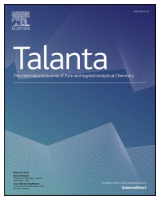
Talanta 289 (2025) 127720

Contents lists available at ScienceDirect

Talanta

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



Disentangling shared and unique variation in multiplatform hazelnut volatilomics using JIVE

Maria Mazzucotelli [a,c], Iuliia Khomenko [a], Emanuela Betta [a], Elena Gabetti [b], Luca Falchero [b], Eugenio Aprea [c], Andrea Cavallero [b], Franco Biasioli [a,*], Pietro Franceschi [a]

[a] Research and Innovation Centre, Fondazione Edmund Mach, San Michele all'Adige, Trento, Italy
[b] Soremartec Italia srl, Alba, Cuneo, Italy
[c] C3A - Center Agriculture Food Environment, University of Trento, San Michele all'Adige, Trento, Italy

ARTICLE INFO

Keywords:
VOCs
Volatilome
Roasting
GC-IMS
GC-MS
PTR-ToF-MS

ABSTRACT

In food science, volatile metabolites play a crucial role in determining sensory quality, acceptability and traceability. Fully characterizing the volatilome often requires combining multiple analytical techniques. However, reliably integrating the outcomes of these independent analyses to identify shared and unique information remains a significant challenge.

In this paper, we illustrate how the multivariate Joint and Individual Variation Explained (JIVE) approach could be used to face this problem on a multiplatform VOC dataset obtained characterizing the volatilome of hazelnut pastes with GC-MS, PTR-ToF-MS and GC-IMS. While standardized data processing strategies were applied for GC-MS and PTR-ToF-MS, an automated pipeline was developed for GC-IMS to extract untargeted peak tables. The samples, representing three geographical origins, were collected during roasting to capture a wide range of intensities, offering a challenging case study for the proposed approach. The results showed that JIVE effectively separated the variability of each dataset into joint and individual components. A high-level comparison of the three analytical methods, based on variation decomposition and variable distribution, confirmed their complementarity. Additionally, identifying latent variables facilitated the visualization of analytical patterns - both shared and platform-specific - and the selection of related key variable trends, supporting the chemical interpretation of the results. This unsupervised data exploration strategy, based on JIVE, provides clearer interpretation of both shared and technique-specific insights. It supports an objective evaluation of the potential of a multiplatform analysis while offering guidance for selecting the most suitable analytical method in studies constrained to a single technique.

1. Introduction

Within the expansion of -omics, volatilomics has emerged as a branch of metabolomics targeting the characterization of the pool of all volatile organic compounds (VOCs) present in a biological system, defined as the volatilome. In food science, volatile metabolites play a crucial role in determining sensory quality and acceptability as they significantly contribute to flavour [1]. The volatile profile of food matrices is also influenced by their geographical and botanical origin and undergoes changes during storage, processing, and shelf-life [2]. Therefore, food volatilomics is a valuable tool for assessing food quality, ensuring traceability, and monitoring processes.

Since the volatilome of food samples is a complex mixture of VOCs belonging to different chemical classes and covering a wide range of concentrations, a single analytical platform might not provide complete characterization. The typical approach to achieve a more comprehensive profiling, then, is to integrate data from diverse platforms, but the price to pay is an increase in time and cost of the analysis. In order to assess the need of this so-called "multi-platform analysis" it would be useful to be able to objectively decouple the shared analytical information from the one provided by each individual technique, analyzing the multiple matrices corresponding to the distinct analytical techniques employed (data sources) as a multi-block dataset [3].

Traditional exploratory multivariate methods, such as Principal

* Corresponding author.
E-mail address: franco.biasioli@fmach.it (F. Biasioli).

https://doi.org/10.1016/j.talanta.2025.127720
Received 3 December 2024; Received in revised form 21 January 2025; Accepted 10 February 2025
Available online 11 February 2025

Component Analysis (PCA), can be used to separately analyze the matrices of a multi-block dataset. However, as single-block methods, they do not address the integration of data across sources, hindering the identification of shared and unique structures, and failing to reveal crucial associations between the data sources [4]. In contrast, multi-block methods enable the extraction of complementary information from data generated using a multi-platform approach. The Joint and Individual Variation Explained (JIVE) method is an unsupervised multivariate data fusion approach, developed as a multi-source extension of PCA [5], and is part of the multi-block exploratory data analysis methods [3]. JIVE allows for the integrated analysis of multiple high-dimensional data sources from a common sample set [6], identifying structured variation across the data sources, and separating it into common and distinct variation [7]. The common variation (joint) corresponds to the multivariate patterns that are shared among the data sources, while the distinct variation (individual) refers to the source-specific structure. In a multi-platform study, the distinct variation can be attributed to analytes detectable only with a specific mode of measurement or to technical artefacts. According to the original paper proposing JIVE [6], this exploratory method quantifies the amount of joint and individual variation among data sources, reduces dimensionality and allows for visual exploration of joint and individual structures.

In this study, we demonstrate the usefulness of such an approach by analyzing a multi block dataset obtained during an investigation of the evolution of hazelnut volatilome during roasting. In this case, samples were characterized for their VOCs using Gas Chromatography coupled with Mass Spectrometry (GC-MS), Gas Chromatography coupled with Ion Mobility Spectrometry (GC-IMS), and Proton Transfer Reaction - Time of Flight - Mass Spectrometry (PTR-ToF-MS). Our investigation flanked two established analytical techniques as GC-MS and PTR-ToF-MS to the recently popular GC-IMS. This technique is particularly promising due to the relatively low complexity of the analytical set-up, but its application as untargeted profiling tool is still an active research area, both for the complexity of the phenomena occurring during ionization, and the lack of a harmonized bioinformatic pipeline to automatically generate peak tables for an entire sample set. This process often still relies on manual peak selection. To overcome this limitation, we developed a tailored approach for peak detection which allowed us to extract untargeted peaks from a set of 72 injections.

The evolution of hazelnut volatilome during roasting provided a challenging case study, characterized by a significant complexity in terms of the range of VOC chemical classes and concentrations. Moreover, hazelnut is a tree nut holding substantial industrial importance due to its economic value, fragmentation of the supply chain and extensiveness of industrial processing. The extensive use of hazelnuts in the confectionery industry leads to an increasing demand for raw materials and a complex supply chain concerning the geographical and botanical origin of the kernels. While Turkey is the leading producer, accounting for 70 % of world production, hazelnut cultivation is increasing in many countries located in different geographical areas (i.e. Chile, Serbia, China). Numerous varieties are cultivated in commercial orchards, each characterized by distinct morphological traits and slightly different chemical composition [8]. Regarding the industrial processing, up to 90 % of hazelnut utilization is based on products obtained from roasted kernels, which can be used whole, ground, or turned into paste [8]. The roasting step dramatically modifies the volatile fraction composition and is responsible for developing the characteristic flavor of roasted hazelnuts [9]. This makes the roasting step crucial in determining the final quality of food products containing hazelnuts. It is, therefore, important to identify characteristic trends of VOC formation and release and differences and similarities between cultivars and geographical origins concerning industrial processing.

The paper is organized as follows.

- We first present a pipeline for the extraction of the untargeted peak table from the set of GC-IMS runs, and we review the basic ideas behind the JIVE approach, detailing the data analysis steps carried out for model interpretation. Both sections are included in the Material and Methods.
- In the Results and Discussion, we start with a high-level comparison between the three techniques based on the JIVE results. Subsequently, we zoom into the chemical details of each technique discussing the variables (GC-MS compounds, GC-IMS peaks, PTR-MS mass peaks) which were more influential on the joint component and on the three individual ones.

2. Materials and methods

2.1. Samples

Hazelnut paste samples, supplied by Soremartec Italia Srl (Alba, Cuneo, Italy), were obtained processing raw kernels (*Corylus avellana* L.) from different geographical and botanical origins: Tonda Gentile Romana (RO) from Lazio region (Italy), Tonda Gentile delle Langhe (TGTP) from Piemonte region (Italy), and Akçakoca hazelnuts (AK) from Turkey. The harvesting year was 2021 for all the samples. The raw kernels were stored in sealed bags at 4 °C and in controlled humidity conditions for six months prior to roasting.

The roasting process was conducted using a pilot-scale infrared roaster Brovind RI OB-800 (Cortemilia, Cuneo, Italy) set at 140 °C. Throughout the roasting process, small amounts (150–200 g) of kernels were collected at various time intervals (5, 10, 12.5, 15, 17.5, 20, 22.5, and 25 min), resulting in samples with increasing levels of roasting intensity ranging from under-roasted to over-roasted (t1, t2, t3, t4, t5, t6, t7, t8). Each collected aliquot of kernels was processed to produce hazelnut paste samples.

2.2. VOC analysis

The VOC measurements were performed in parallel using the three analytical techniques. The headspace sampling was carried out on 1 g of hazelnut paste placed in a 20 ml HS vial. Three analytical replicates were analyzed for each analytical technique. A reference roasted hazelnut paste was used to prepare quality control (QC) samples, which were included in the analytical batches for all the three platforms to monitor the stability of the analytical response over time. All sample vials were stored at −20 °C until the time of measurement. Blanks, consisting of vials containing only air, were also included to assess for any potential carryover effects and to exclude background signals. The instrumental setup and method parameters were optimized to obtain a representative volatilome profile with each technique. Headspace solid phase microextraction (HS-SPME) was used as the sampling method for GC-MS analysis, while the higher sensitivity of PTR-ToF-MS and GC-IMS enabled the use of static headspace (SHS), eliminating the need for preconcentration during sampling.

2.3. SHS-GC-IMS

2.3.1. SHS-GC-IMS instrumental set-up

SHS-GC-IMS analyses were carried out on a FlavourSpec GC-IMS system (^{3}H-IMS) (G.A.S., Dortmund, Germany) equipped with a HT2000H headspace autosampler (HTA, Brescia, Italy) and a polar column MXT-Wax 30 m, 0.53 mm *dc*, 0.5 μm *df* (Restek Corporation, Bellefonte, US). Nitrogen (99.999 % purity grade) was used as carrier gas for chromatographic separation and as drift gas. The IMS operated in positive ionization mode. Instrument control and data acquisition were performed with the Sequence Designer software version 1.1 (G A S., Dortmund, Germany). The instrumental parameters were set according to our previous study [10].

2.3.2. SHS-GC-IMS parameters

- SHS sampling

The vials were incubated at 60 °C for 20 min under constant agitation. HS syringe temperature was 80 °C. The injection volume was 0.5 mL.

- GC-IMS analysis

GC-IMS analyses were carried out under the following conditions: injector and transfer line (inj-oven) temperature 80 °C; GC and transfer line (oven-IMS) temperature 60 °C, IMS temperature 45 °C. GC column flow program: 2 mL/min (constant flow) for 6 min, from min 6 the flow was gradually increased up to 12 mL/min at 16 min, then up to 50 mL/min at 19.5 min, up to 75 mL/min at 22.5 min, up to 124 mL/min at 27 min, up to 150 mL/min at 27 min, ending with 3 min at 150 mL/min constant flow to avoid carryover effect. The total GC runtime was 30 min. IMS drift flow: 150 mL/min.

2.3.3. GC-IMS data processing

An automated workflow for untargeted GC-IMS data processing was developed in R (software version 4.4.1) [11]. The "reticulate" R package [12] was used to integrate Python code with the main R code in RStudio IDE [13].

- Importing data and preprocessing

The output files of GC-IMS (.mea file format) were imported into R. To optimize memory usage and exclude irrelevant background regions, only the spectra region containing the peaks was imported. For hazelnut paste samples analyzed as described in Section 2.3.4, the selected region corresponded to a retention time range from 170 s to 1800 s and a drift time range from 4.5 ms to 8.5 ms. Data reduction via binning was employed, resulting in 100 ion mobility bins (im_bins) and 250 retention time bins (rt_bins) for the selected area. The bin dimensions obtained from the binning step were im_bin width = 0.04 ms and rt_bin width 6.52 s.

This approach effectively reduced computational load and resulted in a substantial improvement of signal to noise ratio without losses in retention time separation. The intensity of each bin corresponded to the maximum intensity of the binned scans. For each imported GC-IMS data file, a matrix containing the coordinates and the intensities (expressed in mV) for all the bins of the selected region was obtained.

- Peak detection

Peak detection was performed using the "findpeaks" Python package [14]. The peak detection algorithm called "topology", implemented in this package, enables a non-parametric detection of 2D peaks based on persistent homology, a method used in topological data analysis [15]. This approach enables the detection of peaks without requiring assumptions about their shape or width, providing a suitable strategy for GC-IMS data automated untargeted processing. A denoising step prior to peak detection is already implemented when using the *"findpeaks"* function. "topology" was set as detection algorithms, while "median" filtering was chosen for noise reduction. For each sample, the coordinates (expressed in bins) of the peaks were obtained.

- Peak clustering

Once the peaks in each GC-IMS measurement are detected, the peak locations across all the measurements must be aligned to ensure that detected peaks correspond to the same signals in the 2D GC-IMS data and thus represent the same ionized species. This step was performed using the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm, implemented with the *"dbscan"* function from the "dbscan" R package [16,17], with parameters set to eps = 1 and minPts = 3. As reported by Horsch et al. [18], DBSCAN is one of the clustering methods suitable for GC-IMS peak clustering. This method was chosen due to its ability to form clusters based on the density of points, which is particularly useful for GC-IMS data where peaks may vary in intensity and distribution. The algorithm effectively grouped peaks that are closely located in both drift time and retention time dimensions, treating noise points (peaks that do not belong to any cluster) appropriately. After clustering, the result was a list of peak clusters, the untargeted features, with their coordinate ranges. These clusters represented consensus peak locations (coordinate ranges) across samples (Supplementary Table 1). Based on these consensus coordinates, the corresponding peak intensities were extracted from the matrices containing the signal intensity for all the bins. If a peak was not detected in a particular sample, a missing data imputation step was performed by extracting the intensity of the bin in the consensus coordinate ranges.

The outcome of this automated workflow was a peak table with the intensity of the untargeted features for each sample. The obtained GC-IMS features were identified with an increasing number corresponding to the cluster number of the clustering step.

2.4. HS-SPME-GC-MS

2.4.1. HS-SPME-GC-MS instrumental set-up

HS-SPME-GC-MS analyses were carried out on a 7890B GC system and coupled to a 5977A MS detector (single quadrupole analyzer and EI Extractor ion source) (Agilent, Santa Clara, US), and equipped with a MPS Multipurpose Sampler (GERSTEL, Mülheim an der Ruhr, Germany). The GC system was configured with an SSL injector and a 0.75 mm id liner for SPME injections. The chromatographic separation was performed using a polar column DB-Wax 30 m, 0.25 mm *dc*, 0.25 μm *df* (Agilent, Santa Clara, US), and helium as carrier gas. The MS was operated in EI mode at 70 eV. The HS-SPME sampling was carried out with using a 2 cm 50/30 μm df divinylbenzene/Carboxen®/polydimethylsiloxane (DVB/CAR/PDMS) fibre (Supelco-Merck, Darmstadt, Germany). Instrument control and data acquisition were performed with Agilent MassHunter GC/MS Data Acquisition version 10.0.368 (Agilent, Santa Clara, US).

2.4.2. HS-SPME-GC-MS parameters

- HS-SPME sampling

The vials were incubated at 60 °C for 10 min before SPME sampling. Extraction time for sampling was set at 20 min. The collected VOCs were desorbed in the SSL injector exposing the fiber with a desorption time of 5 min. To avoid carryover effects, after each a fiber bakeout step in the conditioning station was included (150 °C - 5 min).

- GC-MS analysis

GC-MS analyses were performed under the following conditions: inlet - mode: splitless, injector temperature: 240 °C, splitless time: 5 min (50 mL/min), gas saver on: 10 min (15 mL/min). Column - constant flow, column flow: 1.5 mL/min. Oven – equilibration time 0.5 min, temperature ramp: the oven starting temperature was set at 40 °C for 3 min, then increased up to 220 °C with a rate of 4 °C/min, lastly a ramp of 15 °C/min was used up to a temperature of 240 °C, which was maintained for 2 min. MSD - transfer line temperature: 240 °C, MS source temperature: 230 °C, MS quadrupole temperature: 150 °C, acquisition type: scan, scan range 33.00 u – 350.00 u, scan rate 22.7 scan/s.

2.4.3. GC-MS data processing

GC-MS data were processed using Agilent MassHunter Qualitative Analysis (version 10.0) (Agilent, Santa Clara, US) for peak identification

and Agilent MassHunter Quantitative Analysis for GC-MS (version 10.2) for extracting peak areas. Peaks were identified based on their MS spectra from the NIST/EPA/NIH Mass Spectral Library 2014 and their linear retention indices [19] (Supplementary Table 2). Peak areas were calculated based on the response of a target ion, with up to two qualifier ions monitored to increase the reliability of the peak area extraction. The *"imputePCA"* function from the "missMDA" R package [20] was employed for compound-wise imputation of missing data. This function imputes values in a way that imputation has no effect on the PCA results, and consequently, on the JIVE results.

2.5. SHS-PTR-ToF-MS

2.5.1. SHS-PTR-ToF-MS instrumental set-up

SHS-PTR-ToF-MS analyses were carried out on a PTR-ToF-MS 8000 (IONICON Analytik, Innsbruck, Austria) equipped with an MPS Multi-purpose Sampler (GERSTEL, Mülheim an der Ruhr, Germany) and a static headspace module (IONICON Analytik, Innsbruck, Austria).

2.5.2. SHS-PTR-ToF-MS instrumental set-up

- SHS sampling

The vials were incubated at 40 °C for 20 min. HS syringe temperature was 110 °C. The injection volume was 2.5 mL and injection speed was 100 μL/s. The SHS module and inlet were kept at 110 °C. The constant flow of 90 sccm of purified air generated by the compressor with the scrubber (IONICON Analytik, Innsbruck, Austria) was applied.

- PTR-ToF-MS analysis

PTR-ToF-MS analyses were performed under the following conditions: inlet system: inlet temperature 110 °C; drift tube: temperature 110 °C, pressure 2.80 mbar, voltages 628 V, ion funnel on. This led to an E/N ratio of about 140 Townsend (Td), with E corresponding to the electric field strength and N to the gas number density (1 Td = 10^{-17} Vcm^2). The sampling time per channel of ToF acquisition was 0.1 ns, amounting to 350,000 channels for a mass spectrum ranging up to m/z = 350 which resulted in the acquisition rate of 1 spectrum/s.

2.5.3. PTR-ToF-MS data processing

The internal calibration was performed to reach a good mass accuracy (up to 0.001 Th) according to a procedure described by Cappellin et al. [21]. Noise reduction, baseline removal and peak intensity extraction were performed according to Cappellin et al. [21], using modified Gaussians to fit the peaks. Absolute headspace VOC concentrations expressed in ppbv (parts per billion by volume) were calculated from peak intensities according to Cappellin et al. [22]. Imputation was performed on a compound-wise basis with a random value between 0 and the corresponding minimum value. Limits of detection (LODs) were calculated from blank measurements to determine if the detected concentration of a compound in samples was significantly higher than that observed in blank samples. Mass peaks with intensities lower than the LODs were excluded, as they were considered not reliably distinguishable from blank samples or related to the instrumental background [23,24]. Mass assignment was performed to derive molecular formulas for the detected mass peaks, using a literature database [25] and an in-house library. Annotated mass peaks are reported in Supplementary Table 3.

2.6. Data analysis

Principal Component Analysis was initially performed separately on each dataset as a single-block data exploration method, utilizing the "*PCA*" function of the R package "FactoMineR" [27]. Before PCA, each dataset was log-transformed to address the expected non-normal distribution of metabolomics data. Subsequently, each variable was mean-centered and scaled to unit variance in order to equalize the contribution of all variables within each data source [3].

2.6.1. JIVE decomposition

The three untargeted datasets (GC-IMS, GC-MS, and PTR-MS) were investigated as a multi-source dataset and jointly analyzed applying JIVE with the "*jive*" function from the R package "R.JIVE" [5]. Before applying JIVE, the datasets underwent the same pretreatment as used for PCA, including log-transformation, mean-centering, and scaling to unit variance. Since JIVE performs a multiblock decomposition an additional normalization step was necessary to account for the differences in the number of variables measured in the individual dataset, in particular to reduce the impact of the largest dataset (PTR-MS) on the final JIVE model solution. In order to do that each data matrix was divided by its Frobenius norm [4].

The JIVE model decomposes the multi-source data set into three terms: one joint matrix (J) (capturing shared variation across sources), three individual matrices (I) (capturing variation individual to each source), and residual noise. The number of components (ranks) for joint and individual matrices was determined by permutation testing as part of the model optimization [4]. For our dataset, the ranks were the following: r = 1 for the joint structure, r = 3 for GC-IMS individual structure, and r = 2 for GC-MS and PTR-MS individual structures. For each matrix, the percentage of variation explained was extracted from the model using the "*summary.jive*" function of "R.JIVE".

2.6.2. Calculation of variables partitioning

The JIVE decomposition partitions the variation of the vector of intensities of each feature ($\boldsymbol{F}$) into the joint part ($\boldsymbol{J}$), the individual part ($\boldsymbol{I}$) and the error ($\boldsymbol{R}$). When the error is small then the variance is decomposed in individual and joint components.

$$F = J + I + R \approx J + I$$

In order to quantify the contribution of the joint ($\boldsymbol{J_c}$) and individual ($\boldsymbol{I_c}$) components in each variable we considered the norm of the $\boldsymbol{J}$ and $\boldsymbol{I}$ term of the decomposition which were normalized by the norm of the initial vector ($\boldsymbol{F}$):

$$J_c = \frac{\|J\|^2}{\|F\|^2} \; I_c = \frac{\|I\|^2}{\|F\|^2}$$

This ratio can be considered as a reliable measure of contribution due to the orthogonality constraint of the JIVE model, and it can be represented as a percentage. A variable mostly represented in the individual component will have, for example, $\boldsymbol{I_c}$ close to one. For each variable, $\boldsymbol{I_c}$, $\boldsymbol{J_c}$ and $\boldsymbol{R_c}$ can be visualized in a ternary plot, which summarizes the contribution of the three terms of the decomposition on each variable. Ternary plots were obtained by using the R package "ggtern" [26].

2.6.3. Identification of latent variables of JIVE model

In order to identify the "latent" components associated to the individual terms of the JIVE model, the joint and individual matrices were factorized by Principal Component Analysis by using the "*PCA*" function of the R package "FactoMineR" [27]. Similarly to what happens for ordinary PCA, the loading matrices provide information on the contribution (and correlation) of the original variables on the principal component, while score matrices can be used to identify characteristic patterns in the data. In the specific case of JIVE, loadings and scores were obtained for the joint matrix - which captures shared structure - and for the three individual matrices. It is worth mentioning that, to preserve the variable partition resulting from the JIVE decomposition, the data were not scaled before applying PCA. As an additional model diagnostic, PCA was also applied on the residual noise matrices to confirm that they do not contain residual structures related to the experimental design (Supplementary Fig. 1). For data visualization of

score and loading plots the R package "ggplot2" [28] was used. To facilitate the chemical interpretation of the results the loadings plot will show only the variables which were predominantly contributing to the joint and individual structures. The selection was performed based on the variable partitioning described in sections 2.6.2 restricting the analysis to variables showing a contribution greater than 70 % either in the joint or individual components.

3. Results and discussion

Single-block PCA was conducted as an initial exploration of each dataset individually. The resulting score and loading plots for the first two principal components are presented in Supplementary Fig. 2. In all three datasets, PC1 (explaining approximately 60 % of the variance) captured the trend associated with the roasting process, while PC2 distinguished the samples based on their geographical origins. The score plots also revealed noticeable differences in the multivariate patterns of the samples, particularly in the relative positioning of the origin groups along PC2. Although this individual analysis provided valuable preliminary insights, interpreting the similarities and differences between datasets proved challenging due to the separate, unintegrated nature of the analysis.

3.1. Interpretation of JIVE results: variation decomposition and variable partitioning

The plot in Fig. 1 (adapted from Ref. [5]) illustrates the decomposition of variation and the variation explained for each data source, showing the results of the JIVE model in terms of quantification of joint and individual variation.

The plot clearly shows that the shared and individual variation across the three data sources were comparable, with the joint structure accounting for approximately 35 % of the variation and the individual structures for around 50–55 %. This predominant role of the individual component suggests the complementarity of these techniques. To estimate the variability associated with the variation explained by the joint and individual structures, resampling was conducted. Three scenarios were evaluated, resampling 23, 21, and 18 samples, with 100 random sampling iterations performed for each scenario. The resulting box plots, shown in Supplementary Fig. 3, provide insights into the variability associated with each resampling scenario. As expected, variability increased as the number of samples decreased. However, even in the scenario with 18 samples, the median values remained consistent with the percentages of variation explained by the JIVE model using the original 24-samples dataset.

This high-level comparison between the three data sources, however, did not provide elements to assess the relative contribution of the individual variables to the variation within each data source. In order to investigate this aspect, we rely on the ternary plots of the individual variable partitioning (Section 2.6.2) which are reported in Fig. 2. In this type of plot, each side of the triangle corresponds to the relative importance of J, I and R to each individual variable and each variable is represented by a point in the plot. In order to highlight the most intense features, the transparency of the points reflects the intensity of the variables in the original data. A more intense or less transparent point suggests a higher intensity in the original data, while a more transparent point indicates a lower intensity.

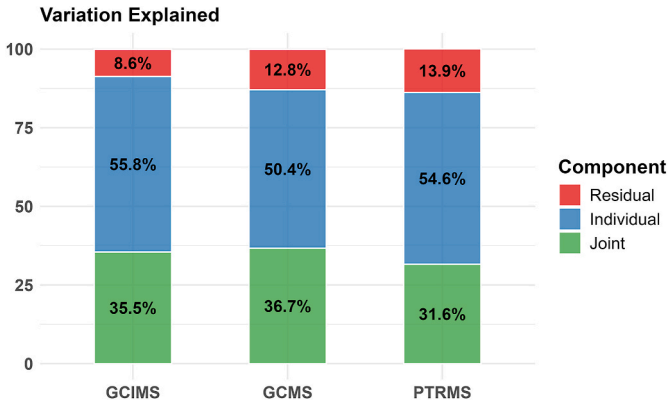


Fig. 1. Decomposition of variation for each data source. For each analytical technique, the percentages of variation explained by the estimated joint structure, individual structure and residual noise are reported.

The first insight from the ternary plots is that the number of features measured by the three analytical techniques varies: PTR-MS yielded the highest number of variables, with 288 features, followed by 131 for GC-IMS and 128 for GC-MS (based on the processing of hazelnut paste samples as outlined in sections 2.3, 2.4, and 2.5). For PTR-MS data a fingerprinting approach was employed that limited the preprocessing step to the exclusion of mass peaks with intensities lower than the calculated LODs, without further peak selection. Regarding GC-IMS, despite its capability to detect analytes at lower concentrations compared to GC-MS and the formation of multiple ionized species from a single compound, the number of detected features using the method described in section 2.3.3 was comparable to those detected with GC-MS. One possible explanation for this unexpected similarity in coverage is the sampling technique: HS-SPME sampling employed with GC-MS is more sensitive than the SHS used with GC-IMS, compensating for GC-MS's lower sensitivity compared to GC-IMS. Another factor potentially affecting the number of peaks (and therefore features) detected by these two techniques is their differing chromatographic resolutions, which arise from variations in the commercial GC platforms used. For GC-MS, the instrumental setup ensures appropriate peak resolution and efficiency through temperature programming and the use of a conventional column (30 m, 0.25 mm diameter, 0.25 μm film thickness). In contrast, the commercial GC-IMS employed in this study prioritized simplicity, cost, and speed (isothermal operation and a wide-bore column with programmed nitrogen flow as the carrier gas) at the expense of separation and efficiency. This compromise led to a higher number of coelutions, particularly for early eluting peaks, and hindered the elution of late-eluting peaks. As a result, the number of detected peaks was comparable despite the higher sensitivity of IMS detection.

The ternary plots clearly show that most variables are effectively divided between the joint and individual components, with only a few showing a high residual component. Notably, the variables with higher residuals tend to have lower intensity, as indicated by the more transparent points. To confirm and quantify this observation, we divided the intensity medians of the variables into quartiles for each analytical technique. We then verified that most variables with a predominant residual component in the ternary plots (greater than 40 %) belonged to the first and second quartiles. The calculation revealed that 60 % of GC-IMS, 80 % of GC-MS, and 87.4 % of the variables with residuals greater than 40 % had median intensities lower than the overall median intensity of the data source.

The majority of variables have a residual component lower than 20 %, a threshold highlighted in the plots by the green dashed line (Fig. 2). Additionally, for all the techniques, there is a noticeable skew toward higher individual component values. This observation suggests that for the majority of variables, the individual component is predominant, further speaking of a general complementarity of the three techniques. Interestingly, the variable distribution in the ternary plots is relatively consistent for the two techniques involving chromatographic separation (GC-IMS and GC-MS), whereas the PTR-MS plot exhibits a slightly different distribution. In PTR-MS, there are more variables with residuals greater than 20 %, likely due to the presence of clusters and common fragments, but also a larger number of features skewed toward the individual component.

Ternary plots effectively illustrate the distribution of variables for each data source, offering insights into identifying key variables of

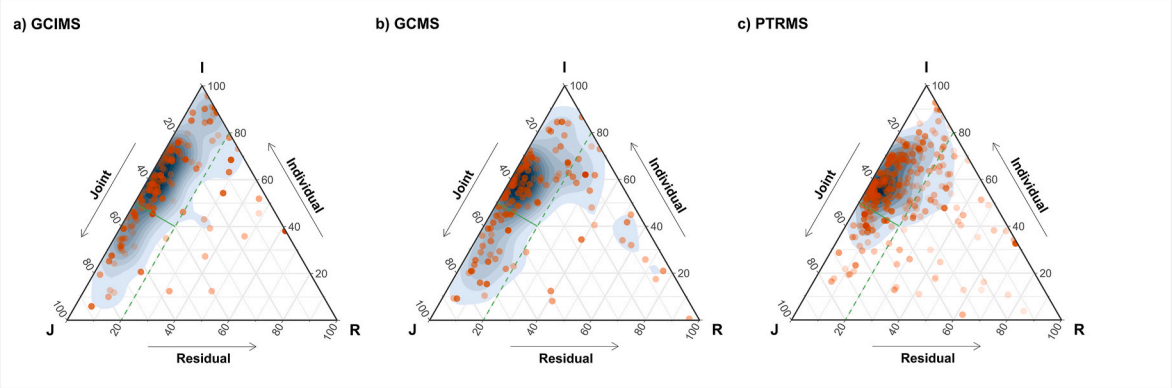


Fig. 2. Ternary plots of variable partitioning following JIVE decomposition for each analytical technique. Each variable is represented as a point within the equilateral triangle, with its position reflecting the ratio of the three components (J, I, R). The transparency of points corresponds to the median intensity of the variable, with more transparent points indicating lower-intensity variables. The green dashed lines highlight the 20 % threshold on the residual axis and the 50 % threshold between the joint and individual axes. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

interest. Specifically, by examining how variables are partitioned among the three components, it is possible to identify those that contribute most to the joint structure (with high values in the joint component) and those that predominantly influence the individual structure (with high individual component values) for each analytical technique.

3.2. *Identification of latent variables and key features*

The interpretation of JIVE model results and the identification of its latent variables enable the data exploration of the multi-source dataset, revealing multivariate patterns in the shared and source-specific structures. Additionally, once key variables contributing to joint and individual variations are identified, they can be further examined by analyzing the original data. In fact, if the variation decomposition achieved through JIVE effectively partitions the variables, it is expected that the patterns observed in the score plots will also be reflected in the original variables, i.e. in the profile of the VOCs in the samples. This approach is useful for identifying characteristic trends in VOC formation

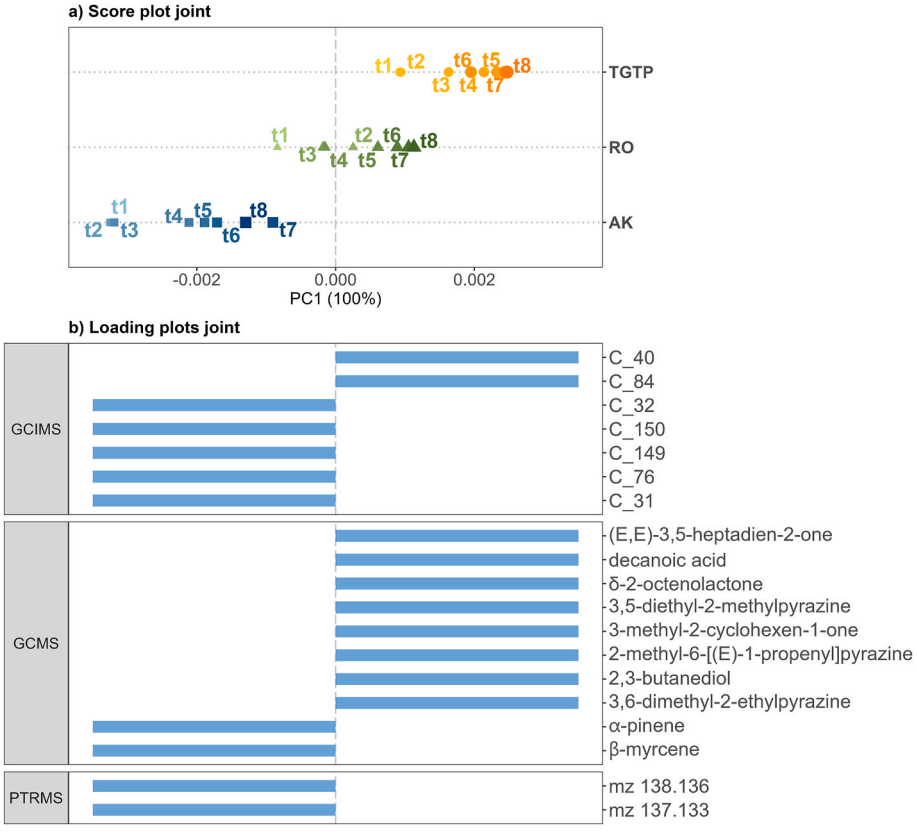


Fig. 3. a) Score plot of the joint common score matrix, showing sample separation by geographical origin and roasting intensity. Points colored according to geographical origin (AK blue, RO green, TGTP orange) of the samples, with label indicating the roasting point from t1 to t8; **b)** Loading plots of key variables for GC-IMS, GC-MS and PTR-MS on PC1 (joint component > 70 %, residual < 20 %). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

and release based on factors such as roasting level and geographical origin of the samples and for determining whether specific chemical classes are associated with common or unique variations.

3.2.1. Joint

The joint matrix has a rank of r = 1, meaning that only PC1, accounting for 100 % of the explained variance, was obtained. The score plot, showing the projections of the samples on PC1, is presented in Fig. 3a. The geographical origins (sample types: AK, RO, and TGTP) are vertically shifted to facilitate interpretation. The three origins groups are separated along PC1 with minimal overlap. Additionally, within each sample type, the temporal evolution related to the roasting process is evident. This multivariate pattern is shared across the three data sources, indicating that this information (origin and roasting) is detectable by all three analytical techniques. For the loading sub-matrices, the contribution to PC1 (positive or negative) of the variables mostly contributing to the joint structure (J > 70 % and R < 20 %) is shown in the loading plot (Fig. 3b). GC-MS is the analytical technique with the highest number of variables identified as contributing significantly to the joint structure (10 variables), followed by GC-IMS (7 variables). Most GC-MS features contribute positively to PC1 (8/10), while most GC-IMS features contribute negatively (5/7). For PTR-MS, only two variables had a joint component value above 70 %, and both contribute negatively to PC1. This is consistent with the distribution observed in the ternary plots (Fig. 2), where PTR-MS variables were more skewed toward higher values of the individual component.

The profiles of the VOCs identified as relevant features for the joint structure were examined plotting the intensity of the original variables according to geographical origin and roasting intensity, as reported in Fig. 4.

From the GC-MS joint loading plot, monoterpenes (α-pinene and β-myrcene) were identified as important compounds contributing to the common variation, with a negative contribution on PC1. The two features reported in the PTR-MS joint loading plot are mass peaks *m/z* 137.133 and *m/z* 138.136, corresponding to a common fragment of monoterpenes ($C_{10}H_{17}^+$) and its C13 isotope, respectively. These results confirm that the JIVE decomposition was effective in identifying variables related to the same chemical class across different data sources and analytically correlated variables (isotopes) within the same data source. Fig. 4a and 4b shows the trends for β-myrcene and *m/z* 137.133. Higher amounts of monoterpenes were found in AK samples, followed by RO samples, with the lowest amounts in TGTP. A slight decrease was observed during roasting for TGTP, but overall, monoterpenes are not significantly influenced by the thermal treatment. A similar trend was observed in the GC-IMS variable C_32 (Fig. 4c), suggesting this peak might belong to the same chemical class. Decanoic acid (GC-MS) shows an opposite yet similar trend (Fig. 4d), as its intensity is not influenced by roasting but is primarily dependent on the origin, with lower levels in AK and higher levels in RO and TGTP. Among the variables with a positive contribution to PC1, three pyrazines can be identified in the GC-MS joint loading plot (3,5-diethyl-2-methylpyrazine (Fig. 4e), 2-methyl-5-[(E)-1-propenyl]pyrazine, and 2-ethyl-3,6-dimethylpyrazine (Fig. 4f)). Pyrazines are compounds known to be generated by the Maillard reaction during roasting, and they significantly contribute to the aroma of roasted foods. In particular, 2-ethyl-3,6-dimethylpyrazine is a key aroma compound for roasted hazelnuts [29]. Their amounts increase during the roasting process, but their intensity varies across the three sample origins, with higher values in TGTP and lower in AK. A similar trend was observed for the two variables from GC-IMS reported in Fig. 4g and 4h. The variation pattern of the common structure, as shown in the joint score plot in Fig. 3a, is clearly recognizable in the trend of these compounds. Finally, looking at the remaining GC-IMS variables, it is identifiable the presence of GC-IMS peaks showing an intensity decrease during roasting (Fig. 4i and 4l), mostly evident for AK.

3.2.2. Individual

Each individual structure matrix had a rank larger than one. As in the case of the joint component only variables with an individual component higher than 70 % (and residual component <20 %) are shown in the loading plots.

PTR-MS is the analytical technique with the highest number of variables identified as contributing significantly to the individual structure (43 variables), followed by GC-IMS (29 variables), and lastly GC-MS (11 variables). This result is consistent with the distribution observed in the ternary plots (Fig. 2), where PTR-MS variables were more skewed toward the individual component.

The individual matrices resulting from the decomposition have ranks r = 3 for GC-IMS, r = 2 for GC-MS and r = 2 for PTR-MS. Therefore, performing factorization using PCA, 3 PCs were obtained from the GC-IMS individual matrix, while 2 PCs were obtained for GC-MS and PTR-MS. First the information provided by the first two principal components was analyzed and compared for the three analytical techniques. Fig. 5 shows the score and loading plots of PC1 and PC2 of the individual score and loading matrices of GC-IMS, GC-MS and PTR-MS.

In the score plots shown in Fig. 5a, the temporal evolution related to the roasting process is evident along PC1 for all three analytical techniques. Additionally, the separation of RO samples from AK and TGTP samples can be observed along PC2. However, since these plots represent individual structures, the relative positions of the samples in the score plots—and consequently the identifiable multivariate patterns—differ, suggesting the presence of VOC trends specific to each technique. The loading plots in Fig. 5b illustrate the relationship between the original variables and the principal components unique to each data source. For GC-MS and PTR-MS, only two PCs were obtained, meaning the variables are fully represented in the PC1-PC2 plane. As the data were not scaled post-JIVE to preserve variation decomposition, the length of the arrows reflects the vector magnitude of the individual component for each variable. Shorter arrows indicate smaller individual components. For GC-IMS, which includes a third principal component (PC3), the quality of representation of the variables in the PC1-PC2 plane (squared cosine) is conveyed by the transparency or shade of grey in the arrows. More transparent arrows indicate that a variable is not well represented by PC1 and PC2, while dark, short arrows suggest that a variable is well represented but has smaller individual components. Comparing arrow lengths within the loading plots, it is notable that variables correlated with PC1 tend to have shorter arrows compared to those correlated with PC2 (and PC3 in GC-IMS). This observation suggests that variables associated with PC1, which reflect the roasting-related pattern, have a portion of their variation attributed to the joint structure, even though they still retain an individual component above the 70 % threshold.

The temporal evolution associated with the roasting process is captured by PC1 for all the three analytical techniques, however, by examining the variables with the highest contributions to PC1 in Fig. 6, different peak intensity profiles can be identified. The first insight emerging from these intensity profiles is that, although the roasting profile is also a recognizable pattern in the joint structure, the temporal evolution trends differ significantly from those observed for the key variable in the joint component. Specifically, for key variables in the joint component - such as pyrazines, C_40 and C_84 (Fig. 4) - an increase in intensity was observed during the roasting process. However, these variables exhibit intensity variations across the three sample origins, indicating clear varietal separation. On the contrary, for key variables for individual components the intensity trends did not show relevant varietal differences.

In the GC-IMS individual loading plot (Fig. 5b.1), three variables (C_39, C_85, and C_133) are well represented on PC1 with negative coordinates. These coordinates correspond to decreasing peak intensity profiles (e.g., C_39 in Fig. 6a) throughout the thermal treatment, suggesting the detection of VOCs that are gradually eliminated during roasting, with no major varietal differences observed. Fig. 6b shows the

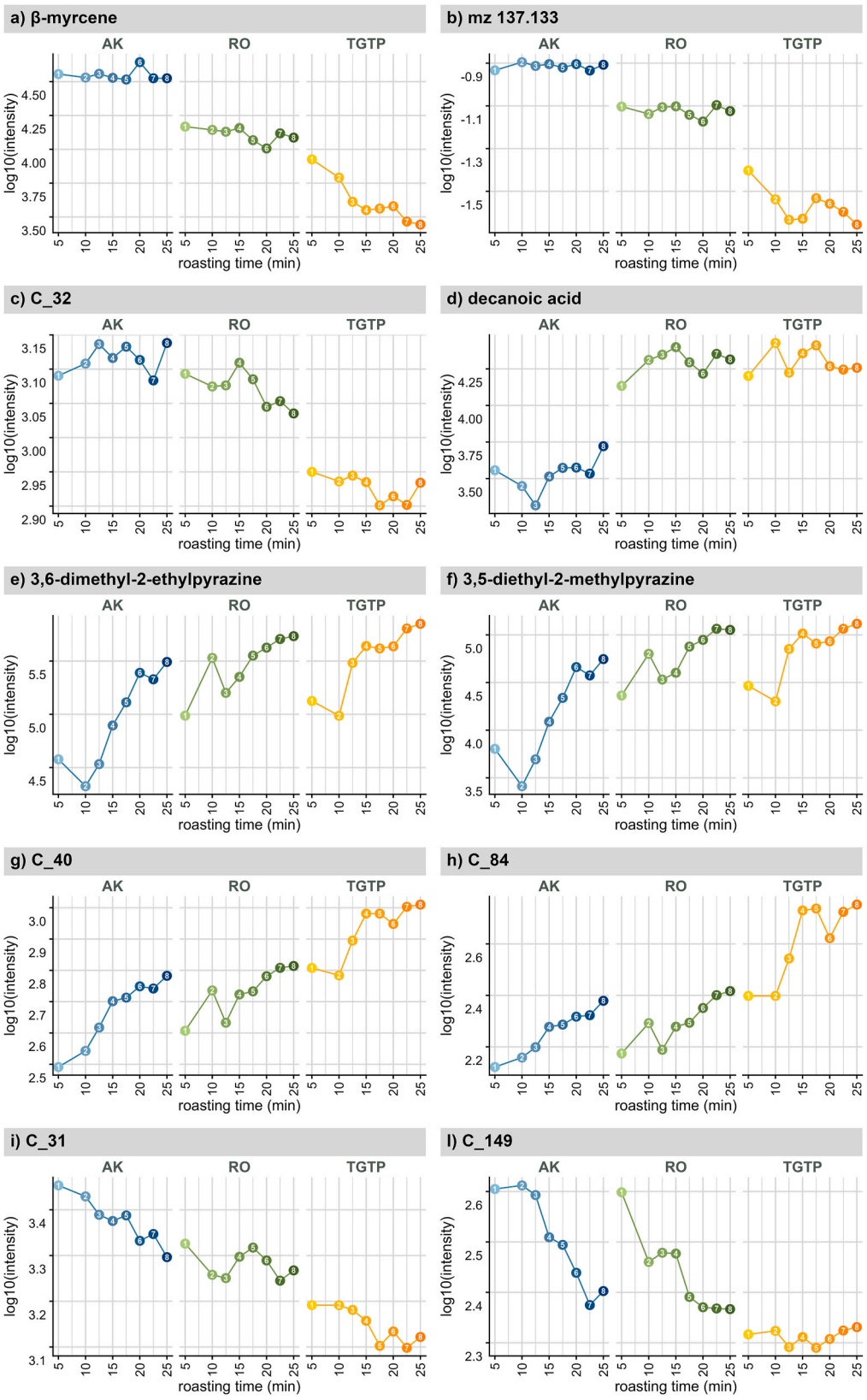


Fig. 4. Trends of VOC profiles associated with joint variation according to geographical origin (AK blue, RO green, TGTP orange). The points represent log10-transformed peak intensities corresponding to roasting points from t1 to t8. Solid lines highlight the VOC trends. **a)** β-myrcene (GC-MS); **b)** m/z 137.133 (PTR-MS); **c)** C_32 (GC-IMS); **d)** decanoic acid (GC-MS); **e)** 3,6-diethyl-2-methylpyrazine (GC-MS); **f)** 2-ethyl-3,6-dimethylpyrazine (GC-MS); **g)** C_40 (GC-IMS); **h)** C_84 (GC-IMS); **i)** C_31 (GC-IMS); **l)** C_149 (GC-IMS). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

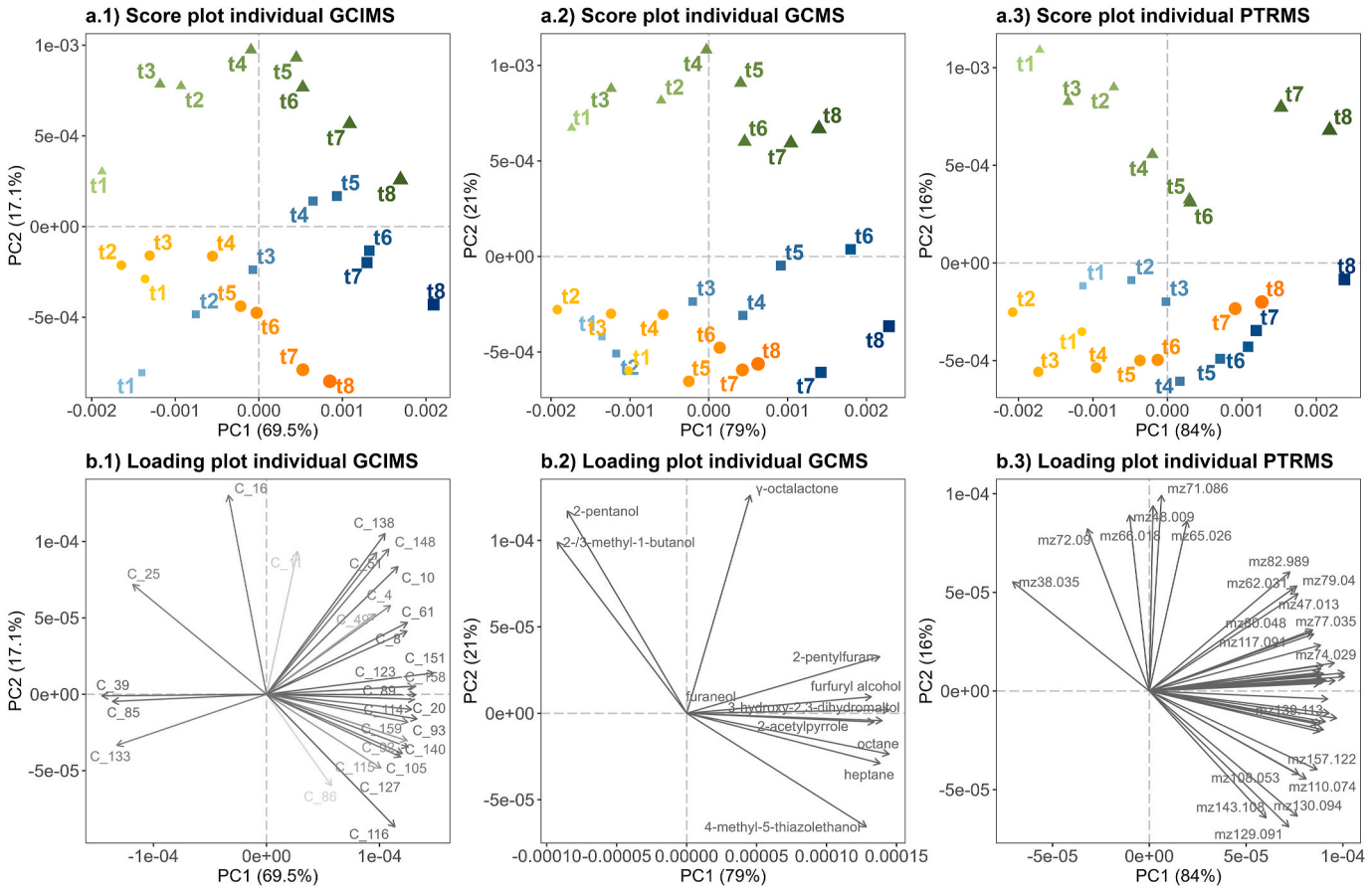


Fig. 5. a) Score plots of the first two components of the individual score matrices of GC-IMS (a.1), GC-MS (a.2) and PTR-MS (a.3). Points colored according to geographical origin (AK blue, RO green, TGTP orange) of the samples, with label indicating the roasting point from t1 to t8; **b)** Loading plots of the first two components of the individual loading matrices of GC-IMS (b.1), GC-MS (b.2) and PTR-MS (b.3). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

intensity profile of C_151, another key variable in the GC-IMS individual component with a high contribution and negative coordinates on PC1. For this peak, the intensity increase range over the roasting process was higher in the AK sample, whereas in RO and TGTP, the intensity ranges were more limited.

In the GC-MS individual component, most key variables are heterocyclic compounds, including furans, thiazoles, pyrroles, and pyrans. Fig. 6c and 6d shows the trends of 3-hydroxy-2,3-dihydromaltol (2,3-dihydro-3,5-dihydroxy-6-methyl-4H-pyran-4-one) and 4-methyl-5-thiazoleethanol. 3-hydroxy-2,3-dihydromaltol, a Maillard reaction product with antioxidant activity [30], was identified in roasted hazelnut by Stilo et al. and is reported as a chiral marker of the roasted hazelnut volatilome [31]. 4-methyl-5-thiazoleethanol, a sulfur-containing heterocyclic compound formed from thiamine degradation, is a flavor compound that has been detected in nuts [32,33] and seeds [34]. The intensity profiles for these compounds showed an earlier plateau in RO samples compared to AK and TGTP samples, with a slower increase observed after 15 min of thermal treatment (roasting point t4).

Considering the intensity profiles for PTR-MS, a trend emerges that differentiates some key features of the individual structures in this technique, such as mass peaks *m/z* 95.0839 (C7H11+) (see Fig. 6e) and *m/z* 119.1067 (C6H15O2+) (not shown). On the contrary to what was observed for C_151, for these peaks, the range of intensity increase over the roasting process was higher in the TGTP sample, whereas in the AK sample, the intensity increase from the initial to final roasting stages was more limited. Finally, Fig. 6f shows the trend for *m/z* 96.0832 (C6H10N+), tentatively identified as dimethyl pyrrole or ethyl pyrrole, which was not detected by GC-MS. This mass peak exhibited similar intensity trends and ranges across the three origins, with high variability in the early stages and a steep increase after 12.5–15 min (t3/t4), corresponding to standard roasting intensity. This suggests the detection of VOCs whose formation is highly dependent on roasting time and whose intensity may be linked to over-roasting.

Moving to PC2, for all techniques it shows a separation between RO samples and AK and TGTP samples, with no observable temporal evolution along this component, particularly in the GC-MS and PTR-MS data. Therefore, variables correlated with PC2 are expected to be more abundant in RO samples and appear to be origin-related rather than strongly influenced by the roasting process. In Fig. 7 are presented the VOC profiles of variables with the highest contribution to PC2 for each analytical technique.

Reflecting the pattern observed in the GC-IMS individual score plot, peak C_16 (Fig. 7a) was more abundant in RO samples, with only a slight decreasing trend related to thermal treatment. In contrast, the intensity of peak C_148 (Fig. 7b) increased sharply with roasting intensity. While the distribution of C_16 is relatively similar to variables detected by the other techniques (despite some differences), the profile observed for C_148 was detectable only by GC-IMS. For GC-MS, the identified VOCs were γ-octalactone (Fig. 7c), 2-pentanol (not shown) and 2-/3-methyl-1-butanol (Fig. 7d). Both short chain alcohols and lactones are products of lipid degradation, and their content is linked to post-harvest processes [9]. 2-/3-methyl-1-butanol and 2-pentanol showed a similar decreasing trend, suggesting they are present in raw hazelnuts and are eliminated during roasting due to their volatility. In contrast, no decreasing trend was observed for γ-octalactone, which is less volatile. These VOC profiles might indicate different susceptibility of the kernels to oxidation

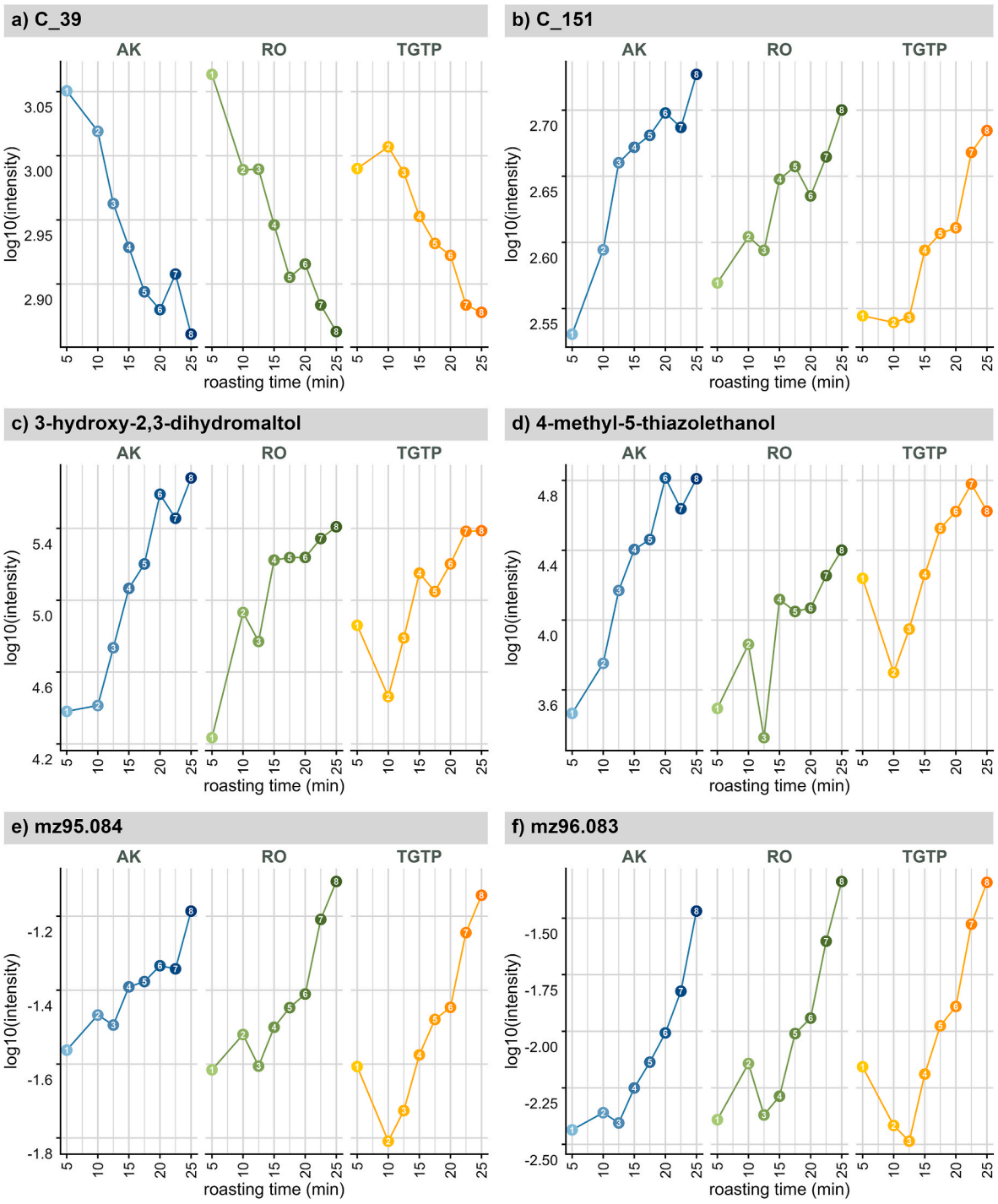


Fig. 6. Trends of VOC profiles associated with PC1 of individual variation, separated and colored according to geographical origin (AK blue, RO green, TGTP orange). The points represent log10-transformed peak intensities corresponding to roasting points from t1 to t8. Solid lines highlight the VOC trends. **a)** C_ 39 (GC-IMS); **b)** C_151 (GC-IMS); **c)** 3-hydroxy-2,3-dihydromaltol (GC-MS); **d)** 4-methyl-5-thiazolethanol (GC-MS); **e)** m/z 95.084 (PTR-MS); **f)** m/z 96.083 (PTR-MS). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

depending on their origin or differences in storage conditions prior to roasting. For PTR-MS, the mass peaks with the highest contribution to PC2 were *m/z* 71.086 (Fig. 7e), m/z 48.009 (Fig. 7f), and *m/z* 66.018 (not shown). The *m/z* 71.086 peak (C5H11+) is generated from the fragmentation of various ionized species, such as C5-alcohols, C5-aldehydes, and acids [25]. This result partially aligns with the GC-MS findings, as this fragment is also produced by 2-pentanol and 2-/3-methyl-1-butanol [35]. However, since PTR-MS is a DIMS technique, isomers are detected as a single feature (*m/z*). Therefore, the PTR-MS feature *m/z* 71.086 includes information from other VOCs generating this fragment. As a result, while it contributes to the separation of RO samples, the profile of *m/z* 71.086 is not directly correlated with the short-chain alcohols identified by GC-MS. For the mass peaks *m/z* 48.009 and *m/z* 66.018 (which had similar profiles), it was not possible to attribute any molecular formula of the protonated ions. Although annotation was not possible, the variable profiles clearly indicate an origin-related distribution, which was detected only by PTR-MS.

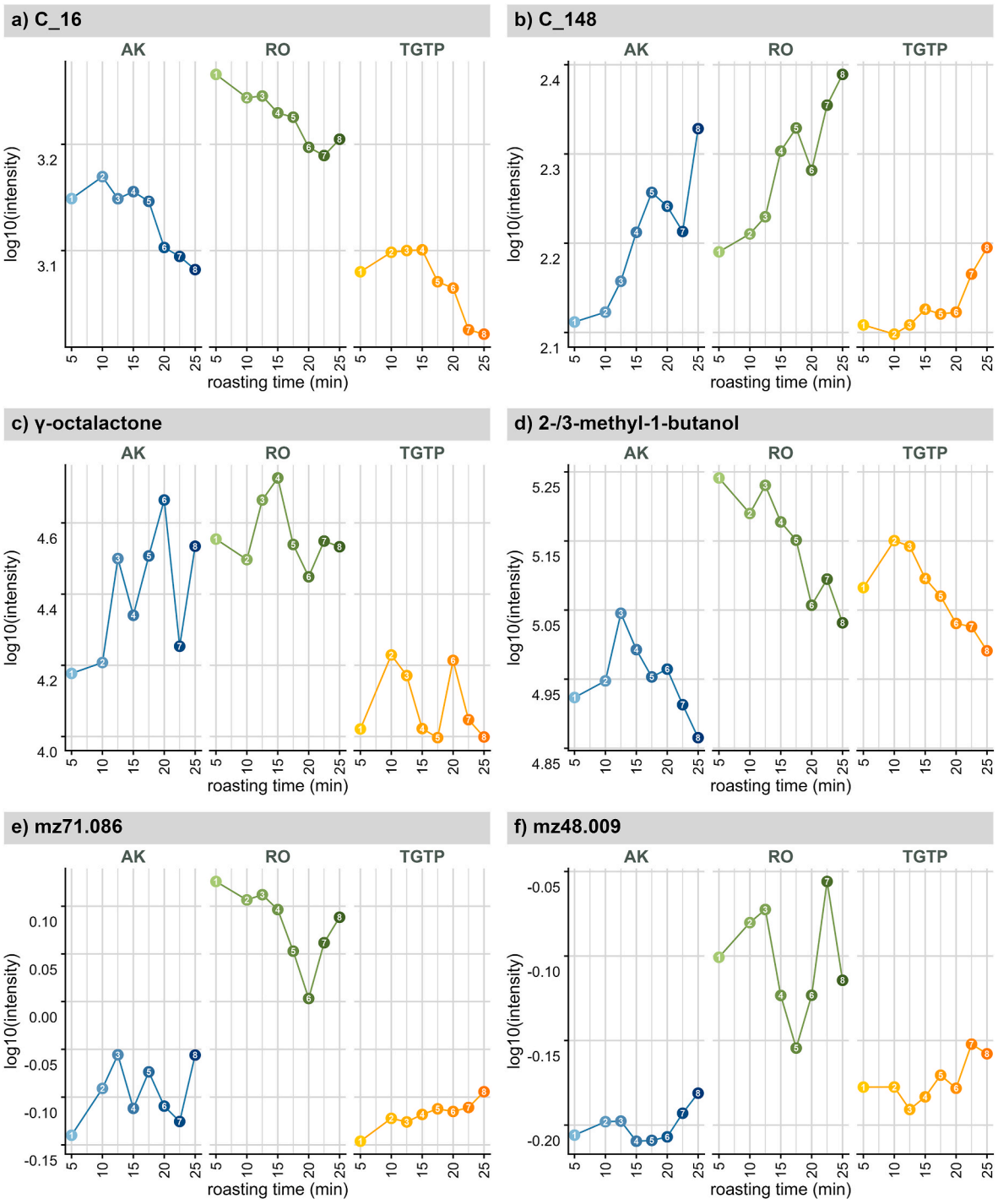


Fig. 7. Trends of VOC profiles associated with PC2 of individual variation, separated and colored according to geographical origin (AK blue, RO green, TGTP orange). The points represent log10-transformed peak intensities corresponding to roasting points from t1 to t8. Solid lines highlight the VOC trends. **a)** C_16 (GC-IMS); **b)** C_148 (GC-IMS); **c)** γ-octalactone (GC-MS); **d)** 2-/3-methyl-1-butanol (GC-MS); **e)** m/z 71.086 (PTR-MS); **f)** m/z 48.009 (PTR-MS). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

As mentioned at the beginning of this section, factorization of the GC-IMS individual matrix yielded three principal components (PCs). Fig. 8 illustrates the structure associated with the third principal component (PC3) of the GC-IMS individual component, showing score and loading plots for PC1 vs. PC3 and PC2 vs. PC3.

In the score plots, a curved trend along PC3 is evident in relation to roasting time. Samples with intermediate roasting intensity (t4 and t5) are positioned at higher PC3 values, while those with low and high roasting intensities are positioned at lower PC3 values within the PC planes. This pattern is most pronounced in the AK and RO samples and was observed exclusively in the GC-IMS dataset. Based on the loading plots and the contributions of variables to PC3, GC-IMS peaks C_11, C_49 and C_86 were identified as the primary contributors to this pattern and their intensity profiles are shown in Fig. 9.

These peaks were further analyzed by examining the GC-IMS raw data as a 2D topographic plot (chromatogram × mobilogram) in the commercial software VOCal. Peaks C_11 and C_49 were identified as protonated monomers based on their coordinates and the presence of a

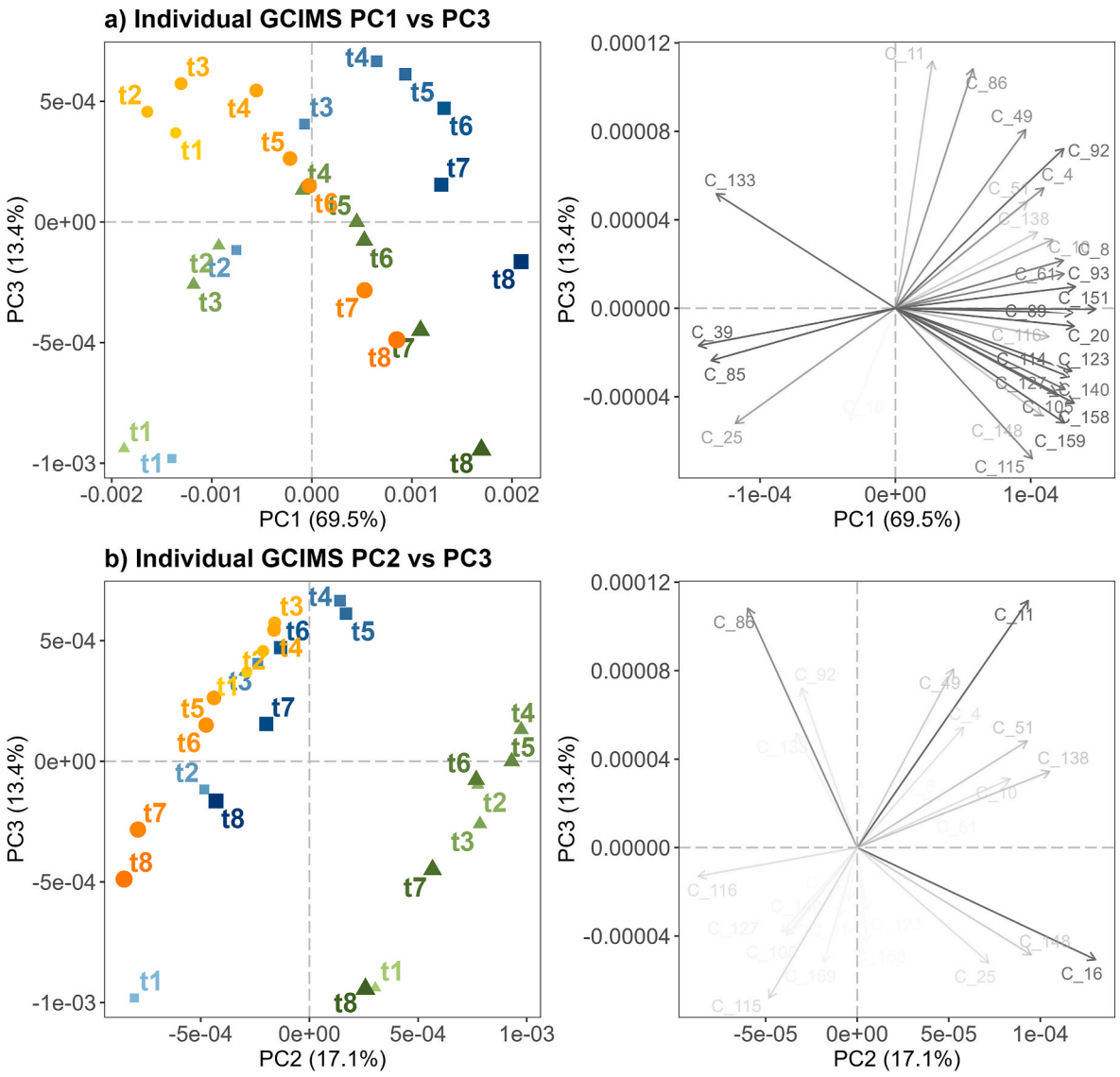


Fig. 8. Score and loading plots of the individual score and loading matrices of GC-IMS, showing the pattern observable for the third principal component: **a)** PC1 vs PC3; **b)** PC2 vs PC3. In the score plots the points are colored according to the geographical origin (AK blue, RO green, TGTP orange) of the samples, with label indicating the roasting point from t1 to t8. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

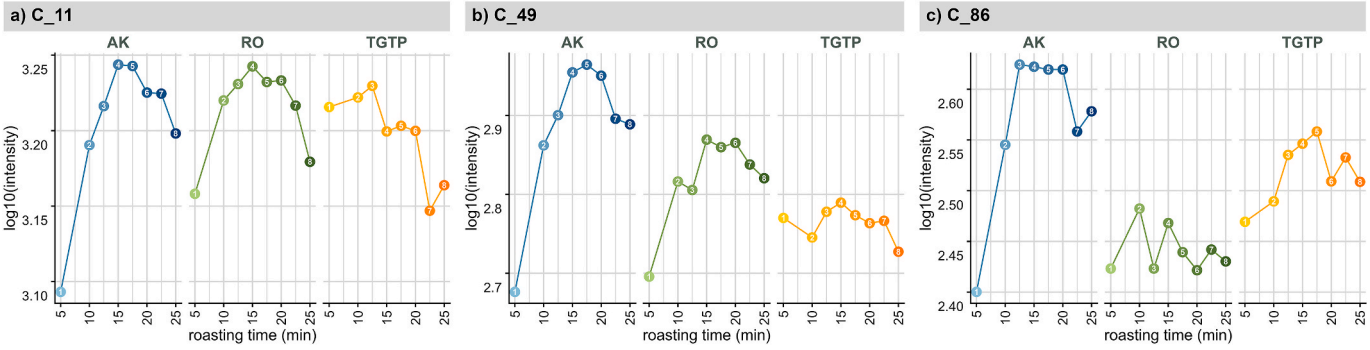


Fig. 9. Trends of GC-IMS peak intensity profiles associated with PC3 of individual variation, separated and colored according to geographical origin (AK blue, RO green, TGTP orange). The points represent log10-transformed peak intensities corresponding to roasting points from t1 to t8. Solid lines highlight the VOC trends. **a)** C_11 (GC-IMS); **b)** C_49 (GC-IMS); **c)** C_86 (GC-IMS). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

second peak perfectly aligned along the retention time axis but shifted to a higher drift time, indicating the proton-bound dimer (Supplementary Figure 4a). This observation provides a possible explanation for the curved trend: at low VOC concentrations, the monomer is the most stable ion, and its peak height increases with volatile concentration until the proton-bound dimer forms from the same compound, causing a decrease in monomer peak intensity [36]. Thus, the observed trend does not reflect an actual variation in the samples but is instead linked to the ionization mechanism of IMS, accurately attributed by the JIVE model to individual variation. The peak C_86 was detected at relatively high drift time (6.90 ms) suggesting it is a proton bound dimer, however identifying the corresponding protonated monomer was not possible due to the presence of many other peaks eluting within a narrow retention time range (Supplementary Figure 4b). As visible in Fig. 9c, the trend for C_86 differs from those of C_11 and C_49, primarily indicating differences among the origins, with a notably lower peak intensity in RO samples, as confirmed by inspection of the 2D topographic plot.

4. Conclusions

The evolution of the hazelnut volatilome during roasting was studied through the combination of multi-platform analysis, utilizing three analytical techniques commonly employed for volatilomic studies (GC-IMS, GC-MS, PTR-ToF-MS), and advanced multiblock data modelling.

The results demonstrate the potential of a multivariate data fusion approach, such as JIVE, for the integrated analysis of a volatilomic dataset obtained through a multi-platform analysis. The described data exploration strategy enabled a high-level comparison of the three techniques, which revealed the predominant contribution of the individual component to the variation for each data source, suggesting the complementarity of the three analytical approaches. This observation was further supported by the visualization of the contribution of the individual variables in the ternary plots, which also revealed distinct patterns. Notably, PTR-MS exhibited a higher number of variables skewed toward greater individual contributions. As far as the chemical interpretation of the results is concerned, the identification of the latent components of the joint and individual matrices allowed us to discuss the trend of the most influential key variables (GC-MS compounds, GC-IMS peaks, PTR-MS mass peaks), highlighting the information shared across the dataset and individually present in the three techniques. This type of analysis also allowed us to relate the observed trends to thermal treatment and/or sample origin.

While this study underscores the complementarity of these analytical techniques and the value of a multi-platform approach for volatilome characterization in food matrices, separating common from individual variations can be beneficial also to design studies that, for budget or time reasons, are constrained to a single analytical technique. In these cases, the assessment of common patterns detectable by all the analytical techniques considered, alongside VOC trends unique to each, offers valuable guidance for selecting the most suitable analytical method based on the study's focus and budget constraints.

CRediT authorship contribution statement

Maria Mazzucotelli: Writing – review & editing, Writing – original draft, Visualization, Software, Investigation, Formal analysis, Data curation. **Iuliia Khomenko:** Writing – review & editing, Data curation. **Emanuela Betta:** Data curation. **Elena Gabetti:** Writing – review & editing, Resources. **Luca Falchero:** Resources. **Eugenio Aprea:** Supervision. **Andrea Cavallero:** Supervision, Project administration, Funding acquisition. **Franco Biasioli:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Pietro Franceschi:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Methodology, Conceptualization.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT in order to improve the readability and language of the manuscript. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

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.

Acknowledgement

This study was carried out within the Interconnected Nord-Est Innovation Ecosystem (iNEST) and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.5 – D.D. 1058 June 23, 2022, ECS00000043).

Maria Mazzucotelli would like to express her gratitude to Andrea Dell'Olio and Antonia Corvino for their support throughout the research process, and to Soremartec Italia srl for the financial support for her PhD project.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.talanta.2025.127720.

Data availability

Data will be made available on request.

References

[1] A.E. Lytou, E.Z. Panagou, G.-J.E. Nychas, Volatilomics for food quality and authentication, Curr. Opin. Food Sci. 28 (Aug. 2019) 88–95, https://doi.org/10.1016/j.cofs.2019.10.003.

[2] S. Squara, et al., Artificial Intelligence decision-making tools based on comprehensive two-dimensional gas chromatography data: the challenge of quantitative volatilomics in food quality assessment, J. Chromatogr. A 1700 (Jul. 2023) 464041, https://doi.org/10.1016/j.chroma.2023.464041.

[3] P. Mishra, et al., Recent trends in multi-block data analysis in chemometrics for multi-source data integration, TrAC, Trends Anal. Chem. 137 (Apr. 2021) 116206, https://doi.org/10.1016/j.trac.2021.116206.

[4] J. Kuligowski, et al., Analysis of multi-source metabolomic data using joint and individual variation explained (JIVE), Analyst 140 (13) (2015) 4521–4529, https://doi.org/10.1039/C5AN00706B.

[5] M.J. O'Connell, E.F. Lock, R.JIVE for exploration of multi-source molecular data, Bioinformatics 32 (18) (Sep. 2016) 2877–2879, https://doi.org/10.1093/bioinformatics/btw324.

[6] E.F. Lock, K.A. Hoadley, J.S. Marron, A.B. Nobel, Joint and individual variation explained (JIVE) for integrated analysis of multiple data types, Ann. Appl. Stat. 7 (1) (Mar. 2013), https://doi.org/10.1214/12-AOAS597.

[7] I. Måge, A.K. Smilde, F.M. van der Kloet, Performance of methods that separate common and distinct variation in multiple data blocks, J. Chemom. 33 (1) (Jan. 2019), https://doi.org/10.1002/cem.3085.

[8] K. Król, M. Gantner, Morphological traits and chemical composition of hazelnut from different geographical origins: a review, Agriculture 10 (9) (Aug. 2020) 375, https://doi.org/10.3390/agriculture10090375.

[9] S. Squara, et al., Corylus avellana L. Aroma Blueprint: potent odorants signatures in the volatilome of high quality hazelnuts, Front. Plant Sci. 13 (Mar) (2022), https://doi.org/10.3389/fpls.2022.840028.

[10] M. Mazzucotelli, et al., Monitoring alkyl pyrazines in roasted hazelnuts by SHS-GC-IMS: IMS response assessment and standardization, Talanta 259 (Jul. 2023) 124568, https://doi.org/10.1016/j.talanta.2023.124568.

[11] R Core Team, R: A Language and Environment for Statistical Computing, 2024. Vienna, Austria. [Online]. Available: https://www.R-project.org/.

[12] K. Ushey, J.J. Allaire, Y. Tang, reticulate: interface to 'Python [Online]. Available: https://rstudio.github.io/reticulate/, 2024.

[13] Posit team, RStudio: Integrated Development Environment for R, 2024. Boston, MA. [Online]. Available: http://www.posit.co/.
[14] E. Taskesen, Findpeaks Is for the Detection of Peaks and Valleys in a 1D Vector and 2D Array (Image), Jul. 2023, https://doi.org/10.5281/zenodo.8145101. Zenodo.
[15] H. Parastar, J. Christmann, P. Weller, Automated 2D peak detection in gas chromatography-ion mobility spectrometry through persistent homology, Anal. Chim. Acta 1289 (Feb. 2024) 342204, https://doi.org/10.1016/j.aca.2024.342204.
[16] M. Hahsler, M. Piekenbrock, D. Doran, Dbscan: fast density-based clustering with R, J. Stat. Software 91 (1) (2019) 1–30, https://doi.org/10.18637/jss.v091.i01.
[17] M. Hahsler, M. Piekenbrock, Dbscan: density-based spatial clustering of applications with noise (DBSCAN) and related algorithms [Online]. Available: https://CRAN.R-project.org/package=dbscan, 2024.
[18] S. Horsch, D. Kopczynski, E. Kuthe, J.I. Baumbach, S. Rahmann, J. Rahnenführer, A detailed comparison of analysis processes for MCC-IMS data in disease classification—automated methods can replace manual peak annotations, PLoS One 12 (9) (Sep. 2017) e0184321, https://doi.org/10.1371/journal.pone.0184321.
[19] H. van Den Dool, P. Dec Kratz, A generalization of the retention index system including linear temperature programmed gas—liquid partition chromatography, J. Chromatogr. A 11 (1963) 463–471, https://doi.org/10.1016/S0021-9673(01)80947-X.
[20] J. Josse, F. Husson, missMDA: a package for handling missing values in multivariate data analysis, J. Stat. Software 70 (1) (2016) 1–31, https://doi.org/10.18637/jss.v070.i01.
[21] L. Cappellin, et al., On data analysis in PTR-TOF-MS: from raw spectra to data mining, Sensor. Actuator. B Chem. 155 (1) (Jul. 2011) 183–190, https://doi.org/10.1016/j.snb.2010.11.044.
[22] L. Cappellin, et al., PTR-ToF-MS and data mining methods: a new tool for fruit metabolomics, Metabolomics 8 (5) (Oct. 2012) 761–770, https://doi.org/10.1007/s11306-012-0405-9.
[23] The fitness for purpose of analytical methods A laboratory guide to method validation and related topics the fitness for purpose of analytical methods A laboratory guide to method validation and related topics second edition 2014 i Eurachem guide the Fitness for purpose of analytical methods A laboratory guide to method validation and related topics second edition acknowledgements [Online]. Available: www.eurachem.org, 2014.
[24] D.A. Armbruster, T. Pry, Limit of Blank, Limit of Detection and Limit of Quantitation, 2008.
[25] A.M. Yáñez-Serrano, et al., GLOVOCS - master compound assignment guide for proton transfer reaction mass spectrometry users, Atmos. Environ. 244 (Jan. 2021) 117929, https://doi.org/10.1016/j.atmosenv.2020.117929.
[26] N.E. Hamilton, M. Ferry, Ggtern: ternary diagrams using ggplot2, Journal of Statistical Software, Code Snippets 87 (3) (2018) 1–17, https://doi.org/10.18637/jss.v087.c03.
[27] S. Lê, J. Josse, F. Husson, FactoMineR: an R package for multivariate analysis, J. Stat. Software 25 (1) (2008) 1–18, https://doi.org/10.18637/jss.v025.i01.
[28] H. Wickham, et al., ggplot2: create elegant data visualisations using the grammar of graphics [Online]. Available: https://ggplot2.tidyverse.org, 2024.
[29] J. Kiefl, P. Schieberle, Evaluation of process parameters governing the aroma generation in three hazelnut cultivars (Corylus avellana L.) by correlating quantitative key odorant profiling with sensory evaluation, J. Agric. Food Chem. 61 (22) (Jun. 2013) 5236–5244, https://doi.org/10.1021/jf4008086.
[30] Z. Chen, et al., Effect of hydroxyl on antioxidant properties of 2,3-dihydro-3,5-dihydroxy-6-methyl-4 *H* -pyran-4-one to scavenge free radicals, RSC Adv. 11 (55) (2021) 34456–34461, https://doi.org/10.1039/D1RA06317K.
[31] F. Stilo, M. Cialiè Rosso, S. Squara, C. Bicchi, C. Cordero, C. Cagliero, Corylus avellana L. Natural signature: chiral recognition of selected informative components in the volatilome of high-quality hazelnuts, Front. Plant Sci. 13 (Apr. 2022), https://doi.org/10.3389/fpls.2022.844711.
[32] N. Artik, S. Akan, Y. Okay, N. Durmaz, A.İ. Köksal, Volatile aroma component of natural and roasted hazelnut varieties using solid-phase microextraction gas chromatography/mass spectrometry, Acta Scientiarum Polonorum Hortorum Cultus 20 (5) (Oct. 2021) 85–96, https://doi.org/10.24326/asphc.2021.5.8.
[33] W. Chiu, T. Muramatsu, Y. Zhou, T. Miyakawa, M. Tanokura, Comparison of peanut compounds during roasting and the effect of peanut shells, ACS Food Science & Technology 2 (4) (Apr. 2022) 691–702, https://doi.org/10.1021/acsfoodscitech.2c00015.
[34] X.-L. Ma, et al., A study of flavor variations during the flaxseed roasting procedure by developed real-time SPME GC–MS coupled with chemometrics, Food Chem. 410 (Jun. 2023) 135453, https://doi.org/10.1016/j.foodchem.2023.135453.
[35] M. Di Guardo, et al., Genetic characterization of an almond germplasm collection and volatilome profiling of raw and roasted kernels, Hortic. Res. 8 (1) (Dec. 2021) 27, https://doi.org/10.1038/s41438-021-00465-7.
[36] R. Brendel, S. Schwolow, S. Rohn, P. Weller, Comparison of PLSR, MCR-ALS and Kernel-PLSR for the quantification of allergenic fragrance compounds in complex cosmetic products based on nonlinear 2D GC-IMS data, Chemometr. Intell. Lab. Syst. 205 (Oct. 2020) 104128, https://doi.org/10.1016/j.chemolab.2020.104128.