Volatile fingerprint of food products with untargeted SIFT-MS data coupled with mixomics methods for profile discrimination: application case on cheese

Marine Reyrolle, Mylène Ghislain, Noëlle Bru, Germain Vallverdu, Thierry Pigot, Valérie Desauziers, Mickael Le Béchec

▶ **To cite this version:**

Marine Reyrolle, Mylène Ghislain, Noëlle Bru, Germain Vallverdu, Thierry Pigot, et al.. Volatile fingerprint of food products with untargeted SIFT-MS data coupled with mixomics methods for profile discrimination: application case on cheese. Food Chemistry, 2022, 369, pp.130801. 10.1016/j.foodchem.2021.130801 . hal-03318841

HAL Id: hal-03318841

https://imt-mines-ales.hal.science/hal-03318841

Submitted on 30 Aug 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Volatile fingerprint of food products with untargeted SIFT-MS data coupled with mixOmics methods for profile discrimination: application case on cheese

Marine Reyrolle,[a] Mylène Ghislain,[a], Noëlle Bru,[b] Germain Vallverdu,[a] Thierry Pigot,[a] Valérie Desauziers,[a] Mickael Le Bechec[a]*

[a] Universite de Pau et des Pays de l'Adour, E2S UPPA, CNRS, IMT Mines Ales, IPREM, Pau, France, Institut des sciences analytiques et de Physicochimie pour l'environnement et les Matériaux, UMR5254, Hélioparc, 2 avenue Président Angot, 64053, PAU cedex 9, France

[b] Universite de Pau et des Pays de l'Adour,E2S UPPA, CNRS, LMAP, Anglet, France, Laboratory of Mathematics and its Applications of Pau – MIRA, UMR5142, Allée du Parc Montaury, Anglet, France

Correspondence to : Mickael Le Bechec, Universite de Pau et des Pays de l'Adour, E2S UPPA, CNRS, IMT Mines Ales, IPREM, Pau, France, Institut des sciences analytiques et de Physicochimie pour l'environnement et les Matériaux, UMR5254, Hélioparc, 2 avenue Président Angot, 64053, PAU cedex 9, France

e-mail: mickael.lebechec@univ-pau.fr

Abstract:

Volatile organic compounds (VOC) emitted by food products are decisive for the perception of aroma and taste. The analysis of gaseous matrices is traditionally done by detection and quantification of few dozens of characteristic markers. Emerging direct injection mass spectrometry technologies offer rapid analysis based on a soft ionization of VOC without previous separation. The recent increase of selectivity offered by the use of several precursor ions coupled with untargeted analysis increases the potential power of these instruments. However, the analysis of complex gaseous matrix results in a large number of ion conflicts, making the quantification of markers difficult, and in a large volume of data. In this work, we present the exploitation of untargeted SIFT-MS volatile fingerprints of ewe PDO cheeses in real farm model, using mixOmics methods allowing us to illustrate the typicality, the manufacturing processes reproducibility and the impact of the animals' diet on the final product.

Keywords: SIFT-MS, mixOmics, volatile fingerprints, typicality of agro-food products, ewe cheeses

I. Introduction

Volatile Organic Compounds (VOCs) are low molecular weight organic molecules able of evaporating or sublimating at room temperature thanks to their high vapour pressure and low boiling point. Depending on the concentration and the nature of the exposure, VOCs may have positive effect (pleasant smell) or negative impact (pollutants). In food, the VOCs composition contributes to the development of odour and flavour, which mainly affects the food acceptability by the consumer (Lytou et al., 2019). Thus, a wide variety of volatile chemical compounds with different concentration ranges are emitted by biological processes in food products (enzymatic or metabolic pathways, . Recent trends in VOCs analysis in food industry, supported by improved analytical techniques, have focused on three main areas:

(i) The aroma profiling: allowing a better understanding of factors that give rise to the aroma profiles (Sousa et al., 2020),

(ii) The food safety: tracking the origin of food deterioration or contaminations (Castro-Puyana & Herrero, 2013),

(iii) The food consistency and quality: ensuring production-line consistency of products from batch to batch (Cecchi et al., 2018).

In addition, depending on the product, the aroma profile can be correlated to several parameters: geographic origin, seasonal variations or storage conditions (Santos & Oliveira, 2017). For example, with cheeses, various factors can impact aromatic profile of cheeses (Boltar et al., 2019), such as the animal's breed (Ferreira et al., 2009) and their diet (Ianni et al., 2020), the ferments or enzymes used during processing (Feutry et al., 2012), the pasteurization of milk (Jiang et al., 2019), the ripening conditions and the season variation (Boltar et al., 2015). Contamination during the steps of the manufacturing process with bacteria or fungi may also alter the flavour of cheeses.

To assess the authenticity of food products, a wide range of analytical approaches are currently used such as mass spectrometry techniques coupled with liquid and gas chromatography (Ch et al., 2021), spectroscopic techniques (Infra-red, NMR) or sensory analysis (Luykx & van Ruth, 2008). The differentiation of products is mostly achieved with identified and quantified variables (markers / compounds) in a three steps methodology: 1) sample preparation with solvent extraction or headspace techniques often associated with solid-phase microextraction for VOC pre-concentration; 2) separation of markers; 3) identification and quantification of markers.

However, there is a growing interest in the notion of volatile fingerprints, where the goal is not to identify every molecule of the matrix, but to compare profiles of products (Balkir et al., 2021). This leads to the definition of the term "volatilome" to describe all the volatile compounds emitted by an organism, an ecosystem or a product (such as food) and of the term volatilomics (Cumeras, 2017) as a subdomain of metabolomics.

Recent years have also seen the development of some direct injection mass spectrometer instruments (DIMS) (Biasioli et al., 2011; Deuscher et al., 2019) such as proton-transfer reaction mass spectrometer (PTR-MS) or selected ion flow tube mass spectrometry (SIFT-MS) (Smith & Španěl, 2005). Such technologies use a soft chemical ionization of the VOCs, leading to product ions detected and quantified without any previous separation step. The selectivity of these instruments is based on the chemical reactivity between a precursor ion and an analyte and no longer requires separation of the latter. These two technologies (PTR-MS and SIFT-MS) can be used for targeted detection and

quantification of markers (H. Z. Castada et al., 2014) with Single Ion Monitoring mode (SIM) or non-targeted analysis with a SCAN detection mode, which gives the signal of all the product ions according to their mass to charge (m/z) ratio. Depending on the number of precursor ions used and the number of VOCs in the sample, these technologies generate a large amount of data that can be sometimes difficult to interpret.

A recent SIFT-MS instrument Voice 200Ultra (SYFT ®) with a dual polarity plasma source offers up to 8 different precursor ions (H_3O^+, $O_2^{\bullet+}$, NO^+, $O^{\bullet-}$, OH^-, $O_2^{\bullet-}$, NO_2^- et NO_3^-), thus increasing the potential selectivity of the technique and increasing the data volume in the SCAN mode (up to 3 080 values of product ions with m/z ratio from 15 to 400). The dataset obtained by SIFT-MS analysis in full scan mode constitutes a real volatile fingerprint of the product. Samples can then be classified according to their volatile profile reflecting differences between groups. However, the substantial volume of data requires dedicated chemometric methods already applied in many domains of omics (Bajoub et al., 2018) and food science (Farag et al., 2020). Chemometric analysis is the application of mathematical and statistical tools to analyse and extract the maximum information from chemical data (Bergamaschi et al., 2020).

This study therefore proposes an original approach by coupling a DIMS analysis (SIFT-MS) and an exploitation of the results by adapted statistical methods. The volatile profile emitted by ewe cheeses have been measured with SIFT-MS in positive and negative full scan detection modes and the resulted profiles have been analysed with chemometric tools to highlight the sample variability and to discriminate samples based on producers. Precisely, the multivariate mixOmics methods (Rohart et al., 2017) were used on the scan dataset as tools for representing the dispersion and the discrimination of samples. MixOmics is a user-friendly R package dedicated to the exploration, mining, integration and visualisation of large data sets. It provides attractive functionalities such as (i) insightful visualisations with dimension reduction, (ii) identification of molecular signatures and (iii) improved usage with common calls to all visualisation and performance assessment methods.

II. Material and Method

1. Samples

Ossau-Iraty PDO cheese, a famous uncooked pressed cheese from French Pyrenees, were provided by the partners of agricultural research project BioNAchol (financially supported by the Conseil Régional Nouvelle Aquitaine, France). The specifications of the Ossau-Iraty PDO include six main rules: 1) the milk has to be produced in Bearn and French Bask Country (south west of France on the Pyrenees mountain), 2) by the only 3 local ewe breeds (Manech ginger head, Manech blach head and Basco-Bearnaise), 3) fed in stables or in fields with local fodders and small amount of cereals but without GMO. 4) The ewes are milked only part of the year (from december to august), 5) the manufacturing has to respect the traditional methodology and finally 6) the ripening step has to be longer than 80 days. These cheeses take the form of a generously shaped tomme with a natural rind, the colour of which varies from orange-yellow to grey depending on the maturing conditions. The aromas are present and varied and the texture is still supple to firm without being either sticky or dry. The taste of Ossau-Iraty is delicately ewe-like, with a slight hazelnut flavour. The objective of this project was to study the impact of plants with secondary bioactive metabolites on lactating ewe's health (limitation of parasite development) and on the organoleptic qualities of cheeses. Thus, two farm models were selected with two ewe breeders (E1 and E2) have been selected: E1 mostly fed his animals in farm stable and E2 mostly fed his animal in fields. A feeding protocol was applied: two weeks with control diet, two weeks with modified diet (supplied with bioactive secondary metabolites plants; sainfoin or chicory) and finally two weeks again with control diet. The composition of the diets was detailed in

table SI 1 & 2. Ossau-Iraty cheeses were then produced with the last collected milk of each period and ripened for 6 to 9 months. This protocol was repeated over two years (Table 1). The samples were identified with a code E_ _ with the first number corresponding to the producer and the second number to the sample (6 samples per producer).

SAMPLE	BATCH/PRODUCERS	DIET	YEAR
E11	E1	Control	1
E12	E1	Modified (sainfoin)	1
E13	E1	Control	1
E14	E1	Control	2
E15	E1	Modified (sainfoin)	2
E16	E1	Control	2
E21	E2	Control	1
E22	E2	Modified (chicory)	1
E23	E2	Control	1
E24	E2	Control	2
E25	E2	Modified (chicory)	2
E26	E2	Control	2

Table 1 : Samples description

2. Sample preparation

15 g of cheese were cut into pieces of one centimeter per side and then introduced into a 1 litre bottle. The bottle was fitted with a polypropylene screw cap with 2 tight connection ports fitted with 0.6 cm PFA Tube. The first one was connected to SIFT-MS and the second one to a 1L Tedlar © (Supelco, Bellefonte, PA, USA) bag, filled with zero dry air (ZeroAir Alliance ZA1500, F-DGSi, Evry France) to compensate the volume taken by the SIFT-MS analysis. The closed bottle was incubated for 2 hours at 22 ± 2 ° C before performing the positive and negative SIFT-MS full scan analysis. This optimized preparation step allows the emission of volatile compounds from the solid sample to the gas phase. Thus SIFT-MS measurements were performed on the gas phase.

3. SIFT-MS analysis

A SIFT-MS Voice 200 Ultra (SYFT Technologies, Christchurch, New Zealand) equipped with a dual source producing positive and negative soft ionizing precursor ions (H_3O^+, $O_2^{\bullet+}$, NO^+, $O^{\bullet-}$, OH^-, $O_2^{\bullet-}$, NO_2^- et NO_3^-) in a single scan was used (Hera et al., 2017; Ghislain et al., 2019; Smith et al., 2020). This instrument uses Nitrogen as carrier gas, the sample was introduced through a temperature (110 °C) and flow controlled (20 mL min^{-1}) sample line (High Performance Inlet HPI®). The instrument was daily calibrated with a standard gas (Scott™ gas mixtures, Air Liquid USA; composition in Table SI 3) containing standards at 2.0 ppmV in nitrogen (Air Liquide, Alphagaz 2). A blank experiment with an empty bottle was performed before each triplicate. The full scan raw data files containing product ion intensities with 15-250 m/z range were collected for a compilation and pre-processing step.

4. Compilation and pre-processing data

The compilation of the full scan raw data was carried out using python programming language in which the product ions were coded into numerical variables of the XAB type with X a constant letter, A ranging from 1 to 8 indicating the precursor ion used and B the value of m/z ranging from 15 to 250. The plasma source of the used instrument produces a very small amount of precursor ion NO_3^- making the ion-analyte reaction inefficient with this ion. As a consequence, we limited the exploitation of the full scan data to other seven precursor ions, that is 1,652 variables.

The data pre-processing consisted first in subtracting the background noise (signal of empty bottle) for every replicate, then in averaging the triplicates for each sample to obtain a single value of signal intensity for each ion. The final dataset was organised in matrix denoted X with n rows (n = 12) corresponding to the analysed cheeses samples and p columns (p = 1,652) corresponding to the intensities (mean values of the triplicate signals for each sample) of the product ions. Qualitative variables associated to the experimental design (batch, diet, years…) were also added in the matrix denoted Y.

To perform statistical analysis and to highlight differences between samples, the dataset was cleaned by suppressing the quantitative variables, which are constant among all samples (variance equal to zero). Negative values of the dataset, corresponding to higher signals with the blank than with a sample, were not supressed. These negatives values mostly correspond to precursor ions and water clusters of precursor ions and give also information on the sample. In addition, information about the origin (producer, diet and year) are associated to each sample as qualitative variables.

Finally, after the pre-processing step and the cleaning step, the dataset can be analysed with statistical tools. The matrix contains two groups of variables: quantitative variables (denoted X) corresponding to signal intensities and qualitative variables (denoted Y) corresponding to product descriptions. One can notice the dimensionality (number of variable) is greater than the number of samples. This dataset is then a wide matrix of omics data, and require specially developed tools.

5. Statistical analysis

The main objective of this work was to draw the volatile fingerprint of cheese samples according to intensity levels (ion count per second) of product ions with SIFT-MS full scan analysis. From these volatile fingerprints, we looked for common profiles correlated to descriptors such as producer or animal diet. From a statistical point of view, it is widely accepted that the number of samples has to be significantly greater than the number of variables, but this is no more the case with high throughput analytical technics as for omics (genomic, proteomic or, in our case, volatilomic). According to Cunningham, with an issue called “big p small n”, the analysis requires dimension reduction to improve efficiency and accuracy of data analysis (Cunningham, 2008). Thus, the R language (R. Core Team, 2018) was used with “MixOmics” package (Rohart et al., 2017) to represent the dispersion and discrimination of the samples. So, we used non-supervised method (Sparse Principal Component Analysis, Sparse PCA) and a supervised one (sparse Partial Least Squares-Discriminant Analysis, sPLS-DA).

a. Representation of a volatile fingerprint of Ossau-Iraty Cheese

The volatile fingerprint corresponds to the global response of a cheese analysis with SIFT-MS Voice 200-ultra in full scan mode after blank subtraction. The figure 1 includes the signal (ion counts per second) of all ions to visualise precursor ions and product ions. All the information is useful since it represents both the molecules present in the matrix and the overall concentration of analytes reacting with the precursor ions. However, the comparison of volatile fingerprints requires adapted statistical tools to identify correlations.

b. Reducing the dimensionality using an unsupervised method: sparse PCA

Sparse Principal Component Analysis (Sparse PCA) is a non-supervised method used to show the dispersion of the sample according to a large number of variables. Derived from classical PCA, this method is an efficient tool in the case of high-dimensional dataset and low sample size for the reduction of data dimensionality by introducing sparse structures to the input variables. Indeed, Johnstone & Lu showed that if p / n does not converge to zero, the classical PCA is not consistent, but the sparse PCA remains consistent even if $p \gg n$ (Johnstone & Lu, 2009).

Sparse PCA is devoted to find linear combinations, called components, that contains just a few meaningful input variables to explain the variability between samples. This method performs dimension reduction by projecting the data into a smaller subspace while capturing and highlighting the largest sources of data variation, resulting in a powerful visualization of the system under study (Rohart et al., 2017).

This method was applied to reduce the dimension of the X matrix and the information of Y was added as an illustrative information to detect potential heterogeneity of the cheese samples.

As in all the multivariate descriptive methods, the results consist mainly in graphical outputs: one highlighting the individual variability and the other the links between variables:

- The “individual plot” represents the samples as points placed according to their projection in the smallest subspace covered by the components. This representation makes it possible to visualise the similarities and the dissimilarities between samples through the distance between points.
- The "Correlation Circle Plot" shows the relationships between variables. In this graph, the coordinates of the variables are obtained by calculating the correlation between each original variable and the components. However, in our case, the large volume of data makes this representation unreadable if we consider all the variables.
- The contribution of each variable for each component is represented in a barplot where each bar length corresponds to the importance of the variable in the construction of the given component, which can be positive or negative.

Reducing the dimension of the dataset involves choosing a small number of components that capture as much of the variability in the data (called inertia) as possible. The choice of the number of components is therefore logically made by looking at the percentage of inertia explained by each. In a graph where two components are used, the inertia of each is cumulated in order to identify the informative power of the graph: the closer this cumulated percentage is to 100% the better the interpretation of the phenomenon. The number of components presented in this paper is limited arbitrarily to 2.

c. Discrimination of sub-populations using a supervised method: sPLS-DA

The objective of supervised methods is mainly to define rules making it possible to classify individuals from a labelled data set, the label coming from a target qualitative variable. Supervised learning consists of input variables (X) and an output variable (Y). In this approach, the algorithm makes iterative predictions on the learning data to discriminate labelled groups (Guerra et al., 2011).

Sparse partial least squares discriminant analysis (sPLS-DA) is an adaptation of PLS regression methods to the problem of supervised clustering and allows identification and quantification of the discrimination relevance (Lê Cao et al., 2011). The first step of sparse PLS-DA is the application of a sparse PLS regression model on variables which are indicators of the groups. PLS is used to find the fundamental relations between two matrices (X and Y), *i.e*. a latent variable approach to model the covariance structures in these two spaces. The second step of sparse PLS-DA is to classify observations from the results of sparse PLS regression on indicator variables (Chevallier et al., 2006). The sparse PLS-DA analysis was applied to the datasets where X is the matrix containing n lines (n = 12) corresponding to the sample and p columns (p = 1,652) corresponding to the ion signals and Y a single qualitative variable such as producers, feeding, etc.
The sparse PLS-DA analysis aims to identify a small subset of variables that best discriminate the classes. A cross-validation procedure of 3-fold CV repeated 10 times was performed to determine the number of components to retain and the optimal number of explanatory variables for each.

This supervised method allowed to obtain the same graphical outputs as the sparse PLS-DA with the "Correlation Circle Plot" and the plot of individuals. For each component of sparse PLS-DA analysis, the importance of each variable was represented in a barplot named the LoadingsPlot where each bar value corresponds to the coefficient affected to the corresponding variable to construct the corresponding component (as a linear combination). It can be positive or negative (Trendafilov & Adachi, 2015). For a discriminant analysis, the colour of the bars in the bar plot corresponds to the sample group in which the element is most "abundant".

The results can also be represented in “Clustered Image Maps” (CIM) or “heatmap” which graphically corresponds to a two-dimensional coloured image. CIM is based on hierarchical clustering operating simultaneously on the rows and columns of a matrix. Dendrograms indicate proximity between variables or samples and represent hierarchical grouping for samples (left) and for variables (top). The colour of the heat map indicates the nature of the correlation between the subsets of variables (positive, negative, strong, or weak). The advantage of this representation is that it carries a classification to discuss the similarity between samples and groups of variables. CIM is a visualization tool to observe (highlight) correlations between groups of subsets of different types of variables: between for example quantitative variables (ion) and qualitative variables (sample).

III. Results

After the pre-processing treatment of the full scan data, an example of the volatile fingerprint of a cheese is presented in Figure 1. The negative values in the radar graphic illustrate the consumption of precursor ions by sample composition (high precursor ion consumption resulting in a high negative signal value) whereas the positive ones correspond to product ions. This graphical representation gives also a global information on the sample and the instrument. In this case, some general comments can be highlighted: i) it appears clearly that $O_2^{\bullet+}$ is the main precursor ion consumed, ii) the most intense signal is obtained from the reaction of H_3O^+ precursor ion with water and iii) the ion production rate

of the instrument source is lower in negative mode than in positive one, as a consequence the intensities are much lower for detected anions (note that NO_3^- data were not presented in this work).

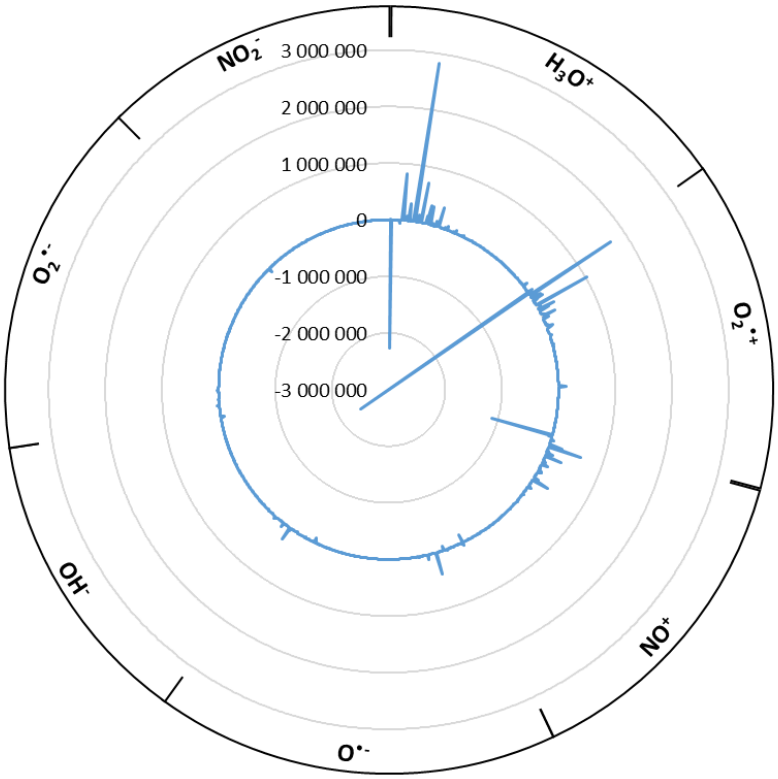

Figure 1 : graphical representation of volatile fingerprint of the sample E11.

1. Sample distribution according to volatile profile

The principal component sparse analysis was first applied on the whole dataset to assess the variability between the different samples. The sparse PCA individual plot (Figure 2) shows that Principal Component 1 (PC1) and Principal Component 2 (PC2) account for 37% of the total variability (1,652 variables). With both these dimensions, a slight differentiation between the two producers (E1 and E2) appears: E1_ samples are located close together in the lower left corner whereas E2_ samples are more to the right and also more distributed on vertical axis. This means that PC1 reflects the maximum variability between producers in the data set, whereas the PC2 highlights the dispersion of the cheeses of producer E2. The volatile fingerprints of control experiments (E11 and E13, E14 and E16) are very close for the producer 1 and no clear trend was observed with the modified experiments (E12 and E15). For producer E2 ‘s cheeses, the control experiments (E21 and E23, E24 and E26) are too dispersed to allow a measurement of diet modification effect (E23 and E25). This first non-supervised analysis indicates that the global impact of animal diet seems to be less predominant than the variability between two producers or the variability between two production years. This observation can be linked to several causes: the lactation period of ewes (the second cheese control is made 4 weeks after the first), the grass evolution during time (and year), the reproducibility of cheese making and the ripening conditions. The homogeneity of cheeses from producer E1 and the heterogeneity of cheeses from producer E2 can furthermore be explained by the rearing conditions and by the ripening parameters. Indeed, producer E1 mostly breeds his ewes in a stable with dry hay, and ripens the cheeses in a controlled chamber, whereas the producer 2 usually grazes his ewes in pastures and uses a home-made ripening chamber in his farm. These results are consistent with the diversity of cheeses from Protected Designation of Origin (PDO) Ossau-Iraty: raw milk or thermised milk, from ewes in

stables or in fields from December to August and with different ripening time. Cheese consumers appreciate this diversity of typicality within this PDO: there is a cheese for every taste.

Sparse PCA shows the dispersion and the proximity of the cheese samples with each other. However, the use of different plants with secondary bioactive metabolites (sainfoin or chicory) by the both farmers prevents a clear identification of markers. We therefore applied the principal component sparse analysis on a dataset containing only the samples of producer E1 (figure 2B), for which a better process control allows to study more precisely the variations in the composition of the volatile fingerprints. First, a year effect on volatile fingerprints is observed: the first year's samples are located to the left and the second year's sample to the right. Second, the samples produced with modified diet are located in the upper side of the figure 2B. Note that this diet effect appears more significant for the first year than the second. To go further in highlighting differences between samples according to labels (*e.g*. producer / diet) other statistical tools were applied on the dataset.

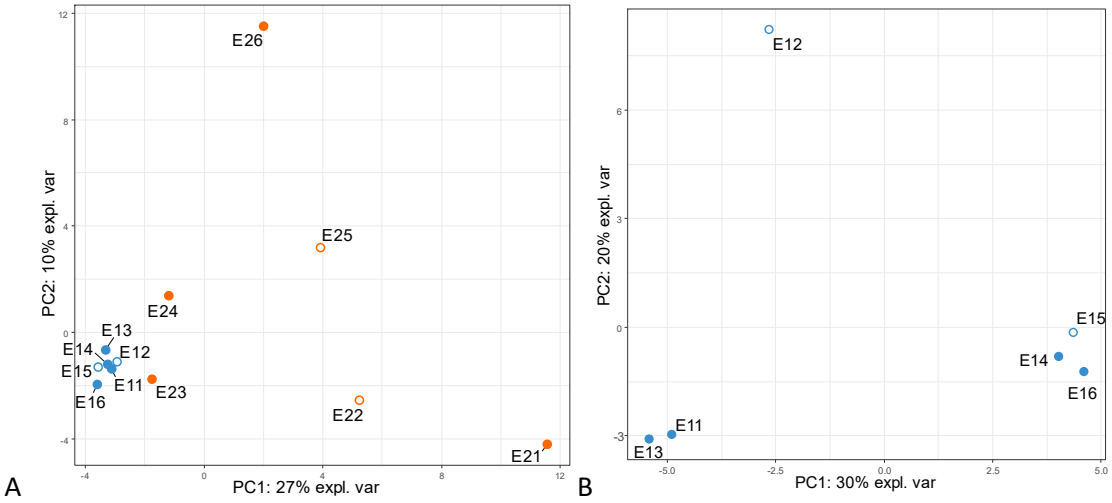

Figure 2: (A) Individual plot of sparse PCA of volatile fingerprints of both producer's samples. (B) Principal component sparse analysis on producer E1 samples only. (Blue: producer 1, orange: producer 2; filled circles: control diet; open circles: modified diet). For a better readability, only mean values of samples (n=3) are represented.

2. Producers discrimination according to the volatile profile

The supervised analysis with sparse PLS-DA was applied on the dataset to discriminate cheeses from the two different producers. According to cross-validation, the number of components selected for the sparse PLS-DA discriminating producers was 2 and the optimal number of variables to select was 70 for component 1 and 40 for component 2. The individual plot (Figure 3A) of PC1 and PC2 shows a clear separation of the two sample's groups. PC1 (26% of total variability) allows a clear discrimination of a producer from the other, whereas PC2 (13% of total variability) also illustrates the dispersion of producer E2 cheeses and the clustering of producer E1 cheeses. The difference with the previous non-supervised analysis rests on the distinction of groups with the label "producer" leading to identify different variables permitting this discrimination.

The 40 most discriminant variables (ions) between the two producers were represented in the plot loading (Figure 4A). The orange bars correspond to variables discriminating producer E2 and the blue bars correspond to variables which mostly discriminate E1. According to the product ion coding, 7 product ions in the top 40 were obtained with H_3O^+ precursor ion, 13 with NO^+, 13 with $O_2^{\bullet+}$, 1 with $O^{\bullet-}$, 5 with OH^- and 1 with NO_2^-. This distribution illustrates the importance of chemical selectivity provided

by the SIFT-MS instrument. H_3O^+ is the most used precursor ion as a quantification marker with DIMS instruments (PTR-MS and SIFT-MS), due to its prior art, but is not the principal precursor ion allowing the discrimination between producers.

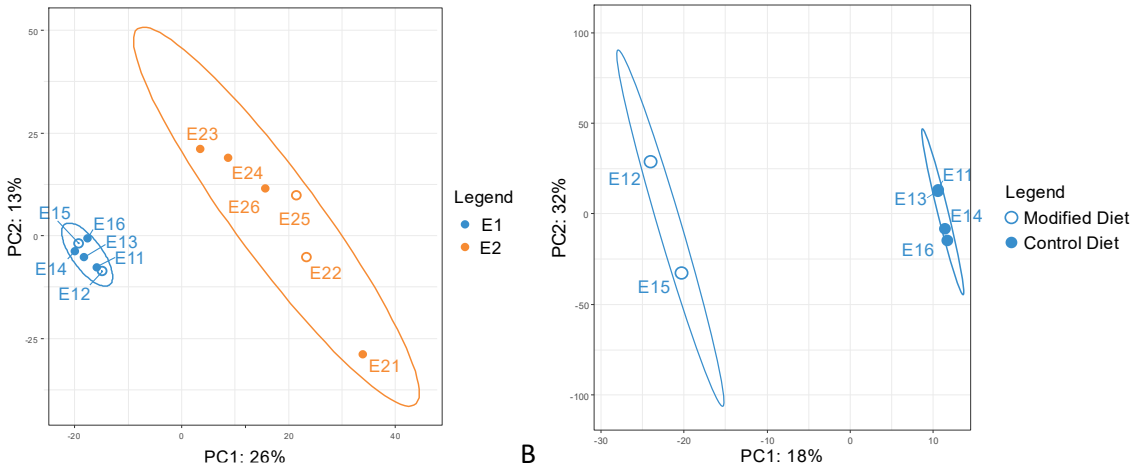

Figure 3 : Individual plot of sparse PLS-DA for the full scan mode data with the first two principal components. The different colours indicate the different batch/producers (Blue: E1 and orange: E2) A: samples distribution of both producers according sparse PLS-DA. B: focus on sample distribution of the Producer 1 according sparse PLS-DA.

As presented in the previous section, the impact of animal diet was not predominant on the whole volatile fingerprint distribution. Accordingly, a second supervised analysis was conducted only on the volatile fingerprints of the cheeses of the producer E1, in order to determine the variables that discriminate the impact of sainfoin on cheese (figure 3B). The impact of year and diet appears clearly with this statistical analysis, where PC1 discriminates the diet and PC2 discriminates the year of cheeses from producer E1. With the same strategy, a plot loading containing the 40 most discriminant ions along PC1 was presented in figure 4B. The significant product ions come also from both positive and negative precursor ions, as for the discrimination between producers (7 with H_3O^+ precursor ion, 8 with NO^+, 2 with $O_2^{\bullet+}$, 3 with $O^{\bullet-}$, 4 with OH^-, 12 with $O_2^{\bullet-}$ and 1 with NO_2^-). However, we can observe that in this test, the ions produced with $O_2^{\bullet-}$ are more represented and that the m/z ratios are also higher (7 product ions with m/z < 100, 18 between 100 < m/z < 200 and 15 between 200 < m/z < 250). This suggests that the most discriminant ions for the animal diet involve high molar mass molecules.

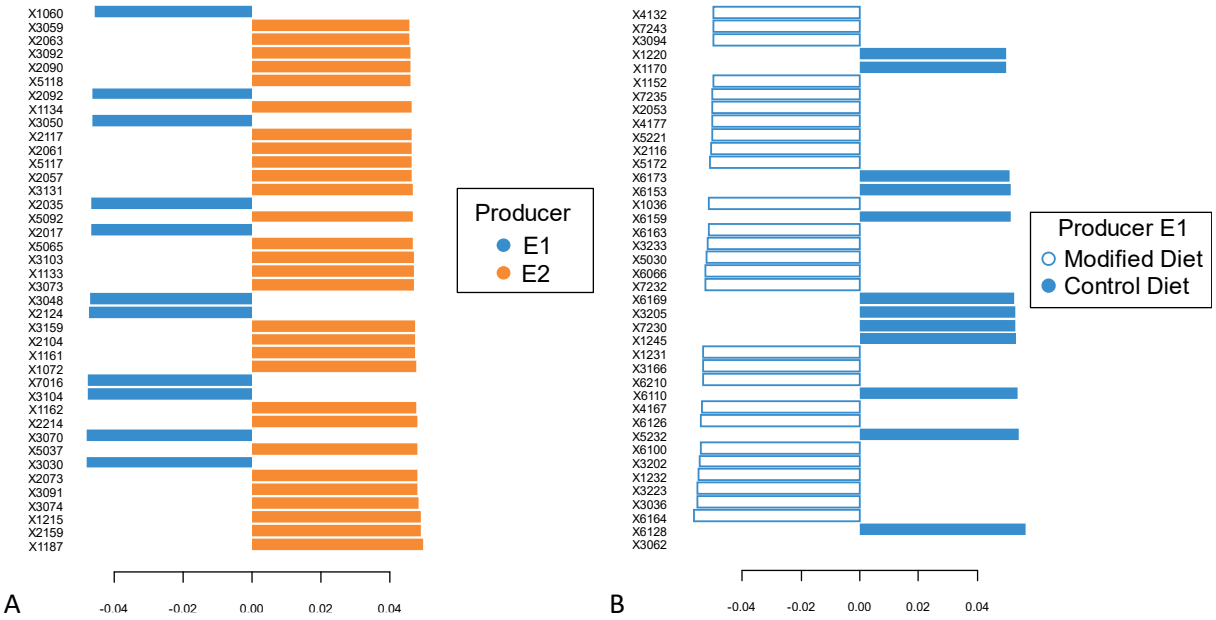

Figure 4 : A: Plot Loadings of PC1 from sparse PLS-DA method to discriminate the two producers. The colours indicate the different producers (Blue: E1 and orange: E2). B: Plot Loadings of PC1 from sparse PLS-DA method to discriminate the animal diet on cheeses of producer 1 samples (coloured bars : control diet, empty bars: modified diet)

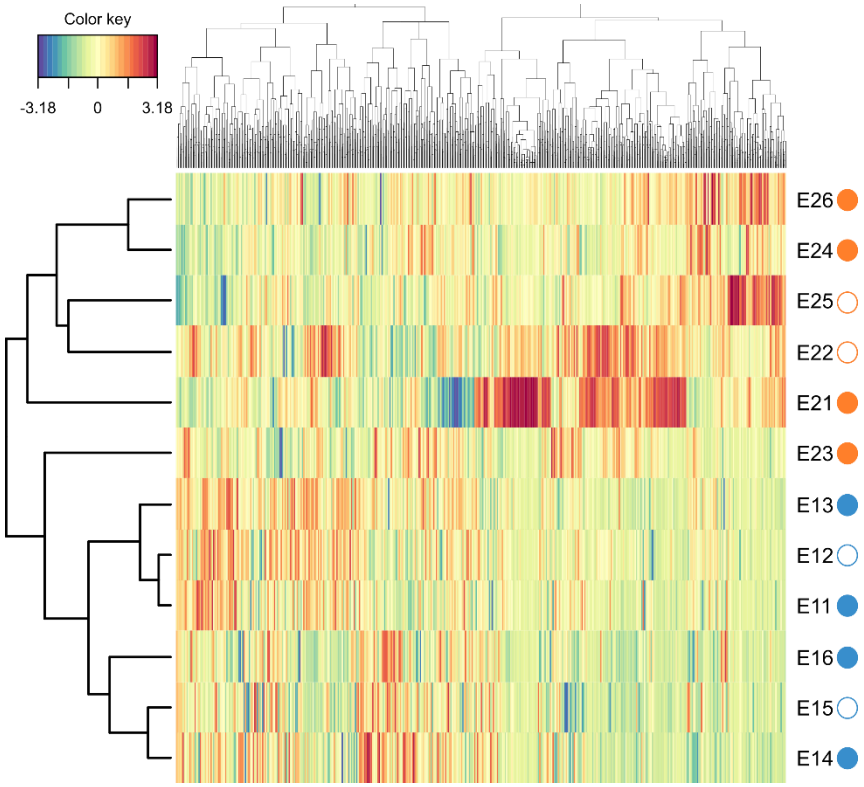

Figure 5 : Clustered Image Maps (CIM) on the dataset with sPLS-DA method to discriminate the producers (legend on right) according to the most discriminated variables (ions) with principal component 1. The dendrograms represent hierarchical grouping for samples (left) and for variables (top). The color key indicates the negative (in blue scale) or positive (in red scale) intensity of the coefficients of each sample on the first component.

The CIM representation of the volatile fingerprint (Figure 5) aims to illustrate the sample classification and, in the same time, the most relevant ions classification with the principal component 1 of the sparse PLS-DA. A clear distinction between cheeses from producers E1 and E2 with colour differences appears within the heatmap indicating that some variables are more expressed in one group than in

the other. The variables on the upper right corner are stronger (red) for the E2 cheeses than the E1 ones and conversely for the variables on the lower left which are stronger in E1 than in E2. The dendrogram of samples (on the left), shows that the sample E1 cheeses are relatively close (short lines), which confirms the homogeneity of this producer samples compared to E2 ones. In addition, for producer E1, we observe a clustering by year of production: year 1 (E11-E12-E13) and year 2 (E14-E15-E16).

IV. Discussions

The non-targeted analysis with SIFT-MS Voice200ultra enabled drawing the volatile fingerprint of cheese products with 1,652 variables. A direct analysis of volatile fingerprint (Figure 1) with such volume of data is quite difficult to achieve. Thus, to compare these complex volatile fingerprints and to address similarity and distribution issues, multivariate methods of the mixOmics package in R software was applied on a dataset of twelve ewe cheeses (analysed in triplicate) from two producers. Unsupervised method Sparse PCA, adapted to such dataset where variables are more important than the number of samples, made it possible to visualize the distribution and the variability of the samples in relation to each other. A different distribution of volatile fingerprints of cheeses appeared, according to the producers. The cheeses from the producer 1 form a cluster meaning a high similarity between samples whereas cheeses from the producer 2 spread out over the map, and that is an indicator for higher heterogeneity. On one hand, according to the principal component 1 and 2 of the sparse PCA analysis, samples E23 and E24 are close to the E1 cluster. On the other hand, it appears from this data set that differences between producers or between years of production have more impact on the overall volatile fingerprints than on the effect of the change in ewe diet. This is consistent with the diversity of ewe uncooked pressed cheeses from French Pyrenees (Feutry et al., 2012; Millet & Casabianca, 2019).

For an accurate discrimination of samples according to the qualitative variable “producer” and thus for looking for the origin of their typicity, a supervised sparse PLS-DA method was applied on the dataset (figure 3A). 26% of variables in the first Principal Component (PC1) led to an excellent separation between producers and allowed the extraction of the most important variables (product ions) in a Plot Loading figure (figure 4A). Hence, it is possible to identify the cheese’s typicity with these 430 variables. The PC2 led to a discrimination between samples and illustrated the diversity in the same batch.

The clustered Image Map (CIM, Figure 5) with sparse PLS-DA shows the classification of samples and the classification of the most relevant variables combined with an ion strength colour code. This tool was very useful to represent the volatile fingerprint of all cheese and to understand the trends. In fact, the distinction between the E1 group and the E2 group was clearly observed with a high homogeneity of the E1 cheese.

The comparison of the twelve cheeses produced by 2 breeders and with diet supplementations with two different plants do not permit identification of a global effect on cheese. However, when focusing only on samples of producer E1, a clear trend was observed according to the diet and the year of production. This is a first positive preliminary result encouraging breeders to study the secondary metabolites of plants on ewes' welfare. Nevertheless, such a study on real farm model should be consolidated by increasing the number of samples and a better control of experiment design, based on two homogenous groups of ewes: one control and one with modified diet in a laboratory farm model.

The most relevant variables for the classification of samples correspond to product ions from the reaction between a precursor ion and an analyte. The aim of this study was to demonstrate that untargeted SIFT-MS analysis coupled with mixOmics supervised method allowed the discrimination of cheese samples. However, to go further, it could be very interesting to identify molecules responsible of a sample's specificity and to go back to the involved enzymatic pathways, especially in the case of food products. Actually, the quadrupole mass detector of the SIFT-MS, is not a high-resolution mass detector and does not allow identification of the exact chemical formula of every product ion. A chemical compound emitted by a cheese sample may react with a precursor ion and give several product ions. In addition, two different chemical compounds may give rise to the same m/z product ion that makes the interpretation of SIFT-MS analysis challenging in complex matrices. It is all the truer because in a food product, the metabolic pathways give rise to families of chemical compounds such as alcohols, acids and esters that have a similar reactivity with precursor ions and then lead automatically to conflict ions, complicated to elucidate (Ghislain et al., 2021).

In most of the publications in the scientific field of DIMS analysis of food products, the authors (H. Castada et al., 2019; Taylor et al., 2013) based their strategies on the quantification of a few dozen of known compounds. This hard work can be reinforced by our approach which demonstrated that there are more discriminating ions for classification. Moreover, the conflict ions with known and unknown compounds of the matrix make such quantifications difficult.

Knowing these experimental limitations, we have nevertheless tried to identify key compounds for the classification of sheep cheeses based on two post-treatment strategies: (i) first, by looking for product ions of known compounds, already identified in the literature: are these ions relevant in the ~~top~~ list of 40 most relevant ions? (ii) second, by attempting attribution of product ions: which molecule could correspond to the most relevant ions?

i) We have established a SIM method with 58 main compounds derived from the literature on volatile compounds released by cheeses, (table 4 SI) and present in the software database (Labsyft®) for the positive ionisation (with H_3O^+, $O_2^{\bullet+}$ and NO^+). These 58 compounds are known to produce 410 different product ions useful for the quantification among which 329 are conflict ions. For example, $O_2^{\bullet+}$ precursor ion reacts with 8 different compounds from the list giving rise to a product ion with the same m/z 43 (Table 1 SI). Thus, the precise determination of a compound concentration from a conflict ion is quite challenging. The remaining 81 ions with no conflict allow the measurements of the concentration of 36 compounds of the list without being confident on the accuracy of these calculations due to the possible presence of unexpected compounds. From the 40 most discriminating product ions obtained during the sPLS-DA analysis differentiating the cheeses producers (Figure 4A), 13 product ions from the 58 compounds of the SIM method were found. This clearly shows that other compounds participate in the discrimination and that the volatile fingerprint provides more information than the SIM method. This SIM method is only based on targeted compounds, usually identified by GC-MS and excluding inorganic compounds however detected by SIFT-MS.

ii) We approached the problem from another angle to identify all the key compounds. The 40 most relevant product ions providing discrimination of the producers VOC, were searched in the instrument database to identify the already known compounds. Approximately 300 compounds were thus identified, corresponding to at least 1 of these 40 ions. Some of these compounds can correspond to several product ions. For example, 3 product ions (m/z 92 with NO^+, m/z 61 with $O_2^{\bullet+}$, m/z 117 with $O_2^{\bullet+}$) discriminating the producer E2 may correspond to propyl hexanoate, an already identified compound in cheeses (Bertolino et al., 2011; Bosset & Gauch, 1993; Di Cagno et al., 2003). This indicates that propyl hexanoate, an odorous compound, can be a more expressed marker in producer E2 cheeses than in producer E1 ones. Additional SPME-GC-MS experiments were carried out on the

same samples (data not shown) and confirmed that propyl hexanoate was overexpressed in producer E2 cheeses.

The same data mining strategy was used to search differences between modified and control diets within E1 cheeses. The discrimination between samples categories was very clear in figure 3B and a large number of the 40 most relevant product ions in Figure 4B have a high m/z value. One of these ions generated from the H_3O^+ precursor ion at m/z 137, could be a very interesting candidate because of its involvement in a large number of terpene molecules. However, other characteristic ions of terpenes were not found in the top 40 ions list. This could be explained by a high level of interferences in product ions of terpenes. Several negative product ions were also highlighted by this analysis but the lack of a SIFT negative ionization database will not enable this research to further develop. The obvious separation of samples in SIFT-MS full scan mode confirmed our hypothesis about the volatile fingerprints: untargeted analysis enhances the ability to detect unexpected compound. Nevertheless, the complementarity of this approach with more traditional techniques can allow the identification of key molecules.

CONCLUSION

The main objective of this work was to show the great potential of non-supervised SIFT-MS analysis in full scan mode, to study the volatile compounds released by cheeses. The use of adapted statistical tools, dedicated to "omics" Data set, allowed some characterization of cheese producers and the effect of animal diet modifications. As for sensory analysis, where several molecules may fix on receptors and give different synergic or antagonist nervous signals, the raw measurement of all product ions (without any determination of molecule concentration) allows obtaining a fingerprint of the matrix. The construction of large database of volatile fingerprints will enable the instrument to correlate a profile to a product specificity. The high frequency mass spectrometers like those in SIFT-MS generate a large data base with a large number of variables and opens a new way of chemical analysis. In this work, the data mining and the statistical tools were adapted to generate a data set for a small number of ewe's cheeses. It was demonstrated that the volatile fingerprints of samples make it possible to identify the cheese of each producer and highlighted the diversity of samples in the same batch. Cheese is a well-known matrix in food analytical science but a large number of parameters and manufacturing steps can impact the final product: the breed of animals, their lactation period, the composition of the nutritional part, the conditions of milk collection and at last, the transformation and maturation processes. To address some issue like the impact of nutrition on the final product, it requires a perfect control of all the steps/parameters. Changing the mind by looking through big-data science will allow researchers and producers to find correlations between fingerprints and products' typicality. To improve the present work, a construction of an Ossau-Iraty PDO SIFT-MS database is in underway, which might help producers to solve question about typicality, diversity and agricultural practices.

Aknowledgments

Authors thank Conseil Régional Nouvelle Aquitaine (CRNA) and Communauté d'Agglomération Côte Basque (CAPB) for their financial support. Authors thank also the partners of BioNAchol project and in particular Xavier Recondo from Institute *Jean Errecart* (Saint Palais, France), the chamber of agriculture of Pyrenées Atlantique (France), the Protected Designation of Origin (PDO) Ossau-Iraty and the producers: *la ferme Oros Mirassou* (Arette, France) and *la ferme Bacquet* (Orègue, France).

Bibliographie

Bajoub, A., Medina-Rodríguez, S., Ajal, E. A., Cuadros-Rodríguez, L., Monasterio, R. P., Vercammen, J., Fernández-Gutiérrez, A., & Carrasco-Pancorbo, A. (2018). A metabolic fingerprinting approach based on selected ion flow tube mass spectrometry (SIFT-MS) and chemometrics: A reliable tool for Mediterranean origin-labeled olive oils authentication. *Food Research International*, *106*, 233–242. https://doi.org/10.1016/j.foodres.2017.12.027

Balkir, P., Kemahlioglu, K., & Yucel, U. (2021). Foodomics: A new approach in food quality and safety. *Trends in Food Science & Technology*, *108*, 49–57. https://doi.org/10.1016/j.tifs.2020.11.028

Bergamaschi, M., Cipolat-Gotet, C., Cecchinato, A., Schiavon, S., & Bittante, G. (2020). Chemometric authentication of farming systems of origin of food (milk and ripened cheese) using infrared spectra, fatty acid profiles, flavor fingerprints, and sensory descriptions. *Food Chemistry*, *305*, 125480. https://doi.org/10.1016/j.foodchem.2019.125480

Bertolino, M., Dolci, P., Giordano, M., Rolle, L., & Zeppa, G. (2011). Evolution of chemico-physical characteristics during manufacture and ripening of Castelmagno PDO cheese in wintertime. *Food Chemistry*, *129*(3), 1001–1011. https://doi.org/10.1016/j.foodchem.2011.05.060

Biasioli, F., Yeretzian, C., Märk, T. D., Dewulf, J., & Van Langenhove, H. (2011). Direct-injection mass spectrometry adds the time dimension to (B)VOC analysis. *TrAC Trends in Analytical Chemistry*, *30*(7), 1003–1017. https://doi.org/10.1016/j.trac.2011.04.005

Boltar, I., Čanžek Majhenič, A., Jarni, K., Jug, T., & Bavcon Kralj, M. (2015). Volatile compounds in Nanos cheese: Their formation during ripening and sesonal variation. *Journal of Food Science and Technology*, *52*(1), 608–623. https://doi.org/10.1007/s13197-014-1565-6

Boltar, I., Majhenič, A. Č., Jarni, K., Jug, T., & Kralj, M. B. (2019). Research of volatile compounds in cheese affected by different technological parameters. *J. Food Nutr. Res.*, *58*, 10.

Bosset, J. O., & Gauch, R. (1993). Comparison of the volatile flavour compounds of six european 'AOC' cheeses by using a new dynamic headspace GC-MS method. *International Dairy Journal*, *3*(4–6), 359–377. https://doi.org/10.1016/0958-6946(93)90023-S

Castada, H., Hanas, K., & Barringer, S. (2019). Swiss Cheese Flavor Variability Based on Correlations of Volatile Flavor Compounds, Descriptive Sensory Attributes, and Consumer Preference. *Foods*, *8*(2), 78. https://doi.org/10.3390/foods8020078

Castada, H. Z., Wick, C., Taylor, K., & Harper, W. J. (2014). Analysis of Selected Volatile Organic Compounds in Split and Nonsplit Swiss Cheese Samples Using Selected-Ion Flow Tube Mass Spectrometry (SIFT-MS): Analysis of selected volatile organic…. *Journal of Food Science*, *79*(4), C489–C498. https://doi.org/10.1111/1750-3841.12418

Castro-Puyana, M., & Herrero, M. (2013). Metabolomics approaches based on mass spectrometry for food safety, quality and traceability. *TrAC Trends in Analytical Chemistry*, *52*, 74–87. https://doi.org/10.1016/j.trac.2013.05.016

Cecchi, T., Sacchini, L., & Felici, A. (2018). First Investigation on the Shelf life of Mediterranean Mussels (Mytilus galloprovincialis) on the Basis of Their Volatiles Profiles. *Food Analytical Methods*, *11*(5), 1451–1456. https://doi.org/10.1007/s12161-017-1129-2

Ch, R., Chevallier, O., McCarron, P., McGrath, T. F., Wu, D., Nguyen Doan Duy, L., Kapil, A. P., McBride, M., & Elliott, C. T. (2021). Metabolomic fingerprinting of volatile organic compounds for the geographical discrimination of rice samples from China, Vietnam and India. *Food Chemistry*, *334*, 127553. https://doi.org/10.1016/j.foodchem.2020.127553

Chevallier, S., Bertrand, D., Kohler, A., & Courcoux, P. (2006). Application of PLS-DA in multivariate image analysis. *Journal of Chemometrics*, *20*(5), 221–229. https://doi.org/10.1002/cem.994

Cumeras, R. (2017). Volatilome Metabolomics and Databases, Recent Advances and Needs. *Current Metabolomics*, *5*(2), 79–89. https://doi.org/10.2174/2213235X05666170502103408

Cunningham, P. (2008). Dimension Reduction. In M. Cord & P. Cunningham (Eds.), *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval* (pp. 91–112). Springer. https://doi.org/10.1007/978-3-540-75171-7_4

Deuscher, Z., Andriot, I., Sémon, E., Repoux, M., Preys, S., Roger, J.-M., Boulanger, R., Labouré, H., & Le Quéré, J.-L. (2019). Volatile compounds profiling by using proton transfer reaction-time of

flight-mass spectrometry (PTR-ToF-MS). The case study of dark chocolates organoleptic differences. *Journal of Mass Spectrometry*, *54*(1), 92–119. https://doi.org/10.1002/jms.4317

Di Cagno, R., Banks, J., Sheehan, L., Fox, P. F., Brechany, E. Y., Corsetti, A., & Gobbetti, M. (2003). Comparison of the microbiological, compositional, biochemical, volatile profile and sensory characteristics of three Italian PDO ewes' milk cheeses. *International Dairy Journal*, *13*(12), 961–972. https://doi.org/10.1016/S0958-6946(03)00145-6

Farag, M. A., Hegazi, N., Dokhalahy, E., & Khattab, A. R. (2020). Chemometrics based GC-MS aroma profiling for revealing freshness, origin and roasting indices in saffron spice and its adulteration. *Food Chemistry*, *331*, 127358. https://doi.org/10.1016/j.foodchem.2020.127358

Ferreira, I. M. P. L. V. O., Pinho, O., & Sampaio, P. (2009). Volatile fraction of DOP "Castelo Branco" cheese: Influence of breed. *Food Chemistry*, *112*(4), 1053–1059. https://doi.org/10.1016/j.foodchem.2008.06.048

Feutry, F., Torre, P., Arana, I., Garcia, S., Desmasures, N., & Casalta, E. (2012). Lactococcus lactis strains from raw ewe's milk samples from the PDO Ossau-Iraty cheese area: Levels, genotypic and technological diversity. *Dairy Science & Technology*, *92*(6), 655–670. https://doi.org/10.1007/s13594-012-0084-3

Ghislain, M., Costarramone, N., Sotiropoulos, J.-M., Pigot, T., Berg, R. V. D., Lacombe, S., & Bechec, M. L. (2019). Direct analysis of aldehydes and carboxylic acids in the gas phase by negative ionization selected ion flow tube mass spectrometry: Quantification and modelling of ion–molecule reactions. *Rapid Communications in Mass Spectrometry*, *33*(21), 1623–1634. https://doi.org/10.1002/rcm.8504

Ghislain, M., Reyrolle, M., Sotiropoulos, J.-M., Pigot, T., & Le Bechec, M. (2021). Chemical ionization of carboxylic acids and esters in negative mode selected ion flow tube – Mass spectrometry (SIFT-MS). *Microchemical Journal*, *169*, 106609. https://doi.org/10.1016/j.microc.2021.106609

Guerra, L., McGarry, L. M., Robles, V., Bielza, C., Larrañaga, P., & Yuste, R. (2011). Comparison between supervised and unsupervised classifications of neuronal cell types: A case study. *Developmental Neurobiology*, *71*(1), 71–82. https://doi.org/10.1002/dneu.20809

Hera, D., Langford, V., McEwan, M., McKellar, T., & Milligan, D. (2017). Negative Reagent Ions for Real Time Detection Using SIFT-MS. *Environments*, *4*(1), 16. https://doi.org/10.3390/environments4010016

Ianni, A., Bennato, F., Martino, C., Grotta, L., & Martino, G. (2020). Volatile Flavor Compounds in Cheese as Affected by Ruminant Diet. *Molecules*, *25*(3), 461. https://doi.org/10.3390/molecules25030461

Jiang, Y., Li, N., Wang, Q., Liu, Z., Lee, Y.-K., Liu, X., Zhao, J., Zhang, H., & Chen, W. (2019). Microbial diversity and volatile profile of traditional fermented yak milk. *Journal of Dairy Science*, S0022030219309427. https://doi.org/10.3168/jds.2019-16753

Johnstone, I. M., & Lu, A. Y. (2009). On Consistency and Sparsity for Principal Components Analysis in High Dimensions. *Journal of the American Statistical Association*, *104*(486), 682–693. https://doi.org/10.1198/jasa.2009.0121

Lê Cao, K.-A., Boitard, S., & Besse, P. (2011). Sparse PLS discriminant analysis: Biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*, *12*(1), 253. https://doi.org/10.1186/1471-2105-12-253

Luykx, D. M. A. M., & van Ruth, S. M. (2008). An overview of analytical methods for determining the geographical origin of food products. *Food Chemistry*, *107*(2), 897–911. https://doi.org/10.1016/j.foodchem.2007.09.038

Lytou, A. E., Panagou, E. Z., & Nychas, G.-J. E. (2019). Volatilomics for food quality and authentication. *Current Opinion in Food Science*, S2214799319300724. https://doi.org/10.1016/j.cofs.2019.10.003

Millet, M., & Casabianca, F. (2019). Sharing Values for Changing Practices, a Lever for Sustainable Transformation? The Case of Farmers and Processors in Interaction within Localized Cheese Sectors. *Sustainability*, *11*(17), 4520. https://doi.org/10.3390/su11174520

R. Core Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. *Www.r-Project.Org*.

Rohart, F., Gautier, B., Singh, A., & Lê Cao, K.-A. (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. *PLOS Computational Biology*, *13*(11), e1005752. https://doi.org/10.1371/journal.pcbi.1005752

Santos, J., & Oliveira, M. B. P. P. (2017). Chromatography: Introduction to Chromatography - Techniques. In C. A. Georgiou & G. P. Danezis (Eds.), *Food Authentication* (pp. 199–232). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118810224.ch7a

Smith, D., McEwan, M. J., & Španěl, P. (2020). Understanding Gas Phase Ion Chemistry Is the Key to Reliable Selected Ion Flow Tube-Mass Spectrometry Analyses. *Anal. Chem.*, 13.

Smith, D., & Španěl, P. (2005). Selected ion flow tube mass spectrometry (SIFT-MS) for on-line trace gas analysis. *Mass Spectrometry Reviews*, *24*(5), 661–700. https://doi.org/10.1002/mas.20033

Sousa, A., Vareda, J., Pereira, R., Silva, C., Câmara, J. S., & Perestrelo, R. (2020). Geographical differentiation of apple ciders based on volatile fingerprint. *Food Research International*, *137*, 109550. https://doi.org/10.1016/j.foodres.2020.109550

Taylor, K., Wick, C., Castada, H., Kent, K., & Harper, W. J. (2013). Discrimination of Swiss Cheese from 5 Different Factories by High Impact Volatile Organic Compound Profiles Determined by Odor Activity Value Using Selected Ion Flow Tube Mass Spectrometry and Odor Threshold: Swiss cheese factory VOC variability…. *Journal of Food Science*, *78*(10), C1509–C1515. https://doi.org/10.1111/1750-3841.12249

Trendafilov, N. T., & Adachi, K. (2015). Sparse Versus Simple Structure Loadings. *Psychometrika*, *80*(3), 776–790. https://doi.org/10.1007/s11336-014-9416-y