensory analysis. *Journal of Agricultural and Food Chemistry, 51*(17), 5067–5072. https://doi.org/10.1021/jf0300846

Alasalvar, C., Shahidi, F., Ohshima, T., Wanasundara, U., Yurttas, H. C., Liyanapathirana, C. M., & Rodrigues, F. B. (2003). Turkish Tombul hazelnut (Corylus avellana L.). 2. Lipid characteristics and oxidative stability. *Journal of Agricultural and Food Chemistry, 51*(13), 3797–3805. https://doi.org/10.1021/jf021239x

Alygizakis, N., Lestremau, F., Gago-Ferrero, P., Gil-Solsona, R., Arturi, K., Hollender, J., & Thomaidis, N. S. (2023). Towards a harmonized identification scoring system in LC-HRMS/MS based non-target screening (NTS) of emerging contaminants. *TrAC Trends in Analytical Chemistry, 159*, Article 116944. https://doi.org/10.1016/j.trac.2023.116944

Azcarate, S. M., Ríos-Reina, R., Amigo, J. M., & Goicoechea, H. C. (2021). Data handling in data fusion: Methodologies and applications. *TrAC - Trends in Analytical Chemistry, 143*. https://doi.org/10.1016/j.trac.2021.116355

Balkir, P., Kemahlioglu, K., & Yucel, U. (2021). Foodomics: A new approach in food quality and safety. *Trends in Food Science and Technology*. https://doi.org/10.1016/j.tifs.2020.11.028

Ballabio, D. (2015). A MATLAB toolbox for Principal Component Analysis and unsupervised exploration of data structure. *Chemometrics and Intelligent Laboratory Systems, 149*, 1–9. https://doi.org/10.1016/j.chemolab.2015.10.003

Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: Linear models. PLS-DA. *Analytical Methods.*. https://doi.org/10.1039/c3ay40582f

Ballabio, D., Robotti, E., Grisoni, F., Quasso, F., Bobba, M., Vercelli, S., & Marengo, E. (2018). Chemical profiling and multivariate data fusion methods for the identification of the botanical origin of honey. *Food Chemistry, 266*(September 2017), 79–89. https://doi.org/10.1016/j.foodchem.2018.05.084

Bottone, A., Cerulli, A., Durso, G., Masullo, M., Montoro, P., Napolitano, A., & Piacente, S. (2019). Plant Specialized Metabolites in Hazelnut (Corylus avellana) Kernel and Byproducts: An Update on Chemistry, Biological Activity, and Analytical Aspects. *Planta Medica*. https://doi.org/10.1055/a-0947-5725

Caratti, A., Squara, S., Bicchi, C., Liberto, E., Vincenti, M., Reichenbach, S. E., & Cordero, C. (2024). Boosting comprehensive two-dimensional chromatography with artificial intelligence: Application to food-omics. *TrAC Trends in Analytical Chemistry, 174*, Article 117669. https://doi.org/10.1016/j.trac.2024.117669

Caratti, A., Squara, S., Bicchi, C., Tao, Q., Geschwender, D., Reichenbach, S. E., & Cordero, C. (2023). Augmented visualization by computer vision and chromatographic fingerprinting on comprehensive two-dimensional gas chromatographic patterns: Unraveling diagnostic signatures in food volatilome. *Journal of Chromatography A, 1699*, Article 464010. https://doi.org/10.1016/j.chroma.2023.464010

Cialiè Rosso, M., Liberto, E., Spigolon, N., Fontana, M., Somenzi, M., Bicchi, C., & Cordero, C. (2018). Evolution of potent odorants within the volatile metabolome of high-quality hazelnuts (Corylus avellana L.): Evaluation by comprehensive two-dimensional gas chromatography coupled with mass spectrometry. *Analytical and Bioanalytical Chemistry, 410*(15), 3491–3506. https://doi.org/10.1007/s00216-017-0832-6

Cialiè Rosso, M., Mazzucotelli, M., Bicchi, C., Charron, M., Manini, F., Menta, R., & Cordero, C. (2020). Adding extra-dimensions to hazelnuts primary metabolome fingerprinting by comprehensive two-dimensional gas chromatography combined with time-of-flight mass spectrometry featuring tandem ionization: Insights on the aroma potential. *Journal of Chromatography A, 1614*(460739), 1–11. https://doi.org/10.1016/j.chroma.2019.460739

Cialiè Rosso, M., Stilo, F., Bicchi, C., Charron, M., Rosso, G., Menta, R., & Cordero, C. (2021). Combined Untargeted and Targeted Fingerprinting by Comprehensive Two-Dimensional Gas Chromatography to Track Compositional Changes on Hazelnut Primary Metabolome during Roasting. *Applied Sciences, 11*(2), 525. https://doi.org/10.3390/app11020525

Cialiè Rosso, M., Stilo, F., Mascrez, S., Bicchi, C., Purcaro, G., & Cordero, C. (2021). Shelf-Life Evolution of the Fatty Acid Fingerprint in High-Quality Hazelnuts (Corylus avellana L.). *Harvested in Different Geographical Regions. Foods, 10*(3), 685. https://doi.org/10.3390/foods10030685

Collins, S. L., Koo, I., Peters, J. M., Smith, P. B., & Patterson, A. D. (2021). Current Challenges and Recent Developments in Mass Spectrometry-Based Metabolomics. *Annual Review of Analytical Chemistry*. https://doi.org/10.1146/annurev-anchem-091620-015205

Cordero, C., Guglielmetti, A., Bicchi, C., Liberto, E., Baroux, L., Merle, P., & Reichenbach, S. E. (2019). Comprehensive two-dimensional gas chromatography coupled with time of flight mass spectrometry featuring tandem ionization: Challenges and opportunities for accurate fingerprinting studies. *Journal of Chromatography A, 1597*, 132–141. https://doi.org/10.1016/j.chroma.2019.03.025

Dankowska, A., & Kowalewski, W. (2019). Tea types classification with data fusion of UV–Vis, synchronous fluorescence and NIR spectroscopies and chemometric analysis. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy, 211*, 195–202. https://doi.org/10.1016/j.saa.2018.11.063

de Juan, A., & Tauler, R. (2019). Data Fusion by Multivariate Curve Resolution. *Data Handling in Science and Technology, 31*, 205–233. https://doi.org/10.1016/B978-0-444-63984-4.00008-9

Di Anibal, C. V., Callao, M. P., & Ruisánchez, I. (2011). 1H NMR and UV-visible data fusion for determining Sudan dyes in culinary spices. *Talanta, 84*(3), 829–833. https://doi.org/10.1016/j.talanta.2011.02.014

Dudzik, D., Barbas-Bernardos, C., García, A., & Barbas, C. (2018). Quality assurance procedures for mass spectrometry untargeted metabolomics. a review. *Journal of Pharmaceutical and Biomedical Analysis, 147*, 149–173. https://doi.org/10.1016/j.jpba.2017.07.044

Dunkel, A., Steinhaus, M., Kotthoff, M., Nowak, B., Krautwurst, D., Schieberle, P., & Hofmann, T. (2014). Nature's chemical signatures in human olfaction: A foodborne perspective for future biotechnology. *Angewandte Chemie - International Edition, 53*(28), 7124–7143. https://doi.org/10.1002/anie.201309508

Fanali, C., Tripodo, G., Russo, M., Della Posta, S., Pasqualetti, V., & De Gara, L. (2018). Effect of solvent on the extraction of phenolic compounds and antioxidant capacity of hazelnut kernel. *Electrophoresis*. https://doi.org/10.1002/elps.201800014

Fernández, A., Lombardo, A., Rallo, R., Roncaglioni, A., Giralt, F., & Benfenati, E. (2012). Quantitative consensus of bioaccumulation models for integrated testing strategies. *Environment International, 45*(1), 51–58. https://doi.org/10.1016/j.envint.2012.03.004

Ghirardello, D., Zeppa, G., Rolle, L., Gerbi, V., Contessa, C., Valentini, N., … Griseri, G. (2014). Effect of different storage conditions on hazelnut quality. In *Acta Horticulturae* (Vol. 1052, pp. 315–318). https://doi.org/10.17660/ActaHortic.2014.1052.44.

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*(2), 500–508. https://doi.org/10.1002/jsfa.9998

Giardina, M., McCurry, J. D., Cardinael, P., Semard-Jousset, G., Cordero, C., & Bicchi, C. (2018). Development and validation of a pneumatic model for the reversed-flow differential flow modulator for comprehensive two-dimensional gas chromatography. *Journal of Chromatography A, 1577*, 72–81. https://doi.org/10.1016/j.chroma.2018.09.022

Hassani, S., Dackermann, U., Mousavi, M., & Li, J. (2024). A systematic review of data fusion techniques for optimized structural health monitoring. *Information Fusion, 103*(June 2023), Article 102136. https://doi.org/10.1016/j.inffus.2023.102136

Tsugawa, H., Cajka, T., Kind, T., Ma, Y., Higgins, B., Ikeda, K., Kanazawa, M., VanderGheynst, J., Fiehn, O., & M. a.. (2015). MS-DIAL: Data Independent MS/MS Deconvolution for Comprehensive. *Nat Methods, 12*(6), 523–526. https://doi.org/10.1038/nmeth.3393.MS-DIAL

Jacobs, D. M., van den Berg, M. A., & Hall, R. D. (2021). Towards superior plant-based foods using metabolomics. *Current Opinion in Biotechnology*. https://doi.org/10.1016/j.copbio.2020.08.010

Jakopic, J., Petkovsek, M. M., Likozar, A., Solar, A., Stampar, F., & Veberic, R. (2011). HPLC-MS identification of phenols in hazelnut (Corylus avellana L.) kernels. *Food Chemistry, 124*(3), 1100–1106. https://doi.org/10.1016/j.foodchem.2010.06.011

Kiefl, J., Pollner, G., & Schieberle, P. (2013). Sensomics analysis of key hazelnut odorants (Corylus avellana L. 'Tonda Gentile') using comprehensive two-dimensional gas chromatography in combination with time-of-flight mass spectrometry (GC×GC-TOF-MS). *Journal of Agricultural and Food Chemistry, 61*(22), 5226–5235. https://doi.org/10.1021/jf400807w

Lai, Z., Tsugawa, H., Wohlgemuth, G., Mehta, S., Mueller, M., Zheng, Y., & Fiehn, O. (2018). Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nature Methods, 15*(1), 53–56. https://doi.org/10.1038/nmeth.4512

Li, S., Tian, Y., Jiang, P., Lin, Y., Liu, X., & Yang, H. (2021). Recent advances in the application of metabolomics for food safety control and food quality analyses. *Critical Reviews in Food Science and Nutrition. Taylor & Francis.*. https://doi.org/10.1080/10408398.2020.1761287

Mack, C., Wefers, D., Schuster, P., Weinert, C. H., Egert, B., Bliedung, S., & Kulling, S. E. (2017). Untargeted multi-platform analysis of the metabolome and the non-starch polysaccharides of kiwifruit during postharvest ripening. *Postharvest Biology and Technology, 125*, 65–76. https://doi.org/10.1016/j.postharvbio.2016.10.011

Márquez, C., López, M. I., Ruisánchez, I., & Callao, M. P. (2016). FT-Raman and NIR spectroscopy data fusion strategy for multivariate qualitative analysis of food fraud. *Talanta, 161*, 80–86. https://doi.org/10.1016/j.talanta.2016.08.003

Ngouta, J. E. O., Backes, M., Albrecht, A., Hempel, K., Schäfer, U., Paetz, S., … Kirschning, A. (2021). Giffonins as contributors to the bitter off-taste in hazelnuts, 3–6. https://doi.org/10.5281/zenodo.5541292.

Ortega-Gavilán, F., Squara, S., Cordero, C., Cuadros-Rodríguez, L., & Bagur-González, M. G. (2023). Application of chemometric tools combined with instrument-agnostic GC-fingerprinting for hazelnut quality assessment. *Journal of Food Composition and Analysis, 115*(September 2022), Article 104904. https://doi.org/10.1016/j.jfca.2022.104904

Pedrosa, M. C., Lima, L., Heleno, S., Carocho, M., Ferreira, I. C. F. R., & Barros, L. (2021). Food metabolites as tools for authentication, processing, and nutritive value assessment. *Foods, 10*(9), 2213. https://doi.org/10.3390/foods10092213

Pedrotti, M., Khomenko, I., Genova, G., Castello, G., Spigolon, N., Fogliano, V., & Biasioli, F. (2021). Quality control of raw hazelnuts by rapid and non-invasive fingerprinting of volatile compound release. *LWT, 143*(February), Article 111089. https://doi.org/10.1016/j.lwt.2021.111089

Pelvan, E., Olgun, E.Ö., Karadağ, A., & Alasalvar, C. (2018). Phenolic profiles and antioxidant activity of Turkish Tombul hazelnut samples (natural, roasted, and roasted hazelnut skin). *Food Chemistry, 244*(June 2017), 102–108. https://doi.org/10.1016/j.foodchem.2017.10.011

Rodionova, O., & Pomerantsev, A. (2023). Multi-block DD-SIMCA as a high-level data fusion tool. *Analytica Chimica Acta, 1265*(April), Article 341328. https://doi.org/10.1016/j.aca.2023.341328

Romo-Pérez, M. L., Weinert, C. H., Häußler, M., Egert, B., Frechen, M. A., Trierweiler, B., & Zörb, C. (2020). Metabolite profiling of onion landraces and the cold storage effect. *Plant Physiology and Biochemistry, 146*, 428–437. https://doi.org/10.1016/j.plaphy.2019.11.007

Schmid, C., Sharma, S., Stark, T. D., Günzkofer, D., Hofmann, T. F., Ulrich, D., & Dawid, C. (2021). Influence of the abiotic stress conditions, waterlogging and drought, on the bitter sensometabolome as well as agronomical traits of six genotypes of daucus carota. *Foods*. https://doi.org/10.3390/foods10071607

Schrimpe-Rutledge, A. C., Codreanu, S. G., Sherrod, S. D., & McLean, J. A. (2016). Untargeted Metabolomics Strategies Challenges and Emerging Directions. *Journal of the American Society for Mass Spectrometry, 27*(12), 1897–1905. https://doi.org/10.1007/s13361-016-1469-y

Shahidi, F., Alasalvar, C., & Liyana-Pathirana, C. M. (2007). Antioxidant Phytochemicals in Hazelnut Kernel (Corylus avellana L.) and Hazelnut Byproducts. *Journal of Agricultural and Food Chemistry, 55*(4), 1212–1220. https://doi.org/10.1021/jf062472o

Silvestri, M., Elia, A., Bertelli, D., Salvatore, E., Durante, C., Li Vigni, M., & Cocchi, M. (2014). A mid level data fusion strategy for the Varietal Classification of Lambrusco PDO wines. *Chemometrics and Intelligent Laboratory Systems, 137*, 181–189. https://doi.org/10.1016/j.chemolab.2014.06.012

Singldinger, B., Dunkel, A., Bahmann, D., Bahmann, C., Kadow, D., Bisping, B., & Hofmann, T. (2018). New Taste-Active 3-(O-β- d -Glucosyl)-2-oxoindole-3-acetic Acids and Diarylheptanoids in Cimiciato-Infected Hazelnuts. *Journal of Agricultural and Food Chemistry*. https://doi.org/10.1021/acs.jafc.8b01216

Smolinska, A., Engel, J., Szymanska, E., Buydens, L., & Blanchet, L. (2019). General Framing of Low-, Mid-, and High-Level Data Fusion With Examples in the Life Sciences. *Data Handling in Science and Technology, 31*, 51–79. https://doi.org/10.1016/B978-0-444-63984-4.00003-X

Squara, S., Caratti, A., Fina, A., Liberto, E., Spigolon, N., Genova, G., & Cordero, C. (2023). Artificial Intelligence decision-making tools based on comprehensive two-dimensional gas chromatography data: The challenge of quantitative volatilomics in food quality assessment. *Journal of Chromatography A, 1700*, Article 464041. https://doi.org/10.1016/j.chroma.2023.464041

Squara, S., Caratti, A., Gavilan, F. O., Bolzoni, P., Spigolon, N., Genova, G., & Cordero, C. (2022). Validation of a high-throughput method for the accurate quantification of secondary products of lipid oxidation in high-quality hazelnuts (Corylus avellana L.): A robust tool for quality assessment. *Journal of Food Composition and Analysis, 114*, Article 104766. https://doi.org/10.1016/j.jfca.2022.104766

Squara, S., Manig, F., Henle, T., Hellwig, M., Caratti, A., Bicchi, C., & Cordero, C. (2023). Extending the breadth of saliva metabolome fingerprinting by smart template strategies and effective pattern realignment on comprehensive two-dimensional gas chromatographic data. *Analytical and Bioanalytical Chemistry, 0123456789*. https://doi.org/10.1007/s00216-023-04516-x

Squara, S., Stilo, F., Cialiè Rosso, M., Liberto, E., Bicchi, C., & Cordero, C. E. I. (2022). Exploring food volatilome by advanced chromatographic fingerprinting based on comprehensive two-dimensional gas chromatographic patterns. *In Comprehensive Analytical Chemistry*. https://doi.org/10.1016/bs.coac.2021.11.008

Squara, S., Stilo, F., Cialiè Rosso, M., Liberto, E., Spigolon, N., Genova, G., & Cordero, C. (2022). Corylus avellana L. Aroma Blueprint: Potent Odorants Signatures in the Volatilome of High Quality Hazelnuts. *Frontiers in Plant Science, 13*(March), 1–25. https://doi.org/10.3389/fpls.2022.840028

Stilo, F., Cialiè Rosso, M., Squara, S., Bicchi, C., Cordero, C., & Cagliero, C. (2022). Corylus avellana L. Natural Signature: Chiral Recognition of Selected Informative Components in the Volatilome of High-Quality Hazelnuts. *Frontiers in Plant Science, 13*(April), 1–15. https://doi.org/10.3389/fpls.2022.844711

Stilo, F., Liberto, E., Reichenbach, S. E., Tao, Q., Bicchi, C., & Cordero, C. (2019). Untargeted and Targeted Fingerprinting of Extra Virgin Olive Oil Volatiles by Comprehensive Two-Dimensional Gas Chromatography with Mass Spectrometry: Challenges in Long-Term Studies. *Journal of Agricultural and Food Chemistry, 67*(18), 5289–5302. https://doi.org/10.1021/acs.jafc.9b01661

Stilo, F., Liberto, E., Reichenbach, S. E., Tao, Q., Bicchi, C., & Cordero, C. (2021). Exploring the Extra-Virgin Olive Oil Volatilome by Adding Extra Dimensions to Comprehensive Two-Dimensional Gas Chromatography and Time-of-Flight Mass Spectrometry Featuring Tandem Ionization: Validation of Ripening Markers in Headspace Linearity Conditio. *Journal of AOAC International, 104*(2), 274–287. https://doi.org/10.1093/jaoacint/qsaa095

Stilo, F., Liberto, E., Spigolon, N., Genova, G., Rosso, G., Fontana, M., & Cordero, C. (2021). An effective chromatographic fingerprinting workflow based on comprehensive two-dimensional gas chromatography – Mass spectrometry to establish volatiles patterns discriminative of spoiled hazelnuts (Corylus avellana L.). *Food Chemistry, 340*, Article 128135. https://doi.org/10.1016/j.foodchem.2020.128135

Stilo, F., Tredici, G., Bicchi, C., Robbat, A., Morimoto, J., & Cordero, C. (2020). Climate and Processing Effects on Tea (Camellia sinensis L. Kuntze) Metabolome: Accurate Profiling and Fingerprinting by Comprehensive Two-Dimensional Gas Chromatography/Time-of-Flight Mass Spectrometry. *Molecules, 25*(10), 2447. https://doi.org/10.3390/molecules25102447

Tsugawa, H., Ikeda, K., Takahashi, M., Satoh, A., Mori, Y., Uchino, H., & Arita, M. (2020). A lipidome atlas in MS-DIAL 4. *Nature Biotechnology, 38*(10), 1159–1163. https://doi.org/10.1038/s41587-020-0531-2

Tsugawa, H., Kind, T., Nakabayashi, R., Yukihira, D., Tanaka, W., Cajka, T., & Arita, M. (2016). Hydrogen Rearrangement Rules: Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software. *Analytical Chemistry, 88*(16), 7946–7958. https://doi.org/10.1021/acs.analchem.6b00770

Tsugawa, H., Nakabayashi, R., Mori, T., Yamada, Y., Takahashi, M., Rai, A., & Saito, K. (2019). A cheminformatics approach to characterize metabolomes in stable-isotope-labeled organisms. *Nature Methods, 16*(4), 295–298. https://doi.org/10.1038/s41592-019-0358-2

Ulaszewska, M. M., Weinert, C. H., Trimigno, A., Portmann, R., Andres Lacueva, C., Badertscher, R., & Vergères, G. (2019). Nutrimetabolomics: An Integrative Action for Metabolomic Analyses in Human Nutritional Studies. *Molecular Nutrition & Food Research, 63*(1), 1800384. https://doi.org/10.1002/mnfr.201800384

Wang, Q., Xiao, J., Li, Y., Lu, Y., Guo, J., Tian, Y., & Ren, L. (2023). Mid-level data fusion of Raman spectroscopy and laser-induced breakdown spectroscopy: Improving ores identification accuracy. *Analytica Chimica Acta, 1240*(October 2022), 340772. https://doi.org/10.1016/j.aca.2022.340772.

Westerhuis, J. A., van der Kloet, F., & Smilde, A. K. (2019). Data Fusion in Metabolomics. *In Metabolomics*. https://doi.org/10.1201/9781315370583-7